

Lecture Notes in Civil Engineering

Rishi Gupta · Min Sun · Svetlana Brzev ·  
M. Shahria Alam · Kelvin Tsun Wai Ng ·  
Jianbing Li · Ashraf El Damatty ·  
Clark Lim *Editors*

# Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022

Volume 1

# **Lecture Notes in Civil Engineering**

Volume 363

## **Series Editors**

Marco di Prisco, Politecnico di Milano, Milano, Italy

Sheng-Hong Chen, School of Water Resources and Hydropower Engineering,  
Wuhan University, Wuhan, China

Ioannis Vayas, Institute of Steel Structures, National Technical University of  
Athens, Athens, Greece

Sanjay Kumar Shukla, School of Engineering, Edith Cowan University, Joondalup,  
WA, Australia

Anuj Sharma, Iowa State University, Ames, IA, USA

Nagesh Kumar, Department of Civil Engineering, Indian Institute of Science  
Bangalore, Bengaluru, Karnataka, India

Chien Ming Wang, School of Civil Engineering, The University of Queensland,  
Brisbane, QLD, Australia



**Lecture Notes in Civil Engineering (LNCE)** publishes the latest developments in Civil Engineering—quickly, informally and in top quality. Though original research reported in proceedings and post-proceedings represents the core of LNCE, edited volumes of exceptionally high quality and interest may also be considered for publication. Volumes published in LNCE embrace all aspects and subfields of, as well as new challenges in, Civil Engineering. Topics in the series include:

- Construction and Structural Mechanics
- Building Materials
- Concrete, Steel and Timber Structures
- Geotechnical Engineering
- Earthquake Engineering
- Coastal Engineering
- Ocean and Offshore Engineering; Ships and Floating Structures
- Hydraulics, Hydrology and Water Resources Engineering
- Environmental Engineering and Sustainability
- Structural Health and Monitoring
- Surveying and Geographical Information Systems
- Indoor Environments
- Transportation and Traffic
- Risk Analysis
- Safety and Security

To submit a proposal or request further information, please contact the appropriate Springer Editor:

- Pierpaolo Riva at [pierpaolo.riva@springer.com](mailto:pierpaolo.riva@springer.com) (Europe and Americas);
- Swati Meherishi at [swati.meherishi@springer.com](mailto:swati.meherishi@springer.com) (Asia—except China, Australia, and New Zealand);
- Wayne Hu at [wayne.hu@springer.com](mailto:wayne.hu@springer.com) (China).

**All books in the series now indexed by Scopus and EI Compendex database!**

Rishi Gupta · Min Sun · Svetlana Brzev ·  
M. Shahria Alam · Kelvin Tsun Wai Ng ·  
Jianbing Li · Ashraf El Damatty · Clark Lim  
Editors

# Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022

Volume 1

*Editors*

Rishi Gupta  
Department of Civil Engineering  
University of Victoria  
Victoria, BC, Canada

Min Sun  
Department of Civil Engineering  
University of Victoria  
Victoria, BC, Canada

Svetlana Brzev  
Department of Earthquake Engineering  
The University of British Columbia  
Vancouver, BC, Canada

M. Shahria Alam  
The University of British  
Columbia-Okanagan Campus  
Kelowna, BC, Canada

Kelvin Tsun Wai Ng  
University of Regina  
Regina, SK, Canada

Jianbing Li  
Environmental Engineering Program  
University of Northern British Columbia  
Prince George, BC, Canada

Ashraf El Damatty  
The University of Western Ontario  
London, ON, Canada

Clark Lim  
Department of Civil Engineering  
The University of British Columbia  
Vancouver, BC, Canada

ISSN 2366-2557

ISSN 2366-2565 (electronic)

Lecture Notes in Civil Engineering

ISBN 978-3-031-34592-0

ISBN 978-3-031-34593-7 (eBook)

<https://doi.org/10.1007/978-3-031-34593-7>

© Canadian Society for Civil Engineering 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Conference Committee Members

## Volume 1

Dr. Rishi Gupta (Chair)  
Dr. Svetlana Brzev (Technical Co-chair)  
Dr. Min Sun (Technical Co-chair)  
Dr. Jianbing Li (Cold Regions Specialty Chair)  
Dr. Kelvin Tsun Wai Ng (Environmental Specialty Chair)

## Volume 2

Dr. Rishi Gupta (Chair)  
Dr. Svetlana Brzev (Technical Co-chair)  
Dr. Min Sun (Technical Co-chair)  
Dr. Ashraf El Damatty  
Dr. Ahmed Elshaer

## Volume 3

Dr. Rishi Gupta (Chair)  
Dr. Svetlana Brzev (Technical Co-chair)  
Dr. Min Sun (Technical Co-chair)  
Dr. Ashraf El Damatty (Structural Specialty Co-chair)  
Dr. Ahmed Elshaer (Structural Specialty Co-chair)  
Dr. Shahria Alam (Material Specialty Chair)  
Dr. Clark Lim (Transportation Specialty Chair)

**Volume 4**

Dr. Rishi Gupta (Chair)

Dr. Svetlana Brzev (Technical Co-chair)

Dr. Min Sun (Technical Co-chair)

Dr. Ashraf El Damatty (Structural Specialty Co-chair)

Dr. Ahmed Elshaer (Structural Specialty Co-chair)

Dr. Shahria Alam (Material Specialty Chair)

Dr. Jianbing Li (Cold Regions Specialty Chair)

Dr. Kelvin Tsun Wai Ng (Environmental Specialty Chair)

# Contents

<b>Construction Management: Construction Management</b>	
<b>Framework for Assessing the Usability of Augmented Reality Applications for Construction Work .....</b>	<b>3</b>
Mayank Arvindbhai Patel, Krithikashree Lakshminarayanan, Zia Din, and Lingguang Song	
<b>Current and Future Trends of Augmented and Mixed Reality Technologies in Construction .....</b>	<b>19</b>
Mahsa Rezvani, Zhen Lei, Jeff Rankin, and Lloyd Waugh	
<b>Stochastic Modeling of Tag Installation Error for Robust On-Manifold Tag-Based Visual-Inertial Localization .....</b>	<b>41</b>
Navid Kayhani, Brenda McCabe, and Angela P. Schoellig	
<b>Virtual Reality-Based Expert Demonstrations for Training Construction Robots via Imitation Learning .....</b>	<b>55</b>
Lei Huang, Weijia Cai, and Zhengbo Zou	
<b>The Effect of Human Body Blockage on UWB Tracking Accuracy in Construction Sites .....</b>	<b>69</b>
Pegah Behvarmanesh and Farnaz Sadeghpour	
<b>Enhanced Activity-on-Arrow Network Diagramming Method for Construction Planning and Scheduling Applications .....</b>	<b>83</b>
Badhon Das Shuvo and Ming Lu	
<b>Integrating Smart 360-Degree Photography and QR Codes for Enhancing Progress Reporting .....</b>	<b>95</b>
Ahmed Bahakim, Ibrahim Abotaleb, and Ossama Hosny	
<b>Envisioning Digital Twin-Enabled Post-occupancy Evaluations for UVic Engineering Expansion Project .....</b>	<b>109</b>
Ishan Tripathi, Thomas Froese, and Shauna Mallory-Hill	

<b>Application of Compressed Sensing on Crowdsensing-Based Indirect Bridge Condition Monitoring</b> .....	123
Qipei Mei	
<b>Adoption and Implementation of Common Data Environments in the Province of Quebec: Barriers, Challenges, and Trends</b> .....	133
Erik Andrew Poirier and Margaux Soyez	
<b>Enwave’s Western Expansion Project—Key Challenges and Solutions</b> .....	153
Mark Bruder, Ahmed Elsherif, James Scharbach, and Kris Landon	
<b>Enhancing Bridges’ Safety Training Using Augmented Reality and Virtual Reality</b> .....	173
M. El Rifae, S. Bader, I. Abotaleb, O. Hosny, and K. Nassar	
<b>Integrating Virtual Reality into IOSH Safety Training</b> .....	197
Y. Elhakim, S. Bader, M. Elrifae, S. Ibrahim, A. Sorour, M. Soliman, M. Sherif, I. Abotaleb, O. Hosny, and K. Nassar	
<b>Semantic Segmentation of Synthetic Images into Building Components for Automated Quality Assurance</b> .....	215
H. X. Zhang, L. Huang, W. Cai, and Z. Zou	
<b>Lessons Learned from Developing and Testing an Augmented Reality Application for Just-in-Time Information Delivery to Improve Construction Safety</b> .....	229
Krithikashree Lakshminarayanan, Mayank Arvindbhai Patel, Zia Din, and Lingguang Song	
<b>The Effectiveness of Data Augmentation in Construction Site-Related Image Classification</b> .....	247
Mansoor Asif, Shuai Liu, Ghulam Muhammad Ali, Ahmed Bouferguene, and Mohamed Al-Hussein	
<b>An Integrated Approach Combining Virtual Environments and Reinforcement Learning to Train Construction Robots for Conducting Tasks Under Uncertainties</b> .....	259
Weijia Cai, Lei Huang, and Zhengbo Zou	
<b>Investigating the Feasibility of Using Virtual Reality Devices to Present Construction Information in Both Mixed Reality and Virtual Reality Environments</b> .....	273
Hardith Suvarna Murari, Zia Din, and Christiane Spitzmueller	
<b>The Incorporation of Learning Theories in VR-Based Safety Training Programs Within the Construction Industry</b> .....	285
S. Bader, I. Abotaleb, O. Hosny, and K. Nassar	

<b>Resources Deployment Optimization in Scattered Repetitive Projects</b> .....	307
Heba Kh. Gad, Mostafa H. Ali, Khaled Nassar, Yasmeen A. S. Essawy, and Abdelhamid Abdullah	
<b>Pull-Based Simulation Modelling for Modular Construction Supply Chain Analysis: A Case Study in Northern Canada</b> .....	319
Keagan Hudson Rankin, Zhuo Cheng, Zhen Lei, Samira Rizaee, Brandon Searle, Cynthia Ene, and Solomon Amuno	
<b>Construction Payment Problems: A Critical Review of Payment Problems, Prompt Payment Legislation, and Challenges</b> .....	333
Dalia H. Dorrah and Brenda Y. McCabe	
<b>Gap Analysis and Areas of Improvement for the CCDC 30: 2018 Integrated Project Delivery Contract</b> .....	347
Audrey Provost, Érik A. Poirier, and Daniel Forgues	
<b>Considerations for Site Layout Planning Decision-Making</b> .....	367
A. Marcano Pina and F. Sadeghpour	
<b>Irregular Dynamic Site Layout Optimization Model</b> .....	391
Heba Kh. Gad, Mostafa H. Ali, Aya Eldesouky, Alshimaa Abdullatif, Mos. Serry, Hosny Ossama, and Yasmeen A. S. Essawy	
<b>Gordie Howe International Bridge Managing a Pile of Paper on a Paperless Project</b> .....	403
Randy Pickle	
<b>Low Regret-Based Design and Corrosion Management for Steel Roadway Bridges</b> .....	421
M. Barkhori, S. Walbridge, and M. Pandey	
<b>Agent-Based Modeling and Simulation of Congested Sites</b> .....	439
Raghda M. Moharram, Yasmeen A. S. Essawy, Abdelhamid Abdullah, and Khaled Nassar	
<b>Selecting Most Appropriate Delay Analysis Technique Using Quantitative Approach</b> .....	451
Mostafa Farouk and Ossama Hosny	
<b>Axiomatic Design-Based Optimization Framework for Factory Logistics Design in Precast Concrete Construction</b> .....	473
Shuai Liu, Asif Mansoor, Ghulam Muhammad Ali, Ahmed Bouferguene, and Mohamed Al-Hussein	



<b>Agent-Based Modeling and Simulation of Project Schedule Risk Analysis in the Construction Industry</b> .....	493
Mohamed ElGindi, Sara Harb, Abdelhamid Abdullah, Yasmeen A. S. Essawy, and Khaled Nassar	
<b>Automated Resource Scheduling for Construction Projects Using Genetic Algorithm</b> .....	513
Raghda M. Moharram, Yasmeen A. S. Essawy, and Osama S. Hosny	
<b>Construction Management: Sustainable Construction</b>	
<b>Optimal Planning of Renewable Energy Integration for Off-grid Residential Buildings in Northern Regions</b> .....	525
Don Rukmal Liyanage, Kasun Hewage, and Rehan Sadiq	
<b>The Impact of Occupancy Pattern on Energy-Efficient Building System Selection: A Case Study of a Living Laboratory in Okanagan (BC)</b> .....	541
S. R. Sultana, M. R. Kamal, M. F. A. Khan, M. Kamali, A. Rana, M. S. Alam, K. Hewage, and R. Sadiq	
<b>An Approach for Teaching the Design of Net Zero-Energy Buildings</b> .....	563
Mokhtar Ahmed	
<b>A Framework Approach to Utilize Real-Time Spatial Labor Data from Construction Sites</b> .....	583
Hoda Author Abouorban, Khaled Nassar, and Elkhayam Dorra	
<b>Construction Management: Building Information Modeling</b>	
<b>Robotic Additive Manufacturing Using a Visual Programming BIM Environment</b> .....	605
Tayeb Boualam Allah, Walid Anane, and Ivanka Iordanova	
<b>An Integrated BIM-GIS Dashboard to Improve BIM Coordination</b> .....	623
Anouar El Haite and Conrad Boton	
<b>Overcoming Obstacles in BIM-Based Multidisciplinary Coordination: A Literature Overview</b> .....	637
Tabassum Mushtary Meem and Ivanka Iordanova	
<b>Adopting Ecolabels in the Construction Industry via Blockchain</b> .....	655
Dilusha Kankanamge and Rajeev Ruparathna	

<b>Role of Electronic Document Management Systems in the Design Change Management Process</b> .....	673
Oussama Ghnaya, Hamidreza Pourzareei, Louis Rivest, and Conrad Boton	
<b>An Interactive Decision-Support Tool to Improve Construction Cost Management with Building Information Modeling</b> .....	687
Nicolas Strange, Daniel Forgues, and Conrad Boton	
<b>Construction Management: Water Resources</b>	
<b>Machine Learning Approximation for Rapid Prediction of High-Dimensional Storm Surge and Wave Responses</b> .....	701
Saeed Saviz Naeini and Reda Snaiki	
<b>Predicting Pumps-as-Turbine Characteristics with the Use of Machine Learning Applications</b> .....	711
Alex Brisbois and Rebecca Dziedzic	
<b>Large Eddy Simulation of Near-Bed Flow Over Bottom Roughness in Open Channel</b> .....	731
Bowen Xu and S. Samuel Li	
<b>Real-Time Water Distribution System Calibration Using Genetic Algorithm</b> .....	749
Ziyuan Cai, Rebecca Dziedzic, and S. Samuel Li	
<b>Flow Field Characteristics of Particle-Laden Swirling Jets</b> .....	763
F. Sharif and A. H. Azimi	
<b>A Detailed 2D Hydraulic Model for the Lower Fraser River</b> .....	775
Junying Qu	
<b>Construction Management: History of Civil Engineering</b>	
<b>A Brief History of the Pattullo Bridge</b> .....	801
Jivan Johal and F. Michael Bartlett	
<b>Britannia Mine: A Canadian Innovation with a Lasting Environmental Impact</b> .....	819
Ali A. Mahmood and F. Michael Bartlett	
<b>A Brief Historical Review of the Defunct La Colle Falls Hydropower Project Near Prince Albert, Saskatchewan</b> .....	835
Jim Kells and Van Pul Paul	
<b>Biennial Update of the Activities of the CSCE National History Committee</b> .....	851
F. Michael Bartlett	

<b>Cariboo Wagon Road—A User’s Perspective</b> .....	867
W. C. Sexsmith	
<b>Ballantyne Pier History</b> .....	877
Willy Yung	
<b>Construction Management: Miscellaneous Topics</b>	
<b>Post-fire Damage Assessment of Buildings at the Wildland Urban Interface</b> .....	893
Ahmad Abo-El-Ezz, Faten AlShaikh, Azarm Farzam, Marc-Olivier Côté, and Marie-José Nollet	
<b>Fire Building Codes in Developed and Developing Countries: A Case Study of Canada and Costa Rica</b> .....	903
Sara Guevara Arce, Hannah Carton, and John Gales	
<b>Coastal Hotels and Resorts: Infrastructure Asset Management System Model</b> .....	917
Athnasious Ghaly, Mahmoud Amin, Tesfu Tedla, Ossama Hosny, and Hatem Elbehairy	
<b>Protection of Concrete Surface from the Canadian Standard, ICRI, and ACI Perspectives</b> .....	933
Claudiane M. Ouellet-Plamondon	
<b>Environmental Specialty: Water and Wastewater Treatment</b>	
<b>Let Us Talk About Microplastic Pollution in Drinking Water Treatment</b> .....	941
Jinkai Xue, Seyed Hesam-Aldin Samaei, and Jianfei Chen	
<b>Selecting a Hybrid Treatment Technology for Upgrading a Lagoon-Based WWTP</b> .....	949
Jeremy Enarson	
<b>Treatment of Aqueous Arsenite Using Modified Biomass-Based Sorbent</b> .....	961
Khaled Zoroufchi Benis, Kerry McPhedran, and Jafar Soltan	
<b>A Framework for the Economic Assessment of a More Sustainable Wastewater Management System</b> .....	977
Bibhas B. Tanmoy and M. Abdel-Raheem	
<b>Economic Analysis of the Utilization of a Greywater System in Residential Dwellings</b> .....	993
Bibhas B. Tanmoy and M. Abdel-Raheem	

<b>Adsorption of Sulfamethoxazole by Dried Biomass of Activated Sludge Collected from Biological Nutrient Removal (BNR) Systems</b> .....	1007
S. Minaei, K. N. McPhedran, and J. Soltan	
<b>H<sub>3</sub>PO<sub>4</sub> and NaOH Treated Canola Straw Biochar for Arsenic Adsorption</b> .....	1019
Julia Norberto, Khaled Zoroufchi Benis, Jafar Soltan, and Kerry McPhedran	
<b>Evaluation of Performance of Pilot-Scale Engineered Permeable Bio-barriers for Removal of Nitrogenous Compounds from Waters Contaminated with Manure Slurry</b> .....	1033
Ali Ekhlasi Nia, Kharazm Khaledi, Bernardo Predicala, Terry Fonstad, and Mehdi Nemati	
<b>Wetland Water Discharge Remediation Using On-Site Non-woven Geotextile Filtration</b> .....	1043
Antonio C. Pereira, Dileep Palakkeel Veetil, Mathew Cotton, Catherine N. Mulligan, and Sam Bhat	
<b>Environmental Specialty: Ecohydrology and Environmental Hydraulics</b>	
<b>On Developing Extreme Rainfall Intensity–Duration–Frequency Relations for Canada: A Comparative Study of Different Estimation Methods</b> .....	1059
Van-Thanh-Van Nguyen and Truong-Huy Nguyen	
<b>Artificial Neural Networks and Extended Kalman Filter for Easy-to-Implement Runoff Estimation Models</b> .....	1071
Arash Yoosefdoost, Syeda Manjia Tahsien, S. Andrew Gadsden, William David Lubitz, and Mitra Kaviani	
<b>On Instantaneous Behaviour of Microplastic Contaminants in Turbulent Flow</b> .....	1101
Arefeh Shamskhany and Shooka Karimpour	
<b>Influence of Biochar Amendment on Runoff Retention and Vegetation Cover for Extensive Green Roofs</b> .....	1117
Jad Saade, Samantha Pelayo Cazares, Wenxi Liao, Giuliana Frizzi, Virinder Sidhu, Liat Margolis, Sean Thomas, and Jennifer Drake	
<b>A Non-stationary Stochastic Model of Extreme Rain Events in the Changing Climate</b> .....	1133
Rituraj Bhadra and Mahesh Pandey	

## **Environmental Specialty: Environmental Sustainability**

<b>The Elephant in the Room: Engaging with Communities About Climate Change Uncertainty</b> .....	1149
---	------

J. A. Daraio

<b>Prioritization of Barriers for Photovoltaic Solar Waste Management in Saskatchewan</b> .....	1171
---	------

Monasib Romel, Golam Kabir, and Kelvin Tsun Wai Ng

<b>Forecasting of Solar Installation Capacity in Canada</b> .....	1185
---	------

Monasib Romel, Golam Kabir, and Kelvin Tsun Wai Ng

<b>The Meadoway: Urban Ecosystem Restoration at a City-Level Scale Providing Enhanced Regulating and Supporting Services</b> .....	1199
--	------

Ke Qin, Marney Isaac, and Jennifer Drake

<b>Investigation of Climate Risks Within the St. Lawrence Marine Corridor Supported by Ultra-High-Resolution Climate Modelling</b> .....	1221
--	------

Bernardo Teufel, Keihan Kouroshnejad, Laxmi Sushama, Enda Murphy, and Julien Cousineau

<b>Development of Performance Index for Small and Medium-Sized Drinking Water Systems</b> .....	1235
---	------

Sarin Raj Pokhrel, Gyan Chhipi-Shrestha, Haroon Mian, Kasun Hewage, and Rehan Sadiq

<b>Improved NASM Framework for Food Processing Wash-Water and Solid Residuals</b> .....	1249
---	------

Richard G. Zytner, Connor Dunlop, and Bassim Abbassi

<b>Settling and Rising Hydrodynamics of Microplastic Pollutants: A Numerical Study</b> .....	1263
--	------

Zihe Zhao and Shooka Karimpour

<b>Effects of Amendments on Bioretention Systems: The Field and Laboratory Investigations</b> .....	1277
---	------

Yihui Zhang, Anton Skorobogatov, Jianxun He, Caterina Valeo, Angus Chu, Bert van Duin, and Leta van Duin

## **Cold-Regions Specialty**

<b>Cold Temperature Effects on Reinforced Concrete Structural Behavior</b> .....	1293
--	------

William T. Riddell, Douglas B. Cleary, Gilson R. Lomboy, Shahriar Abubakri, Benjamin E. Watts, Danielle E. Kennedy, Brian Berry, Amelia Chan, Nicholas Giagunto, Joseph Goodberlet, Maximilian Husar, Joseph Kayal, and Christopher McCormick

<b>Development of Effective and Low-Cost Water Treatment Method for First Nations and Rural Communities in British Columbia, Canada</b> .....	1307
Zawad Abedin, Jianbing Li, and Sayed Mohammad Nasiruddin	
<b>Biorefinery Paradigm in Wastewater Management: Opportunities for Resource Recovery from Aerobic Granular Sludge Systems</b> .....	1319
Oliver Terna Iorhemen and Sandra Ukaigwe	
<b>Effects of Microplastic Size on Oil Dispersion in Oceans</b> .....	1335
Min Yang, Baiyu Zhang, Hemeihui Zhao, Chushi Wang, and Bing Chen	

## About the Editors

**Dr. Rishi Gupta** is a Professor in the Department of Civil Engineering at the University of Victoria. He leads the Facility for Innovative Materials and Infrastructure Monitoring (FIMIM) at UVic. He received both a masters and a Ph.D. in Civil Engineering from the University of British Columbia. His current research is focused on studying smart self healing cement-based composites containing supplementary cementitious materials and fiber reinforcement. His areas of interest include development of sustainable construction technologies, structural health monitoring, and non-destructive evaluation of infrastructure. He has more than 20 years of combined academic and industry experience. His industry experience includes working as the Director of Research of Octaform Systems Inc in Vancouver.

Rishi is a Fellow of Engineers Canada, the Canadian Society of Senior Engineers, and a past chair of the EGBC's Burnaby/New West branch. He is the past Chair of the international affairs committee of the Canadian Society of Civil Engineering (CSCE). He is a long standing member of the American Concrete Institute and is also a voting member of several subcommittees of ASTM C 09.

**Dr. Min Sun** is an Associate Professor and the Director of Undergraduate Program in the Civil Engineering Department at the University of Victoria (UVic). His research interests lie primarily in the field of structural engineering. His past research activity has included the development of design rules for hollow structural section connections and fillet welds, which have been incorporated into Canadian and American national steel design standards. From 2018 to 2020, he was the Western Region VP of the Canadian Society for Civil Engineering (CSCE). In 2019, he received the Faculty Award for Excellence in Teaching at UVic. Before joining UVic, he worked as a structural designer at Read Jones Christoffersen.

**Dr. Svetlana Brzev** is currently an Adjunct Professor at the Department of Civil Engineering, University of British Columbia. She has more than 35 years of consulting and research experience related to structural and seismic design and retrofitting of masonry structures in Canada and several other countries. Her research has been focused on seismic behaviour and practical design and construction issues related

to masonry structures. She has served as a member of the Technical Committee responsible for developing current Canadian masonry design standard CSA S304 since 2009 and is a member of the Technical Committee 250/SC8/WG1 responsible for developing Eurocode 8 provisions for seismic design of masonry buildings. She has been actively involved in several international initiatives related to promoting safe housing in seismically active regions, such as the EERI-sponsored World Housing Encyclopedia and Confined Masonry Network. Dr. Brzev served as a Director and Vice-President of EERI, a Director of the Masonry Society, and is currently a Director of the International Association for Earthquake Engineering. Dr. Brzev has co-authored two books related to design of masonry structures and a textbook on design of reinforced concrete structures and numerous papers and publications.

**Dr. M. Shahria Alam** is a Professor of Civil Engineering and the Tier 1 Principal's Research Chair in Resilient and Green Infrastructure in the School of Engineering at The University of British Columbia (UBC)'s Okanagan campus. He is serving as the founding Director of the Green Construction Research and Training Center (GCRTC) at UBC. Dr. Alam is the Vice President (Technical Program) of the Canadian Society for Civil Engineering (CSCE) and Chair of the Engineering Mechanics and Materials Division of CSCE. He received his Ph.D. in Civil/Structural Engineering from Western University in 2008. His research interests include smart and recycled materials and their structural engineering applications. He has published more than 350 peer-reviewed articles in these areas. He is the recipient of more than forty national and international awards including three best paper awards. Currently, Dr. Alam is serving as an Associate Editor of *ASCE's Journal of Bridge Engineering* and *Journal of Materials in Civil Engineering*.

**Dr. Kelvin Tsun Wai Ng** major fields of interest are in sustainable waste management system, disposal facility design, and data-driven waste policy. Kelvin has over 120 publications, and his projects have been supported by NSERC, Mitacs, CFI, Innovation Saskatchewan, Ministry of Environment, and other national and provincial sponsors. Kelvin has received both research and teaching awards, including six RCE Education for Sustainability Recognition Awards (2018-2022), Elsevier's Top Reviewer Award—Waste Management (2021), Saskatchewan Innovation Challenge (2019), McMaster Engineering Top 150 Alumni for Canada's Sesquicentennial (2017), and the University of Regina President's Award for Teaching Excellence (2017). Provincially, he has been appointed by the Ministry of Environment in Saskatchewan to serve on the Solid Waste Management Advisory Committee. Kelvin has been serving on the Association of Professional Engineers and Geoscientists of Saskatchewan (APEGS) Award Committee since 2017. Currently, he is chairing the APEGS Awards Committee. Kelvin is the Environmental Division Chair for Canadian Society for Civil Engineering and has organized and chaired/co-chaired a



number of conferences, including the CSCE General Conference in 2015, and four CSCE Environmental Specialty Conferences in 2019, 2020, 2021, 2022. Currently, he is organizing the 2023 CSCE Environmental Specialty Conference at Moncton.

**Dr. Jianbing Li** is a professor and professional engineer in the Environmental Engineering program at the University of Northern British Columbia (UNBC). He received his Ph.D. degree in environmental systems engineering from the University of Regina. He has research interests in environmental pollution control, petroleum waste management, contaminated soil and groundwater remediation, environmental modelling and decision analysis, environmental risk assessment, and oil spill response. He has produced more than 300 publications in international journals and conferences, including over 200 refereed journal publications, with an h-index of 45 (Google Scholar). His research has been supported by various organizations, including Natural Science and Engineering Research Council of Canada (NSERC), Fisheries and Oceans Canada (DFO), Natural Resources Canada (NRCan), BC Ministry of Forests, and Geoscience BC. He obtained the 2013 Northern BC Business and Technology Award (Collaborative Research Award with Husky Energy), the 2010, 2014 and 2019 UNBC Research Excellence Award, and the 2013 UNBC Achievement Award in Professional Practice and Mentorship. Dr. Li was nominated for the 2022 and 2023 CUFA BC Ehor Boyanowsky Academic of the Year Award. He served as a member of NSERC's Research Tool and Instruments Selection Committee for Civil, Industrial and Systems Engineering in 2016–2021, and the Committee Chair for the 2021 competition. He is also a member of NSERC's Discovery Grant Evaluation Group for Civil, Industrial and Systems Engineering (EG1509) (2022–2025). He has been a member of the Academic Examiners Subcommittee with Engineers and Geoscientists BC since 2017. He has served as the co-director of the UNBC/UBC environmental engineering program for 4 years (2013–2017). He served as a guest editor for seven international journals, and is currently an Associate Editor or Editorial Board member of seven journals. He is the inaugural Chair of the Northern British Columbia Section of Canadian Society for Civil Engineering (CSCE). Dr. Li was named a CSCE Fellow in 2022.

**Dr. Ashraf El Damatty**, Professor and Chair of the Department of Civil and Environmental Engineering, Western University. He is a Fellow of the Canadian Academy of Engineering, the Engineering Institute of Canada, and the Canadian Society of Civil Engineering (CSCE). He is a Research Director at the WindEEE Research Institute and Editor-in-Chief of the *Journal of Wind and Structures*. He holds honorary Professorship titles at four international universities. He obtained B.Sc. and M.Sc. from Cairo University, Egypt, Ph.D. from McMaster University, and MBA from University College London, UK. He is the founder of the CSCE Steel Structures Committee and served for five years as the Chair of the CSCE Structures Division. He has written more than 250 publications, supervised more than 60 graduate students and has been invited as keynote speaker in 14 countries. He received several awards including the Alan Yorkdale Award by ASTM, Best Paper Award at the Canadian Conference on Effective Design, Honourable Mention in 2014 Casimir

Gzowski Medal, 2015 CSCE Whitman Wright Award, 2016 CSCE Horst Leipholz Medal, Western University Faculty Scholar Award, 2018 Professional Engineers of Ontario Medal of Research and Development, 2021 Pratley Award for Best Paper on Bridges, and the 2021 Western Engineering Award for Excellence in Research, Western University. He is an international leader in the interdisciplinary field of Wind and Structural Engineering. His research has influenced the international codes and the Engineering practice worldwide. It resulted in the first specifications in the world for downburst and tornado loading on transmission line structures that was recently incorporated into the guidelines of the ASCE.

**Clark Lim** has over three decades of experience in public, private, and academic sectors, specializing in analytical methods and information systems for transportation applications. As a consultant, he advises senior officials on policy, technology, and governance matters, where he utilizes an evidence-based and technically progressive approach to establish sound policy frameworks. In the mid-1990's, he was part of the team that established TransLink, the Greater Vancouver Transportation Authority, where he was also the Project Manager of the Evergreen Rapid Transit Line planning and consultation process. At UBC Clark is currently an Adjunct Professor in the Department of Civil Engineering where he has taught transportation engineering and planning to senior undergraduate and graduate students since 2006. His previous research at UBC focused on intelligent transportation systems for freight, and the impact quantification of the 2010 Winter Olympic Games. Currently he is researching the effects of hybrid working on transportation policies, the impacts of ride-hailing trips through big data methods, and developing tools to measure sustainability and diversity-equity-inclusion indices for corporate boards.

# **Construction Management: Construction Management**

# Framework for Assessing the Usability of Augmented Reality Applications for Construction Work



Mayank Arvindbhai Patel, Krithikashree Lakshminarayanan, Zia Din, and Lingguang Song

**Abstract** Visualizing real-world features along with digital textual data and virtual objects has made augmented reality (AR) technology attractive for commercial uses such as sales, manufacturing, and construction. AR application developers are designing new applications for a variety of purposes. New AR applications related to construction are becoming increasingly available to industry practitioners. It can be challenging for them to select a truly useful one from many existing AR applications through a trial-and-error approach that is time-consuming and costly. Therefore, a systematic evaluation procedure is desired to allow users to evaluate AR applications during their trial period without having to commit significant financial resources. This study aims to develop a framework that evaluates the usability of off-the-shelf AR applications. The authors proposed a framework for assessing the usability of construction-focused augmented reality applications based on the International Organization for Standardization 9241-11:2018 standard and other relevant works of literature. Using the proposed criteria, the authors examined the five most popular AR construction apps with the highest number of downloads and at least four-star reviews to validate the proposed framework. To validate the study, participants were recruited to evaluate applications and rank them in predefined categories, including effectiveness, efficiency, and satisfaction, using the proposed framework. Out of five applications tested, two applications, VisualLive and Gamma AR, were ranked as the first and second-best applications. The ranking is consistent with the expert review of the applications. The proposed framework helps in the selection of the most useful applications from a pool of off-the-shelf software. It enables users in the construction industry to make an informed decision about AR application selection before committing time and money to acquisition and implementation.

**Keywords** Augmented reality applications · Construction work

---

M. A. Patel · Z. Din (✉) · L. Song

Department of Construction Management, University of Houston, Houston, TX 77004, USA

e-mail: [uziauddi@central.uh.edu](mailto:uziauddi@central.uh.edu)

K. Lakshminarayanan

Department of Computer Science, University of Houston, Houston, TX 77004, USA

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_1](https://doi.org/10.1007/978-3-031-34593-7_1)

# 1 Introduction

Historically, the construction industry has been reluctant to adopt new technology [1–3]. One of the reasons for not adopting new technology quickly could be the inability to decide which product is most suitable. Several emerging technologies, such as three-dimensional (3D) modeling tools [4], augmented reality tools [5], and cloud-based real-time progress tracking and analysis systems [6], are available to assist the construction industry with various challenges ranging from safety to productivity. Given the fast pace in technology development and a growing number of competing product offerings, it can be difficult for construction professionals to decide which technologies and products meet their needs. When construction industry professionals are persuaded to use specific technology for an intended purpose, the next logical question is which product or application to choose to best deliver the desired outcomes. For example, if a company is convinced about advanced visualization technology, such as augmented reality, it must decide which competing products best meet its unique requirements. Many AR tools offer a range of visualization capabilities, from simple static 3D model presentations to advanced functions such as interactive 3D model visualization that allows filtering of model components, measuring of the physical environment, and access to plans and specifications. It can be mentally draining for construction professionals to select the best application for their unique needs from a potentially large number of commercially available products.

Easy-to-use guidelines can help construction industry professionals select the most appropriate application tools to keep competitive in the market. Such guidelines are especially beneficial to small contractors who lack the necessary resources to implement new technologies.

AR technology is used in almost all industries, including medical, industrial, military, retail, and marketing [7, 8]. AR's capabilities include combining computer-generated details in the physical world, interactively in real-time, and displaying virtual elements that are naturally aligned with real-world perspectives [9]. This technology uses mobile phones, tablets, head-mounted displays, or computers to improve human sense capabilities. AR applications such as Augment, Dalux Viewer, Kuity Go, and others are currently available on the market. These AR applications primarily focus on augmenting visualization to provide additional information interactively.

AR-based applications in the construction sector have tremendous opportunities [10] to overcome information communication challenges posed by the industry's complex environment, where accurate and timely information availability remains a challenge [11]. A typical application of AR is to improve the safety and productivity of construction workers by training and providing information in an immersive environment [12, 13]. AR software developers continue to add new capabilities to improve user experience to better perform a specific function for which the tool is created, such as construction or design visualization. However, as with any new technology in the evolution process, many AR applications do not have all the features, such as a virtual measurement feature. The lack of such a fundamental function can

limit the utility of the AR application in the real world, where construction workers must constantly take simple measurements. Thus, selecting software during the trial period of the application by conducting a systematic evaluation of its function is critical to avoiding time and money wasted on purchasing and implementing new software and hardware.

This paper focuses on developing a framework for evaluating augmented reality (AR) applications developed for handheld devices such as mobile phones and tablets in construction. The research objectives are: (1) identify AR applications with a four-star or higher rating in construction areas; (2) develop assessment criteria; (3) evaluate the identified applications; and (4) based on the evaluation, draw conclusions.

## 2 Literature Review

Augmented reality is a technology that overlays information such as two-dimensional images, three-dimensional models, video, and audio on top of the real environment and allows for real-time interaction [14]. AR is different from virtual reality technology, which provides a computer-generated, fully digital environment similar to a video game for visualization [15].

### 2.1 *Augmented Reality in the Construction Industry*

Various researchers have studied the usage of AR in construction work, such as for design reviews [16], training [17], and construction-related assemblies [18]. AR is frequently highlighted for its ability to improve construction crew communication and BIM integration. For example, users can interact with 3D elements embedded in the actual environment using the Trimble application and Microsoft's AR hardware, which combines HoloLens with 3D modeling tools [19]. Many AR-based solutions improve construction work that can be used with mobile devices, glasses, or HMD technology [20].

Although it is good news that new AR tools are emerging, their adoption in the construction industry is minimal [15]. This explosion of information can cause construction industry professionals to remain undecided. Establishing a framework for application evaluation that will allow users to subjectively assess the characteristics of the product, including effectiveness, efficiency, and user satisfaction, is critical. Although the evaluation cannot be definitive, it can be used to determine the appropriate use of the application from the user's perspective.

## **2.2 *User Experience (UX)***

User experience (UX) aims to increase user satisfaction with a computerized system, applications, and websites [21, 22]. The UX fundamentals are generic and can be applied to any physical or digital interaction with any object design, even a simple coffee cup. The UX design process includes product definition, research analysis, design, and testing. This method works for any design category, digital or not [23]. Under UX, the usability of a product or application comes into play for testing purposes.

Usability is a set of criteria that may be used to assess the feasibility of a product. Usability is not limited to ease of use; instead, it is a usability component; ease of use does not always imply high usability [24].

## **2.3 *Measuring Usability of AR Applications***

The International Organization for Standardization (ISO) has defined ergonomic standards for human–system interaction in standard 9241-11:2018 [25]. This standard focuses on designing and evaluating systems, products, and services for usability. Usability is measured as the degree to which intended users can use an application to perform functions or tasks effectively, efficiently, and satisfactorily for the particular context for which it is designed [25, 26]. Effectiveness is the precision and completeness of various functions in the application that enable users to achieve their goals. Efficiency is the consumption of resources such as time, human effort, cost, and material to achieve an intended goal. The last part of the usability evaluation measures how satisfied users are with how the application makes them feel physically, cognitively, and emotionally when using it.

## **2.4 *The Benefit of Application Evaluation***

There are various levels of evaluation. Evaluation can be performed during application development. It can be used to guide application designers and developers in improving the application before publishing it. However, the evaluation framework has a different purpose for the end-users. End-users can choose an application from a pool of standard applications.

The ISO 9241-11 standard specifies the goals of the program and the user environment to evaluate the usability of an application. The evaluation framework comprises three key components: (1) the objective of the application, (2) the context of usage, that is, design or construction, and (3) usability measurements. The authors divided the usability evaluation of AR applications into three areas: effectiveness, efficiency, and satisfaction.

### 3 Methodology

Researchers proposed tools to assess ISO-9241-11 compliance of various products, such as websites and software [27–29]. However, they did not discuss the specifics of AR applications for construction-related use. As a result, the authors used ISO-9241-11 to develop a new evaluation framework to rate AR applications for the construction industry. The authors adapted the published evaluation frameworks to evaluate AR applications in different industries to evaluate AR applications in the construction sector. There are a large number of AR applications available. The following criteria were used to choose the top five AR applications for further consideration: (1) the aim of the application, whether it is focused on construction in general or on a specific area of construction, such as design visualization or construction information delivery, (2) a four-star rating or higher on an online store such as the Google Play Store or the Apple App Store, as well as qualitative analysis of user comments; (3) the number of applications downloads. Table 1 shows a list of all applications considered for this study. The authors selected the top five to demonstrate the proposed evaluation method.

The authors studied publicly available applications on application store platforms such as the Google Play Store and the Apple App Store. After installing and testing all the applications listed in Table 1, it was found that Morpholio AR Skechwalk, Augment, Dalux view, and Kubity Go focus more on the visualization of architectural design. The authors were unable to evaluate five applications out of 15 in total for various reasons. An AR application, Trimble Connect AR, is not available for the trial version; therefore, it was not evaluated. Two applications, Simblab AR/VR and Fuzor, are VR rather than AR applications. The other two applications, XYZ reality and Argyle Build, are unavailable on the Google Play Store or App Store, and both are for HMD devices. Therefore, the top five applications that met the pre-established criteria are Arki, Fologram, Gamma AR, Visual Live, and Augin (Fig. 1).

#### 3.1 *Development of AR Application Evaluation Framework for Construction*

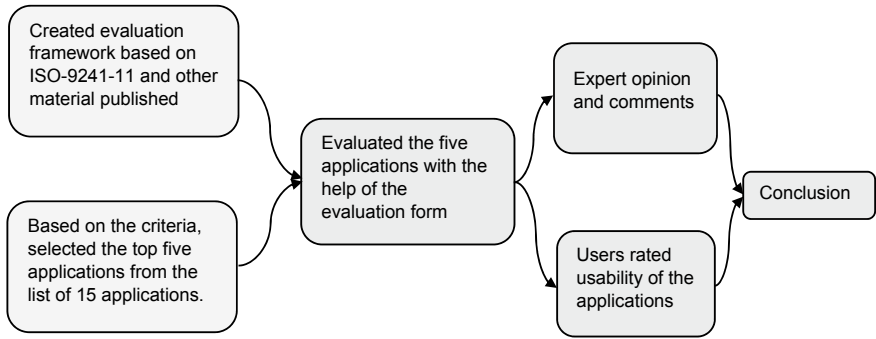
The authors conducted literature research on the usability requirements of AR applications in the construction sector. Typical construction-specific applications should include the following features: basic interaction options such as rotation, scale, and move to interact with the 3D model; task-specific 2D drawings; video simulations for understanding construction steps; options to interact with the provided information such as taking a screenshot, doodling options, making notes, downloading notes as pdf files, and so on.

For instance, the Quick Response code option can be used to fulfill the goal of efficiency in augmented reality applications. Quick Response (QR) codes enable quick access to digital models and supporting data, making access to information quick



**Table 1** List of augmented reality applications related to the construction industry

No.	Application	Rating (out of five)	Number of downloads	References
1.	Morpholio AR Sketchwalk	4.7	Not available	Morpholio LLC [30]
2.	Augment	3.8	10,00,000+	Augment [31]
3.	Fologram	4.4	1000+	Fologram [32]
4.	Gamma AR	3.8	1000+	GAMMA Technologies [33]
5.	ARki	4.2	Not available	Darf Design LTD [34]
6.	Dalux View	4.6	50,000+	Dalux aps [35]
7.	Trimble connects AR	3.8	500+	Trimble Inc. [36]
8.	Simlab AR/VR view	3.3	10,000+	Simulation Lab Software [37]
9.	Fuzor	4.5	1000+	Kallos Studios [38]
10.	VisualLive	Not available	1000+	Visual Live 3D LLC [39]
11.	Argyle Build	Not available	Not available	Argyle Inc. [40]
12.	XYZ Reality	Not available	Not available	XYZ Reality [41]
13.	Augmentecture	3.4	10,000+	Augmentecture Inc. [42]
14.	Kubity Go	3.9	100,000+	Kubity [43]
15.	Augin	4.1	100,000+	Augin Softwares Ltda. [44]



**Fig. 1** Steps for AR applications’ evaluation

and straightforward. A QR code can provide construction job-specific information to workers, increasing efficiency and reducing overall waste with minimal effort [45]. The effectiveness factor of an application can be added by including features that enable the intended work to be performed. When an augmented reality application superimposes a virtual model on top of the real environment, the model must integrate seamlessly with the physical world with respect to scale, texture, color, and location. The application should enable users to do all the tasks they want to accomplish with its help. Finally, their satisfaction is high when users can do their jobs swiftly with the application.

The authors adopted statements from seven different sources, including ISO-9241-11, and tested the statements with commercially available applications. The authors modified the statements regarding the need of construction industry users and concluded the statements considered to measure the system's effectiveness, efficiency, and satisfaction.

The authors included information on AR application assessments from seven sources (sources listed in Table 2), including ISO-9241-11, and tested the proposed evaluation form using commercially available AR applications. The evaluation framework proposed in this research, as shown in Table 1, is pragmatic for professionals in the construction industry. The framework contains statements that address the ISO-9241-11 criteria and can be used to evaluate AR applications in the construction industry. The authors identified and prioritized the usability challenges that needed to be resolved. The evaluation framework may be used on a completed application or on one still under development. Essential features are evaluated on five levels, allowing for the weakness of various features to be identified. The authors proposed the following scale: 1 = strongly disagree; 2 = disagree; 3 = no opinion; 4 = agree; 5 = strongly agree. In total, three sets of statements were prepared. Each set of statements is placed in a separate section of the framework. Table 2 contains statements about effectiveness, efficiency, and user satisfaction. This assessment framework helps users to determine the suitability of an application.

## 4 AR Application Evaluation

The authors created a framework to help select AR applications designed for the construction industry. Without thoughtful application evaluation, technology and product adoption can be highly risky. Therefore, a formal evaluation is intended to ensure consistency in evaluating applications by different evaluators, eliminating errors, and ensuring that the evaluation's main factors are not overlooked. The evaluation framework was derived from a review of seven sources, including ISO-9241-11, which the researchers adapted and modified to evaluate AR applications in the construction sector. The researchers created, tested, and used the evaluation framework on the top five AR applications to improve the accuracy of the usability measure.

**Table 2** Rating the effectiveness, efficiency, and user satisfaction characteristics of construction-related applications

No.	Sources	Statement
<i>Rating the effectiveness characteristics of construction-related AR applications</i>		
1.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47], Johari [48], Dünser and Billinghamurst [49] and Sutcliffe and Gault [50]	The application achieves the goal
2.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47] and Johari [48]	The application is mentally demanding
3.	De Paiva Guimarães and Martins [46], Sutcliffe and Gault [50] and Johari [48]	The application provides natural engagement of end-users
4.	De Paiva Guimarães and Martins [46]	The commands “redo” and “undo” are simple to use (i.e., go back to an earlier state without the virtual object)
<i>Rating the efficiency characteristics of construction-related AR applications</i>		
1.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47], Dünser and Billinghamurst [49], Sutcliffe and Gault [50] and Parente da Costa and Dias Canedo [51]	The virtual object in the scene loads quickly
2.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47], Dünser and Billinghamurst [49] and Sutcliffe and Gault [50]	Scanning a marker (QR code) with a smartphone or tablet is simple
3.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47], Sutcliffe and Gault [50] and Parente da Costa and Dias Canedo [51]	Are virtual objects adequately blended with the physical realm? <i>Explanation: Virtual objects can be scaled, have realistic texture and color, and be placed in the required position</i>
4.	De Paiva Guimarães and Martins [46] and Hensely-Schinking et al. [47]	The learning curve is high for unskilled users
5.	De Paiva Guimarães and Martins [46] and Hensely-Schinking et al. [47]	Expert users will be able to use the program effectively <i>Explanation: Is it possible for them to skip the descriptions of commands?</i>
6.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47], Sutcliffe and Gault [50] and Parente da Costa and Dias Canedo [51]	During interactions, the user is instructed on what to do <i>Explanation: The user is instructed to perform steps with textual/audio guidance</i>

(continued)

**Table 2** (continued)

No.	Sources	Statement
7.	De Paiva Guimarães and Martins [46] and Hensely-Schinking et al. [47]	There are specific requirements other than hardware <i>Explanation: Print a Quick Response (QR) code to access relevant information or install a program to access additional information</i>
8.	Dünser and Billingham [49], Sutcliffe and Gault [50] and Parente da Costa and Dias Canedo [51]	There are clear entry and exit points
9.	Dünser and Billingham [49] and Parente da Costa and Dias Canedo [51]	The application crashed quite often
10.	Sutcliffe and Gault [50]	When the user turns around, signals or directions are provided
<i>Rating the satisfaction characteristics of construction-related AR applications</i>		
1.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47] and Johari [48]	Users are satisfied with the interaction solution. <i>Explanation: The app allows the user to move freely during the interaction</i>
2.	De Paiva Guimarães and Martins [46], Hensely-Schinking et al. [47] and Parente da Costa and Dias Canedo [51]	The number of interaction options is satisfactory
3.	Johari [48] and Sutcliffe and Gault [50]	Feedback from all users is positive

Four researchers and a professor (with more than six years of experience in AR/VR). All users have backgrounds in construction management and computer science. These users evaluated the application from different perspectives. Each application was tested for a maximum of 10 min by the evaluators. The authors followed Nielsen's five-user testing approach [52]. According to Nielsen, to test usability, only five users are sufficient, as they can identify 75% or more of the problem in the application. The proposed evaluation framework is designed to achieve a quick and cost-effective application evaluation process. The authors evaluated five applications, which are (1) Arki, (2) Fologram, (3) Augin, (4) Visual Live, and (5) Gamma AR, as discussed previously. The evaluation used trial versions of these applications and the demo models provided by the applications.

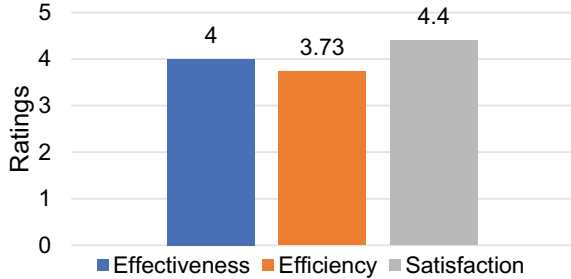
Table 2 presents the usability measures evaluated using a Likert scale and expert comments. The authors used the following scales: 1 = strongly disagree; 2 = disagree; 3 = no opinion; 4 = agree; 5 = strongly agree. The authors attempted to simplify the rating scale to be used by all users, thereby improving the evaluation framework's speed and ease of use.

## 4.1 Data Analysis

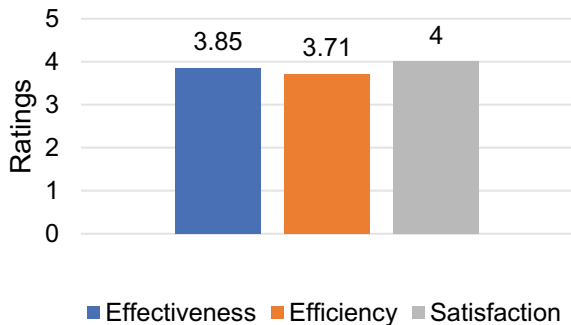
After inspecting five different applications with the four researchers, the authors aggregated the ratings for each statement. Additionally, the expert tested all five applications, provided comments, and evaluated the final ratings. In addition, rating means were used as the final rating of an application. This evaluation framework was perceived as efficient, easy to use, and relevant to the construction industry. Any user can utilize it to evaluate an application related to construction. Figures 2, 3, 4, 5 and 6 provide the overall ratings for the applications with respect to three characteristics based on the final evaluation framework. The final scores for all applications are average scores of effectiveness, efficiency, and satisfaction as follows: (1) Visual Live = 4.04, (2) Gamma AR = 3.85, (3) Akri = 3.67, (4) Augin = 3.59, (5) Fologram = 3.28.

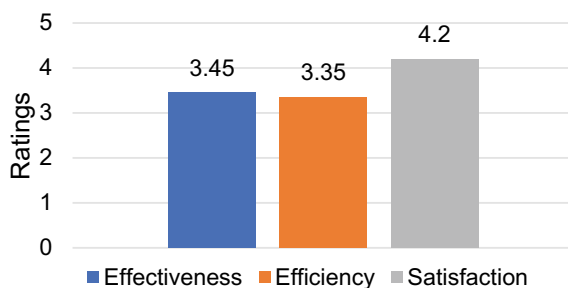
The evaluation focuses on selecting an application that can meet the industry's needs. The authors evaluated applications with three critical measures: effectiveness, efficiency, and level of satisfaction. Two AR applications, VisualLive and Gamma AR, were rated 4.04 and 3.85. Both advanced-level applications offer functionalities that can be very helpful in the construction process. For example, they require aligning a 3D model with the physical world before starting to interact with the model. This allows the user to visualize the task on a 1:1 scale to execute the task using a handheld device. These applications provide additional features such as presenting information

**Fig. 2** Rating of VisualLive  
(score = 4.04)

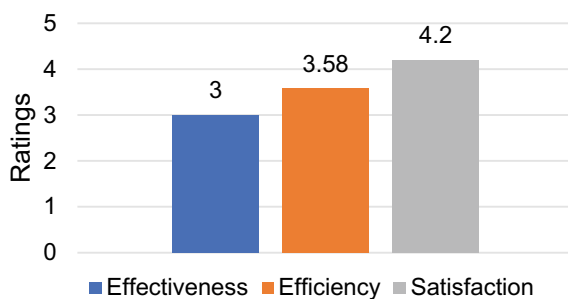


**Fig. 3** Rating of Gamma  
AR (score = 3.85)

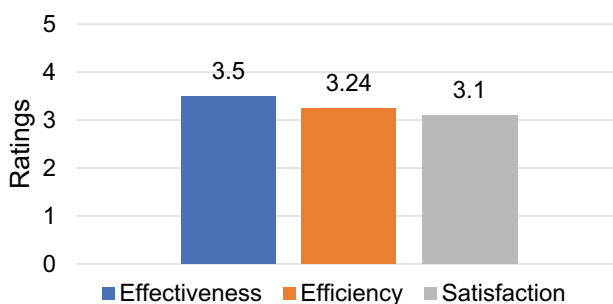




**Fig. 4** Rating of Arki (score = 3.67)



**Fig. 5** Rating of Augin (score = 3.59)



**Fig. 6** Rating of Fologram (score = 3.28)

in layers with the ability to toggle them on and off, transparency of the virtual model so that the underlying physical world can be seen by the user, virtual measurements, anchors to hold the virtual model in place, issue reporting, status tracking, to name a few. Gamma AR and VisualLive have a more significant learning curve than other applications due to the features and complex user interface to perform specific tasks. However, it will be much more effective for trained personnel to use.

Arki provides an excellent usability rating of 3.67 points out of 5 with minimalistic visualization features. However, it lacks effectiveness and efficiency in terms of using construction-specific tasks. For instance, Arki does not provide the features needed to execute construction assembly. Fologram and Augin, with total scores of 3.28 and 3.59, allow construction tasks to be performed by using the application to some extent. However, Fologram has lower ratings in efficiency and satisfaction, mainly due to fewer interaction features available to the user. Augin lacks in providing material textures and offers fewer interaction options. Therefore, it affected the effectiveness and efficiency rating.

## ***4.2 Limitations of the Framework***

This research focused on developing an evaluation framework for construction-focused AR applications. The evaluation framework is easy to implement, faster, and relevant for the construction sector. There may be some limitations, such as fewer users who tested the applications and fewer applications tested. Some applications were unavailable at the time of this study through the Google Play Store or the Apple App Store, so they could not be evaluated. Other applications did not offer trial periods, so they were not included in the study. Most of the applications in the trial period offer limited functionality, so not all functions can be tested. This probably reflects the real-world situation where potential users must test multiple applications using their trial version to select the most appropriate ones.

## **5 Conclusion and Suggestions**

This study focuses on developing an evaluation framework for measuring construction AR applications' effectiveness, efficiency, and user satisfaction. The usability evaluation framework used ISO-9241-11 as a basis and was further refined based on seven additional studies focused on human-technology interfaces and AR/VR technology interactions. However, these studies do not include specific usability requirements for AR applications in the construction industry. As a result, the authors developed the entire framework and evaluation framework to assess the user experience of AR applications for the construction industry. It simplifies the determination of the usefulness of a construction AR application. The framework primarily identifies AR applications' advantages and weaknesses using rating statements. Users can rate an AR application's effectiveness, efficiency, and user satisfaction using the proposed evaluation framework. The authors adapted the statements from different sources for implementation in the construction sector. Five different AR applications focused on construction were used to test the framework. Two AR applications, VisualLive and Gamma AR, were rated highest by users, and two applications, Augin and Fologram, were ranked at the fourth and fifth levels. The top applications perform

better in effectiveness and user satisfaction, with some efficiency issues most likely due to their complex nature. A comprehensive user interface is expected to lead to learning challenges to meet various information requirements. These include aligning the virtual model and the physical world where actual construction will occur and filtering through different model layers to obtain the required information. However, these applications are at the top due to their overall value relative to the other three applications, which were efficient to operate. Still, their level of effectiveness was lower hence lower satisfaction.

**Acknowledgements** This study was supported by the McElhattan Foundation. Its contents are solely those of the authors and do not necessarily represent the official opinions and views of the McElhattan Foundation.

## References

1. Agarwal R, Chandrasekaran S, Sridhar M (2016) Imagining construction's digital future. McKinsey & Company
2. Hossain MA, Nadeem A (2019) Towards digitizing the construction industry: state of the art of construction 4.0. *Proc ISEC*
3. Xu H, Feng J, Li S (2014) Users-orientated evaluation of building information model in the Chinese construction industry. *Autom Constr* 39:32–46
4. Trimble (2022) 3D modeling on the Web-SketchUp
5. Trimble (2021a) Trimble connects AR
6. Autodesk (2022) Construction management software—autodesk construction cloud
7. Klein A, De Assis GA (2013) A markerless augmented reality tracking for enhancing the user interaction during virtual rehabilitation. In: *Proceedings—2013 15th symposium on virtual and augmented reality, SVR 2013*, pp 117–124
8. Valjus V, Järvinen S, Peltola J (2012) Web-based augmented reality video streaming for marketing. In: *Proceedings of the 2012 IEEE international conference on multimedia and expo workshops, ICMEW 2012*, pp 331–336
9. Nincarean D, Alia MB, Halim NDA, Rahman MHA (2013) Mobile augmented reality: the potential for education. *Proc Soc Behav Sci* 103:657–664
10. Wang P, Wu P, Wang J, Chi HL, Wang X (2018) A critical review of the use of virtual reality in construction engineering education and training. *Int J Environ Res Publ Health* 15(6)
11. Gamil Y, Rahman IA (2017) Identification of causes and effects of poor communication in construction industry: a theoretical review. *Emerg Sci J* 1(4):239–247
12. Li X, Yi W, Chi H-L, Wang X, Chan APC (2018) A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Autom Constr* 86:150–162
13. Linares Garcia DA, Dongre P, Roofigari-Esfahan N, Bowman DA (2021) The mobile office: a mobile AR systems for productivity applications in industrial environments. In: *International conference on human-computer interaction*. Springer, pp 511–532
14. Azuma RT (1997) A survey of augmented reality. *Presence Teleop Virt Environ* 6(4):355–385
15. Davila Delgado JM, Oyedele L, Beach T, Demian P (2020) Augmented and virtual reality in construction: drivers and limitations for industry adoption. *J Constr Eng Manag* 146(7):04020079
16. Maftai L, Harty C (2015) Designing in caves: using immersive visualisations in design practice, pp 53–75
17. Bosché F, Abdel-Wahab M, Carozza L (2016) Towards a mixed reality system for construction trade training. *J Comput Civ Eng* 30(2):04015016



18. Wang X, Truijens M, Hou I, Wang Y, Zhou Y (2014) Integrating augmented reality with building information modeling: onsite construction process controlling for liquefied natural gas industry. *Autom Constr* 96–105
19. Trimble Buildings (2015) Trimble and Microsoft HoloLens: the next generation of AEC-O Technology. *Int J Wildland Fire*
20. Jones K (2014) Five ways the construction industry will benefit from augmented reality
21. Hassenzahl M, Tractinsky N (2006) User experience—a research agenda. *Behav Inf Technol* 25(2):91–97
22. Kuusinen K, Mikkonen T, Pakarinen S (2012) Agile user experience development in a large software organization: good expertise but limited impact. In: *International conference on human-centred software engineering*. Springer, pp 94–111
23. Hillmann C (2021) UX for XR: user experience design and strategies for immersive technologies. *Design thinking*. Apress
24. Hassenzahl M, Diefenbach S, Göritz A (2010) Needs, affect, and interactive products—facets of user experience. *Interact Comput* 22(5):353–362
25. ISO (2018) ISO 9241-11:2018(en) Ergonomics of human-system interaction—part 11: usability: definitions and concepts
26. Weichbroth P (2020) Usability of mobile applications: a systematic literature study. *IEEE Access* 8:55563–55577
27. Arthana IKR, Pradnyana IMA, Dantes GR (2019) Usability testing on website wadaya based on ISO 9241-11. *J Phys Conf Ser* 1165:12012
28. Green D, Pearson JM (2006) Development of a web site usability instrument based on ISO 9241-11. *J Comput Inf Syst* 47(1):66–72
29. Stigall J, Sharma S (2017) Virtual reality instructional modules for introductory programming courses. Institute of Electrical and Electronics Engineers Inc., Bowie State University, United States, pp 34–42
30. Morpholio LLC (2020) Morpholio AR Sketchwalk. <https://apps.apple.com/us/app/morpholio-trace-sketch-cad/id547274918>
31. Augment (2021) Augment—3D augmented reality. [https://play.google.com/store/apps/details?id=com.ar.augment&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.ar.augment&hl=en_US&gl=US)
32. Fologram (2020) Fologram. [https://play.google.com/store/apps/details?id=com.fologram.android&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.fologram.android&hl=en_US&gl=US)
33. GAMMA Technologies (2021) GAMMA AR. [https://play.google.com/store/apps/details?id=gamma.ar.com&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=gamma.ar.com&hl=en_US&gl=US)
34. Darf Design LTD (2019) ARki. <https://apps.apple.com/us/app/arki/id700695106>
35. Dalux View (n.d.) [https://play.google.com/store/apps/details?id=com.dalux.freebim&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.dalux.freebim&hl=en_US&gl=US)
36. Trimble Inc. (2021b) Trimble connects AR. [https://play.google.com/store/apps/details?id=com.trimble.connectAR&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.trimble.connectAR&hl=en_US&gl=US)
37. Simulation Lab Software (2022) Simlab AR/VR view. [https://play.google.com/store/apps/details?id=com.Simlab.SimlabViewer&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.Simlab.SimlabViewer&hl=en_US&gl=US)
38. Kalloc Studios (2021) Fuzor. [https://play.google.com/store/apps/details?id=com.kalloc.fuzor&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.kalloc.fuzor&hl=en_US&gl=US)
39. Visual Live 3D LLC (2022) VisualLive. [https://play.google.com/store/apps/details?id=com.VisualLive.MobiLive&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.VisualLive.MobiLive&hl=en_US&gl=US)
40. Argyle Inc. (2021) Argyle. <https://www.microsoft.com/en-us/p/argyle-build/9pp8zb4zw5wb#activetab=pivot:overviewtab>
41. XYZ Reality (n.d.) XYZ reality. <https://www.xyzreality.com/>
42. Augmentecture Inc. (2022) Augmentecture. [https://play.google.com/store/apps/details?id=com.production.augmentecture\\_new&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.production.augmentecture_new&hl=en_US&gl=US)
43. Kubity (2021) Kubity Go. Google play store. [https://play.google.com/store/apps/details?id=com.kubity.player&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.kubity.player&hl=en_US&gl=US)
44. Augin Softwares Ltda. (2022) Augin. [https://play.google.com/store/apps/details?id=com.VisualJoy.AugePauluzzi&hl=en\\_US&gl=US](https://play.google.com/store/apps/details?id=com.VisualJoy.AugePauluzzi&hl=en_US&gl=US)

45. Hammad A, Khabeer B, Mozaffari E, Devarakonda P, Bauchkar P (2005) Augmented reality interaction model for mobile infrastructure management systems. In: Proceedings, annual conference—Canadian society for civil engineering, pp 1–10
46. De Paiva Guimarães M, Martins VF (2014) A checklist to evaluate augmented reality applications. In: Proceedings—2014 16th symposium on virtual and augmented reality, SVR 2014, Feb 2016, pp 45–52
47. Hensely-Schinking S, de Carvalho AFP, Glanznig M, Tellioglu H (2015) The definition and use of personas in the design of technologies for informal caregivers. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)
48. Johari MFM, Majid NAA, Yunus F, Arshad H (2016) Development framework based on MAR4 preliteracy kit 10(8):2905–2908
49. Dünser A, Billinghamurst M (2011) Handbook of augmented reality. In: Handbook of augmented reality
50. Sutcliffe A, Gault B (2004) Heuristic evaluation of virtual reality applications. *Interact Comput* 16(4):831–849
51. Parente da Costa R, Dias Canedo E (2019) A set of usability heuristics for mobile applications. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 11566. LNCS, pp 180–193
52. Nielsen J (2000) Why you only need to test with 5 users

# Current and Future Trends of Augmented and Mixed Reality Technologies in Construction



Mahsa Rezvani, Zhen Lei, Jeff Rankin, and Lloyd Waugh

**Abstract** Augmented reality (AR) and mixed reality (MR) technologies have gained significant interest throughout the past two decades in the Architecture, Engineering, and Construction (AEC) Industry. However, despite the rapid growth of these technologies, their effective implementation in the AEC industry is still in its infancy. Therefore, a comprehensive investigation of the state-of-the-art applications and categories of AR/MR in the construction industry can guide researchers and industry experts to choose the most suitable AR/MR solution for research and implementation. This paper provides a comprehensive overview of 103 AR/MR articles published in credible journals in the field of the AEC industry within the years 2013–2021. Typically, review-type papers assess articles primarily based on their application areas. However, this classification approach overlooks some other critical dimensions, such as the article's technology type, the maturity level of technology used in the research, and the project phase in which technology is implemented. Accordingly, this paper classifies articles based on ten dimensions and their relevant categories: research methodology, improvement focus, industry sector, target audience, project phase, stage of technology maturity, application area, comparison role, technology type, and location. The results reveal that AR/MR literature has increasingly focused on simulation/visualization applications during construction and maintenance/operation phases of the project, emphasizing improving the performance of workers/technicians. Additionally, the increasing trend in AR/MR articles was identified as using self-contained headsets (e.g., Microsoft HoloLens). Markerless tracking systems show a significant trend among the articles. Moreover, the target location of implementing AR/MR primarily found to be in on-site and in outdoor spaces. The trend indicates an increase in immersive and mobile AR/MR applications in outdoor job sites such as construction sites to aid workers/technicians in assembly works during the construction phase.

**Keywords** Augmented and mixed technologies · Construction

---

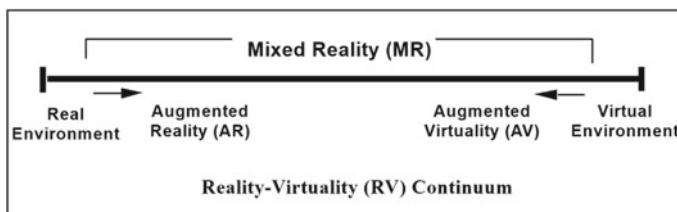
M. Rezvani (✉) · Z. Lei · J. Rankin · L. Waugh  
University of New Brunswick, Fredericton, Canada  
e-mail: [mahsa.rezvani@unb.ca](mailto:mahsa.rezvani@unb.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_2](https://doi.org/10.1007/978-3-031-34593-7_2)

# 1 Introduction

Extended reality (XR) technologies such as augmented reality (AR), augmented virtuality (AV), virtual reality (VR), and mixed reality (MR) have influenced many industries. The Architecture, Engineering, and Construction (AEC) industry has no exception [18]. A clear definition of MR, AR, and AV is provided by [16], where AR refers to augmenting the real background with virtual contents (e.g., text, images, videos, and virtual objects). In this definition, AV, as opposed to AR, uses computer graphic images or videos as the background behind the real-world elements. However, MR is considered as any environment that consists of a blending of real and virtual objects. As shown in Fig. 1, AR and AV are both subcategories of MR and are mutually exclusive. For VR lying at the right end of the RV continuum, the environment is considered one in which the operator is totally immersed in a completely synthetic world [3]. In recent years, due to equipment updates and mature technology, the use of AR has vastly increased [24]. However, the number of fields that have adopted AV is much less compared with AR, indicating that AR is more popular than AV [4]. Aside from that, it usually is ambiguous to distinguish AR from MR, as the primary function for both is to superimpose virtual information onto the real world. In this paper, articles focusing primarily on AR and MR are investigated. However, articles with a significant focus on a mere virtual environment, i.e., VR, are excluded.

AR and MR development history refers to the first see-through head-mounted AR display developed by Ivan Sutherland at Harvard [25]. More recently, Google introduced the Glass [9] to the market with AR function. Microsoft released a head-mounted device named HoloLens [14] that connects users with remote colleagues in real-time and provides hand tracking tools in addition to its head-mounted display. In recent years, researchers have also investigated the use of AR on various mobile computing interfaces, including smartphones [10], laptops [22], and tablets [21]. This paper focuses on the AEC industry since its practical implementation in this industry is still in its infancy despite the rapid growth of the technologies. It aims to provide a comprehensive investigation of the state-of-the-art applications and categories of AR and MR technologies in the construction industry that is required to guide researchers and industry experts to choose the most suitable AR/MR solution for research and implementation.



**Fig. 1** Milgram's reality–virtuality continuum. Adapted from [16]



**Fig. 2** Research methodology

Several comprehensive literature review publications have been done throughout the recent years. They classified the articles based on various categories and applications in the AEC, AECO (Architecture, Engineering, Construction, and Operation), and AEC/FM (Facility Management) industries. Chi et al. [5] focused on four technologies—localization, natural user interface (NUI), cloud computing, and mobile devices in a literature review of 101 research efforts. The author outlined the research trends and opportunities for applying AR in AEC/FM. Rankohi and Waugh [20] reviewed and categorized a group of 133 research articles within eight well-known journals in the (AEC/FM) industry until the end of 2012. More recently, Cheng et al. [4] reviewed academic journals within the domain of the AECO industry and classified a group of 87 journal papers by the end of 2018 into four application categories: architecture and engineering, construction, operation, and multiple stages. All in all, there is still a need to review the articles based on their application areas and a group of comprehensive dimensions such as the technology type, construction phase, and the improvement focus in the AEC industry. Compared to the review articles that are mentioned above, Rankohi and Waugh [20] included a more comprehensive system for classifying the articles (i.e., improvement focus, industry sector, target audience, project phase, stage of technology maturity, application area, comparison role, technology, and location). Therefore, this paper majorly follows the proposed system in that article, reviews the articles from 2013 to the end of 2021, and extends the proposed approach by adding some categories to the existing categories within the “Technology Type” and “Location” dimensions. The following sections describe the dimensions and the categories in the reference system and the extended one.

The general structure of this paper is described below. First, the criteria for selecting the journals and articles are expressed. The following section demonstrates the distribution of the articles by year and by journal. Then, we provide an overview of the selected categories to classify the AR and MR articles and introduce the extended categories used for the classification practice. Finally, we discuss the results and elaborate on the current and future trends of AR and MR technologies in the AEC industry. Figure 2 illustrates the methodology of this review paper.

## 2 Selection of Journals and Articles

This paper reviews the articles from well-known academic journals from 2013 to 2021 by extending the literature reviews carried out to the end of 2012 by [20] to reflect the current development and the future trend of AR and MR technologies

**Table 1** Selected journals for this review paper

	Journal title	Title abbreviations
1	Automation in Construction	AIC
2	Journal of Computing in Civil Engineering	CCE
3	Advanced Engineering Informatics	AEI
4	Journal of Construction Engineering and Management	CEM
5	Journal of Information Technology in Construction	ITCon
6	Journal of Architectural Engineering	AE
7	Construction Robotics	CR
8	Engineering, Construction and Architectural Management	ECAM
9	Construction Innovation	CI
10	Professional Issues in Engineering Education and Practice <sup>a</sup>	CEE

<sup>a</sup> Effective January 1, 2020, the title changed to Journal of Civil Engineering Education

in construction. Therefore, the methodology is similar to the review article by [20]. The selected journals are included in the Science Citation Index Expanded (SCIE) database and are determined using SCOPUS and Google Scholar search engines. The articles were selected in two phases. In phase I, a total of 118 articles (from 2013 to 2021) were found in these journals using a combination of key phrases, including “augmented reality”, “mixed reality”, “construction”, and “AEC industry”, all of which separated with the “OR” Boolean operator. Then in phase II, the authors read the abstract of each article to ensure that the primary focus is within the AEC industry. The articles were limited to only research articles, literature reviews, and case studies within each journal, which finalized this phase with 103 articles. Table 1 shows the selected journals for this purpose.

### 3 Review and Identification of the Article Characteristics

The distribution of articles by journal and by year is presented in Fig. 3. AIC with 46 articles (45% of the articles) focused more on AR and MR articles compared to other journals such as CR and CEE, covering three articles (or 3%) and two articles (or 2% of the articles), respectively. The maximum number of articles published in a single year is 25 in 2021. Moreover, 17 articles published in 2013 in AIC include a group of 13 articles published in a special issue (August 2013) of the AIC journal with a focus on AR technology. It is entitled “Augmented Reality in Architecture, Engineering, and Construction” [11].

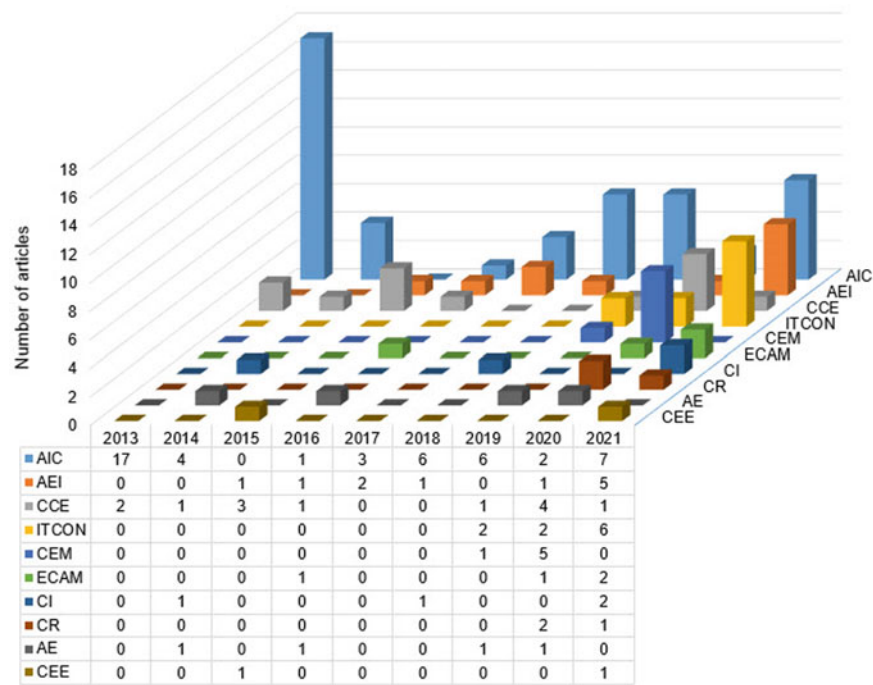


Fig. 3 Distribution of articles by journal and year of publication

4 Identifying the Existing Categories

In order to better comprehend and further segregate the literature, this paper first adopted the dimensions and categories defined by [20] for classifying the articles. Then, by doing the review and classification, a few other categories in addition to the existing ones are identified to provide this review paper with a comprehensive system of classification and to pinpoint the current and future trends among the AR and MR articles.

5 Proposing Additional Categories

Firstly, from the technology-type perspective, the delivery type of AR and MR hardware systems [20] was divided into two categories web-based and standalone. Web-based AR was defined as technologies that can deliver project information to remote locations instead of standalone individual delivery of AR. However, Craig [6, 7] indicates that web-based AR means that the tool the participant uses to interact with the application is their web browser. On this basis, not all delivery types of AR devices

connected to a server to deliver the project to remote locations can be considered web-based. In other words, there are some AR applications that run on devices such as smartphones or tablets, and depending on the project requirement, they might be connected to a server [12] or not connected to a server [23]. In some online platforms, users only need internet access and run the application by logging into an Android or iOS application [19] but not necessarily run on the web browsers. Therefore, from the “technology type” perspective, the categories are redefined within the delivery type, as listed in Table 3 and further discussed in the following sections.

Moreover, from the “Technology Type” perspective, spatial registration, also known as AR/MR tracking system, is a critical technology type. It is defined as the ability to combine the virtual world and the real world through a proper relationship of the relative positions [5]. This is because accuracy was identified as one of the current challenges of AR/MR applications [4]. However, there are minimal literature review papers that investigate and compare the articles based on the type of spatial registration technology they used (2 from 103 articles) [4, 5]. Therefore, the “spatial registration” category is added within the “technology type” dimension, including two markerless and marker-based registration subcategories. In summary, the reference system of classification and the new suggested categories are presented in Table 2.

From the “Location” perspective, augmented reality technologies can be implemented in different locations during a construction project defined as (a) field and (b) home-office [20]. However, the category of “field” includes a variety of locations in AR/MR implementations. For instance, Koch et al. [13] assessed the possible solutions for indoor navigations during the facility maintenance projects and presented a natural marker-based AR framework that can digitally support facility maintenance (FM) operators. Moreover, some articles considered outdoor light conditions to be a crucial parameter in the success of their developed AR/MR technology [17]. Some other articles showed a focus on both indoor and outdoor built environments throughout their research in the field of AR/MR technology [8]. Therefore, as proposed in Table 2, the category of “field” is divided into two subcategories of “indoor” and “outdoor” environments to comprehensively categorize the articles from the location perspective.

## 6 Classification of Articles

In this section, 103 articles found within nine years from 2013 to 2021 are classified based on the number and percentage of articles in each category.



**Table 2** Defined dimensions and their relevant categories, including both referenced and proposed additional categories

Dimensions	Referenced categories <sup>a</sup>	Proposed additional categories
Research methodology	Case study, experimental/empirical study, proof of concept, questionnaire, literature review	N/A
Improvement focus	AEC industry, organization (facility owner, contractor, designer), projects, individuals	N/A
Industry sector	Building commercial, municipal/infrastructure, heavy/highway, residential, industrial	N/A
Target audience	Design team, project manager, worker/technician, inspector, project end-user, building systems' engineers, students, others	N/A
Project phase	Initiation, design, procurement, construction, operation/maintenance	N/A
Stage of technology maturity	Theory, framework, sub-system technical issues, system development, system application	N/A
Application area	Simulation/visualization, communication/collaboration, information modeling, information access/evaluation, progress monitoring, education/training, safety/inspection	N/A
Comparison role	Comparison mode (model vs. model, model vs. reality, reality vs. reality) Comparison purpose (progress monitoring, defect detection, evaluating the model, updating the model, validating the model)	N/A
Technology type	User perspective (immersive, non-immersive), device (mobile, non-mobile), delivery (web-based, standalone)	Delivery (handheld, server-based handheld, desktop, server-based desktop, self-contained headsets, web-based, cloud-based) Tracking system (marker-less, marker-based)
Location	Field, home-office	Field (indoor, outdoor), home-office

<sup>a</sup> Source [20]

**Table 3** Categories with the current focus of AR and MR technologies in construction

Dimensions	Categories with current focus	Total no. (or percentage) of articles from 2013 to 2021
Journals	AIC	46 (45%)
Research methodology	Experimental/empirical	42 (41%)
Improvement focus	Individuals	45 (44%)
Industry sector	Building/commercial	37 (36%)
Target audience	Worker/technicians	29%
Project phase	Construction	39 (39%)
Stage of technology maturity	System application	55 (53%)
Application area	Simulation/visualization	22%
Comparison role	Comparison mode (model vs. model)	29 (50% of 58 articles only)
	Comparison purpose (evaluating the model)	28 (48% of 58 articles only)
	User (non-immersive)	47 (46%)
Technology type	Device (mobile)	74 (72%)
	Delivery (self-contained headset)	27 (26%)
	Spatial registration system (marker-based)	43 (42%)
Location	Field (outdoor)	39 (38%)

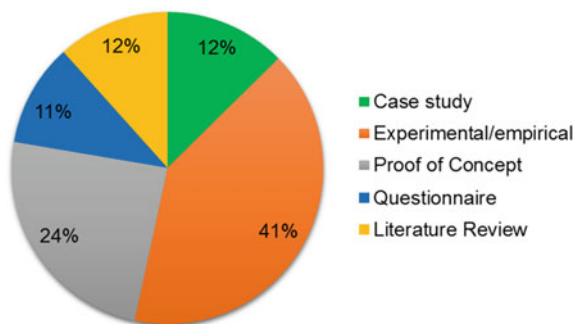
## 6.1 Research Methodology

As expressed in Table 2, five categories are identified in this dimension. Figure 4 illustrates the percentage of articles based on their research method. As shown, experimental/empirical methodology with 41% of total articles is the dominant category among AR and MR articles. Next is the proof-of-concept methodology, with 24% of the total articles. Literature review, questionnaires, and case study methodologies encompass around 12% of the articles and are among the least dominant categories in this dimension.

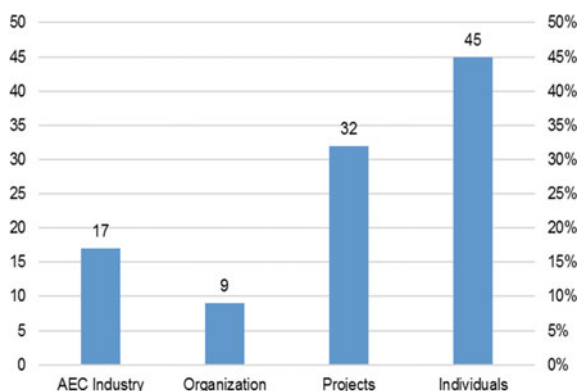
## 6.2 Improvement Focus

There are four categories to identify where the improvement of AR/MR technologies may occur, including the AEC industry, organization, projects, and individuals. Figure 5 depicts the number of articles within each improvement focus category. It is shown that 45 articles (44%) focus on individuals, while 32 articles (31%) have a

**Fig. 4** Percentage of articles based on their research methodology



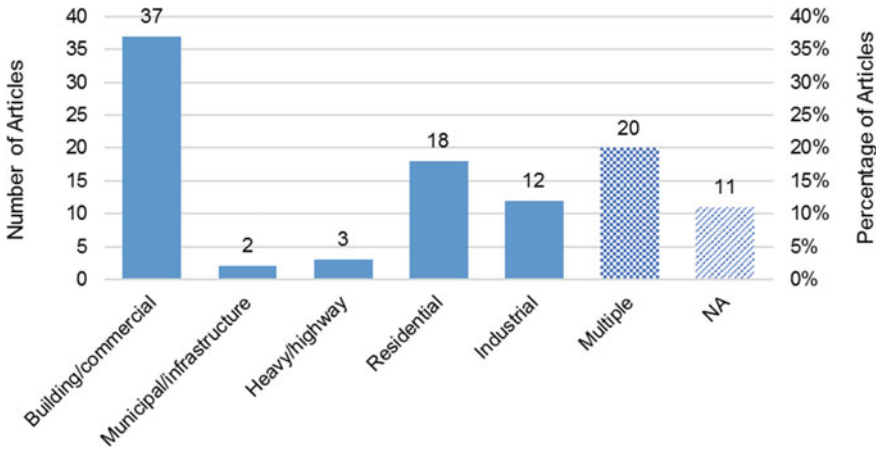
**Fig. 5** Number of articles by improvement level



principal focus on projects. Additionally, 17 articles (17%) and 9 articles (9%) focus on the AEC industry and organization levels, respectively.

### 6.3 Industry Sector

AR and MR technologies can facilitate various project types in the construction industry. As indicated in Table 2, the categories comprise Municipal/infrastructure, Residential, Building/commercial, Heavy/highway, and Industrial. Figure 6 illustrates the number of articles within each industry-type category. As shown, 37 articles have a principal focus on the building/commercial sector of the construction industry using AR/MR technologies. Residential, industrial, heavy/highway, and municipal/infrastructure have 18 articles (17%), 12 articles (12%), 3 articles (3%), and 2 articles (2%), respectively. Twenty articles (19%) cover multiple categories, and 11 (11%) articles were not applicable in any category.



**Fig. 6** Number of articles by industry sector

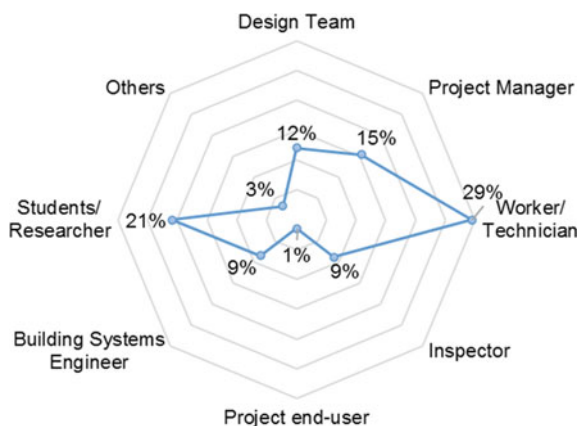
## 6.4 Target Audience

An extensive range of target audiences benefits from AR and MR technologies. These target audiences are divided into eight categories, including (1) design team (e.g., architects, interior and exterior designers), (2) project manager (e.g., schedule and budget professionals, site manager), (3) worker/technician (e.g., machine operators and technicians, assembly operators), (4) inspector (e.g., project safety officers, facility manager), (5) project end-user (e.g., building occupants, office employees), (6) building systems engineer (e.g., structural, mechanical, and electrical engineers), (7) students/researchers (e.g., engineering students, researchers), and (8) other stakeholders (e.g., clients, building owners). Since each article may refer to more than one category, the percentage of articles in each category is reported. For instance, for the articles in which AR and MR technologies proposed a change in the work of three of these audiences, each audience category counted as one-third of a sole category. Then, the total sum of numbers is reported as a percentage for the contribution of each category. Figure 7 depicts the percentage of articles in each category type of target audience. As shown, workers/technicians with 29% of the articles are the dominant target audiences in AR and MR articles.

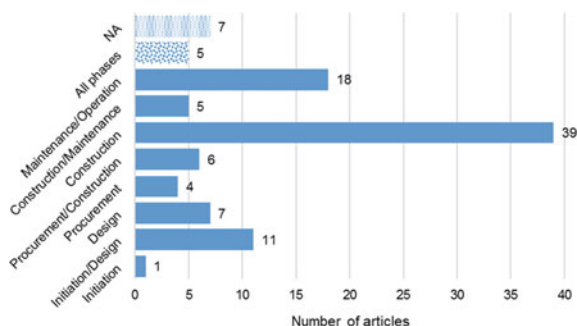
## 6.5 Project Phase

Each project consists of various steps and phases throughout its whole lifecycle. As mentioned in Table 2, these phases start from (1) initiation and outline design, (2) design development, (3) procurement, contract, and pre-construction, (4) construction, and finally ends in (5) maintenance and operation. Figure 8 shows the number of

**Fig. 7** Percentage of articles by target audience



**Fig. 8** Number of articles by project phase



articles by project phase. As shown, 39 articles (38%) mainly focus on the construction phase, and 18 articles (17%) primarily focus on the operation and maintenance phase of the projects. Figure 9 illustrates the number of articles for each project phase by year of publication. This diagram excludes articles that focus on multiple phases (which reduced the number of articles to 69).

Both Figs. 8 and 9 show that construction and operation/maintenance are the dominant categories among the articles published from 2013 to 2021. Moreover, as shown in Fig. 9, the highest number of articles occurred in 2021, with 15 articles. Additionally, a growing trend in articles focusing on the construction, operation, and maintenance phases can be interpreted.

## 6.6 Stage of Technology Maturity

AR and MR technologies are leveraged in different maturity levels, based on what is described in Table 2, and they include five categories. Figure 10 presents the number of articles within each stage-of-technology-maturity category. In this perspective,

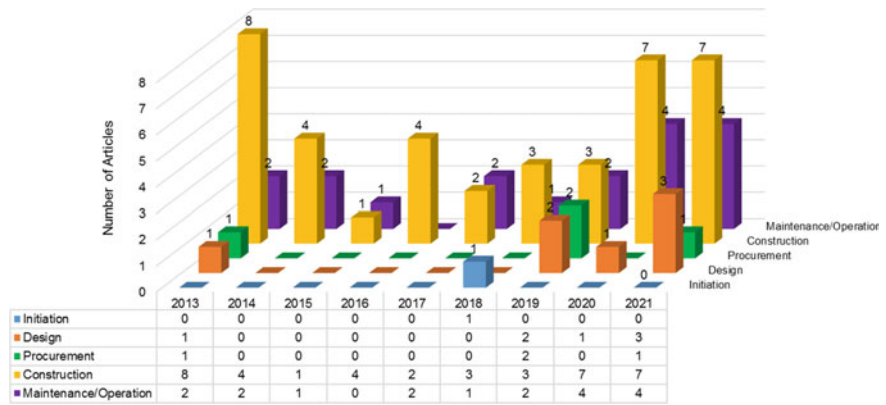
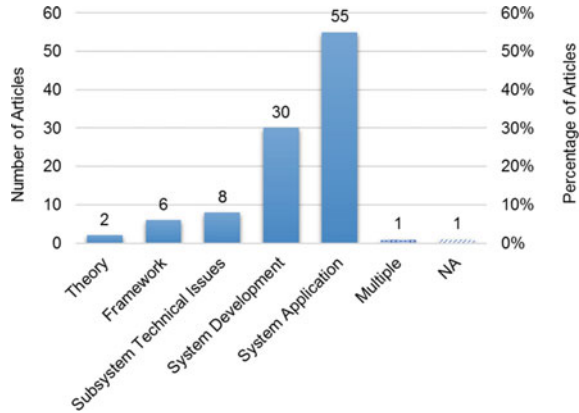


Fig. 9 Distribution of articles by year and project phase

Fig. 10 Number of articles by stage of technology maturity

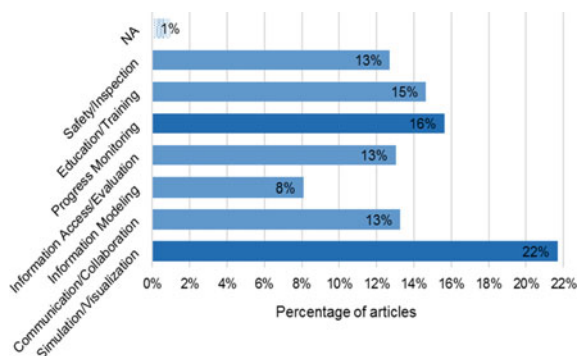


only two articles focus on the theory of AR/MR technologies. In comparison, the most significant number of articles focus on the system application (55 articles or 53%) and system development (30 articles or 29%) in this dimension.

6.7 Application Area

There are various application areas for augmented and mixed reality technologies in the AEC industry. As indicated in Table 2, this dimension is classified as simulation/visualization, communication/collaboration, information modeling, information access/evaluation, progress monitoring, education/training, and safety/inspection. Figure 11 illustrates the percentage of articles considered in each category from the application area perspective. Due to the overlapping nature of application areas, some

**Fig. 11** Percentage of articles by application area

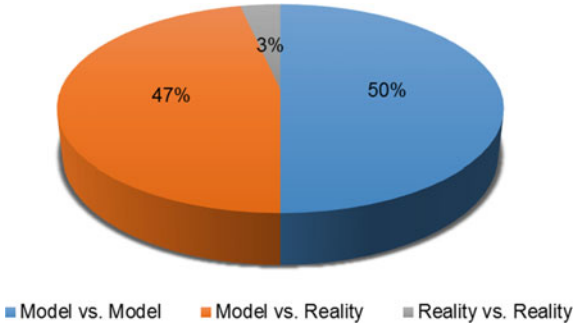


articles were applicable in several categories. Therefore, the classification results are not presented by their numbers but based on the percentage of articles identified in each category. As shown in Fig. 11, about 22% of the articles focus on simulation/visualization as the primary application area of AR and MR technologies, while 16% of them mainly focus on the progress monitoring area. Information modeling shows the least focus (8%) of AR and MR articles from the application area perspective.

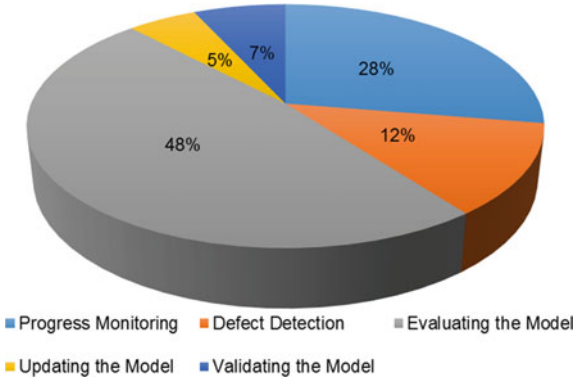
## 6.8 Comparison Role

Construction practitioners use different modes of comparison in implementing the AR/MR technologies, and as mentioned in Table 2, they are divided into three categories of Model vs. Model, Model vs. Reality, and Reality vs. Reality. Each comparison mode pursues a comparison purpose using the AR and MR technologies, including: (a) progress monitoring, (b) defect detection, (c) validating the model, (d) updating the model, and evaluating the model. Of 103 articles, 58 articles carry out a comparison practice in their work, about 56% of all. The rest do not show any interest in comparison modes and purposes of implementing AR and MR. Of these 58 articles, 29 articles (50%) focus on comparing model versus model, primarily to evaluate the model (28 articles or 48%). Comparing model versus reality (27 articles or 47%) with the purpose of progress monitoring (16 articles or 28%) is another primary focus of AR and MR articles in this category. Figures 12 and 13 depict the percentage of these 58 articles in both comparison modes and purposes.

**Fig. 12** Percentage of articles based on comparison modes



**Fig. 13** Percentage of articles based on comparison purpose



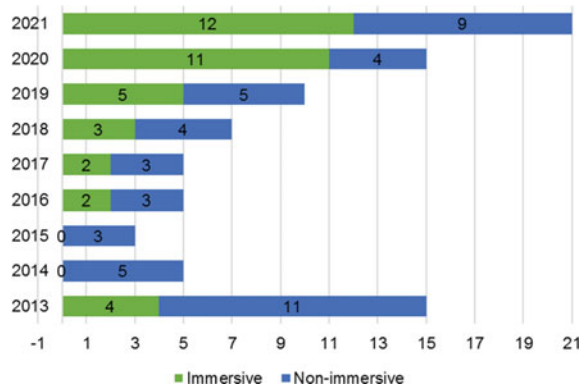
**6.9 Technology Type**

From a technology perspective, the articles can be classified into the following categories: (1) User experience, (2) Device, and (3) Delivery.

1. User: (a) Immersive and (b) Non-immersive. Immersive environments allow participants to feel as though they are inside the environment. Examples include HMDs, data gloves, and AR glasses. In contrast, non-immersive environments only allow participants to see the contents based on how the device in use—PC, smartphone, or tablet—is held and moved [27]. From a user perspective, 47 articles (46%) show a principal focus on non-immersive technologies, while 39 articles (38%) have a primary focus on immersive technologies. Additionally, 17 articles (17%) were not applicable in this category. However, as depicted in Fig. 14, the growing trend (from zero articles in 2014 to 12 articles in 2021) among the articles is using immersive forms of AR and MR technologies.
2. Device: (a) mobile and (b) non-mobile. With mobile augmented reality, the hardware required to perform an AR application is something that you can take with you wherever you go [7], while non-mobile augmented reality uses an ordinary desktop PC equipped with a Webcam (desktop AR) where the fusion between



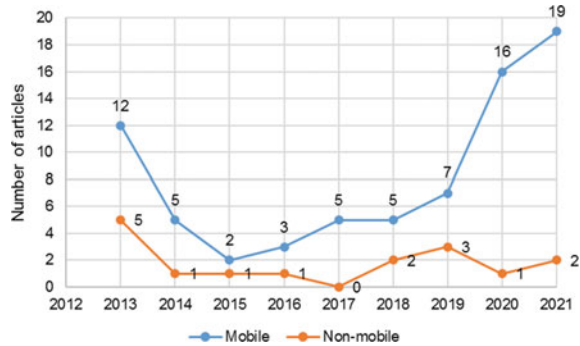
**Fig. 14** Immersive and non-immersive AR technology by year



real-world and its digital augmentation is displayed on the computer screen [2]. Figure 15 illustrates the number of articles focusing primarily on mobile and non-mobile augmented reality. A total of 74 articles (72%) have a primary focus on mobile AR technologies, while only 16 articles (16%) focus on non-mobile AR. Moreover, as shown in Fig. 15, a growing trend among the articles focusing on mobile augmented reality is recognized (from 2 articles in 2015 to 19 articles in 2021). Thirteen articles (13%) were not applicable in these categories.

3. Delivery: In terms of delivery perspective, there are several configurations/architectures used for augmented reality applications: (a) application run on handheld systems such as smartphones, (b) application run on handheld systems connected to remote server(s), (c) application run on desktop/laptop computers, (d) application run on desktop/laptop computers connected to remote server(s), (e) application run as a web application, and (f) application run on a cloud with a thin client [6]. Additionally, there are some examples of applications run on lightweight wearable devices such as Microsoft HoloLens 2 (a self-contained computer with Wi-Fi connectivity) [15] or Google Glass which is a small, lightweight wearable computer with a transparent display for hands-free work [9]. Accordingly, in this section, the subcategory of (g) self-contained headsets is

**Fig. 15** Number of articles by mobile/non-mobile AR technology and by year



also defined for classifying articles. From the delivery point of view, 26 articles (27%) have a primary focus on self-contained headsets, while only three articles (3%) focus on cloud-based augmented reality. Figure 16 illustrates a distribution of articles by year and delivery perspective. As shown, the largest number of articles in a single year belong to the self-contained headsets' category in 2021. Moreover, since there are always combinations of local and remote systems of AR and MR applications, 18 articles (18%) are considered in this category, while six articles were not applicable from the delivery perspective.

4. Spatial Registration: One of the most important requirements for AR systems is tracking. Visual tracking attempts to calculate the trajectory of an object in the image plane as it moves around a scene through features detected in a video stream. There are two primary tracking systems in AR implementations: (1) marker-based and (2) marker-less. Marker-based AR relies on placing fiducial markers (such as barcodes, QR codes, to name a few) in the real world, which is captured by a camera, thus creating an AR experience. In contrast, markerless AR does not depend on fiducial markers; however, the systems rely on natural features to execute tracking [1]. The articles are also classified based on their tracking system into two categories: marker-based and markerless systems. Figure 17 shows the number of articles by these two categories. Forty-three articles (42%) focus on marker-based tracking methods in AR and MR, while 38 articles (37%) focus on markerless methods. Two articles used both methods

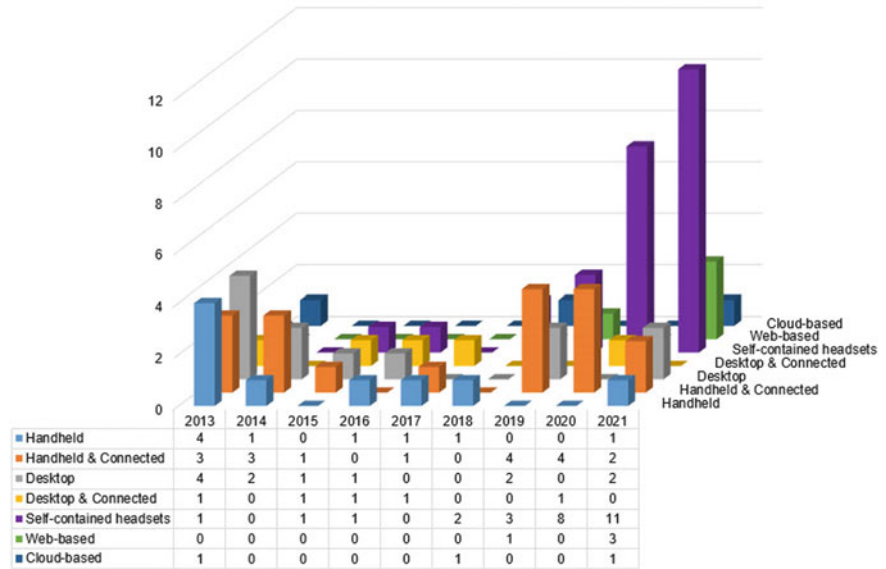


Fig. 16 Distribution of articles by the year and by delivery perspective

**Fig. 17** Distribution of articles by spatial registration system in AR and MR technologies and by year



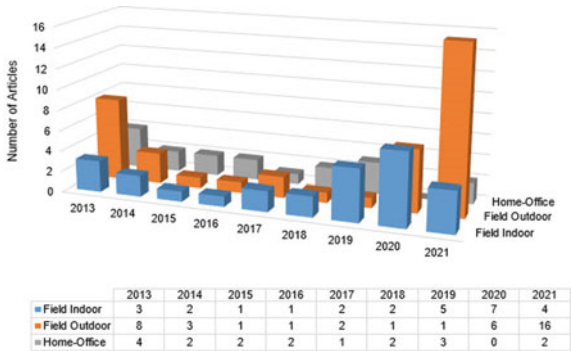
(mainly for comparison purposes), and 20 articles were not applicable in these categories. As shown in Fig. 17, there is a rising trend in using markerless AR and MR technologies among the articles.

6.10 Location

Augmented reality technologies can be implemented in different locations during a construction project. The categories include (a) Field: (a-1) Field-Indoor: e.g., building facility maintenance [13], manufacturing plants for off-site construction [26], (a-2) Field-Outdoor: e.g., urban area infrastructure, buildings construction sites, and (b) Home-office.

Figure 18 shows the number of articles by location and by year. It can be inferred that 39 articles focus on implementing AR and MR in on-site outdoor environments. In comparison, 27 and 18 articles focus on implementing AR and MR for on-site indoor and home-office purposes, respectively. Moreover, a rising trend among the articles (from one article in 2016 to 7 articles in 2020) is recognized in using AR and MR for on-site indoor implementations. The increasing trend also appears in the field-outdoor locations (i.e., from one article in 2016 to 16 articles in 2021).

**Fig. 18** Number of articles by location and by year



## 7 Discussion: Current and Future Trends

In recent years, the rapid development of AR and MR technologies revealed the significance of implementing those technologies in construction. Consequently, a higher number of construction practitioners are inclined to leverage the capabilities of AR and MR and will encourage their passion for research, development, and investment in this area. Based on the classification of articles in previous sections, some of the current and future trends of AR and MR in construction are discussed. Tables 3 and 4 summarize the current and future trends, respectively. Some categories are indicated as dominant categories among the articles based on the total number of articles in each category. In contrast, in other categories such as target audience and application area, since most articles target multiple audiences and application areas, they are reported based on the percentage. From the results shown in Table 3, it can be inferred that Automation in Construction published most AR and MR articles. The articles used experimental/empirical research methodology focusing on the individual level. Most articles focused on the building/commercial sector of the industry with an emphasis on improving the work of workers/technicians during the construction phase. The principal stage of technology maturity was leveraging the application of AR and MR systems rather than the theory, the framework, or the sub-system technical issues. The comparison approach appeared to be mainly to compare model versus model to evaluate the model. The major application area that the articles focused on is simulation/visualization, which can aid workers/technicians, project managers, and inspectors to visualize the 3D models for progress monitoring and inspection purposes during construction and maintenance phases.

Although the total number of articles used AR and MR in non-immersive environments, there is a growing trend in using AR and MR in immersive environments such as MS HoloLens and Google Glasses. The articles’ current and growing trends are both on mobile devices such as tablets, smartphones, and head-mounted devices

**Table 4** Categories with expected future trend of AR and MR articles in the construction industry

Dimensions	Categories	Future trend		
		No. of articles in 2017	No. of articles in 2021	Factor
Project	Construction	2	7	3.5
Phase	Maintenance/operation	2	4	2
	User (immersive)	2	12	6
	Device (mobile)	3	19	6.3
Technology type	Delivery (self-contained headset)	0	11	$\infty$
	Spatial registration system (marker-less)	1	11	11
Location	Field (outdoor)	2	16	8

(HMD), as indicated in Tables 3 and 4. Moreover, the articles' primary focus and growing trend show that most researchers and industry experts tend to use self-contained headsets to efficiently run the AR and MR on the device itself without requiring any other devices. Additionally, in terms of the location, most of the articles used AR and MR technologies in outdoor spaces such as construction sites, with a clear rising trend for outdoor areas as well.

## 8 Conclusion and Future Works

In this paper, 103 articles are selected from credited journals in the AEC industry and classified with a comprehensive set of dimensions and categories. The aim is to indicate the current focus of articles in using AR and MR technologies and identify the categories and areas with growing trends to pinpoint the potential areas that need more research and investment. As described in Sect. 7, a few areas need more focus in research and development, which are explained below.

Although worker/technicians are the dominant target audience in AR articles (29% of the articles), little research has been done with a focus on education/training of the workers (of 24 articles with emphasis on improving the work of workers/technicians 7 articles focus on training workers). Therefore, more research needs to be done on training workers and technicians and to develop improved AR-based work instruction to help novice workers quickly become familiar with the correct construction and assembly steps. Little focus is on using AR and MR technology in the procurement phase of the projects (10 articles, 3 of which are in the year 2013). Therefore, more research needs to be done on demonstrating the capabilities of AR technology in procurement management plans, such as quality management plans, which are a crucial part of construction projects and prefabrication plants for construction elements. Little work has been done on information access/evaluation area among the articles (13% of the articles), from which only one article focuses on inspection application and one on progress monitoring. Therefore, more research needs to be done on leveraging mobile AR capabilities and integrating them with BIM to help inspectors retrieve inspection data, checklists, and inspection lots during inspection tasks. The growing trend is toward using self-contained headsets (11 articles only in 2021). However, little work is done comparing the impacts of each delivery type (User Interface) of AR on the workers/technicians' cognitive behavior and task performance during assembly and construction tasks. From the Stage of Technology Maturity Perspective, little work is done with a primary focus on Subsystem Technical Issues (8% of the articles). However, spatial registration accuracy and occlusions are still among the common technical issues in overlaying the AR content on real objects. Therefore, more emphasis on this category needs to be done in future research.

## References

1. Brito PQ, Stoyanova J (2018) Marker versus markerless augmented reality. Which has more impact on users? *Int J Hum Comput Interact* 34(9):819–833. <https://doi.org/10.1080/10447318.2017.1393974>
2. Camba J, Contero M, Salvador-Herranz G (2014) Desktop vs. mobile: a comparative study of augmented reality systems for engineering visualizations in education
3. Chen K, Yang J, Cheng JCP, Chen W, Li CT (2020) Transfer learning enhanced AR spatial registration for facility maintenance management. *Autom Constr* 113, 103135. <https://doi.org/10.1016/j.autcon.2020.103135>
4. Cheng JCP, Chen K, Chen W (2020) State-of-the-art review on mixed reality applications in the AECO industry. *J Constr Eng Manag* 146(2):03119009. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001749](https://doi.org/10.1061/(asce)co.1943-7862.0001749)
5. Chi HL, Kang SC, Wang X (2013) Research trends and opportunities of augmented reality applications in architecture, engineering, and construction. *Autom Constr* 33:116–122. <https://doi.org/10.1016/j.autcon.2012.12.017>
6. Craig AB (2013) Augmented reality hardware. In: *Understanding augmented reality: concepts and applications*. Elsevier Science, Amsterdam (Chapter 3)
7. Craig AB (2013) Mobile augmented reality. In: *Understanding augmented reality: concepts and applications*. Elsevier Science, Amsterdam, pp 209–211 (Chapter 7)
8. Golparvar-Fard M, Ham Y (2014) Automated diagnostics and visualization of potential energy performance problems in existing buildings using energy performance augmented reality models. *J Comput Civ Eng* 28(1):17–29. [https://doi.org/10.1061/\(asce\)cp.1943-5487.0000311](https://doi.org/10.1061/(asce)cp.1943-5487.0000311)
9. Google (2021) Discover glass enterprise edition. Retrieved 10 Aug 2021, from <https://www.google.com/glass/start/>
10. Hakkarainen M, Woodward C, Billingham M (2008) Augmented assembly using a mobile phone. In: *IEEE international symposium on mixed and augmented reality*
11. Kang S-C, Wang X (2013) Augmented reality in architecture, engineering, and construction. Retrieved from <https://www.sciencedirect.com/journal/automation-in-construction/vol/33/suppl/C>
12. Kim C, Park T, Lim H, Kim H (2013) On-site construction management using mobile computing technology. *Autom Constr* 35:415–423. <https://doi.org/10.1016/j.autcon.2013.05.027>
13. Koch C, Neges M, König M, Abramovici M (2014) Natural markers for augmented reality-based indoor navigation and facility maintenance. *Autom Constr* 48:18–30. <https://doi.org/10.1016/j.autcon.2014.08.009>
14. Microsoft (2019) HoloLens 2. Retrieved from <https://www.microsoft.com/en-us/hololens/hardware>
15. Microsoft (2020) About HoloLens 2. Retrieved from <https://docs.microsoft.com/en-us/hololens/hololens2-hardware>
16. Milgram P, Takemura H, Utsumi A, Kishino F (1995) Augmented reality: a class of displays on the reality-virtuality continuum. In: *Telemanipulator and telepresence technologies*. SPIE, vol 2351, pp 282–292. <https://doi.org/10.1117/12.197321>
17. Mitterberger D, Dörfler K, Sandy T, Salveridou F, Hutter M, Gramazio F, Kohler M (2020) Augmented bricklaying. *Constr Robot* 4(3–4):151–161. <https://doi.org/10.1007/s41693-020-00035-8>
18. Osorto Carrasco MD, Chen P-HH (2021) Application of mixed reality for improving architectural design comprehension effectiveness. *Autom Constr* 126, 103677. <https://doi.org/10.1016/j.autcon.2021.103677>
19. Pour Rahimian F, Chavdarova V, Oliver S, Chamo F (2019) OpenBIM-Tango integrated virtual showroom for offsite manufactured production of self-build housing. *Autom Constr* 102(February):1–16. <https://doi.org/10.1016/j.autcon.2019.02.009>
20. Rankohi S, Waugh L (2013) Review and analysis of augmented reality literature for construction industry. *Vis Eng* 1(1). <https://doi.org/10.1186/2213-7459-1-9>

21. Schmalstieg D, Wagner D (2007) Experiences with handheld augmented reality. In: 6th IEEE and ACM international symposium on mixed and augmented reality, 2007, ISMAR 2007. IEEE
22. Shin DH, Dunston PS (2010) Technology development needs for advancing augmented reality-based inspection. *Autom Constr* 19(2):169–182. <https://doi.org/10.1016/j.autcon.2009.11.001>
23. Shin DH, Park J, Woo S, Jang WS (2013) Representations for imagining the scene of non-existing buildings in an existing environment. *Autom Constr* 33:86–94. <https://doi.org/10.1016/j.autcon.2012.09.013>
24. Song Y, Koeck R, Luo S (2021) Review and analysis of augmented reality (AR) literature for digital fabrication in architecture. *Autom Constr* 128
25. Sutherland IE (1968) A head-mounted three dimensional display. In: Fall joint computer conference, vol 5, pp 391–394. [https://doi.org/10.1016/0033-5894\(75\)90039-3](https://doi.org/10.1016/0033-5894(75)90039-3)
26. Tavares P, Costa CM, Rocha L, Malaca P, Costa P, Moreira AP, Sousa A, Veiga G (2019) Collaborative welding system using BIM for robotic reprogramming and spatial augmented reality. *Autom Constr* 106. <https://doi.org/10.1016/j.autcon.2019.04.020>
27. Ventura S (2019) Immersive versus non-immersive experience: exploring the feasibility of memory assessment through 360° technology. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2019.02509>

# Stochastic Modeling of Tag Installation Error for Robust On-Manifold Tag-Based Visual-Inertial Localization



Navid Kayhani, Brenda McCabe, and Angela P. Schoellig

**Abstract** Autonomous mobile robots, including unmanned aerial vehicles (UAVs), have received significant attention for their applications in construction. These platforms have great potential to automate and enhance the quality and frequency of the required data for many tasks such as construction schedule updating, inspections, and monitoring. Robust localization is a critical enabler for reliable deployments of autonomous robotic platforms. Automated robotic solutions rely mainly on the Global Positioning System (GPS) for outdoor localization. However, GPS signals are denied indoors, and pre-built environment maps are often used for indoor localization. This entails generating high-quality maps by teleoperating the mobile robot in the environment. Not only is this approach time-consuming and tedious, but it also is unreliable in indoor construction settings. Layout changes with construction progress, requiring frequent mapping sessions to support autonomous missions. Moreover, the effectiveness of vision-based solutions relying on visual features is highly impacted in low texture and repetitive areas on site. To address these challenges, we previously proposed a low-cost, lightweight tag-based visual-inertial localization method using AprilTags. Tags, in this method, are paper printable landmarks with known sizes and locations, representing the environment's quasi-map. Since tag placement/replacement is a manual process, it is subjected to human errors. In this work, we study the impact of human error in the manual tag installation process and propose a stochastic approach to account for this uncertainty using the Lie group theory. Employing Monte Carlo simulation, we experimentally show that the proposed stochastic model incorporated in our on-manifold formulation improves the robustness and accuracy of tag-based localization against inevitable imperfections in manual tag installation on site.

**Keywords** Stochastic modeling · Tag installation · Manifold tag-based visual-inertial localization

---

N. Kayhani (✉) · B. McCabe

Department of Civil and Mineral Engineering, University of Toronto, Toronto, Canada

e-mail: [navid.kayhani@mail.utoronto.ca](mailto:navid.kayhani@mail.utoronto.ca)

A. P. Schoellig

Institute for Aerospace Studies, University of Toronto, Toronto, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_3](https://doi.org/10.1007/978-3-031-34593-7_3)



## 1 Introduction

Automated monitoring and inspections have been extensively studied in the architecture, engineering, and construction (AEC) community. These methods need frequent and high-quality input data from the job site. Visual data have been one of the dominant data modalities used in these methods due to their information richness and low cost of collection [17]. Site personnel or hired professionals commonly capture the required images using hand-held cameras and mobile phones [14]. Alternatively, automated fixed cameras installed on indoor sites or mounted on tower cranes capture real-time visual feed. Manual data collection is costly and error-prone, while stationary cameras can only cover a limited area and are ineffective indoors [16]. Mobile robots, on the other hand, can be programmed to autonomously move the camera (or other sensors) around and are ideal for automated on-site data collection [16].

Aerial and ground robots with different levels of autonomy have been deployed as automated data collection platforms in construction and built environments. Deploying autonomous ground robots for indoor environmental air quality [9], semantic modeling [1], and building retrofit performance evaluation [15] are examples of the proposed automated robotic data collection solutions in built environments. Above all, camera-equipped compact aerial robots, such as rotary unmanned aerial vehicles (UAVs), can efficiently provide high-resolution images from various locations and fields of view [16]. They have shown promising potential for visual data collection in indoor [8, 16] and outdoor construction environments [7, 19]. However, most of the custom-built prototypes proposed in academia [2, 3, 13] and the cutting-edge platforms deployed in the industry (e.g., Spot [5]) are costly, limiting their scalability and applicability in practice.

One of the critical enablers for autonomy is robust global localization. Autonomous mobile robots may rely on the Global Positioning System (GPS) for outdoor localization (e.g., in [6, 14]). However, reliable localization is challenging in ever-changing, low-texture, and GPS-denied indoor construction environments [10]. The majority of the proposed indoor localization methods rely on maps generated from data gathered with the same sensor modality used in localization. These maps are built by collecting sensory measurements via robot teleoperation. The collected data are then incorporated into a coherent environmental representation using simultaneous localization and mapping (SLAM) techniques, typically offline. However, since construction environments constantly evolve and change with project progress, frequent teleoperated mapping sessions may be required. Another challenge in these ever-changing environments is the potential loss of track due to dynamic, transparent, or temporary objects. Moreover, generating and maintaining large maps require computational and storage resources, which are highly limited, particularly in aerial robots. These technical challenges in localization in indoor construction environments make safe and reliable autonomous data collection missions difficult.

To address these challenges, we previously proposed a low-cost, versatile, lightweight visual-inertial localization method using fiducial markers such as AprilTags [12]. AprilTags are square-shaped payload tags that provide robust data association correspondences [18]. Given that the location and size of the tags are known, this method can globally localize any platform, including inexpensive, off-the-shelf UAVs, with the minimum sensor suite of a camera and an inertial measurement unit (IMU) in real-time. The proposed formulation in [12] is based on an on-manifold extended Kalman filter (EKF), properly addressing the topological structure of rotations and poses in 3D space. The test results showed that our tag-based localization method could estimate the vehicle's 3D position with an accuracy of 2–5 cm. This method leverages construction-specific processes and practices, such as frequent indoor layout surveying and four-dimensional building information models (4D-BIM) [11], to provide robust indoor localization. Tag-based visual-inertial can be applied to enable indoor localization in a wide range of applications where many vision-based techniques often face challenges due to perceptual aliasing and feature scarcity. However, the manual tag placement/replacement process may be subjected to installation errors, which affect localization performance.

This work relaxes the assumption of perfectly known 3D tags' position and orientation (i.e., pose) and considers the errors in the tag installation process using a stochastic approach. We take advantage of our previously proposed on-manifold formulation's capabilities to account for uncertainties in the input tag poses in the inertial frame, i.e., a fixed global reference frame. It is assumed that the underlying uncertainty in the determination of the tags' pose can be stochastically represented as a random variable with multi-variant zero-mean Gaussian distribution  $\epsilon_\tau \sim \mathcal{N}(\mathbf{0}, \Sigma_\tau)$ . In the remainder of this paper, we briefly provide some background regarding the Lie group theory and our proposed on-manifold tag-based EKF. Next, we update our formulation by incorporating uncertainty on the input global tag pose. Finally, we study the impact of incorporating the stochastic tag pose model in localization accuracy using Monte Carlo simulation and the data collected from laboratory and simulation experiments.

## 2 Lie Group Theory and State Estimation

Motivated by the necessity of reliable performance of robotic platforms in practice, considerable effort has been devoted to the proper formulation of state estimation problems that can result in precise, consistent, and stable solutions [20]. The state of a system is a set of those underlying quantities of the system, by which one can describe its intended characteristics at any snapshot [4]. The process of reconstructing these underlying quantities is referred to as state estimation. A state estimation technique assumes a sequence of noisy measurements/observations, a sequence of inputs to the system, and a priori system and measurement models [4]. These estimates are imperfect, therefore uncertain. The uncertainty in estimation can be caused by many factors, including random effects and imperfect sensors, models, and computations,

that must be managed and acknowledged by the estimation framework. Therefore, a robust estimation framework needs to (1) properly formulate the underlying state and (2) account for the sources of uncertainty. Our formulation is an effort to properly consider the manifold structure of the pose and the rotation groups in 3D and carefully deal with the representation and propagation of the state and the associated uncertainties over time. To provide the necessary background, the rest of this section briefly reviews the key concepts and operations in matrix Lie groups that can be leveraged in localization.

Localization is the problem of estimating the pose of a vehicle with respect to a reference frame over time. If we fix the reference frame, the vehicle can be globally localized. Tracking the pose of a vehicle in 3D space involves estimating six degrees of freedom. Pose in 3D space is a member of the 3D special Euclidean group,  $SE(3)$ , represented in the form of valid  $4 \times 4$  transformation matrices [4]:

$$SE(3) = \left\{ \mathbf{T} = \begin{bmatrix} \mathbf{C} & \mathbf{r} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4 \times 4} \mid \mathbf{C} \in SO(3), \mathbf{r} \in \mathbb{R}^3 \right\} \quad (1)$$

where  $\mathbf{r}$  is the 3D translation vector, and  $\mathbf{C}$  is the  $3 \times 3$  matrix in the special orthogonal group of  $SO(3)$  that represents rotations in 3D and is defined as:

$$SO(3) = \{ \mathbf{C} \in \mathbb{R}^{3 \times 3} \mid \mathbf{C}\mathbf{C}^T = \mathbf{1}, \det(\mathbf{C}) = 1 \} \quad (2)$$

Mathematically,  $SE(3)$  and  $SO(3)$  are differentiable and continuous (i.e., smooth) manifolds of matrix Lie groups. A smooth manifold is a curved surface without edges or spikes that can be locally approximated as a linear hyperplane [20]. The smoothness guarantees a unique linear tangent space at each point on the manifold, and linear space allows for applying calculus (i.e., taking derivatives and integrals). Lie groups are smooth manifolds with the nice properties of groups, such as closure, identity, inversion, and associativity. For example, the 3D surface of a unit sphere is a smooth manifold and forms a Lie group. Every point on a unit sphere looks the same, and a unique plane exists at any point on its 3D surface. The matrix Lie groups of  $SE(3)$  and  $SO(3)$  have also their corresponding tangent spaces that are referred to as the Lie algebra of  $se(3)$  and  $so(3)$ , respectively. A matrix Lie group and their corresponding Lie algebra are linked through exponential and logarithmic mappings. Using the exponential mapping from  $se(3)$  to  $SE(3)$ , we have:

$$\mathbf{T} = \exp(\xi^\wedge) = \sum_{n=0}^{\infty} \frac{1}{n!} (\xi^\wedge)^n, \quad \xi = \begin{bmatrix} \rho \\ \phi \end{bmatrix} \in \mathbb{R}^6 \quad (3a)$$

where  $\rho \in \mathbb{R}^3$  and  $\phi \in \mathbb{R}^3$  are the translational components of the pose vector  $\xi$ ,  $\xi^\wedge$  is a  $4 \times 4$  matrix in the Lie algebra of  $se(3)$ , and  $\phi^\wedge \in so(3)$  is the Lie algebra associated with  $SO(3)$  and equivalent to the rotation vector ( $\phi \in \mathbb{R}^3$ ) expressed in the skew-symmetric matrix format. Another useful operator is the dot operator,  $(\cdot)^\odot$ , which is defined as  $\xi^\wedge \mathbf{p} \equiv \mathbf{p}^\odot \xi$  [4], where  $\mathbf{p}$  is expressed in the homogeneous

coordinates:

$$\xi^\wedge = \begin{bmatrix} \rho \\ \phi \end{bmatrix}^\wedge = \begin{bmatrix} \phi^\wedge & \rho \\ \mathbf{0}^T & 1 \end{bmatrix} \in \text{se}(3), \quad \phi^\wedge = \begin{bmatrix} \phi_x \\ \phi_y \\ \phi_z \end{bmatrix}^\wedge = \begin{bmatrix} 0 & -\phi_z & \phi_y \\ \phi_z & 0 & -\phi_x \\ -\phi_y & \phi_x & 0 \end{bmatrix} \in \text{so}(3) \quad (3b)$$

$$\mathbf{p}^\odot = \begin{bmatrix} sx \\ sy \\ sz \\ s \end{bmatrix}^\odot = \begin{bmatrix} \boldsymbol{\varepsilon} \\ \eta \end{bmatrix}^\odot = \begin{bmatrix} \eta \mathbf{1} & -\boldsymbol{\varepsilon}^\wedge \\ \mathbf{0}^T & \mathbf{0}^T \end{bmatrix} \in \mathbb{R}^{4 \times 6} \quad (3c)$$

By defining the inverse of the skew-symmetric operator as  $(\cdot)^\vee$ , the logarithmic mapping can be used to go in the other direction (not uniquely):

$$\xi = \ln(\mathbf{T})^\vee \quad (4)$$

An uncertain rigid body transformation ( $\mathbf{T}$ ) can be expressed as the combination of a noise-free nominal (i.e., mean) component ( $\bar{\mathbf{T}}$ ) and a small, zero mean, noisy, perturbation component ( $\exp(\boldsymbol{\epsilon}^\wedge)$ ) [4]. Given that the perturbation is zero-mean Gaussian, we have:

$$\mathbf{T} = \exp(\boldsymbol{\epsilon}^\wedge) \bar{\mathbf{T}}, \quad \boldsymbol{\epsilon} \in \mathbb{R}^6 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad (5)$$

This on-manifold formulation using Lie group theory allows convenient transformations of distributions through other group elements using a beneficial  $6 \times 6$  linear transform called the adjoint matrix of an element of  $\text{SE}(3)$  [4]:

$$\text{Ad}(\mathbf{T}) = \text{Ad}\left(\begin{bmatrix} \mathbf{C} & \mathbf{r} \\ \mathbf{0}^T & 1 \end{bmatrix}\right) = \begin{bmatrix} \mathbf{C} & \mathbf{r}^\wedge \mathbf{C} \\ \mathbf{0}^T & \mathbf{C} \end{bmatrix} \in \mathbb{R}^{6 \times 6} \quad (6)$$

The following section summarizes our on-manifold tag-based visual-inertial localization formulation without considering the tag installation error [12]. Next, the errors involved in the manual tag installation process are discussed. Finally, we incorporate these errors in our formulation using the capabilities of matrix Lie groups and the theoretical concepts reviewed in this section.

### 3 Tag-Based Visual-Inertial Localization Using On-Manifold EKF

The state ( $\mathbf{x}$ ) to be estimated in the localization (pose tracking) problem can be defined as:

$$\mathbf{x} = \left\{ \left\{ \mathbf{r}_i^{v_0 i}, \mathbf{C}_{v_0 i} \right\}, \left\{ \mathbf{r}_i^{v_1 i}, \mathbf{C}_{v_1 i} \right\}, \dots, \left\{ \mathbf{r}_i^{v_K i}, \mathbf{C}_{v_K i} \right\} \right\} = \{\mathbf{T}_0, \mathbf{T}_1, \dots, \mathbf{T}_K\},$$

$$\mathbf{T}_k \in \text{SE}(3) \quad (7)$$

The tag-based visual-inertial localization uses an on-manifold EKF to estimate the 3D global pose of vehicles with a camera and an IMU, including low-cost, compact UAVs. It assumes that the camera lens parameters and tag poses are known a priori. Furthermore, translational and rotational velocities ( $\boldsymbol{\omega}$ ) and the vehicle's initial state  $\check{\mathbf{T}}_0$  are the system inputs  $\mathbf{v}$ . Following the perturbation scheme in Eq. (5) and given an additive perturbation, the motion model to predict the state and propagate uncertainty in time can be written as [4]:

$$\text{Nominal: } \bar{\mathbf{T}}_k = \Xi_k \bar{\mathbf{T}}_{k-1}, \quad \Xi_k = \bar{\mathbf{T}}_{v_k v_{k-1}} \in \text{SE}(3) \quad (8)$$

$$\text{Perturbation: } \delta \xi_k = \underbrace{\text{Ad}(\Xi_k)}_{\mathbf{F}_{k-1}} \delta \xi_{k-1} + \mathbf{w}_k, \quad \mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q}_k) \quad (9)$$

Measurements from tag detection are then incorporated to update (correct) the predictions. From a single image, it is possible to estimate the relative pose of the detected tag with respect to the camera frame. However, instead of using the relative camera-tag pose as measurements, our formulation considers the four corresponding pixel coordinates as pixel-level measurements in a tightly coupled data fusion approach. We already showed the advantage of using pixel-level tag measurements in [12].

The measurement model  $g(\cdot)$  can be viewed as a combination of two nonlinear functions of the state  $\mathbf{x}$ ,  $z(\cdot)$  and  $s(\cdot)$ . The 3D coordinates of the  $n$ -th corner point of tag  $j$  expressed in the camera frame  $z_k^{\tau_j, n}(\mathbf{x}) = \mathbf{p}_{c_k}^{p_{\tau_j, n} c_k} = [X \ Y \ Z]^T$  can be written as:

$$z_k^{\tau_j, n}(\mathbf{x}) = \mathbf{p}_{c_k}^{p_{\tau_j, n} c_k} = \mathbf{D}^T \mathbf{T}_{cv} \mathbf{T}_{v_k i} \mathbf{p}_{\tau_j, n}, \quad (10)$$

where  $\mathbf{D}^T = [\mathbf{I}_3 | \mathbf{0}_{3 \times 1}]$ ,  $\mathbf{T}_{cv}$  is the vehicle-to-camera transform determined by calibration,  $\mathbf{T}_{v_k i}$  is the state to be estimated ( $\mathbf{x}$ ), and  $\mathbf{p}_{\tau_j, n}$  is the  $n$ -th corner point of tag  $j$  expressed in the inertial coordinate frame in the homogeneous format. Given the pose of tag  $j$  in the inertial frame  $\mathbf{T}_{\tau_j i}$  is known as a priori, we have:

$$\mathbf{p}_{\tau_j, n} = \mathbf{T}_{\tau_j i}^{-1} \mathbf{P}_{\tau_j, n} = \mathbf{T}_{i \tau_j} \mathbf{P}_{\tau_j, n} \quad (11)$$

where  $\mathbf{P}_{\tau_j, n}$  is the homogeneous coordinates of its  $n$ -th corner in the tag frame  $\bar{\mathcal{F}}_{\tau_j}$ . In [12], we had the ideal assumption that  $\mathbf{T}_{i \tau_j}$  is deterministic and subjected to no uncertainty. In this work, however, we relax this assumption and are interested in investigating the impact of introducing uncertainty in  $\mathbf{T}_{i \tau_j}$  on estimation accuracy.

The second nonlinearity  $s(\cdot)$  arises from the sensor model, which is an ideal pinhole camera model. Together, the nonlinear measurement model can be written as:

$$\mathbf{y}_{k,j}^n = g_{k,j}^n(\mathbf{x}) = s(z_k^{\tau_{j,n}}(\mathbf{x})) = \begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{D}_p \mathbf{K} \frac{1}{Z} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \delta \mathbf{n}_{k,j}^n, \quad \delta \mathbf{n}_{k,j}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{k,j}) \quad (12a)$$

where  $\mathbf{y}_{k,j}^n = [u \ v]^T$  is the pixel coordinates of the  $n$ -th corner of tag  $j$ , observed at time  $k$ , projected onto the frontal image plane,  $\mathbf{K}$  is the  $3 \times 3$  camera intrinsic matrix,  $\mathbf{D}_p = [\mathbf{1}_2 | \mathbf{0}_{2 \times 1}]$ , and  $\delta \mathbf{n}_{k,j}^n$  is an additive zero-mean Gaussian measurement noise at the pixel level.

The EKF algorithm involves a prediction and a correction step in a recursive manner. The prediction step projects the current estimate of the state and the covariance (uncertainties) given the previous estimate. The correction step incorporates the measurements in the prior estimates from the prediction step and updates the associated uncertainties. To obtain the recursive steps in EKF, it is necessary to linearize the motion and observation models about the state's mean ( $\bar{\mathbf{x}}$ ), as the operating point. In this formulation, the motion model is already linear in  $\delta \xi$ . Hence, we only need to linearize the measurement model in Eq. (12a). The generic form of the linearized motion model is as follows:

$$\mathbf{y}_{k,j}^n \approx g_{k,j}^n(\bar{\mathbf{x}}) + \mathbf{G}_k^{\tau_{j,n}} \delta \xi_k + \delta \mathbf{n}_{k,j}^n, \quad \mathbf{G}_k^{\tau_{j,n}} = \mathbf{S}_k^{\tau_{j,n}} \mathbf{Z}_k^{\tau_{j,n}} \quad (13)$$

where  $\mathbf{S}_k^{\tau_{j,n}}$  and  $\mathbf{Z}_k^{\tau_{j,n}}$  are the Jacobians of nonlinearities  $s(\cdot)$  and  $z(\cdot)$ . The sensor model's Jacobian is independent of the assumption of uncertainty in  $\mathbf{T}_{i\tau_j}$ . So, we can write:

$$\mathbf{S}_k^{\tau_{j,n}} = \left. \frac{\partial s}{\partial z_k^{\tau_{j,n}}} \right|_{z_k^{\tau_{j,n}}(\bar{\mathbf{x}})} = \mathbf{D}_p \mathbf{K} \mathbf{S}, \quad \mathbf{S} = \begin{bmatrix} \frac{1}{Z} & 0 & -\frac{X}{Z^2} \\ 0 & \frac{1}{Z} & -\frac{Y}{Z^2} \\ 0 & 0 & 0 \end{bmatrix} \quad (14a)$$

Preserving the assumption of deterministic  $\mathbf{T}_{\tau_{ji}}$ , we already showed [12]:

$$\mathbf{Z}_k^{\tau_{j,n}} = \mathbf{D}^T \mathbf{T}_{cv} (\bar{\mathbf{T}}_{vki} \mathbf{T}_{i\tau_j} \mathbf{P}_{\tau_{j,n}})^{\odot} \quad (14b)$$

In the following sections, we derive the Jacobian of the first nonlinearity, that is  $\mathbf{Z}_k^{\tau_{j,n}}$ , for uncertain world-to-tag transforms ( $\mathbf{T}_{\tau_{ji}}$ ).

## 4 Tag Installation Error: Causes and Modeling

We already suggested [11] two main strategies for the global tag pose determination. One strategy is distributing tags in the indoor workspace and surveying their pose. The other strategy is first finding the location and orientation of the tags based on a systematic placement plan that considers factors such as localizability, safety, and cost-efficiency and then placing the tags at the specified locations. The other factor worth considering is that the tags might need to be replaced on a construction site to guarantee their functionality over time. For example, paper-printed tags might be damaged, and those surface sprayed might need redoing. All these manual works are subjected to human error that can be represented as uncertainty in  $\mathbf{T}_{\tau_{ji}}$ . To model the uncertainty, we assume that  $\mathbf{T}_{\tau_{ji}}$  is perturbed by a zero-mean Gaussian distribution noise,  $\boldsymbol{\epsilon}_\tau \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\tau)$ . For example, if the error for placing tags on a wall parallel to  $XZ$  plane follows a zero-mean error Gaussian distribution with a standard deviation of 2 cm in position ( $\sigma_x = \sigma_z = 0.02$  m), assuming uncorrelated errors, we can write:

$$\boldsymbol{\Sigma}_\tau = \text{diag}(0.0004, 0.0, 0.0004, 0.0, 0.0, 0.0) \quad (15)$$

However, the identification and quantification of the covariance matrix,  $\boldsymbol{\Sigma}_\tau$ , are out of the scope of this work and depend on the adopted strategy and many other factors.

## 5 Stochastic Representation of Tag Installation Error in On-Manifold Formulation

Using on-manifold formulation and leveraging the properties of matrix Lie groups allow for incorporating the uncertain pose  $\mathbf{T}_{i\tau_j} = \{\bar{\mathbf{T}}_{i\tau_j}, \boldsymbol{\Sigma}_\tau\}$  into the equations introduced in Sect. 3. Following Eq. (5), the uncertain pose  $\mathbf{T}_{i\tau_j}$  can be represented as:

$$\mathbf{T}_{i\tau_j} = \delta \mathbf{T}_{i\tau_j} \bar{\mathbf{T}}_{i\tau_j} = \exp(\boldsymbol{\epsilon}_\tau^\wedge) \bar{\mathbf{T}}_{i\tau_j}, \quad \boldsymbol{\epsilon}_\tau \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\tau), \quad \boldsymbol{\Sigma}_\tau \in \mathbb{R}^{6 \times 6} \quad (16)$$

where  $\boldsymbol{\epsilon}_\tau \in \mathbb{R}^6$  is a vector random variable. From Eqs. (10) and (11), we know that:

$$\mathbf{z}_k^{\tau_{j,n}}(\mathbf{x}) = \mathbf{D}^T \mathbf{T}_{cv} \mathbf{T}_{v_k i} \mathbf{T}_{i\tau_j} \mathbf{P}_{\tau_{j,n}} = \mathbf{D}^T \mathbf{T}_{cv} \exp(\delta \boldsymbol{\xi}_k^\wedge) \bar{\mathbf{T}}_{v_k i} \exp(\boldsymbol{\epsilon}_\tau^\wedge) \bar{\mathbf{T}}_{i\tau_j} \mathbf{P}_{\tau_{j,n}} \quad (17a)$$

$$\begin{aligned} \mathbf{z}_k^{\tau_{j,n}}(\mathbf{x}) \approx & \underbrace{\mathbf{D}^T \mathbf{T}_{cv} \bar{\mathbf{T}}_{v_k i} \bar{\mathbf{T}}_{i\tau_j} \mathbf{P}_{\tau_{j,n}}}_{\mathbf{z}_k^{\tau_{j,n}}(\bar{\mathbf{x}})} + \underbrace{\mathbf{D}^T \mathbf{T}_{cv} (\bar{\mathbf{T}}_{v_k i} \bar{\mathbf{T}}_{i\tau_j} \mathbf{P}_{\tau_{j,n}})^\odot}_{\mathbf{z}_k^{\tau_{j,n}}} \delta \boldsymbol{\xi}_k \\ & + \underbrace{\mathbf{D}^T \mathbf{T}_{cv} \bar{\mathbf{T}}_{v_k i} (\bar{\mathbf{T}}_{i\tau_j} \mathbf{P}_{\tau_{j,n}})^\odot}_{\mathbf{E}_k} \boldsymbol{\epsilon}_\tau \end{aligned} \quad (17b)$$

$$g_{k,j}^n(\mathbf{x}) = \mathbf{y}_{k,j}^n = s(z_k^{\tau_j,n}(\mathbf{x})) + \delta \mathbf{n}_{k,j}^n = s(z_k^{\tau_j,n}(\bar{\mathbf{x}}) + \mathbf{Z}_k^{\tau_j,n} \delta \boldsymbol{\xi}_k + \mathbf{E}_k \boldsymbol{\epsilon}_\tau) + \delta \mathbf{n}_{k,j}^n \quad (18a)$$

$$\begin{aligned} g_{k,j}^n(\mathbf{x}) &\approx \underbrace{s(z_k^{\tau_j,n}(\bar{\mathbf{x}}))}_{g_{k,j}^n(\bar{\mathbf{x}})} + \underbrace{\mathbf{S}_k^{\tau_j,n} \mathbf{Z}_k^{\tau_j,n}}_{\mathbf{G}_k^{\tau_j,n}} \delta \boldsymbol{\xi}_k + \underbrace{\overbrace{\mathbf{S}_k^{\tau_j,n} \mathbf{E}_k \boldsymbol{\epsilon}_\tau}^{E'_k}}_{\delta \mathbf{N}_{k,j}^n} + \delta \mathbf{n}_{k,j}^n \\ &= g_{k,j}^n(\bar{\mathbf{x}}) + \mathbf{G}_k^{\tau_j,n} \delta \boldsymbol{\xi}_k + \delta \mathbf{N}_{k,j}^n \end{aligned} \quad (18b)$$

By defining  $\delta \mathbf{N}_{k,j}^n \sim \mathcal{N}(\mathbf{0}, \mathcal{R}_{k,j})$  and for  $M$  measurements, we have:

$$\mathcal{R}_{k,j} = E[\delta \mathbf{N}_{k,j}^n \delta \mathbf{N}_{k,j}^{nT}] = \mathbf{E}'_k \boldsymbol{\Sigma}_\tau \mathbf{E}'_k{}^T + \mathbf{R}_{k,j}, \quad \mathcal{R}_k = \text{diag}(\mathcal{R}_{k,j}); \forall j \in [1, M] \quad (19)$$

Finally, the *tag installation error-aware EKF* (TIE-EKF) can be written as [4]:

$$\text{Prediction: } \check{\mathbf{T}}_k = \Xi_k \hat{\mathbf{T}}_{k-1}, \quad \check{\mathbf{P}}_k = \mathbf{F}_{k-1} \hat{\mathbf{P}}_{k-1} \mathbf{F}_{k-1}^T + \mathbf{Q}_k \quad (20a)$$

$$\text{Kalman Gain: } \mathbf{K}_k = \check{\mathbf{P}}_k \mathbf{G}_k^T (\mathbf{G}_k \check{\mathbf{P}}_k \mathbf{G}_k^T + \mathcal{R}_k)^{-1} \quad (20b)$$

$$\text{Correction: } \hat{\mathbf{P}}_k = (1 - \mathbf{K}_k \mathbf{G}_k) \check{\mathbf{P}}_k, \quad \hat{\mathbf{T}}_k = \exp\left((\mathbf{K}_k(\mathbf{y}_k - \check{\mathbf{y}}_k))^\wedge\right) \check{\mathbf{T}}_k \quad (20c)$$

where  $(\hat{\cdot})$  and  $(\check{\cdot})$  denote posterior (estimated) and prior (predicted) quantities, respectively, and  $\mathbf{K}_k$  is the *Kalman gain*. Also,  $\mathcal{R}_k$  can be thought of as the covariance of an augmented additive zero-mean measurement noise that varies with time and the vehicle's prior (predicted) state.

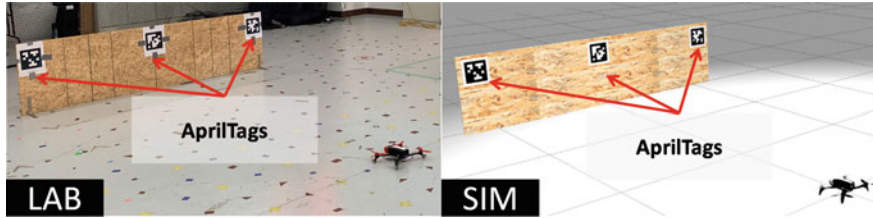
## 6 Experiments

Monte Carlo simulation (MCS) is used to study the performance of our on-manifold tag-based localization with (TIE-EKF) and without (EKF) [12] the stochastic modeling of tag installation errors. Three experiments were conducted in simulation and laboratory settings, as summarized in Table 1. In these experiments, a low-cost, compact UAV (i.e., Parrot Bebop2) was deployed as the aerial robotic platform, and three AprilTags with known size (0.165 m) and pose were used (Fig. 1). The trajectories were planned such that at least one tag would remain in the camera's field of view at any point in time. The UAV's camera lens parameters were previously obtained by calibration. To incorporate the installation error and equivalently tag pose uncertainty, for each iteration in MCS, we perturbed tags' pose by sampling from the pose perturbation distribution  $\boldsymbol{\epsilon}_\tau \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\tau)$  and exponential mapping in



**Table 1** Summary of custom-designed experiments in laboratory and simulation environments

Name	SIM/LAB	Description
SLS	SIM	A straight-line back-and-forth trajectory (3 m × 4)
3DC	LAB	A 3D circular trajectory of radius one meter
SLL	LAB	A straight-line back-and-forth trajectory (3 m × 4)

**Fig. 1** Laboratory and simulation environments and the utilized tag configurations

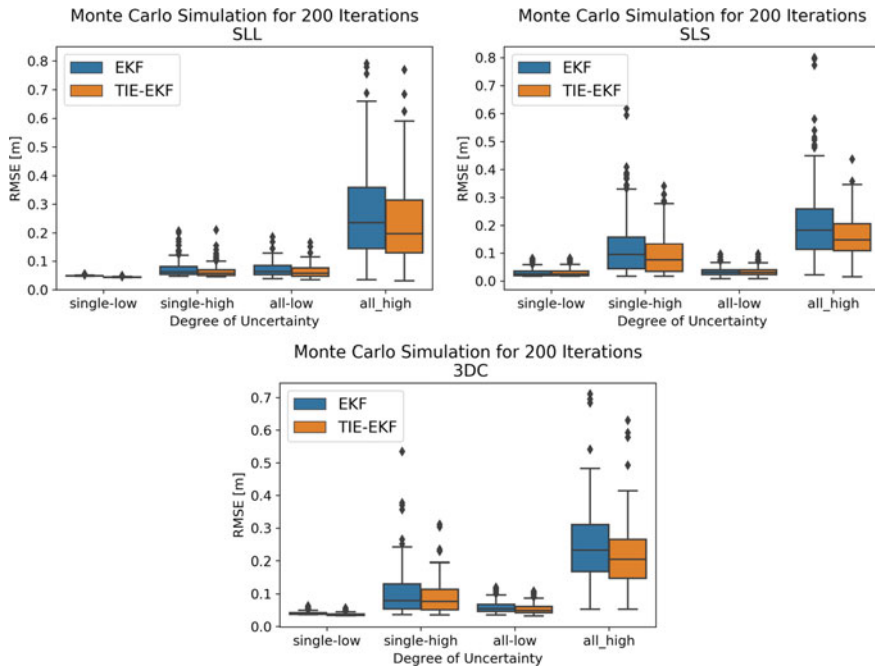
Eq. (16). Then, we estimated the vehicle's pose using TIE-EKF and EKF methods for performance comparisons. The root mean squared error (RMSE) in UAV's 3D position estimates were selected as the metric for the comparisons.

In each experiment, we repeated the MCS for 200 iterations and in scenarios with different levels of uncertainty: (1) single tag with low uncertainty (*single-low*); (2) single tag with high uncertainty (*single-high*); (3) all tags with low uncertainty (*all-low*); (4) all tags with high uncertainty (*all-high*). In Scenarios 1 and 2, only the pose of the middle tag was subjected to in-plane perturbation, while the other two remained untouched. In Scenarios 3 and 4, however, the poses of all three tags were perturbed. Low and high uncertainties correspond to zero-mean Gaussian error distributions of ( $\sigma_x = \sigma_z = 0.01$  m and  $\sigma_{\theta_y} = 1^\circ$ ) and ( $\sigma_x = \sigma_z = 0.05$  m and  $\sigma_{\theta_y} = 5^\circ$ ), respectively.

In a separate case study, all the tag poses in experiment *3DC* were corrupted by extreme uncertainties to push the estimation methods to their limits. It is expected that highly corrupted measurements cause divergence, large errors, and biases. The reason for investigating this extreme case is to stress-test the robustness of the methods against unusual measurement uncertainties caused by errors in tag installations or pose determinations. The extreme uncertainty corresponds to a zero-mean Gaussian error distribution of ( $\sigma_x = \sigma_y = \sigma_z = 0.05$  m and  $\sigma_{\theta_x} = \sigma_{\theta_y} = \sigma_{\theta_z} = 5^\circ$ ).

## 7 Results

Figure 2 shows the distribution of RMSE values in position estimation using TIE-EKF and EKF methods in the three experiments introduced in Table 1 and for the degrees of uncertainty discussed above. The reported results involve 200 iterations



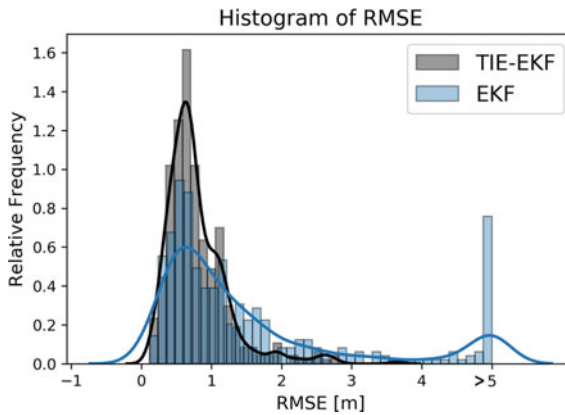
**Fig. 2** Comparison of TIE-EKF and EKF methods with uncertain tag poses

of MCS. In all experiments and almost for all degrees of uncertainty, the TIE-EKF methods resulted in RMSE values with a lower median, max, and spread. However, the minimum value hardly differed regardless of the method used. As seen in Fig. 2, increasing uncertainty resulted in unreliable and biased estimates in all experiments. In two experiments (SLS and 3DC), a single tag with high uncertainty performed worse than when all tags were slightly corrupted, surprisingly. One possible explanation could be that their perturbation had canceled out one another. Overall, the results suggest higher accuracy and improved robustness for TIE-EKF.

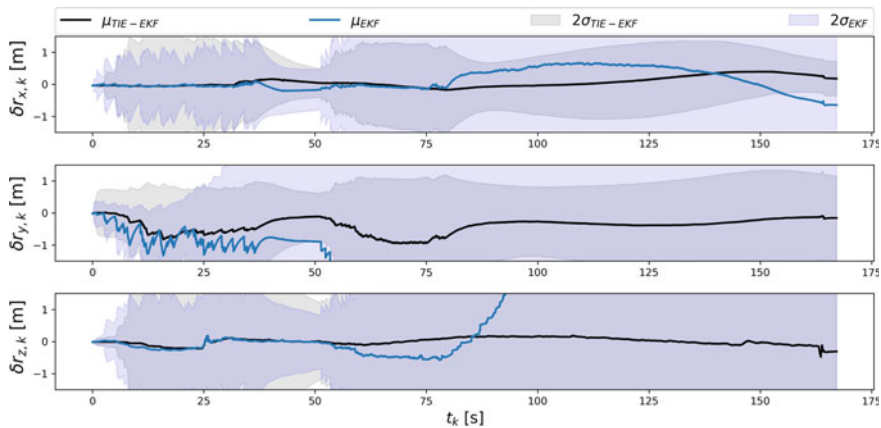
## 7.1 Extreme Case

In *all-high* scenarios discussed above, the tag poses were subjected to significant errors, already resulting in high RMSEs. However, the extreme case investigates the methods' robustness in handling even higher uncertainties in tags' pose. This test explores the estimation outcome when the tag poses are significantly off. Therefore, the high RMSE errors are expected and not a concern here, as the relative performance of the two methods is the objective of this case study.

Figure 3 shows the histogram and density of the RMSE values for TIE-EKF and EKF in an extreme tag installation error scenario. Without considering tag installation errors (EKF), the estimates diverged and resulted in huge errors in position estimates. However, TIE-EKF had a relatively limited spread, with modes and medians closer to zero. Furthermore, Fig. 4 illustrates the errors in  $x$ ,  $y$ , and  $z$  estimates and the  $2\sigma$  envelopes over time, separately. The mean error for EKF, shown in Fig. 4, clearly indicates the divergence of the estimates, whereas the error values for TIE-EKF remained close to zero and did not diverge. Furthermore, the estimation variations for TIE-EKF are significantly lower than those of EKF. The results suggest that TIE-EKF is more robust even in extreme cases.



**Fig. 3** RMSE histogram and density function from 400 iterations of MCS for TIE-EKF and EKF methods in an extreme tag installation error scenario in 3DC [RMSE values greater than 5 m were considered as estimation divergence and consolidated in a single bin ( $> 5$ )]



**Fig. 4** Error in 3D position estimates and  $2\sigma$  envelopes over time (3DC)

In summary, these results show that considering tag installation errors and incorporating the tags' pose uncertainties in the localization formulation can improve the position estimation accuracy by 3–9%, depending on the degree of uncertainty. Furthermore, the extreme case study indicates that an error-aware estimator (TIE-EKF) is more robust against divergence in the presence of large uncertainties. In essence, TIE-EKF first acknowledges that the input tag poses might be subjected to installation errors, making the tag reading measurements uncertain. Therefore, the estimator must rely less on low-quality measurements than when tag poses are more accurate. However, in the case of uncertain tag poses, since they are off from the true pose, the corresponding measurements are off, hence the corrections and consequently the final estimates. In other words, low-quality measurements will result in inaccurate estimates. This study suggests that by only acknowledging this uncertainty, the robustness and accuracy of estimates are relatively improved.

## 8 Conclusion

Autonomous robotic data capture solutions can enhance the efficiency of collection and the quality of required data for downstream automation tasks. Although many solutions have been proposed, they are mainly costly and face technical challenges for localization in indoor construction settings due to perceptual aliasing and feature scarcity. We previously proposed a low-cost, lightweight, versatile, tag-based visual-inertial localization method to tackle these challenges. Tags, in this method, are paper printable landmarks with known locations and sizes. Since tag placement/replacement is a manual process, it is subjected to human errors. This work investigated the impact of human error in the manual tag installation process and the uncertainty in the tags' pose. Using the Lie group theory, a stochastic approach was proposed to account for this uncertainty in indoor localization. Employing Monte Carlo simulation, we experimentally showed that the proposed stochastic model incorporated in our on-manifold formulation improved the robustness and accuracy of tag-based localization against imperfections in manual tag installation on site. However, some limitations should be taken into account. First, tags were always visible in all scenarios and throughout the trajectories. Therefore, the impact of tag-blind zones should be explored in future work. Second, the distribution of tag installation errors was assumed to be zero-mean Gaussian with known statistics. More investigation might be required to determine the underlying distribution of errors and their statistics. We also recommend that future research examines the tag placement problem, considering cost and localizability.

**Acknowledgements** Financial support from the Natural Science and Engineering Research Council (NSERC) grant number RGPIN-2017-06792 is appreciated. The first author is grateful to Prof. Tim Barfoot for his guidance and constructive advice.

## References

1. Adán A, Quintana B, Prieto SA, Bosché F (2020) An autonomous robotic platform for automatic extraction of detailed semantic models of buildings. *Autom Constr* 109:102963
2. Asadi K, Kalkunte Suresh A, Ender A, Gotad S, Maniyar S, Anand S, Noghabaei M, Han K, Lobaton E, Wu T (2020) An integrated UGV-UAV system for construction site data collection. *Autom Constr* 112:103068
3. Asadi K, Ramshankar H, Pullagurta H, Bhandare A, Shanbhag S, Mehta P, Kundu S, Han K, Lobaton E, Wu T (2018) Vision-based integrated mobile robotic system for real-time applications in construction. *Autom Constr* 96:470–482
4. Barfoot TD (2017) State estimation for robotics. Cambridge University Press, Cambridge
5. Boston Dynamics (2021) <https://www.bostondynamics.com/spot>, 17 May 2021
6. Freimuth H, König M (2018) Planning and executing construction inspections with unmanned aerial vehicles. *Autom Constr* 96:540–553
7. Ham Y, Han KK, Lin JJ, Golparvar-Fard M (2016) Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (UAVs): a review of related works. *Vis Eng* 4(1):1
8. Hamledari H, McCabe B, Davari S (2017) Automated computer vision-based detection of components of under-construction indoor partitions. *Autom Constr* 74:78–94
9. Jin M, Liu S, Schiavon S, Spanos C (2018) Automated mobile sensing: towards high-granularity agile indoor environmental quality monitoring. *Build Environ* 127:268–276
10. Kayhani N, Heins A, Zhao W, Nahangi M, McCabe B, Schoellig A (2019) Improved tag-based indoor localization of UAVs using extended Kalman filter. In: Proceedings of the 36th international symposium on automation and robotics in construction, ISARC 2019. Banff, Alberta, Canada, pp 624–631
11. Kayhani N, McCabe B, Abdelaal A, Heins A, Schoellig AP (2020) Tag-based indoor localization of UAVs in construction environments: opportunities and challenges in practice. In: Construction research congress 2020. American Society of Civil Engineers, Reston, pp 226–235
12. Kayhani N, Zhao W, McCabe B, Schoellig AP (2022) Tag-based visual-inertial localization of unmanned aerial vehicles in indoor construction environments using an on-manifold extended Kalman filter. *Autom Constr* 135:104112
13. Kim P, Chen J, Cho YK (2018) SLAM-driven robotic mapping and registration of 3D point clouds. *Autom Constr* 89:38–48
14. Lin JJ, Ibrahim A, Sarwade S, Golparvar-Fard M (2021) Bridge inspection with aerial robots: automating the entire pipeline of visual data capture, 3D mapping, defect detection, analysis, and reporting. *J Comput Civ Eng* 35(2):04020064
15. Mantha BRK, Menassa CC, Kamat VR (2018) Robotic data collection and simulation for evaluation of building retrofit performance. *Autom Constr* 92:88–102
16. McCabe BY, Hamledari H, Shahi A, Zangeneh P, Azar ER (2017) Roles, benefits, and challenges of using UAVs for indoor smart construction applications. In: Computing in civil engineering 2017. American Society of Civil Engineers, Reston, pp 349–357
17. Mostafa K, Hegazy T (2021) Review of image-based analysis and applications in construction. *Autom Constr* 122:103516
18. Olson E (2011) AprilTag: a robust and flexible visual fiducial system. In: Proceedings—IEEE international conference on robotics and automation, pp 3400–3407
19. Siebert S, Teizer J (2014) Mobile 3D mapping for surveying earthwork projects using an unmanned aerial vehicle (UAV) system. *Autom Constr*
20. Solà J, Deray J, Atchuthan D (2018) A micro Lie theory for state estimation in robotics, pp 1–17

# Virtual Reality-Based Expert Demonstrations for Training Construction Robots via Imitation Learning



Lei Huang, Weijia Cai, and Zhengbo Zou

**Abstract** The construction industry faces challenges of skilled labor shortage, low productivity, continuous cost and schedule overruns, and unsafe working conditions for workers. The application of construction robots is a potential solution to alleviate these problems by providing higher productivity, quality, and a safer working environment. Existing construction robotic solutions, such as bricklaying robots and grid drawing robots, are typically programmed to follow specific sequences of instructions designed beforehand. Recent advancements in robot control algorithms like Reinforcement Learning (RL) have enabled robots to adapt to unseen scenarios by learning sequences of optimal actions from videos of expert demonstrations. However, these demonstrations need to be collected on real construction sites, which can be costly and potentially dangerous to experts, especially when multiple demonstrations are required. In this study, we propose a novel approach that leverages Virtual Reality (VR) to collect expert demonstrations, allowing the demonstrator to illustrate the procedure of a construction task using handheld VR controllers. During demonstrations, direct parameters including the robotic arm's joint states, positions, and orientations of the object to be manipulated are extracted from the virtual environment to generate a control policy that imitates the behavior of the expert. We implemented the proposed approach for the task of window installation as validation. The control policy was learned and later applied to a robot arm in a virtual environment. Results show that for all 10 testing cases, the control policy could successfully generate actions given the observed states and lead the robotic arm to first pick up the window and then install it at the target location. These results confirm the effectiveness of the proposed approach in providing virtual demonstrations for the robot to learn a control policy, which eliminates the need for on-site demonstrations from experts, avoiding potentially unsafe scenarios.

**Keywords** Training • Construction robots • Imitation learning

---

L. Huang · W. Cai · Z. Zou (✉)  
University of British Columbia, Vancouver, Canada  
e-mail: [zhengbo@civil.ubc.ca](mailto:zhengbo@civil.ubc.ca)

# 1 Introduction

The construction industry has endured severe labor shortages in recent years [1] and the situation is still being exacerbated by factors such as an aging workforce and the lack of interest from the next generation [2]. It was predicted that the demand for construction workers in the USA would increase by 430,000 in 2021 [3], with around 62% of contractors expressing that they experienced high difficulties in recruiting as of the fourth quarter of 2021 [4]. To accelerate project progress with insufficient labor, construction workers are usually required to be on duty overtime, which causes high pressure and physical fatigue [5] that, in turn, lead to consequences including escalation of safety incidents, poor productivity, and cost and schedule overruns [6]. The adoption of construction robotics is broadly deemed to be a feasible solution to these persistent challenges in the construction industry [7] because robots have the capacity to deliver tasks with potentially improved quality and productivity [8]. Meanwhile, it also has the potential to transform the role of on-site workers from operators to supervisors hence decreasing their exposure to perilous working conditions [9].

Prior research has dived into construction robotics such as bricklaying robots [10] and painting robots [11], some of which have already been put into practice on-site. These robots usually require specialists (e.g., experienced construction workers and mechanical engineers) to pre-define a specific sequence of instructions for each task, which is then translated into an executable program controlling the robot [12]. Although such programs can be executed repeatedly, they lack the flexibility to adapt to unexpected conditions [13]. This is because the carefully designed sequence of instructions lacks the ability to generalize and may fail in new scenarios even if the task is identical [14].

More recently, increasing research has shifted the focus to utilize Reinforcement Learning (RL) to generate more robust and generalized optimal control policies [15], where a robot determines the succeeding action to take based on the current observation (e.g., coordinate, pose, surroundings). To start a training process with a more appropriate initialization, some RL-based methods for construction robots optionally provided their policy with very few demonstrations and then kept optimizing it through trial and error [16], leaving the bottleneck as the training time. To reduce the training time, researchers explored the possibility of collecting a great quantity of videos of expert demonstrations on real sites [17], which were then used as inputs for the RL algorithms to imitate expert's behaviors in the demonstration videos. However, the setup of experiment scenes used to collect video demonstrations is costly, and repeated demonstrations from experts might cause fatigue leading to dangerous situations [18].

To address these issues, we propose an approach that enables expert to demonstrate construction tasks in the Virtual Reality (VR) environment, where the procedure can be illustrated using VR controllers. In the virtual environment, both observations (i.e., states of joints of the robotic arm and the spatial information of objects to manipulate) and actions (i.e., rotation changes of joints) can be conveniently extracted during demonstrations, which are then used to train a control policy that imitates the behavior of the expert based on Imitation Learning (IL). The method is mainly composed of three steps as: virtual environment creation, virtual demonstrations collection, and control policy training and testing.

## 2 Background

In construction, traditional pre-programmed robots have been broadly studied and applied to execute tasks such as bolting steel structures [19] and transporting construction materials [20]. Although practically mature and effective, pre-programmed robots require significant efforts to pre-design sequences of instructions for the robot to execute [9]. For example, Iturralde and Bock [21] proposed a method for robots to assemble timber frames whose trajectories were pre-programmed and provided offline. The main limitation of the pre-programmed robot lies in its tendency to fail in unseen scenarios because it only follows deterministic sequences, hence lacking awareness of the constantly changing surroundings. In addressing this issue, remote-controlled robots through teleoperation were developed to equip robots with flexibility [22]. Nevertheless, teleoperation inherently relies on the intelligence of the human operator and does not provide pathways for robots to carry out tasks independently without human intervention.

To obtain a higher level of automation and flexibility in control, RL-based robots have been introduced to conduct construction tasks in recent years [9]. Instead of following pre-defined instructions, an RL-based robot learns an optimal policy to control itself based on the surroundings (i.e., environment) so that the robot has the capability to adapt to various situation dynamically. Although RL has experienced rapid development in the past decade in regard to designing control policy for robots [23], only limited research explores the feasibility of adopting RL in construction robotics [9]. Due to the difficulty and cost of experimenting directly on real robots on construction sites, existing studies turned to simulation to train their RL control policies, which also applies to our approach.

Among these research efforts, the most widely studied RL-based construction task is assembly (e.g., timber assembly), because of the increasing adoption of off-site modular construction [24], which promises improved quality and higher efficiency [25]. For example, Belousov et al. [16] implemented an RL algorithm (i.e., Deep Deterministic Policy Gradient) [26] to train a control policy in simulation for the placement of a single building block (e.g., rooftop brick), which is a basic subtask of the more complex assembly problem. Apolinarska et al. [27] collected one expert demonstration of the assembly of lap joints in the simulation environment, which was



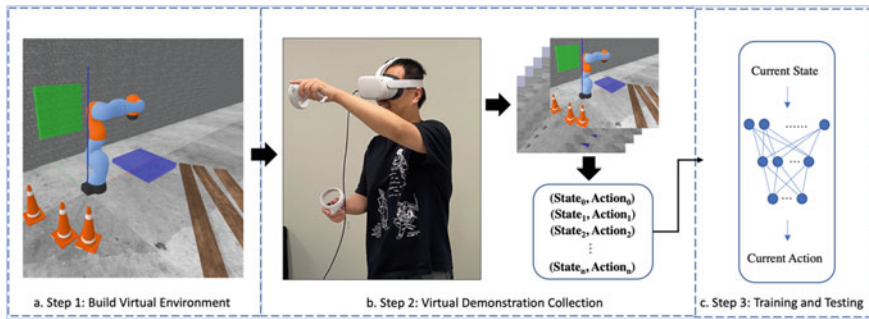
then incorporated into the adopted RL algorithm to train the control policy entirely in simulation. Both of these studies used observations with low dimensions (e.g., robot state, pose of objects, data from sensors, etc.), while the main difference lies in the number of demonstrations needed for the RL agent to learn. The two methods also share common limitations of lengthy training time and instability during training.

Another type of construction task that has been tested with the RL-based robot includes quasi-repetitive tasks such as ceiling installation. For example, Liang et al. [17] collected videos of expert demonstrations both on a real site and in simulation to train a control policy for ceiling installation. However, using images as observations can be problematic since images usually contain irrelevant information such as moving background and changing light [28], and high-dimensional input such as images requires more computation and longer training time.

To overcome time and computation-intensive RL trainings and to avoid irrelevant information in image observations, we propose a novel approach that uses low-dimensional information as observations and adopts the Behavioral Cloning (BC) algorithm [29–31], a simple and efficient form of IL, as the training method. In addition, to avoid exposing the demonstrator to potential dangers resulted from fatigue, we enabled collection of expert demonstrations in the virtual environment for training robotic control policies. Finally, different from prior research that assumes objects which were already picked up and ready for the next step such as placing or installation [16, 17], our method starts with the object (i.e., window) on the ground waiting to be first picked up and then installed.

### 3 Methodology

An IL-based approach is proposed to enable a robotic arm to learn a control policy for conducting a construction task (i.e., window installation) from only virtual demonstrations. The method is composed of three steps. First, a virtual environment containing articles (i.e., a robotic arm, a window, and a target) is built using Pybullet (as shown in Fig. 1a), which is an open-source real-time physics simulation engine that is capable of simulating physical properties such as collision and dynamics [32]. Second, we collect virtual demonstrations in the created virtual environment using a VR headset and a pair of handheld VR controllers (as shown in Fig. 1b). The data are then extracted from virtual demonstrations in the form of state and action pairs. The detailed information about the data is introduced in Sect. 3.2. Third, the state and action pairs are fed into an algorithm (e.g., neural network) that trains a control policy (as shown in Fig. 1c) that can generate actions for the robotic arm given the current state of the robot and surroundings (i.e., the window and the target). The learned control policy is then used for testing in the same environment. The overarching architecture of the proposed method is shown in Fig. 1. Each step is introduced in detail in the following subsections.



**Fig. 1** Overarching architecture of methodology

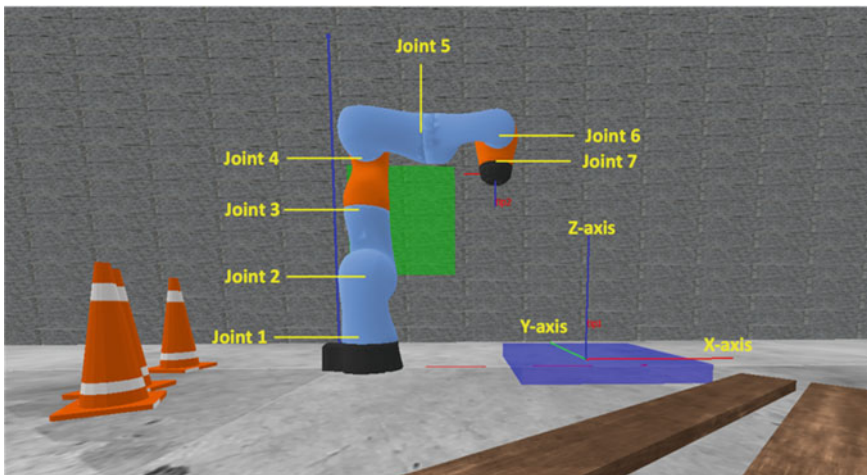
### 3.1 Virtual Environment Creation

The first step is to build a virtual environment of a construction site, with a robotic arm, a window on the ground waiting to be picked up and moved toward the target, and a green transparent cuboid denoting the target. We built the virtual environment in the Pybullet platform [32]. For the robotic arm, we chose the KUKA LBR iiwa robot, as it is one of the most widely used robots in tasks such as object manipulation, painting, and assembly. The KUKA robot has seven axes (i.e., seven joints) and has a load capacity of 7 or 14 kg based on different configurations. The robotic arm was initialized to be located at the origin of the three-dimensional Cartesian world coordinate system.

As shown in Fig. 2, the initial pose of the KUKA robot (i.e., the rotation of each joint) was specified such that the rotation of Joint 4 was 1.57 radians, and the rotation of Joint 6 was  $-1.57$  radians, while the rotations of the other joints were set to zeros. These values allow the robot in the virtual environment to simulate its default initial posture in the real world. The robotic arm can be manipulated by either using the handheld VR controllers during the virtual demonstration collection phase or by taking instructions of each joint's action from the control policy during the testing phase. To ensure that the KUKA robot can operate on the window properly, the window and the target should be within reach of the KUKA robot.

### 3.2 Virtual Demonstration Collection

The second step is to collect virtual demonstrations of experts executing construction tasks in the virtual construction site. The demonstrator was immersed in the virtual environment through a VR headset, viewing the objects including the robotic arm, the window, the target, and two handheld VR controllers (as shown in Fig. 3a). In the virtual environment, the demonstrator first moved a VR controller to the mass center of the end effector of the robotic arm (i.e., the last link denoted in Fig. 3b), which

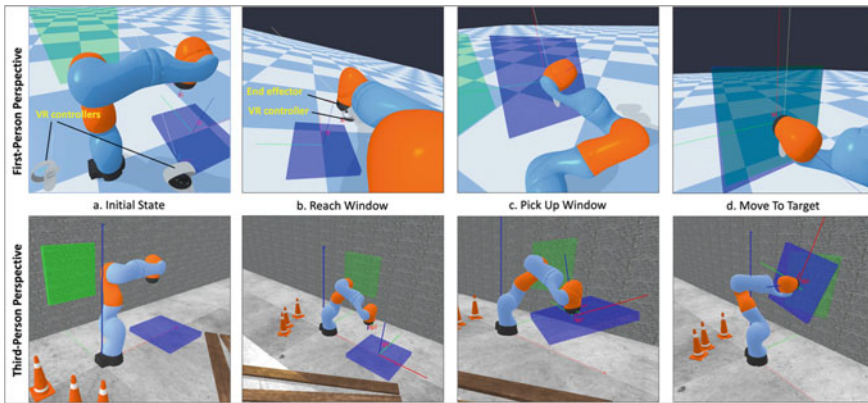


**Fig. 2** KUKA robot and window initial state

automatically attached the controller to the end effector when they overlapped. The robotic arm could then be controlled via the VR controller. To assist this process, there were auxiliary annotations for the demonstrator. In Fig. 3a, b, “tip2” annotates the mass center of the end effector; the red, green, and blue annotate the local  $X$ ,  $Y$ , and  $Z$ -axis of an object respectively. Next, the demonstrator moved the controller to the center of the window, which was annotated with “tip1” (as shown in Fig. 3b). The window can then be picked up by the end effector (as shown in Fig. 3c). The final actions were to move the window steadily toward the target position with an appropriate orientation via the VR controller (as shown in Fig. 3d). For a more intuitive illustration, Fig. 3 shows the whole process from the first-person perspective (i.e., the expert’s perspective) and the third-person perspective, respectively.

The collection process was repeated to collect multiple sequences of data. During each demonstration, the states of all joints of the robotic arm were logged for training the control policy. Each state consists of a seven-dimensional joint reading of the robotic arm, a seven-dimensional position and orientation of the window, and a seven-dimensional position and orientation of the target. To conveniently control the robotic arm in the virtual environment, the position is represented by a three-dimensional coordinate of  $X$ ,  $Y$ , and  $Z$ , and the orientation is represented by a four-dimensional quaternion transformed from the corresponding three-dimensional Euler angles of roll around  $X$ -axis, pitch around  $Y$ -axis, and yaw around  $Z$ -axis.

From the log of states in every virtual demonstration, we retrieved a seven-dimensional action vector at each simulation step (i.e., the rotation changes of all joints of the robotic arm) by subtracting the joint readings of the robotic arm of the current state from those of the next state, essentially allowing the robot to learn an incremental action from its previous state. For the last simulation step of each virtual demonstration, the action was set to all rotation changes being 0 s, indicating the



**Fig. 3** Virtual demonstration collection—first-person perspective versus third-person perspective

end of a demonstration with no further increments. We then fragmented the virtual demonstrations into state-action pairs and combined them into a single large dataset that has 28 columns, with the first 21 columns composing a state, and the other seven columns composing an action. This dataset is the output of the virtual demonstration collection step.

### 3.3 Control Policy Training and Testing

The third step is to train a control policy using the collected dataset from virtual demonstrations. We used IL as the learning framework. The objective of IL is to derive a policy from a set of close-to-optimal demonstrations  $\{\tau_1, \tau_2, \tau_3, \dots\}$ , where each  $\tau$  represents a sequence of state-action pairs in a demonstration  $\{(s_0, a_0), (s_1, a_1), \dots, (s_n, a_n)\}$  [33]. In other words, given the current state  $s$ , the learned policy  $\pi$  can map it to an action  $\pi_\theta(s) \rightarrow a$  that behaves like an expert's policy  $\pi^*$  (i.e., the optimal policy) [30]. The  $\theta$  represents the learned parameters of the model. In our experiment, we assume that the observations are the states (i.e., the first 21 columns of the output dataset from 3.2).

The IL algorithm of choice is BC, which is one of the two categories of IL, along with Inverse Reinforcement Learning (IRL). BC is adopted in our approach because of its high efficiency for training. BC reduces the RL problem to a supervised learning problem [31, 34], where the input is the observed state, and the prediction target is the action. This is because in BC, the state-action pairs are assumed to be independent and identically distributed (i.i.d.). These samples are then stacked together, shuffled, and used as the training dataset for the supervised learning problem.

To solve the supervised learning problem, various models can be used such as Multilayer Perceptron (MLP) and Multi-output Decision Tree Regression (MDTR). In this study, we used a MDTR for the window pick up and a MLP for the

window installation, because of their efficiency in training and ability to consider a vast number of possible actions in the search space. Our algorithm uses 21-dimensional states as inputs and outputs seven-dimensional actions. The loss function implemented during training was Mean Squared Error (1):

$$L(\mathbf{a}^*, \pi_{\theta}(s)) = \frac{1}{n} \sum_{i=1}^n (a_i^* - \pi_{\theta}(s_i))^2, \tag{1}$$

where  $\mathbf{a}^*$  represents seven-dimensional optimal actions from virtual demonstrations, essentially the ground truth of the supervised learning problem;  $s$  represents the 21-dimensional input states; and  $\pi_{\theta}$  represents the control policy with parameters  $\theta$ . After obtaining the trained policy, we applied the policy back to the virtual environment to test its effectiveness.

### 4 Results and Discussion

To validate the proposed approach of using BC to learn from virtual demonstrations collected in VR, we tested the approach for the task of picking up and installing a window on a virtual construction site. After setting up the robotic arm, the window, and the target in the virtual environment, we collected a total of ten virtual demonstrations. The data collection was done as an alpha test within the research group to validate the viability of the approach. The number of state–action pairs in each demonstration is shown in Table 1. The difference in the number of state–action pairs across demonstrations indicates that the demonstrations were conducted considering various possible trajectories of the robot arm, allowing the IL agent to generalize in unseen circumstances.

**Table 1** Sizes of virtual demonstrations datasets

Virtual demonstration index	Number of state–action pairs (rows)
Demo 1	6224
Demo 2	6876
Demo 3	5833
Demo 4	7851
Demo 5	7000
Demo 6	6093
Demo 7	6613
Demo 8	7206
Demo 9	7639
Demo 10	5794
Total	67,129

Following the data processing steps in Sect. 3.2, we obtained a dataset of 67,129 rows (i.e., 67,129 state-action pairs) and 28 columns as input for the BC algorithm. A view of states is shown in Table 2. In Table 2, Joint1–Joint7 represent the states of the seven joints of the robotic arm; W- $x$ , W- $y$ , and W- $z$  represent the coordinate of mass center of the window; W-Ori- $x$ , W-Ori- $y$ , W-Ori- $z$ , and W-Ori- $w$  represent the rotation of the window in the form of quaternion; T- $x$ , T- $y$ , and T- $z$  represent the coordinate of mass center of the target; T-Ori- $x$ , T-Ori- $y$ , T-Ori- $z$ , and T-Ori- $w$  represent the rotation of the target in the form of quaternion. A view of actions is shown in Table 3. In Table 3, Joint1–Joint7 represent the rotation changes of the seven joints of the robotic arm.

Considering that the whole process of the window installation task is composed of two stages: window pick up and window installation, we trained BC using two models for the two stages, respectively. For each stage, we trained both MLP and MDTR models using the scikit-learn library [35], then we chose the model with the better performance. After experimenting, we adopted MDTR for the window pick-up stage and MLP for the window installation stage.

After model selection, we trained the MDTR and MLP on the entire dataset. Then, we applied the trained models to the robotic arm to test the effectiveness of our approach. We repeated the test 10 times and all of them succeeded in the window pick up and the window installation tasks (with some offsets). For each test, we recorded the states of the window in the last 50 simulation steps to calculate the offsets between the window and the target, because the window and the robotic arm were already stabilized during this period. Considering the success rate, the models had a good performance overall. Table 4 displays the average values of the offsets with respect to the position (i.e., coordinates  $X$ ,  $Y$ , and  $Z$ ) and the orientation (i.e., quaternions  $i$ ,  $j$ ,  $k$ , and  $1$ ). The smallest error is in the  $Z$  value of the coordinate, with a value of 0.038. Overall, offsets of the coordinate are smaller than those of the orientation. We expected that the orientation information was less accurate than the coordinate information because, in the virtual demonstration collection step, we observed that applying sequences of data back to the robotic arm could not exactly recover the expert demonstration, especially that the orientation information was corrupted with some noise.

We validated the feasibility of the proposed approach by showing that the robotic arm could successfully pick up the window and install it at the target location. For the offsets shown in Table 4, we consider that it is largely due to the limitations of the BC algorithm and noise in the data from the virtual environment. Regarding the algorithm, BC provides efficiency in our approach, but this algorithm treats all states as i.i.d. samples, which is not true in many scenarios because a state is very likely to be correlated with other states that are one or more steps back. Thus, variants of BC that take earlier states into consideration can be one way to improve the performance of the robotic arm.

Table 2 Examples of states

Sample index	Joint1 (rad)	Joint2 (rad)	Joint3 (rad)	Joint4 (rad)	Joint5 (rad)	Joint6 (rad)	Joint7 (rad)
1	3.054e-7	2.000e-5	- 2.793e-3	- 1.571	2.891e-20	1.571	4.000e-6
2	3.054e-7	2.000e-5	- 2.793e-3	- 1.571	2.891e-20	1.571	4.166e-1
3	3.054e-7	1.733e-3	- 2.793e-3	- 1.570	2.891e-20	1.571	4.166e-1
4	3.054e-7	1.733e-3	- 2.793e-3	- 1.570	2.891e-20	1.571	8.333e-1
Sample index	W-x (m)	W-y (m)	W-z (m)	W-Ori-x (rad)	W-Ori-y (rad)	W-Ori-z (rad)	W-Ori-w (rad)
1	8.000e-1	3.300e-5	2.499e-2	- 9.519e-7	4.010e-8	3.590e-4	1.000
2	8.000e-1	3.300e-5	2.499e-2	- 9.519e-7	4.010e-8	3.590e-4	1.000
3	8.000e-1	3.300e-5	2.499e-2	- 9.519e-7	4.010e-8	3.590e-4	1.000
4	8.000e-1	3.300e-5	2.499e-2	- 9.519e-7	4.010e-8	3.590e-4	1.000
Sample index	T-x (m)	T-y (m)	T-z (m)	T-Ori-x (rad)	T-Ori-y (rad)	T-Ori-z (rad)	T-Ori-w (rad)
1	1.000e-1	6.000e-1	6.000e-1	7.071e-1	0.000	0.000	7.071e-1
2	1.000e-1	6.000e-1	6.000e-1	7.071e-1	0.000	0.000	7.071e-1
3	1.000e-1	6.000e-1	6.000e-1	7.071e-1	0.000	0.000	7.071e-1
4	1.000e-1	6.000e-1	6.000e-1	7.071e-1	0.000	0.000	7.071e-1

**Table 3** Examples of actions

Sample index	Joint1 (rad)	Joint2 (rad)	Joint3 (rad)	Joint4 (rad)	Joint5 (rad)	Joint6 (rad)	Joint7 (rad)
1	0.000	0.000	0.000	0.000	0.000	0.000	4.000e−6
2	− 3.854e−20	1.714e−3	0.000	4.840e−4	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	− 2.435e−21	3.563e−3	0.000	2.500e−3	0.000	3.000e−6	4.167e−1

## 5 Conclusion and Future Work

In this study, we proposed an approach to collect expert demonstrations in the virtual environment by manipulating VR controllers. The collected virtual demonstrations were then used to train an optimal control policy via IL, specifically, using BC. Results showed that the learned policy could successfully control the robotic arm to reach the center of the window, pick it up, and install it, although there existed minor offsets between the final state of the window and the target, which is largely due to the noise of data during virtual demonstration collection and the limitations of BC algorithm despite its efficiency.

The main contribution of this study is that we provided an approach to collect demonstrations in the virtual environment, which eliminated the need for on-site demonstrations from the experts. We also validated the effectiveness of the approach using virtual demonstrations to train a control policy that controls a robotic arm to conduct construction tasks. This study lays a concrete first step toward a higher level of automation in the construction industry using robotics. Researchers and practitioners can use the proposed approach to easily scale up the virtual demonstration and allow the training of various construction tasks that can benefit from a full automated approach, which has the potential to improve the pace of adoption for robotic solutions in construction.

For future work, we aim to collect a more comprehensive set of demonstrations and utilize data augmentation techniques to make the robotic arm more robust to different state scenarios. Another thread of future work is to explore interactions between the sequence of states and actions generated by the control policy; therefore, the robotic arm control agent can correct the trajectory when it deviates from ideal cases. Finally, we will expand the current study by testing with different types of observations/states and algorithms for various construction tasks.



**Table 4** Errors between the final window state and the target

Pos/Ori	Pos- <i>X</i> (m)	Pos- <i>Y</i> (m)	Pos- <i>Z</i> (m)	Ori- <i>X</i> (quaternion)	Ori- <i>Y</i> (quaternion)	Ori- <i>Z</i> (quaternion)	Ori- <i>W</i> (quaternion)
Offset	0.152	0.140	0.038	0.454	0.500	0.575	0.111

## References

1. Kim S, Chang S, Castro-Lacouture D (2020) Dynamic modeling for analyzing impacts of skilled labor shortage on construction project management. *J Manag Eng* 36(1):04019035
2. Tonnon SC, Van Der Veen R, Westerman MJ, Robroek SJ, Van Der Ploeg HP, Van Der Beek AJ, Proper KI (2017) The employer perspective on sustainable employability in the construction industry. *J Occup Environ Med* 59(1):85–91
3. Associated Builders and Contractors (ABC) (2021) Construction spending and employment: history and forecast terms and sources. Retrieved from: <https://www.abc.org/Portals/1/News%20Releases/explainer-2021.pdf>. Accessed on 28 Feb 2022
4. U.S. Chamber of Commerce (USCC) (2021) Q4 2021 Construction commercial construction index. Retrieved from: <https://www.uschamber.com/assets/documents/Q4-2021-CCI-Report.pdf>. Accessed on 28 Feb 2022
5. Karimi H, Taylor TR, Goodrum PM, Srinivasan C (2016) Quantitative analysis of the impact of craft worker availability on construction project safety performance. *Construct Innov*
6. Karimi H, Taylor TR, Dadi GB, Goodrum PM, Srinivasan C (2018) Impact of skilled labor availability on construction project cost performance. *J Constr Eng Manag* 144(7):04018057
7. Pradhananga P, ElZomor M, Santi Kasabdj G (2021) Identifying the challenges to adopting robotics in the US construction industry. *J Constr Eng Manag* 147(5):05021003
8. de Soto BG, Agustí-Juan I, Hunhevicz J, Joss S, Graser K, Habert G, Adey BT (2018) Productivity of digital fabrication in construction: cost and time analysis of a robotically built wall. *Autom Constr* 92:297–311
9. Liang CJ, Wang X, Kamat VR, Menassa CC (2021) Human-robot collaboration in construction: classification and research trends. *J Constr Eng Manag* 147(10):03121006
10. Sklar J (2015) Robots lay three times as many bricks as construction workers. In: MIT technology review
11. Seriani S, Cortellessa A, Belfio S, Sortino M, Totis G, Gallina P (2015) Automatic path-planning algorithm for realistic decorative robotic painting. *Autom Constr* 56:67–75
12. King N, Bechthold M, Kane A, Michalatos P (2014) Robotic tile placement: tools, techniques and feasibility. *Autom Constr* 39:161–166
13. Cully A, Clune J, Tarapore D, Mouret JB (2015) Robots that can adapt like animals. *Nature* 521(7553):503–507
14. Carlson J, Murphy RR (2005) How UGVs physically fail in the field. *IEEE Trans Rob* 21(3):423–437
15. Levine S (2018) Reinforcement learning and control as probabilistic inference: tutorial and review. arXiv preprint [arXiv:1805.00909](https://arxiv.org/abs/1805.00909)
16. Belousov B, Wibranek B, Schneider J, Schneider T, Chalvatzaki G, Peters J, Tessmann O (2022) Robotic architectural assembly with tactile skills: simulation and optimization. *Autom Constr* 133:104006
17. Liang CJ, Kamat VR, Menassa CC (2020) Teaching robots to perform quasi-repetitive construction tasks through human demonstration. *Autom Constr* 120:103370
18. Yu Y, Li H, Yang X, Kong L, Luo X, Wong AY (2019) An automatic and non-invasive physical fatigue assessment method for construction workers. *Autom Constr* 103:1–12
19. Chu B, Jung K, Lim MT, Hong D (2013) Robot-based construction automation: an application to steel beam assembly (part I). *Autom Constr* 32:46–61
20. Jung K, Chu B, Hong D (2013) Robot-based construction automation: an application to steel beam assembly (part II). *Autom Constr* 32:62–79
21. Iturralde K, Bock T (2018) Integrated, automated and robotic process for building upgrading with prefabricated modules. In: ISARC. Proceedings of the international symposium on automation and robotics in construction, vol 35. IAARC Publications, pp 1–8
22. Khasawneh A, Rogers H, Bertrand J, Madathil KC, Gramopadhye A (2019) Human adaptation to latency in teleoperated multi-robot human-agent search and rescue teams. *Autom Constr* 99:265–277

23. Ibarz J, Tan J, Finn C, Kalakrishnan M, Pastor P, Levine S (2021) How to train your robot with deep reinforcement learning: lessons we have learned. *Int J Robot Res* 40(4–5):698–721
24. Yin X, Liu H, Chen Y, Al-Hussein M (2019) Building information modelling for off-site construction: review and future directions. *Autom Constr* 101:72–91
25. Kamali M, Hewage K (2016) Life cycle performance of modular buildings: a critical review. *Renew Sustain Energy Rev* 62:1171–1183
26. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv preprint [arXiv:1509.02971](https://arxiv.org/abs/1509.02971)
27. Apolinarska AA, Pacher M, Li H, Cote N, Pastrana R, Gramazio F, Kohler M (2021) Robotic assembly of timber joints using reinforcement learning. *Autom Constr* 125:103569
28. Zhang A, McAllister R, Calandra R, Gal Y, Levine S (2020) Learning invariant representations for reinforcement learning without reconstruction. arXiv preprint [arXiv:2006.10742](https://arxiv.org/abs/2006.10742)
29. Florence P, Lynch C, Zeng A, Ramirez OA, Wahid A, Downs L, Wong A, Lee J, Mordatch I, Tompson J (2022) Implicit behavioral cloning. In: *Conference on robot learning*. PMLR, pp 158–168
30. Torabi F, Warnell G, Stone P (2018) Behavioral cloning from observation. In: *Proceedings of the 27th international joint conference on artificial intelligence*, pp 4950–4957
31. Pomerleau DA (1988) Alvin: an autonomous land vehicle in a neural network. In: *Advances in neural information processing systems*, vol 1
32. Coumans E, Bai Y (2016) Pybullet, a python module for physics simulation for games, robotics and machine learning
33. Le H, Jiang N, Agarwal A, Dudik M, Yue Y, Daumé H III (2018) Hierarchical imitation and reinforcement learning. In: *International conference on machine learning*. PMLR, pp 2917–2926
34. Ross S, Bagnell D (2010) Efficient reductions for imitation learning. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR workshop and conference proceedings*, pp 661–668
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830

# The Effect of Human Body Blockage on UWB Tracking Accuracy in Construction Sites



Pegah Behvarmanesh and Farnaz Sadeghpour

**Abstract** Statistics reveal that construction has one of the highest fatality rates among all industries. A solution to improve construction sites' safety and avoid accidents is the precise localization of personnel and equipment. Tracking with Ultra-Wideband (UWB) has proved to be suitable for construction site applications due to its high location estimation accuracy, high signal penetration, and low power consumption. However, if the line-of-sight (LOS) between the tag and the anchors is blocked, the location estimation accuracy deteriorates. Due to the dynamic nature of the construction sites, the workers and personnel can easily block the LOS and create non-line-of-sight (NLOS) or quasi-line-of-sight (quasi-NLOS) conditions, which can lead to high errors in positioning accuracy. However, it should be determined how significantly NLOS and quasi-NLOS conditions affect the accuracy. Accordingly, the objective of this study is to investigate the effect of human body blockage on location estimation accuracy by conducting some experiments. The impact of the number of the human blocks and their position with respect to the anchors and the tag is evaluated in NLOS and quasi-NLOS scenarios. 2D and 3D accuracies are calculated to compare the estimated locations and the actual ones. The results indicated that placing the anchors at an elevation higher than the height of human blocks and keeping the human blocks closer to the anchors in comparison with the tag provides more accurate location estimations.

**Keywords** Human body blockage · UWB tracking accuracy

---

P. Behvarmanesh (✉) · F. Sadeghpour  
University of Calgary, Calgary, Canada  
e-mail: [pegah.behvarmanesh@ucalgary.ca](mailto:pegah.behvarmanesh@ucalgary.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_5](https://doi.org/10.1007/978-3-031-34593-7_5)

## 1 Introduction

Real-Time Location Systems (RTLS) are becoming ubiquitous for tracking purposes. There are several applications for RTLS; for instance, RTLSs are employed for monitoring the elderly people and patients in the hospitals [7, 9], tracking the resources in retail, manufacturing, and industrial environments [2, 3, 5], performing time-motion analysis in sports [11], and tracking the equipment, materials, and workers on construction sites [1, 16, 28]. Tracking the resources including workers, materials, and equipment on construction sites can help to improve the safety. Construction sites involve indoor and outdoor environments. The global positioning system (GPS) is the most common RTLS for outdoor areas. However, it is not efficient to use GPS for indoor areas as its location estimation accuracy drops significantly as the signals are blocked by obstacles, such as roofs and walls. There are a number of indoor positioning systems such as Ultra-Wide Band (UWB), Radio Frequency Identification (RFID), ZigBee, and Wi-Fi [14, 17]. Compared to other systems, UWB has shown to be more suitable for applications in construction sites due to its high location estimation accuracy, high signal penetration, and low power consumption [1, 23].

Different methods, such as Angle-of-Arrival (AOA) and Time-Difference-of-Arrival (TDOA), have been used in prior investigations of UWB systems to estimate the location. These algorithms, however, have several drawbacks that prevents precise localization, such as the need for timing cables to connect the network in TDOA [8]. In addition, a tag should be designated as the calibration reference for the AOA and TDOA systems, which must always maintain LOS with the anchors [4, 22]. Time-of-Flight (TOF) is another location estimation method that has been recently made available for UWB systems. TOF is based on two-way ranging which means that it does not require timing cables and reference tags.

Construction sites are usually crowded environments with a large number of workers and other resources present. The materials and equipment can interfere with the UWB signals and cause refraction, diffraction, reflection, and scattering, which will lead to inaccurate estimations [12, 19]. In addition, different obstacles including people can obstruct the direct path of UWB propagation and create non-line-of-sight (NLOS) circumstances. Several studies have investigated the impact of steel, wooden, and concrete obstacles, and other common building materials on UWB signals' propagation [18, 20]. On the other hand, the presence of people between the tags and anchors can increase path loss and error, a phenomenon known as human body shadowing [6]. The objective of this study is to determine the impact of human body blockage on UWB performance in indoor environments. This study investigates the effect of three variables on location estimation accuracy: (1) NLOS and quasi-NLOS conditions, (2) closeness of the human blockage to the tag or the anchors, and (3) number of the human blocks. Experiments will be carried out employing TOF-based UWB under two scenarios with different degrees of obstruction imposed by the human body: NLOS where human blocks completely block the LOS and

quasi-LOS conditions where they partially block the LOS. The location estimation accuracy of the UWB system will be measured. The findings of this study can be used in construction sites to improve the accuracy of location estimation while tracking workers and assets.

## 2 Literature Review

Determination of the direct path of the signal propagation plays an important role in the accurate location estimation of the UWB RTLS. The presence of obstacles can influence the signal propagation as it blocks the LOS as well as the direct path. Refraction, diffraction, reflection, and scattering are the most probable phenomena that can affect the UWB signals. Refraction is the change in the direction of a wave passing from one medium into another with a different speed. Diffraction is when the wave bends around the corners of an object. Reflection occurs when the direction of the wave changes as it hits the boundary between two mediums and the wave returns to the first medium. Scattering is the change in the direction of the wave when it collides with another object. The dominant phenomenon on construction sites is the reflection of UWB signals from metallic surfaces of steel structures or heavy equipment, which will lead to inaccurate location estimation [19, 21].

Different materials on construction sites can block the LOS between the UWB receivers and transmitters, which causes positioning errors. Shahi et al. [18] investigated the effect of steel and wooden obstacles on UWB performance by placing a UWB tag in closed wooden and steel boxes. By comparing the results with the no enclosure case, they found out that the error of location estimation for the wooden box was similar to the no enclosure case and the errors were less than 15 cm. However, the error for the steel box was higher, and it was about 45 cm. It was concluded that blocking the UWB signals by steel influences the accuracy more significantly than wooden materials. In another study, Sudhakar and Madhavi [20] conducted some experiments to determine the UWB signal attenuation when passing through obstacles made of clay brick, plywood, glass, and cloth, which are common materials found on construction sites. They stated that clay brick led to the lowest attenuation among these materials. It could be expressed that the effect of occlusion on UWB accuracy should be considered while tracking the resources on construction sites.

In addition to the equipment and materials, workers and personnel on the construction site can block the UWB signals and influence the location estimation accuracy. The effect of the human body on UWB signals has been previously investigated in the literature. Some studies have inspected the human body shadowing effect while using body-mounted sensors. Trogh et al. [25] confirmed that body-worn nodes (tags) are influenced by the human body. They proposed using more than one body-worn node to decrease the location tracking error as if one node loses the LOS because of the human body orientation, other nodes can still communicate with the anchor. Otim et al. [15] investigated the impact of body wearable sensors' positions on UWB ranging (not location estimation). They conducted experiments with the UWB

sensors on seven parts of the human body and found that placing the wearable sensor on the forehead results in the best ranging accuracy as it can better maintain the LOS with anchors. Similarly, Vorobyov and Yarovoy [26] determined the signal loss in the head, torso, and belt. They indicated that placing the sensor on the head provides the minimum signal loss. Maalek and Sadeghpour [13] indicated that attaching a single tag on top of the head provides the highest accuracy compared with the cases when the tag is attached to the shoulders. In another study, Tian et al. [24] determined how the relative heading angle (RHA) between the tag and the anchor influences the ranging error, and they claimed that the human body does not affect the ranging accuracy while the tag and the anchor maintain the LOS. Zasowski et al. [27] investigated the transmission of UWB signals through the head. They found out that the direct path is lost, and the signal is diffracted. This study focuses on the impact of quasi-NLOS and NLOS conditions, the position of the human blocks, and their number on the location estimation accuracy of the UWB.

### 3 Methodology: Measuring the Effect of Human Body Block on UWB Performance

UWB has the better performance in the LOS condition followed by quasi-NLOS, and it has the highest errors in the NLOS condition. In construction sites, it is hardly possible that the tags maintain the LOS with all the anchors at all times. In this study, the impact of human body blockage is studied under two scenarios of quasi-NLOS and NLOS conditions. This study aims to determine how significantly the accuracy is influenced by the quasi-NLOS condition compared to the NLOS condition. It is speculated that the position of the human block (close to tag or anchor) will affect the accuracy. Accordingly, there are two subsets under each scenario: subset A, where human blocks are standing close to the anchors and subset B, where human blocks are standing close to the tag. Each scenario also includes a benchmark experiment with no blockage in order to employ them as a reference and compare the results of the experiments with them. Each scenario includes 25 experiments: 12 experiments in subset A, 12 experiments in subset B, and one benchmark. There are four anchors and one tag in the experiments as illustrated in the room layout in Fig. 1. Under each subset, anchors are blocked increasingly from one anchor blocked at a time to all four anchors blocked to determine the effect of the number of human blocks. Due to the symmetry in the shape of the room and layout of the anchors, this study investigates four combinations of blocking one anchor, three combinations of blocking two anchors, four combinations of blocking three anchors, and the case where all four anchors are blocked. The details for the 50 experiments are summarized in Table 1.

To assess the accuracy of UWB tracking, the location estimations obtained from the TOF-based UWB system are compared to the ground truth. There are different approaches to evaluate the performance of the UWB positioning in the literature, such as offset, absolute error, and precision. This study employs Distance Root Mean

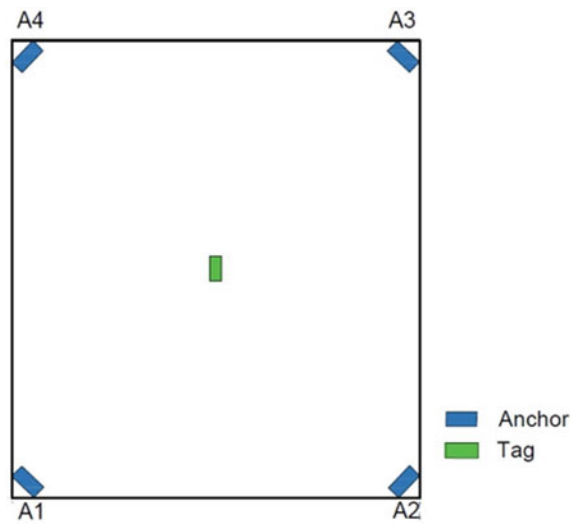


Fig. 1 Experiment layout

Table 1 Experiments details

Experiment No.		Experiment No.						
Scenario 1 (Quasi-NLOS)		Scenario 2 (NLOS)		# of Blocked Anchors	Blocked Anchors			
Subset A (Close to anchor)	Subset B (Close to tag)	Subset A (Close to anchor)	Subset B (Close to tag)		A1	A2	A3	A4
2	14	27	39	1	✓			
3	15	28	40	1		✓		
4	16	29	41	1			✓	
5	17	30	42	1				✓
6	18	31	43	2	✓	✓		
7	19	32	44	2	✓		✓	
8	20	33	45	2	✓			✓
9	21	34	46	3	✓	✓	✓	
10	22	35	47	3	✓	✓		✓
11	23	36	48	3		✓	✓	✓
12	24	37	49	3	✓		✓	✓
13	25	38	50	4	✓	✓	✓	✓



Squared (DRMS) and Mean Radial Spherical Error (MRSE) as they combine offset and precision metrics to provide a single value for 2D and 3D accuracies, respectively [10]. DRMS and MRSE are calculated as follows:

$$\text{DRMS} = \sqrt{\frac{\sum_{i=1}^n (x_i - x_{\text{Actual}})^2}{n} + \frac{\sum_{i=1}^n (y_i - y_{\text{Actual}})^2}{n}} \quad (1)$$

$$\text{MRSE} = \sqrt{\frac{\sum_{i=1}^n (x_i - x_{\text{Actual}})^2}{n} + \frac{\sum_{i=1}^n (y_i - y_{\text{Actual}})^2}{n} + \frac{\sum_{i=1}^n (z_i - z_{\text{Actual}})^2}{n}} \quad (2)$$

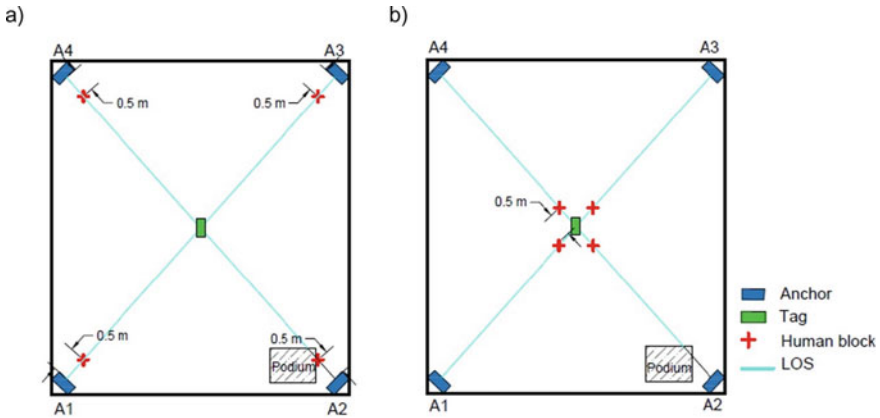
where  $x_i$ ,  $y_i$ , and  $z_i$  are the UWB estimations that are compared with the ground truth; and  $n$  is the number of readings for a specific location point. Around 2000 location estimation readings were recorded for each of the experiments. The data processing was conducted using Python. The next section outlines the experiment settings and data collection.

## 4 Experiments Setup and Data Collection

As mentioned, experiments in this study are conducted under two scenarios: (1) quasi-NLOS and (2) NLOS. There are two subsets in each scenario: subset A, where human blocks are standing 0.5 m away from the anchors on LOS (close to the anchors) and subset B, where human blocks are standing at the distance of 0.5 m from the tag (close to the tag). Figure 2 shows the position of the human blocks in subsets A and B. The experiments are conducted in a room of approximately 6 m  $\times$  7 m. Decawave DWM1001 modules are used to create a network of four anchors and one tag. The DWM1001 modules are transceivers which means that they can perform either as receivers or transmitters. This study defines the transmitter as tags and the receivers as anchors. There are four anchors mounted on the walls in the corners of the room with an angle of 45°. A tag is placed in the middle of the room. The actual location (ground truth) for the anchors and the tag are determined by a Leica TCR 803 Total Station. The room layout and the position of the tag and the anchors are illustrated in Fig. 2.

### 4.1 Scenario 1: Quasi-NLOS Condition

This scenario includes 25 experiments. The four anchors are placed at the height of 2.4 from the floor during this set, and the tag is placed at the middle of the room at the height of 0.80 m from the floor. In this scenario, the height of the anchors is higher than the average height of the human blocks; therefore, they partially block the LOS



**Fig. 2** Position of human blocks: **a** subset A: close to the anchors and **b** subset B: closer to the tag

which is called the quasi-NLOS condition. The human blocks had to climb on a chair or a desk to create the quasi-NLOS condition. However, on a construction site or in a larger room, where human blocks or workers can stand further from the anchors, they would completely block the LOS while standing on the floor. The experiments are designed to determine the effect of the number of human body blockages and their distance from the tag and the anchors on UWB performance. As mentioned before, there are two subsets for this scenario. Experiments 2–13 and 14–25 are included in subset A and subset B, respectively. Experiment 1 is the benchmark as there are no human blocks in the room and all the anchors maintain the LOS with the tag. In experiments 2–5, there is only one human block present in the room at a time, and the LOS of each anchor is blocked separately. Experiments 6–8 and 9–12 investigate different combinations of simultaneously blocking two and three anchors, respectively. In experiment 13, the LOSs of the four anchors are blocked.

## 4.2 Scenario 2: NLOS Condition

This scenario aims to determine the effect of human blocks on location estimation accuracy, while they are completely blocking the LOS. The anchors are placed at a lower elevation of 1.5 m from the floor, so the human blocks can create the NLOS condition. All the settings and conditions of the experiments are similar to the first scenario, except for the elevation of the anchors and the degree of blocking the LOS. The settings of experiments 26–50 are indicated in Table 1. Figure 3 shows the settings for experiment 31 where the LOS of the anchors A1 and A2 is blocked by two human blocks.



**Fig. 3** Two people blocking LOS of two anchors (experiment 31)

## 5 Results and Analysis

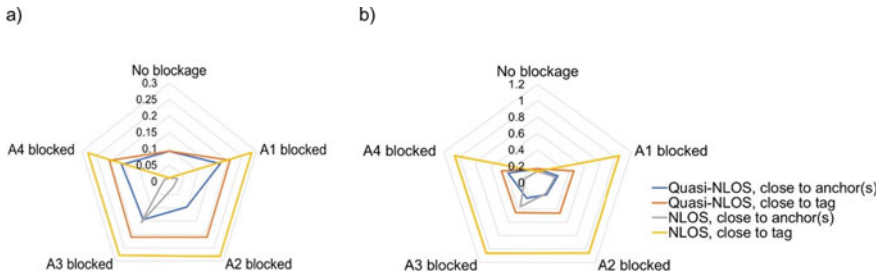
The results of the experiments are analyzed based on the number of the anchors with blocked LOS, and they are presented in four groups as follows:

- Quasi-NLOS where human blocks are placed close to the anchors
- Quasi-NLOS where human blocks are placed close to the tag
- NLOS where human blocks are placed close to the anchors
- NLOS where human blocks are placed close to the tag.

The result of each experiment is compared with the benchmark experiment where there is no blockage in the LOS. Spider diagrams are used to demonstrate the location estimation accuracy of the experiments. The advantage of using the spider diagram is that it could be easily distinguished if the accuracy error is biased toward a specific anchor.

### 5.1 One-Anchor Combinations

There was a total of 16 experiments with one human block combination in this study. Figure 4 summarizes the results of the experiments with one anchor blocked. As shown in the figure, the NLOS cases with a human block standing closer to the tag result in the highest error. On the other hand, the NLOS condition where the human



**Fig. 4** Location estimation accuracy for one anchor blocked **a** DRMS and **b** MRSE

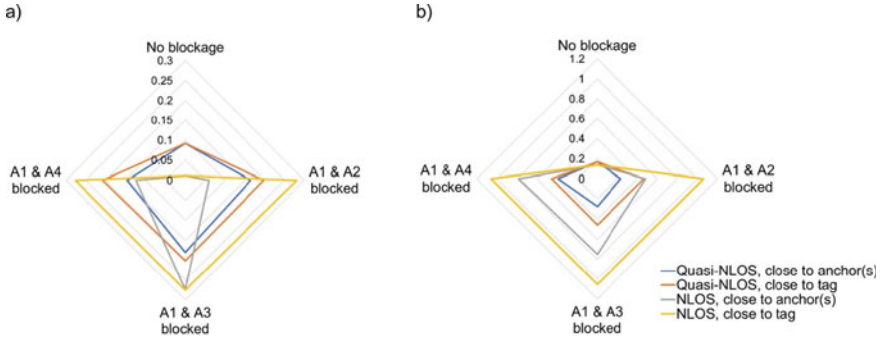
block is placed closer to the anchor has provided the most accurate location estimation in both 2D and 3D. Therefore, it could be stated that placing the block closer to the anchor, which is the receiver even in NLOS condition, provides better accuracy. As expected, 3D accuracy is worse than 2D because UWB has a weakness in the  $z$ -direction. In 2D, the error is less than 30 cm in all the combinations and the error of the benchmark experiment is about 10 cm. Similarly, in 3D, all the combinations except the *NLOS, close to tag* has provided an error of less than 30 cm. *NLOS, close to tag* has resulted in an error of 1.05 m indicating that NLOS condition and placing the block close to the tag is the worst case for location estimation accuracy.

## 5.2 Two-Anchor Combinations

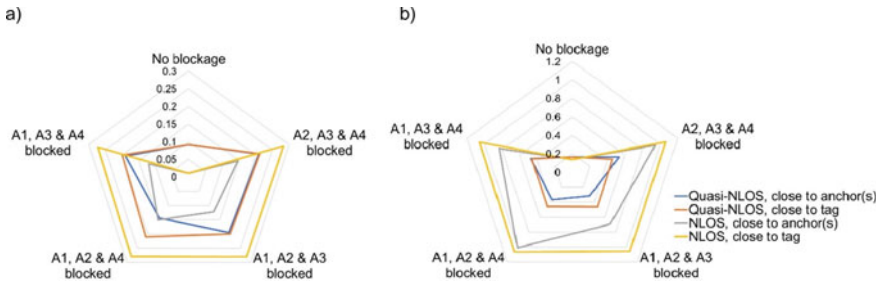
There are 12 experiments with two anchors with blocked LOS. In the cases where A1 and A2 or A1 and A4 are blocked, two adjacent anchors' LOS is blocked. However, the LOS of the two anchors along the diagonal is blocked by placing human blocks between the tag and anchors A1 and A3. As Fig. 5 shows, the NLOS condition where human blocks are closer to the tag leads to the highest error. In 2D analysis, the combination of the lower elevation of the anchors (NLOS) and blocks standing closer to the anchors has the lowest error, except for the case of blocking A1 and A3. This is justifiable as the distance between A1 and A3 is longer than the distance between two adjacent tags. However, for 3D analysis, the quasi-NLOS condition has provided more accurate location estimations. In general, MRSE is more than DRMS, which means that the system has a better performance in 2D.

## 5.3 Three-Anchor Combinations

There are 16 experiments where the LOSs of three anchors are blocked. Figure 6 indicates that the location estimation error is the highest for the NLOS condition



**Fig. 5** Location estimation accuracy for two anchors blocked: **a** DRMS and **b** MRSE



**Fig. 6** Location estimation accuracy for three anchors blocked **a** DRMS and **b** MRSE

where human blocks are standing closer to the tag. For 2D analysis, the experiments with the NLOS condition and the human blocks standing closer to the anchors have provided the most accurate location estimations. However, in 3D, placing the anchors at the height of 2.4 m leads to lower errors. In addition, similar to previous experiments, the 2D accuracy is better than the 3D.

#### 5.4 Four-Anchors Combinations

There are four combinations of the elevation of the anchors and the position of the human blocks when the LOSs of all the anchors are blocked at the same time. As illustrated in Fig. 7, for 2D, placing the anchors at the lower elevation and human blocks closer to the anchors provides the highest accuracy. However, for 3D, placing the anchors at an elevation of 2.4 m improves the accuracy. For both 2D and 3D, combining the lower elevation of the anchors and human blocks standing closer to the tag results in the highest error.

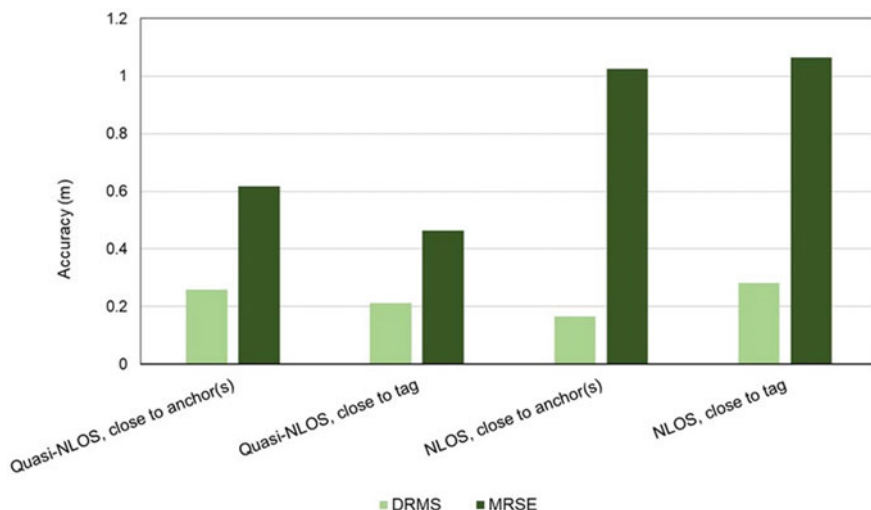
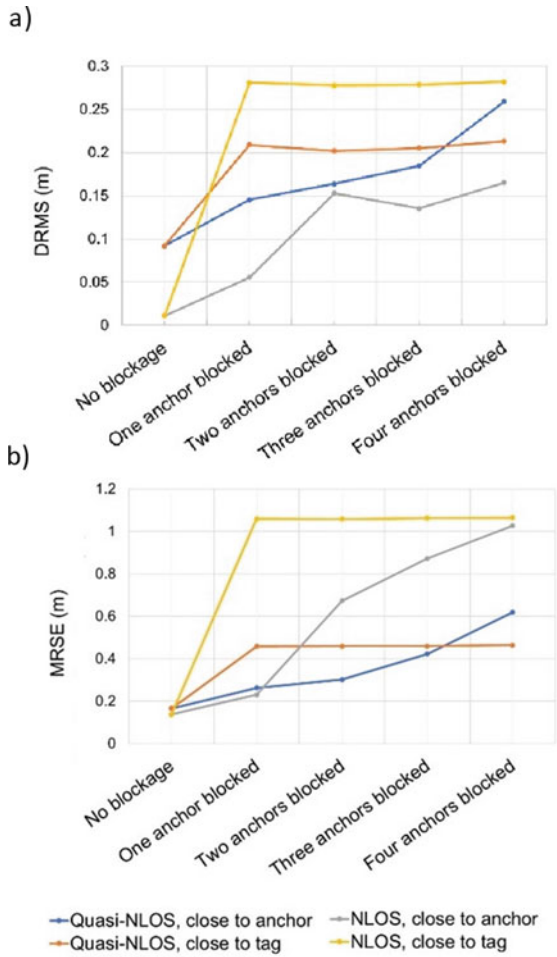


Fig. 7 DRMS and MRSE for four anchors blocked

## 6 Summary and Conclusion

To discuss the results of all the 50 experiments, the DRMS and MRSE of experiments with the same number of blocked anchors are averaged, and they are shown in Fig. 8. First, it was expected that the quasi-NLOS condition provides better accuracy than the NLOS condition. However, the NLOS condition where the human blocks are placed closer to the anchor resulted in the least error in all combinations in 2D. Second, adding the number of anchors with blocked LOS where human blocks are standing closer to the tag does not significantly affect the accuracy (yellow and orange lines). Third, *NLOS, close to tag* has provided the highest error in both 2D and 3D. However, as illustrated in Fig. 8, closeness to the tag is the dominant factor affecting the accuracy in 2D analysis while the degree of obstruction is the factor leading to higher error in 3D. In addition, comparing the accuracy of benchmark experiments in scenario 1 and scenario 2 reveals that placing the anchors at a higher elevation provides more accurate location estimations. Finally, the results demonstrate that partially blocking the LOS of the anchors decreases the 2D and 3D accuracies at least by 57.9% and 58.4%, respectively. And completely blocking the LOS of the decreases the 2D and 3D accuracies at least by 404.1% and 67.9%, respectively.

This study investigated the effect of the human body blockage on the location estimation accuracy of the UWB system by conducting several experiments. It could be concluded that as expected, by increasing the number of human blocks, the accuracy decreases. However, interestingly, when the human blocks are standing closer to the tag, increasing their number does not affect the accuracy significantly. In addition, in all the experiments, MRSE is worse than DRMS; therefore, it is presumed that the UWB has a weakness in the z-direction. Finally, the quasi-NLOS case where the



**Fig. 8** Accuracy average: **a** average DRMS and **b** average MRSE for all the experiments

human blocks are standing closer to the anchors provides the location estimation with lower errors. Therefore, it is recommended to place the anchors at a higher elevation than the height of the workers to improve the performance of the UWB systems on construction sites to track resources.

## References

1. Cheng T, Venugopal M, Teizer J, Vela PA (2011) Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments. *Autom Constr* 20(8):1173–1184. <https://doi.org/10.1016/j.autcon.2011.05.001>
2. Contigiani M, Pietrini R, Mancini A, Zingaretti P (2016) Implementation of a tracking system based on UWB technology in a retail environment. In: MESA 2016—12th IEEE/ASME international conference on mechatronic and embedded systems and applications—conference proceedings. <https://doi.org/10.1109/MESA.2016.7587123>
3. Delamare M, Boutheau R, Savatier X, Iriart N (2020) Static and dynamic evaluation of an UWB localization system for industrial applications. *Science* 2(2):23. <https://doi.org/10.3390/sci2020023>
4. Giretti A, Carbonari A, Naticchia B, De Grassi M (2009) Design and first development of an automated real-time safety management system for construction sites. *J Civ Eng Manag* 15(4):325–336. <https://doi.org/10.3846/1392-3730.2009.15.325-336>
5. Huang S, Guo Y, Zha S, Wang F, Fang W (2017) A real-time location system based on RFID and UWB for digital manufacturing workshop. *Proc CIRP* 63:132–137. <https://doi.org/10.1016/j.procir.2017.03.085>
6. Januszkiewicz Ł (2018) Analysis of human body shadowing effect on wireless sensor networks operating in the 2.4 GHz band. *Sensors (Switz)* 18(10). <https://doi.org/10.3390/s18103412>
7. Jiang L, Hoe LN, Loon LL (2010) Integrated UWB and GPS location sensing system in hospital environment. In: Proceedings of the 2010 5th IEEE conference on industrial electronics and applications, ICIEA 2010, pp 286–289. <https://doi.org/10.1109/ICIEA.2010.5516828>
8. Jiang S, Skibniewski MJ, Yuan Y, Sun C, Lu Y (2011) Ultra-wide band applications in industry: a critical review. *J Civ Eng Manag* 17(3):437–444. <https://doi.org/10.3846/13923730.2011.596317>
9. Kolakowski J, Djaja-Josko V, Kolakowski M, Broczek K (2020) UWB/BLE tracking system for elderly people monitoring. *Sensors (Switz)* 20(6). <https://doi.org/10.3390/s20061574>
10. Leick A (2004) GPS satellite surveying, 3rd edn. Wiley
11. Leser R, Schleindlhuber A, Lyons K, Baca A (2014) Accuracy of an UWB-based position tracking system used for time-motion analyses in game sports. *Eur J Sport Sci* 14(7):635–642. <https://doi.org/10.1080/17461391.2014.884167>
12. Maalek R, Sadeghpour F (2013) Accuracy assessment of ultra-wide band technology in tracking static resources in indoor construction scenarios. *Autom Constr* 30:170–183. <https://doi.org/10.1016/J.AUTCON.2012.10.005>
13. Maalek R, Sadeghpour F (2016) Accuracy assessment of ultra-wide band technology in locating dynamic resources in indoor scenarios. *Autom Constr* 63:12–26. <https://doi.org/10.1016/J.AUTCON.2015.11.009>
14. Mainetti L, Patrono L, Sergi I (2014) A survey on indoor positioning systems. In: 2014 22nd international conference on software, telecommunications and computer networks, SoftCOM 2014, pp 111–120. <https://doi.org/10.1109/SOFTCOM.2014.7039067>
15. Otim T, Bahillo A, Diez LE, Lopez-Iturri P, Falcone F (2019) Impact of body wearable sensor positions on UWB ranging. *IEEE Sens J* 19(23):11449–11457. <https://doi.org/10.1109/JSEN.2019.2935634>
16. Pittokopiti M, Grammenos R (2019) Infrastructureless UWB based collision avoidance system for the safety of construction workers. In: 2019 26th international conference on telecommunications, ICT 2019, pp 490–495. <https://doi.org/10.1109/ICT.2019.8798845>
17. Sahib Hasan H, Hussein M, Mad Saad S, Azuwan Mat Dzahir M (2018) An overview of local positioning system: technologies, techniques and applications. *Int J Eng Technol* 7(3)
18. Shahi A, Aryan A, West JS, Haas CT, Haas RCG (2012) Deterioration of UWB positioning during construction. *Autom Constr* 24:72–80. <https://doi.org/10.1016/j.autcon.2012.02.009>
19. Silva B, Hancke GP (2016) IR-UWB-based non-line-of-sight identification in harsh environments: principles and challenges. *IEEE Trans Industr Inf* 12(3):1188–1195. <https://doi.org/10.1109/TII.2016.2554522>



20. Sudhakar A, Madhavi D (2017) Analysis and measurement of attenuation constants of ultra wideband signal through commonly used building materials. *Progr Electromagnet Res Symp* 1:161–164. <https://doi.org/10.1109/PIERS.2017.8261726>
21. Tanghe E, Gaillot DP, Liénard M, Martens L, Joseph W (2014) Experimental analysis of dense multipath components in an industrial environment. *IEEE Trans Antennas Propag* 62(7):3797–3805. <https://doi.org/10.1109/TAP.2014.2321162>
22. Teizer J, Lao D, Sofer M (2007) Rapid automated monitoring of construction site activities using ultra-wideband. In: *Automation and robotics in construction—proceedings of the 24th international symposium on automation and robotics in construction*, pp 23–28. <https://doi.org/10.22260/isarc2007/0008>
23. Teizer J, Venugopal M, Walia A (2008) Ultrawideband for automated real-time three-dimensional location sensing for workforce, equipment, and material positioning and tracking. *Transp Res Rec* 2081:56–64. <https://doi.org/10.3141/2081-06>
24. Tian Q, Wang KIK, Salcic Z (2019) Human body shadowing effect on uwb-based ranging system for pedestrian tracking. *IEEE Trans Instrum Meas* 68(10):4028–4037. <https://doi.org/10.1109/TIM.2018.2884605>
25. Trogh J, Plets D, Thielens A, Martens L, Joseph W (2016) Body shadowing compensation 16(7):2105–2114
26. Vorobyov A, Yarovoy A (2012) Human body impact on UWB antenna radiation. *Progr Electromagnet Res* 22:259–269
27. Zasowski T, Meyer G, Althaus F, Wittneben A (2005) Propagation effects in UWB body area networks. In: *ICU 2005: 2005 IEEE international conference on ultra-wideband, conference proceedings*, pp 16–21. <https://doi.org/10.1109/icu.2005.1569949>
28. Zhang C, Shen W, Ye Z (2020) Technical feasibility analysis on applying ultra-wide band technology in construction progress monitoring. *Int J Constr Manag*. <https://doi.org/10.1080/15623599.2020.1834928>

# Enhanced Activity-on-Node Network Diagramming Method for Construction Planning and Scheduling Applications



Badhon Das Shuvo and Ming Lu

**Abstract** Occurrences of interruptions of deterministic or probabilistic nature are inevitable on typical construction projects because of unavailability of resources or human and management factors, thereby causing high variability in activity duration and lower labor productivity. This research proposes enhanced activity-on-node network diagramming method (named AON+) for construction planning and scheduling applications to formulate production schedules for repetitive workflows featuring non-uniform work units. In contrast to established project planning techniques, AON+ enables construction managers to represent details in workflows in a streamlined network diagram by sufficiently factoring in logical constraints imposed by both technology and resource. Further, an “AON+” network model readily feeds as the structured problem definition to rapidly generate discrete event simulation models for resource-constrained scheduling and operations simulation analyses. Two application cases based on industry practice are presented, namely a “rebar installation on a bridge deck” problem and a “bored pile concrete pouring” problem.

**Keywords** Node (AON) network diagramming · Construction planning · Scheduling applications

## 1 Introduction

Construction crew often repeats the same work of that activity, moving from one repetitive unit to another completing a set of work packages that are repeated sequentially at different locations. These activities usually maintain a technological driven sequence and are continuously subject to resource constraints imposed by internal technological, managerial, or external causes holding true for the entire life span of the project [7]. Projects of this type are thus considered high risk, which potentially causes delays in overall project completion and budget overruns, thus making

---

B. D. Shuvo (✉) · M. Lu

Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada  
e-mail: [badhonda@ualberta.ca](mailto:badhonda@ualberta.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_6](https://doi.org/10.1007/978-3-031-34593-7_6)

the management of resources a very significant issue [6]. During the last decades a significant number of planning techniques have been proposed for scheduling repetitive construction projects. These techniques can be classified into two basic categories: resource-driven methods (i.e., LOB, RSM) and duration-driven methods (i.e., CPM, PERT). A Line of Balance (LOB) or Repetitive Scheduling Method (RSM) schedule allows the balancing of operations such that each activity is continuously performed from one unit to the other [4]. A LOB schedule is presented graphically as an  $X$ - $Y$  plot where the two axes represent units (or cycles) and time. For houses, stores, apartments, or floors in high-rise construction (vertical construction projects), work progress is generally measured in units completed [3]. As a resource-driven technique, LOB identifies a balanced mix of resources and synchronizes the activities such that these resources are fully employed. The major advantage of a LOB scheduling is that it represents the duration and production rate information in such an easily interpreted graphical format [14], it allows the planner to adjust the rates to meet project deadlines, while maintaining the work continuity of the resources [5]. In contrast to the LOB or RSM methodology, duration-driven methods like critical path method (CPM) assumes the activity durations as functions of the resources required (rather than available) to complete each activity [1].

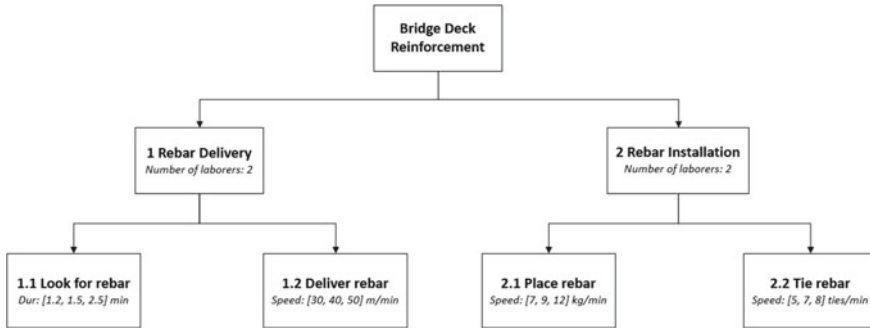
Despite the wide application of both RSM and CPM in construction management, various researchers have challenged their applicability in an attempt to prove their inadequacy for scheduling real-world repetitive projects. Although reduced cost is desired in most of the construction projects [13], this assumption may not be true since extending the project duration may result in delay penalties and lost revenues. Kavanagh [8] pointed out that the LOB techniques are designed to model simple repetitive production processes with a limited degree of complexity and, therefore, practically meaningless in a complex construction environment. CPM-based techniques have been criticized for its inefficiency to describe the repetitive nature of Linear Repetitive Project scheduling (LRP). Network analysis has been characterized by its critics as insufficient to account for neither resource availability nor work continuity [4]. This methodology treats the piecewise execution across the project units as a set of distinct activities connected only through precedence relationships. The size of the corresponding CPM network of the LRP is quickly exploding. Ammar and Mohieldin [2] developed a CPM-based repetitive scheduling model to schedule repetitive activities in an easy nongraphical manner, where the model considered only the most significant resource for each activity and failed to acknowledge multiple crews used by an activity. To tackle the above identified limitations and improve the effectiveness of AON/CPM scheduling under limited resources and activity interruptions, this research study proposes a novel resource scheduling methodology based on the simplified discrete-event simulation approach (SDESA). Two example applications have been presented afterward to illustrate the implementation and features of the proposed model based upon the workflow patterns associated with repetitive work units.

## 2 Scheduling Repetitive Activities and Workflows

Repetitive construction projects undergo a set of activities that are repeated sequentially at different locations or units. However, among these activities there may exist both repetitive and non-repetitive jobs together in a construction project. Vorster et al. [12] divided the repetitive projects into typical and non-typical. Typical repetitive construction projects utilize resources (crews and equipment) having same productivity and consist of activities with same quantity of work needed for each repetitive unit. On the other hand, the non-typical projects have different quantities of work needed for each repetitive unit, which utilize resources operating with different productivity. One of the reasons why AON/CPM is still the most popular scheduling method in scheduling repetitive activities is that it provides a well-established logic in the network analysis phase and permits further computerized applications. The network analysis method involves forward and backward pass calculation, which logically identifies the critical path and performs scheduling analysis to determine the total project duration. Also, the network can be rearranged and redesigned based on planner's intuition and experience to obtain the most desirable result. Although AON can be optimized to develop a well-established logic, its analysis features are limited to tackle the typical and non-typical repetitive construction projects. When the number of repetitive details continues to increase, the resultant CPM schedule gets cluttered with the repetition of information posing it a complex challenge to interpret information for the practitioners.

The enhanced version of AON (AON+) is not to reinvent the wheel from scratch but only to add necessary features to its current format. AON+ versus SDESA simulation analysis is analogous to AON versus CPM scheduling analysis [9]. AON+ results in a cost-effective simulation modeling methodology enabling engineers to capture details in construction operations in basic process dimensions. A SDESA model represents a construction operation by delineating major workflows, where a workflow consists of major activities (or work packages) logically connected with arrows to denote finish to start precedence relationships. Each workflow can handle a set of work units or cycles, each being easily identifiable and measurable unit of work that is processed by activities and easily identifiable and measurable. The work units, also called as flow entities in simulation terms, can also be likened to the materials takeoff in certain unit measures in estimating costs for construction activities. The associated flow entities are generally work packages or production units (e.g., units of material, parts, and subassemblies). In the AON+ model, resource entities are classified into disposable resources and non-disposable resources. Disposable resources represent material units or information units, generated by one activity and requested by another which are associated with two different resource workflows. Disposable resources constitute part of resource-availability constraints in matching resources prior to starting activities. This is contrast with non-disposable resources such as manpower/machinery, or space block, commonly applied in construction.

The AON+ model feeds into the input of a SDESA simulation model. As per SDESA, a central resource pool holds the definitions of all the resources relevant to



**Fig. 1** Work packages for bridge deck reinforcement

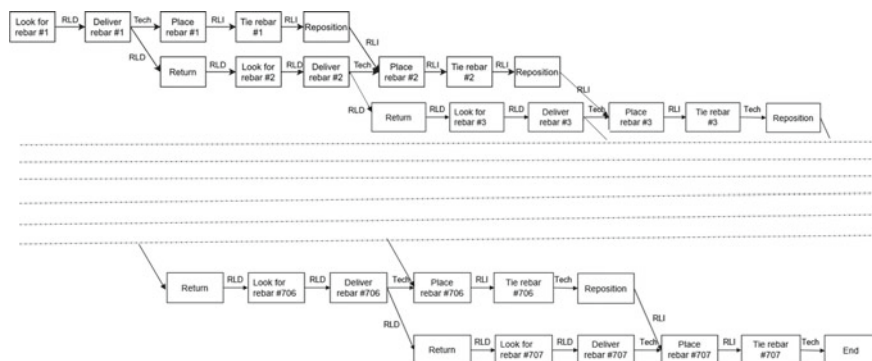
construction planning, regardless non-disposable resources or disposable resources. The resource pool is dynamically managed to control resource-availability status in simulation. AON+ adds a counter to control the number of repetitions for each resource workflow and a “disposable resource” definition to link various resource workflows, keeping the simplicity and flexibility of the original AON.

### 3 Illustrating Case

LOB/RSM is confined to specific applications in construction that fit the above generalized AON patterns. In practice, many construction processes of repetitive nature do not fall in this category, as demonstrated with the rebar installation on a bridge deck problem illustrated in literature [10]. According to the example project, the Creek Deck is 57.5 m in length and 11 m in width. As per design drawings, rebars are placed in two layers and two perpendicular directions, namely top and bottom layers and short and long directions. Rebar stock will be processed in a rebar bending yard next to the site and then delivered to site as cut-to-length rebar segments. The installation operation of reinforcing steel is decomposed into two work packages (WP) as in Fig. 1.

Two laborers work as a team on WP1. They can carry 30 kg on each trip; the distance from the rebar bending shop to the site is 25 m. Besides, two laborers also work on WP2. A total of four reinforcing labors are employed on-site for this job for a total takeoff of 21,191.3 kg steel. Based on the calculations, a total of 707 repetitive cycles/jobs are required to complete this project. AON representation of the planning patterns for this project is shown in Fig. 2.

In Fig. 2, the AON prepared for the project consists of 707 repeating units of work, where each unit consists of six activities (except first unit). The nature of the precedence relationship (resource involved or technology constraint) is marked on each arrow in AON. The solid lines linking the activities within each unit define the technical precedence constraints in the network. For example, activity “Tie rebar



**Fig. 2** AON representation of the rebar case study

#2” cannot be started until activity “Place rebar #2” is completed. The lines linking similar activities from one unit to another unit represent the resource-availability constraints; for example, activity “return” cannot start until the crew from activity “Delivery rebar” is available. RLD represents Rebar Delivery Labor, whereas RLI represents Reinforcement installation Labor. In the horizontal direction, there is a mix of resource constrained and technology constrained precedence relationships. For instance, in handling job #1, after delivering rebar by the RLD to the deck, a technology link exists between “Delivery rebar #1” and “Place rebar #1” by RLI. The ensuing activities (“Tie rebar #1” and “Reposition”) are connected by two arrows involving RLI.

In this case study problem, AON treats the piecewise execution of a repetitive task across the project units as a set of distinct activities connected only through precedence relationships. A network representation of a repetitive project consisting of  $m$  tasks,  $p$  precedence relationships, which is repeated in  $n$  units generally contains  $(m \times n)$  tasks and  $(p \times n) + m \times (n - 1)$  precedence relationships [6]. For this rebar case study with six tasks, repeated in 707 units, there are more than 7770 precedence relationships, which makes the AON to explode due to size and complexity. The objective of AON is the minimization of the duration through the definition of the critical path and the optimum time/cost trade-off of the project by means of crashing the critical activities without guaranteeing work continuity for the repetitive activities. Although resource allocation in repetitive units scheduling require revision of the project schedule in order to comply with resource-availability constraints, AON makes an initial assumption of unlimited availability of resources in the development of the project schedule. The revision of the vertical relationships (resource constraints) becomes even more tedious if the planner has, realistically, deployed more than one crew to work on each typical activity. Thus, AON network model provides a process modeling solution but becomes unscalable for practical application and ineffective to lend decision support for planning. In larger and more complex applications, AON can become too cumbersome and complicated for process modeling and communication. Scheduling without correcting these errors

can produce unworkable schedules. Without AON, CPM thus becomes pointless for scheduling analysis. Thus, it makes the project control impractical for project managers who want to monitor the location and the production rate of each activity as it is progressing through the entire length of the project. The research inquiry is how to enhance AON so to keep its flexibility and ease of use while making it scalable and practical in construction applications.

Meanwhile, in the LOB representation, resource links are not supposed to connect activities in the horizontal direction, but only links denoting technology constraints between activities. LOB is based on the graphical representation on an  $X$ - $Y$  diagram and aims to ensure continuous resource utilization. It employs a pull-system approach, where the finish time of the predecessor activity is pulled forward to meet the start date of the successor to achieve work continuity and uninterrupted resource utilization. Thus, it focuses only on achieving the work continuity while minimizing overall cost without considering the minimization of the project completion time. Although in conventional construction projects, the minimization of cost might be more desirable rather than reducing the project duration, this assumption always does not turn out true, considering there may be delay penalties and legal disputes due to delay in delivery leading to financial losses. Also, the activity duration depends on the shape, complexity of the bar segments, the position on the jobsite, so the same activity can take longer or shorter in processing different batches of rebar. These variations exist in reality and exert significant influence upon the project duration. The impact of such variations cannot be overlooked or avoided in project planning and scheduling; instead, they need to be identified first, clearly specified in the production plan, and analytically modeled to assess their impact on shop scheduling [11].

AON+ can be utilized to facilitate scheduling in repetitive workflows subject to limited resources and activity interruptions. It represents a construction system by defining major activities and workflows; delineating individual activities within each workflow; and then identifying the flow entities and resource entities involved in each workflow. Flow entity is the head of a flow of activities, work package or production units and represents the numbers of work units ( $N1, N2, \dots$ ) to be processed of the workflow. In the AON+ model, resource entities are classified into disposable (DR) and non-disposable (manpower/machinery) resources ( $R1, R2, \dots$ ). Disposable resource entities are material or information units that are generated by one activity and required by another (Fig. 3). They play the key role to setup the interdependent relationships between various activities and constitute part of resource-availability constraints in matching resources prior to starting activities.

For this case study, the first two major activities (Look for rebar, Deliver) can be linked with the “Rebar Delivery” flow entity, where each activity or workflow can be executed no matter how many times the project demands (Fig. 4). Disposable resource “Rebar” is generated at the end of the final activity “Deliver Rebar to Shop,” which logically connects and initiates its successor activities for Rebar installation. With parallel workflows dynamically linked with disposable resources, the overlapping on time or concurrence of various resource workflows can be sufficiently modeled. Then, the forward-pass calculations are performed by executing the model

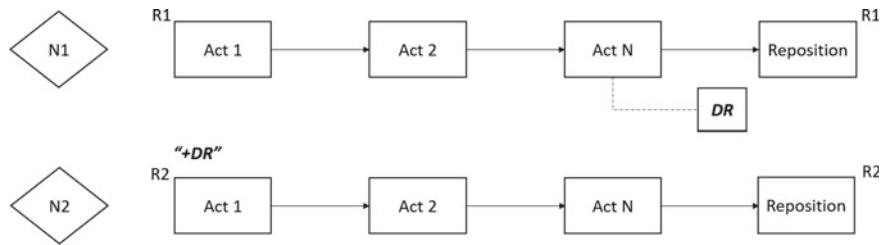


Fig. 3 Generic pattern of resource workflow in AON+

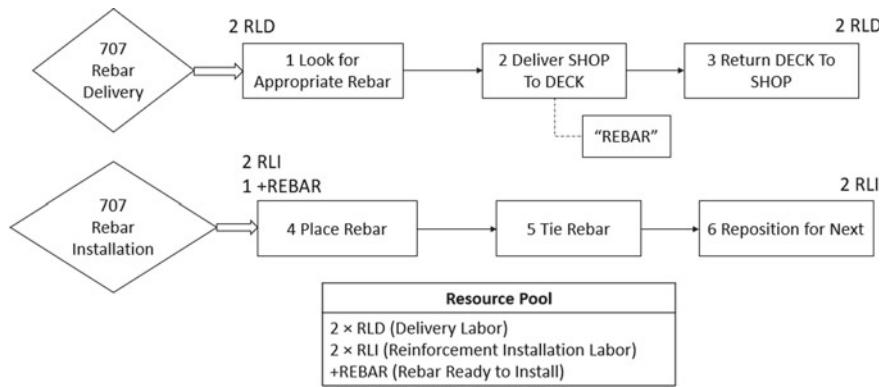


Fig. 4 AON+ workflow planning for the rebar case study

complying with CPM logic, which derives the total project duration accounting for both technological and resource constraints. In this framework, there is no need to use smart relationships such as Finish-to-Finish (FF), Start-to-Start (SS), or Start-to-Finish (SF) with lags. Complying with the simplest dependency, Finish-to-Start (FS), this model defines logical relationships in each workflow; identifies multiple resources or resource workflows with the introduction of disposable resources and controls the times of repetitions for each resource workflow.

It is noteworthy the activity duration for this case study also depends on the shape, complexity of the bar segments, the position on the jobsite, and other factors; so, the same activity could take longer or shorter in processing different batches of rebar. To solve this problem, AON+ sets the activity time by sampling from a particular representative distribution (i.e., Normal, Triangular, Beta, Uniform, etc.) and runs simulation analysis according to the duration variations for activities undergoing defined workflows. For this problem, the formulated AON+ model (Fig. 4) sets triangular distribution (Low, Mode, High) based on the likely duration of all the four activities provided in the WBS (Fig. 1) and linkups the parallel crew workflows logically. Thus, AON+ mimics the common logic and practice of CPM; enhances the scheduling accuracy by accurately quantifying variations and uncertainty with



simulation in order to process identical rebar segments requiring much less effort compared to the conventional scheduling methods.

## 4 Case Study: Bored Pile Concreting

At the beginning, a record laborer records the arrival of the truck mixer carrying concrete from a concrete supplier to the foundation construction site. If the RMC of the truck mixer can pass the Quality Assurance (QA) tests, the truck mixer comes to the unloading bay and climbs up the inclined platform where a hopper is located and prepares for the unloading. When the hopper and crane are ready, the truck mixer starts unloading in hopper until the truck mixer is empty. For this project, each hopper can be filled by one truck mixer concrete. When empty, the truck mixer leaves the unloading bay immediately after unloading; goes to the washing bay for cleaning and then leaves the site. A record labor at the site-exit marks the leave time on the concrete slip and returns the receipt to the driver.

When a hopper is filled with RMC, the hopper is hooked up to the bored pile position by a crane where a journeyman is waiting to pour concrete from the hopper into the pile. The platform that acts as a stand for the concreting on the top of casing is then removed by the crane. When empty, the hopper is hooked back to the unloading bay for the next concreting cycle. A temporary casing will be lifted up and truncated upon pouring ten hoppers of concrete. So, the hopper-concreting cycle needs to be repeated ten times until ten hoppers of concrete are cast. The truncation prevents the casing from being buried in the ground before the concrete hardening. To cut the temporary steel casing, the crane keeps hooking up the casing to hold it staying upright, and when the unscrewed casing section is removed, the working platform will be re-installed to the top of bored pile for another concreting cycle.

A simplified discrete event simulation model will be created complying with AON+ logic in the SDESA platform as follows: Firstly, each activity in the AON network will be represented with a flow entity diamond linked with one Activity Block in SDESA platform, ensuring each activity is executed only once or in multiple repetitive units based on the project requirements. Secondly, the disposable resource entity substitutes for the arrows in the AON and acts as an information unit to enforce the precedence relationship in the network diagram. An activity generates disposable resource entities, which are then requested to initiate its successors. As soon as all its preceding activities are finished, as a result, the required disposable resource entities become available to trigger the start of the current activity. Once the SDESA model is defined, the forward-pass calculations of CPM can be performed by executing the SDESA model, accounting for both technological and resource constraints.

**Table 1** Resource constraints (flow entities) for activities in bored pile case study

Flow entity (number of cycles)			
Activities		Resources	
No.	Name	Required	Released
<i>(I) Truckload (10)</i>			
1	Enter site	1 JM	–
2	QA	2 JM, 1 QA	1 JM, 1 QA
3	Park to unload bay	1 JM, 1 UN_BAY	1 UN_BAY, 1 UNLD_R
4	Go to wash	1 JM, EMP_TRK	–
5	Wash truck	1 JM, 1 WB	1 WB
6	Leave site	1 JM	–
7	Reposition	1 JM	1 JM
<i>(II) Hopper load (10)</i>			
8	Receive concrete	1 JM, 1 CR, 1 UN_BAY, 1 UNLD_R	1 UN_BAY, 1 EMP_TRK
9	Load concrete to pile	2 JM	1 JM, 1 CR, 1 SIG_CUT
10	Reposition	1 JM	1 JM
<i>(III) Pile section (1)</i>			
11	Remove platform	1 CR, 1 JM, 10 SIG_CUT	–
12	Cut casing	1 OS, 1 JM	–
13	Lift casing	2 JM	1 OS
14	Reinstall platform	1 JM	1 CR
15	Reposition	1 JM	2 JM

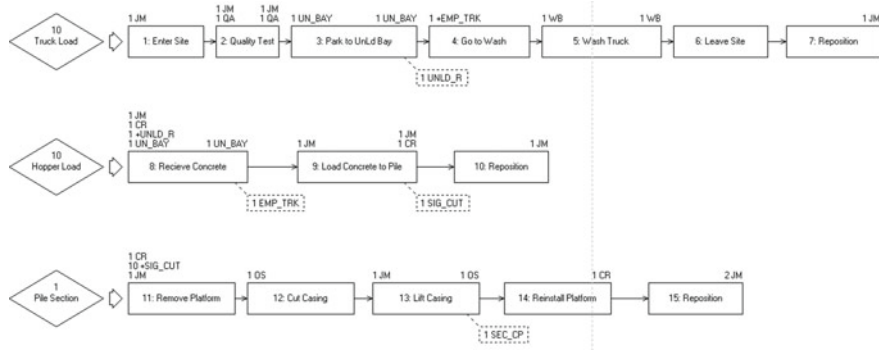
## 5 Results and Analysis

Tables 1 and 2 are the resource workflow summary of the AON+ model using truck load, hopper load, and pile section as flow entities. Each activity in the AON network is represented and linked with one of these three flow entities. The table lists the different work packages the work units undergo, and what their resource-constrained relationships are. The first disposable resource, “UNLD\_R,” is generated at the end of the activity “Park to Unload Bay” in the first crew workflow to represent one truck load has been dumped at the end of that activity. This disposable resource triggers the start of next crew workflow, where the hopper load undergoes different work packages (i.e., receiving concrete, loading concrete to pile).

The next disposable resource, “SIG\_CUT” is generated at the end of “Load Concrete to Pile” activity. Since 10 “SIG\_CUT” is in the resource required list of the activity “Remove Platform” in next crew workflow, it is assumed that the next crew workflow will be carried out when ten hopper loads of aggregates will be accumulated. These ten cycles of aggregates accumulation will further lead to the Platform Removal, Cutting and Lifting Casing, and Reinstalling Platform again to

**Table 2** Disposable and non-disposable resource pool in bored pile case study

Resource pool
Non-disposable: JM (journeyman), QA (quality test), UN_BAY (unloading bay), CR (crane), WB (washing bay)
Disposable: UNLD_R (ready to unload), SIG_CUT (cut section signal), SEC_CP (section complete), EMP_TRK (empty truck)



**Fig. 5** AON+ model formulated in SDESA

complete the one pile section cycle. SDESA can track the movement of the resources (labor, equipment, materials) from one location to another and identify the areas for productivity loss. AON+ prioritizes the processing sequence of production jobs by the first-in-first-out rule. The start time of an activity is delayed until the demanded resources are available on-site. Figure 5 shows the SDESA layout based on AON+ which tracks the movement of resources between different activities in the system.

Since the activity duration in the parallel crew workflows varies based on the material properties, resource availability, and the position on the jobsite, the uncertainty in duration is characterized by defining the distributions for each resource workflow. For example, for this case study, the activity “Park to Unload Bay” follows a Uniform Distribution (low = 0.5, high = 1) from “Site Lab” to “Site-Unload Bay,” whereas the activity “Load Concrete to Pile” follows a triangular distribution (low = 3, mode = 5, high = 6) from “Site-Unload Spot” to “Site-BP Spot.” In the SDESA platform, resource’s location state is initialized by user and automatically tracked by computer algorithm. For this case study, this automated process enables the user to apply these distributions without the need to update each activity individually after each crew workflow.

The make span of a particular activity can be measured as the time difference between the end and start times of that process workflow. According to the simulation results, different units undergo different duration based on the variations in activity duration and crew availability. The truckload unit consists of seven activities and these activities have different quantities of work needed for each repetitive unit and utilize resources operating with different productivity. The duration for each unit

of truckload varies in the range between 13 and 19 h (most likely: 16 h), whereas each unit of hopper load undergoes the duration variation in the range between 7 and 11 h (most likely: 9 h). Also, the idle time (non-value-added time) between different activities in different crew workflows was computed using the AON+ logic in SDESA and they vary in the range between 6 and 7 h. For example, the idle time for the flow entity “hopper load” in Unit 3 and Unit 4 is 7 h, whereas between Unit 8 and Unit 9, the idle time is 6 h. After running the simulation model, the total project duration was calculated as 192 h.

## 6 Conclusion

The research inquires how to enhance AON so to keep its flexibility and ease of use while making it more efficient and scalable considering various resource-constrained relationships and logics. Although AON/CPM is still the most popular scheduling method in scheduling these types of repetitive activities, its inability to accurately model resource-constrained relationships and reflect actual conditions leads to its inadequacy to schedule complex construction environments, posing it a complex challenge to interpret information for the practitioners. Therefore, the study explores a novel resource-constrained scheduling framework named AON+, which is capable of (1) generating robust resource use plans for multiple interdependent concurring workflows, (2) interconnecting and synchronizing schedules while accounting for both technological and resource constraints, and (3) analyzing the crew performance KPIs in regard to resource use, productivity, and “lean” at various levels of granularity. Based on the generalized modeling structure explained with a case study in this paper, this planning approach can be readily scaled up in terms of more repetitive activities with more complex process workflows so as to cope with a construction project of any practical size and complexity.

## References

- Ammar MA (2013) LOB and CPM integrated method for scheduling repetitive projects. *J Constr Eng Manag* 139(1):44–50
- Ammar MA, Mohieldin YA (2002) Resource constrained project scheduling using simulation. *Constr Manag Econ* 20:323–330
- Harris RB, Ioannou PG (1998) Scheduling projects with repeating activities. *J Constr Eng Manag* 124(4):269–278
- Hegazy T (2002) Computer-based construction project management. Prentice Hall, Upper Saddle River, NJ
- Hegazy T, Kamarah E (2008) Efficient repetitive scheduling for high-rise construction. *J Constr Eng Manag* 34(4):253–264
- Ipsilandis PG (2007) Multi-objective linear programming model for scheduling linear repetitive projects. *J Constr Eng Manag* 133(6):417–424

- Kallantzis A, Lambropoulos S (2004) Critical path determination by incorporating minimum and maximum time and distance constraints into linear scheduling. *Eng Constr Archit Manag* 11(3):211–222
- Kavanagh DP (1985) SIREN: a repetitive construction simulation model. *J Constr Eng Manag* 111(3):308–323
- Lu M (2003) Simplified discrete-event simulation approach for construction simulation. *J Constr Eng Manag* 129(5):537–546
- Lu M, Zheng C, Yi C, Hasan M (2017) Synthesizing engineering design, material takeoff and simulation-based estimating on a bridge deck reinforcement case. In: *Proceedings of the 2017 winter simulation conference (WSC)*, pp 2507–2517
- Shuvo BD, Lu M (2020) Lean construction planning subject to variations in detailed features of fabricated bridge girders. In: *Proceedings of the 28th annual conference of the international group for lean construction (IGLC28)*, Berkeley, CA, pp 1–12
- Vorster MC, Beliveau YJ, Bafna T (1992) Linear scheduling and visualization. *Transp Res Rec* 1351:32–39
- Yang IT (2002) Stochastic analysis on project duration under the requirement of continuous resource utilization. In: *Proceedings of the 10th international group for lean construction conference*, Gramado, Brazil, Aug 2002, pp 527–540
- Yang I, Ioannou PG (2004) Scheduling system with focus on practical concerns in repetitive projects. *Constr Manag Econ* 22(6):619–630

# Integrating Smart 360-Degree Photography and QR Codes for Enhancing Progress Reporting



Ahmed Bahakim, Ibrahim Abotaleb, and Ossama Hosny

**Abstract** Periodical construction progress reports are essential in project evaluation and review. They impact stockholder communication, transparency, and trust. While conventional pictures and videos are currently the norm in supporting progress reporting, their use is not always efficient. Therefore, commercial products are now available for integrating 360-degree photography into progress reports. However, there is a shortage of academic studies that actually assess the effectiveness of such tools. The goal of this research is to analyze the methods, benefits, and limitations of using smart 360-degree models and QR codes in progress reporting, and how such use impacts the overall project performance. The keyword “smart” indicates that the 360-degree pictures would show additional info on them such as comments, side-by-side comparisons to previous months, all into a single 360-degree advanced progress model. To this end, information was collected from construction projects and firms to determine the status and used methods of progress reporting. Then, a pilot 3-month study was conducted where smart 360-degree pictures and QR codes are introduced as means of progress reporting for ongoing residential projects and institutional and commercial building. After thorough analysis of minutes of meetings, letters, and interviews before, during, and after using the technology, the results indicate that the choice of the integration platform is important. For example, it must have capabilities of integrating and stitching several 360-degree pictures together into one large model, linking the model to BIM and Procore software, enabling side-by-side comparison, and having a strong back-end enabling multiple users to access it remotely. The results also indicate that the proper use of such technology enhanced

---

A. Bahakim (✉) · I. Abotaleb · O. Hosny  
The American University in Cairo, Cairo, Egypt  
e-mail: [ahmedbahakim@aucegypt.edu](mailto:ahmedbahakim@aucegypt.edu)

I. Abotaleb  
e-mail: [ibrahimsalah@aucegypt.edu](mailto:ibrahimsalah@aucegypt.edu)

O. Hosny  
e-mail: [ohosny@aucegypt.edu](mailto:ohosny@aucegypt.edu)

the overall coordination, transparency, trust, and responsibility between the project parties. It also led to less finger-pointing. Additional studies in longer periods and a larger number of projects are needed to determine the effect of such technology on cost and time.

**Keywords** Smart 360-degree photography · QR codes · Progress reporting

## 1 Introduction

A progress report is used to summarize the site situation in a document or file by including some details like the project name, date, location, pictures. These details enhance credibility and transparency in reports. Through several visiting sites in Egypt and meeting several engineers, we noticed that the document includes many pages of conventional pictures; however, they do not cover the full details on site. While photos and videos are currently in heavy use to enhance and support progress reporting, many construction companies face issues with using job site photos such as firstly, slow turnaround which means attaching the job site photos in the progress report takes time, so the stockholders typically do not receive them that day (or that week). While waiting, photos quickly become outmoded, reflecting the job site inaccurately. Secondly, wasted time and effort because the construction site contains essential details, so covering one room from every side takes more than six normal photos instead of one 360° photo. Finally, it is difficult to locate the image in site layouts. Thus, nowadays, several commercial applications based on 360° cameras are involved in construction management, for example: HoloBuilder, StructionSite, Reconstruct, VisualPlan, CUIPX, OpenSpace OnSiteIQ, and ContextVR [6].

Nevertheless, these applications are costly, and they are not commonly used in the Middle East, North Africa, and other countries. There is a need to test the effectiveness of applying a new project management methodology such as integrating Smart 360-degree model and QR codes for enhancing progress reporting. This paper relied on short feedback cycles, frequent adapting to change, real time and continuous communication between all project stakeholders, conducting post-project reviews and question surveys. The study's primary purpose is to use one page of smart 360-degree models and QR codes in progress reporting instead of many pages of normal photos in the progress reports. Moreover, several criteria were identified in the questionnaire to determine the effectiveness of using this model in progress reports, such as saving effort, time, cost, resolving conflicts and enhancing cooperation between the parties, increasing productivity, reducing errors. The following sections of the paper present a literature review, focusing on related definitions, benefits of using 360-degree photos in the construction industry, approach to implementing our model in progress reports—additionally, the proposed framework provides expected and promised results for future research areas.

## 2 Review of Previous Literature

### 2.1 Historic Significance

During the days of the Roman Empire, painters and artists decorated the interiors of buildings and villas from floor to ceiling to depict scenes of nature, court life, cities or dramatize historical or mythical events, creating a kind of 360° visual experience [3]. In the late eighteenth century, this way of showcasing is developed to create which called Cycloramas by English painter Robert Barker. It is a panoramic image inside a cylindrical platform, designed to provide a viewer standing in the center of the cylinder with a 360° view, as well as a building designed to display a panoramic image. The purpose of the effect is to make the viewer, surrounded by the panoramic image, feel as if they are standing in the middle of the place depicted in the image. And the idea was widely spread in many countries in Europe and North America [4]. Although this viable medium was a great future step in developing the idea of 360 photography, it was expensive, large, and high weight. The Gettysburg Cyclorama, for example, is 377 ft. long, 42 ft. high, and weighs 12.5 tons. At the same time, photographers started developing cameras that could capture several wide shots together to create unbroken overviews of cities and landscapes called Panoramic Photography. The Library of Congress in the USA contains an amazing collection of early panoramic pictures [11]. In 1851, panoramic photography of San Francisco was made with five daguerreotype plates. This process is recognized as the first commercial photographic process to create highly detailed images using silver copper plates, see Fig. 1.

In early-to-mid twentieth century, 360 photography was turned to use a flexible film instead of pressed plates, and new camera technology was appeared such as rotating cameras. But it was still in limited use for scientific or military research. Even in the late twentieth when the panoramic camera has become smaller and more cheap, agile, it needed many setup requirements and training. Thus, it has been limited use. Companies such as Nikon and Canon have introduced digital single lens reflex (DSLR) cameras which inspired new generation of wide camera. After that, GoPro action camera took a high place 360-degree shooting especially with sports lovers



**Fig. 1** San Francisco from Rincon Hill, 1851; c1910. Martin Behrman. Gelatin silver print; 5.5 × 36 in. PAN US GEOG—California, no. 235 (E size) P&P (<https://www.loc.gov/resource/pan.6a01850/>)



due to its features such as small size, and robust with a very wide lens. Nowadays, the 360-degree camera has become available in market with cost-effective and easy to start new era of using and trying in different fields.

## ***2.2 Benefits of Using Photos in the Construction Industry***

In past years, photographs played a significant role in documentation of buildings, as mentioned in the book titled, *The Life and Death of Building on Photography and Time* [12]. However, video and photos serve as evidence in the documentation of actual job progress at a specific time and present the project state, there are several arbitrators preferred photographic instead of video in visual information. And they highly recommended to link photo evidences with date, location, notes [9]. With the development of technology and the merge of cameras into the monitoring and follow-up process, it has become easy to track workflow and present transparent progress reporting. Many researchers have investigated using the camera in monitoring construction sites, such as the research conducted on the benefits of using the application of camera-equipped Unmanned Aerial Vehicles (UAVs) to control the collection, analysis, visualization, and communication of the visual data captured of projects [2]. Moreover, surveillance camera is used for analyzing tower crane operations throughout a workday to understand construction activity [15]. In recent years, the photography field has developed, as the use of 360-degree cameras has appeared in many domains, including education, tourism, and entertainment, in addition to recent uses in the construction field. For example, it has enhanced educational stakeholders' ability to record detailed 360-degree videos to enrich tertiary education in civil engineering by a suitable learning medium [14]. Another example in real estate is matching stakeholders between sellers and buyers by using a 360-degree camera that helps the buyer see the construction quality, workmanship, and defects in real time. Thus, purchasing confidence is increased [7]. Another paper presents and explains the Matterport that uses 360-degree photos as a new platform for marketing which aims at helping real estate businesses survive during the COVID-19 pandemic [13]. On the other hand, several applications that use 360-degree cameras are involved in construction management (HoloBuilder, StructionSite, Reconstruct, VisualPlan, CUPIX, OpenSpace OnSiteIQ, and ContextVR). They have different tools and techniques, but the standard process is capturing and documenting the project with a 360° camera, then, the images are automatically mapped to plans and viewed interactively. These applications offer time savings, more complete documentation, reduced cost, and better collaboration. For example, this study [1] used the HoloBuilder application with the Four-Legged Robot to capture site photos. A 360° camera, called Ricoh Theta V, was connected to the mobile device that runs the JobWalk App with 2D floor plans. The objective was to automate building progress monitoring and learn about the benefits and drawbacks of using this process. The discussions of

this study had provided several chances in automated construction progress monitoring, including integration with 360° camera, HoloBuilder application, and automatic image capturing process. These studies show the possibility of decreasing time and labor in the data collecting procedure. However, many companies and software packages incorporate 360° cameras in progress reporting, academic research in assessing the effectiveness of 360° photos in progress reporting is scarce. Moreover, these applications cost money, and most of them are not available in the Middle East and North Africa.

### ***2.3 The Current Available 360-Degree Cameras***

An omnidirectional camera (from “omni,” meaning all), often known as a 360-degree camera, has a field of vision that covers almost the whole sphere or at least a full circle in the horizontal plane. This camera usually contains two lenses (Dual Fisheye) or more to take two photos or more at the same moment. For example, a type of camera called Insta360 Pro2 has six wide-angle lenses in opposite directions. It captures six images in different angles at the same time. After that, the images are stitched together to create a spherical 360° panorama image [8]. Stitching is known as converting two 180-degree shots to a 360-degree shot. And the stitch line is the line at which two photographs are stitched together. They can be seen on conventional devices like smartphones or laptops, but sometimes 360-degree viewer applications are needed to allow the photo to be freely rotated on the screen. Therefore, the most comprehensive and detailed records of construction site can be presented from one 360° photo. Which it can be an advantage in construction management like dispute resolution proceedings, follow up the workflow if it was well done in photo documentation (Fig. 2).

Regarding expected cameras that can be used, there are several categories of 360° cameras, first, in terms of the brand, such as Insta360, Samsung Gear 360, Garmin VIRB 360, or Ricoh. Second, in terms of use, there are professional cameras like Insta360 Pro2. Furthermore, there are consumer cameras such as Insta360 R, Insta360 ONE X2, or Ricoh Theta Z1 which are highly recommended from some applications like OpenSpace, HoloBuilder, StructionSite, or Matterport. An initial survey was conducted in the Insta360 community from the website that includes around 27,000 members who use or are interested in using 360-degree cameras. This survey gave indicators about which is the best 360° camera that can be easily used in real estate or construction with 115 participants of members. From the survey, it was clear that the majority of participants preferred to use Insta360 ONE X2 by 67% while around 26% of participants preferred to use Insta360 R. Nearly 3% like to use GoPro Fusion 360, 1% with Ricoh Theta Z1, and only a small minority preferred other 360-degree cameras which were around 3%. In general, 360-degree cameras are still an evolving technology, new features, so the high price means cameras with more stabilization, high resolution, storage, and flexible in software. For example, several people preferred the Insta360 camera because of its easy software and application

Photos before stitching process



**Fig. 2** Example of a photo captured with 360-degree camera called Insta360 Pro2

on phone in terms of editing photos or video, uploading, sharing, stitching, and downloading fast and easily. The recommended camera price is around \$600–\$1000 US dollars. Like Insta360 ONE X2 and Insta360 R, the price can be changed depend on where to buy and when [3]. There are certainly advantages to cameras with multiple lenses, but the focus of this study will be on smaller portable devices. There are several features in that you should take into consideration before buying a 360-degree camera. These are presented in no particular order:

1. **Resolution.** This means more resolution more detail in the footage. However, high resolution is in high demand on camera, it needs high storage of memory like large SD Card, and the battery drains fast.
2. **Battery Power** is one of main factor in 360° cameras. And the best option is to get a camera with battery replacement capabilities.
3. **Stabilization.** It means that a lot of movement in the picture makes the viewer uncomfortable. And to avoid that it should use a good stabilization camera or use some tools like a traditional tripod (with three legs), a monopod, or a selfie stick.
4. **Memory.** Some 360° cameras, such as the Ricoh Theta V and the Theta Z1, have fixed memory whit no way to expand that. This could be a disadvantage for those who are shooting many pictures in travel. Consequently, the best recommendation is to buy a camera that has a removable memory card like a microSD card slot.
5. **Other consideration** such as Frames Per Second, waterproof or not, and camera repair price. All these features may differ in importance from one person to another, according to the need.

### 3 Research Methodology

#### 3.1 *Proposed Framework of the Model*

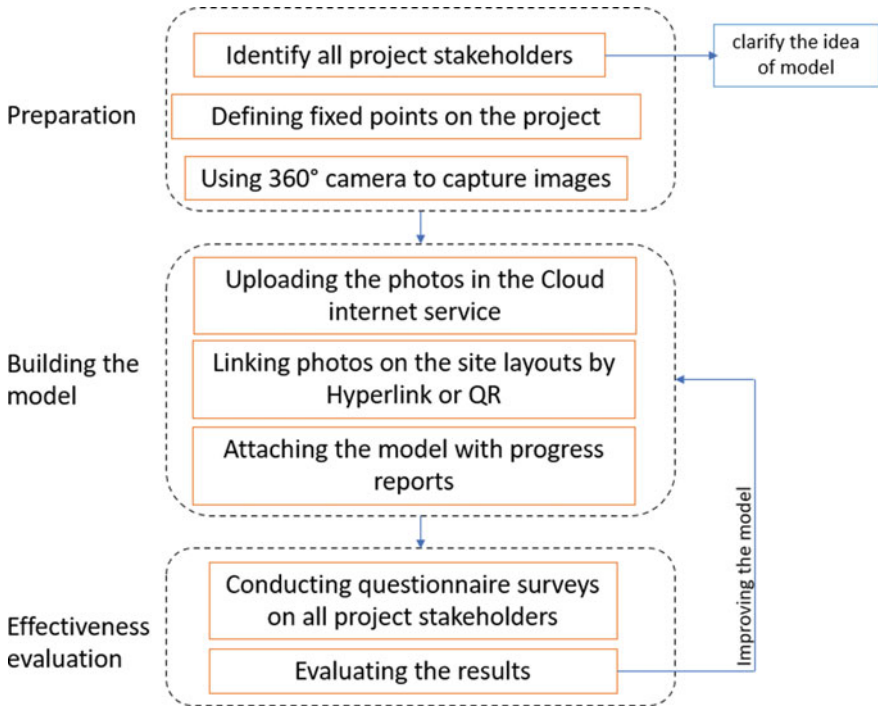
In this model, the proposed framework is based on three certain assumptions:

1. The preparation step requires the identification of the model idea and clarify the main purpose of integrating the model into progress reports to all project stakeholders such as engineers, owners, and contractors and informing them of how they will receive progress reports. Even if the model is available for use in the construction reports, there are challenges of adapting the new management method among stakeholders. Therefore, this step is highly recommended before starting to apply the smart 360-degree model to avoid the challenges and barriers of implement new method, for example, People: resistance to change, insufficient training. Technical: low levels of usability technology. Managerial: low levels of management commitment, a lack of client support, and a low level of utilization [5]. All these obstacles may facilitate the subsequent steps.

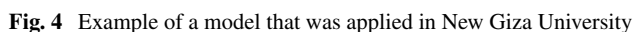
In the project site, fixed points should be identified with the help of a surveyor (if possible). These points help take fast pictures periodically and assist in several future uses, such as the comparison process when putting pictures side-by-side to follow work progress. Through the author's experience of using several 360° cameras and the reliance on the survey, Insta360 One X2 camera was an excellent option to use in this research. It has photo resolution 360: 6080 × 3040 (2:1); Panorama: 4320 × 1440 (3:1), the battery power is until 80 min recoding, Flow-State Stabilization: Better-than-ever stabilization and horizon leveling algorithms keep your shots steady—no gimbal needed, changeable memory one Micro SD up to 1 TB, and waterproof.

2. Building the model aims to set all the 360° panoramic images on one site layout sheet. A 360 panoramic image captures a location's whole 360-degree field of view. It frequently seems distorted when shown as a flat picture. Therefore, to make a 360 panoramic image, it needs at least two 180-degree images, and they are linked together by a technique called stitching images. Mostly, this is done automatically through the system loaded in the 360° camera. After that, captured project images are uploaded to online cloud platform whether it is free like the one that comes with Insta360 camera or any other cloud. This cloud allows access to files anytime and anywhere, all you need the Internet. Next step is to link the images in one sheet layout by using Hyperlink or QR. A symbol, graphic, or text that links to another file or object is known as a hyperlink, for example, in the model, the hyperlink is on the black dots in the layout also number of floors in schedule when you have multiple story building, see Fig. 4. Moreover, the QR is used when the progress reports on a paper. After linking the images in the layout, the model is ready to be used in the reports. And the enhancement of reports with

- images is become only one sheet. In the first time, building the model can be hard compared to apps like OpenSpace or HoloBuilder [10], but the model has several advantages such as less expensive and easier to share as PDF or a paper report.
3. The effectiveness evaluation step is required to measure the effectiveness of this model in progress reports. The questionnaire selected various criteria, including saving effort, time, and cost, resolving conflicts, and enhancing collaboration between the parties, increasing productivity, and reducing errors. The stakeholders involved were in a variety of areas of construction where the focus of impact measurement was on different parties of projects such as owners, engineers, and contractors. The aim of the evaluation is to examine the model by collecting answers from the contributors and analyzing it according to these criteria. This in turn will give an impression of the effectiveness of using the model in enhancing the reports and the possibility of its development in the future (Fig. 3).



**Fig. 3** Proposed framework of model

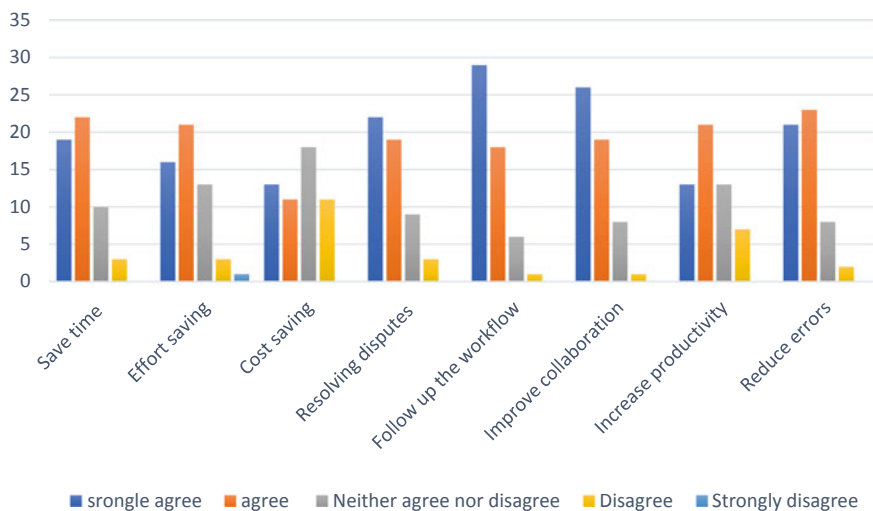


Smart 360-degree model and QR codes in progress reporting is a review and documentation method for the site that integrates 360-degree images through the interactive site layout to present it as a virtual tour. By linking the images through a hyperlink in the case of using a portable document format (PDF) or by using QR codes in the case of using a paper report, the location of images in project site can be found easily through layouts by click on the black dots or the numbers in schedule. This method makes you choose the place you want to see which gives your stakeholders—operations, finance, safety, owners, VDC, architects, trade partners, inspectors, and investors—remote access to your project sites and 360° views of current and past developments. Also, this method led to minimizing the number of reporting pages and covering the whole site from different directions instead of using several conventional photos that provide little context and insight and take up many pages in the report, see Fig. 4.

The first survey was conducted before applying the smart 360-degree model through a group of construction stockholders in the Middle East (Yemen, Saudi Arabia) and North Africa (Egypt). In the first part of the survey, the aim was to know the type of photo documentation used and to what extent they use images to enhance reports. In general, the vast majority of participants agree on using the phone camera as the

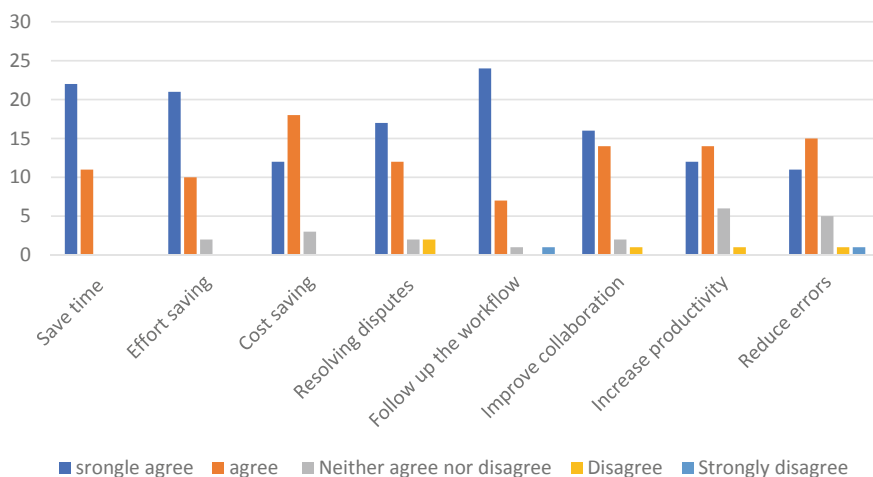
main device in photo documentation more than a normal or professional camera. Also, they agree on the benefits of using the photo documentation in progress reports whether daily, monthly, or annually. In addition, this survey presents that 89.1% of the 55 participants are not familiar with using 360° photos in the construction industry. These results give a great indication of a lack of knowledge about the integration of 360° photos to enhance the progress reporting. On the other hand, it can be noticed that most of them believe using 360° photos instead of conventional photos could play a significant role in enhancing the progress reports in several respects like save time, follow up the workflow and improve collaboration, see Fig. 5. Moreover, there were many suggestions such as using live cameras may improve reporting, using drone cameras in large projects such as pipelines and road projects, or linking 360 photography with some programs such as building information modeling BIM. Likewise, there were some questions from the participants about the cost of cameras, how to use them, and how to integrate 360° photos in the field of construction.

The pilot study to validate previously expected results was in three different projects in Egypt: first project, The New Giza University expansion project, which consists of five educational buildings: building (A) medical, building (B) dental, building (C) behind the house (BOH), building (D) pharmacy, building (E) student center and a total building area (55,000 m). The average number of building floors is four, and the daily average number of workers on the site is 150 workers. Second project, an eight-story residential tower and a total building area (3900 m) Third project, a two-story residential villa with total building area (225 m). All projects,



**Fig. 5** First survey on the expected outcomes to agree on the impact of using 360-degree images instead of conventional images in the preparation of progress reports in several respects





**Fig. 6** Second survey after using model in progress reports

in general, were under structural works. The proposed framework of the model was followed in these projects to test the effectiveness of 360° photos using progress reports (Figs. 3 and 6).

The second survey was conducted on stockholders in construction after using the model (integrating Smart 360-degree model and QR codes in progress reporting). Regarding the participants, they were from different fields of construction, engineers, contractors, owners, and 18.2% of them had less than 5 years of experience and 81.2% had more than 5 years of experience. The results show that 80% of participants of stockholders were satisfied and agreed to apply the proposed model instead of conventional pictures in progress reports. Moreover, most of them agreed that the reason for not applying the model is due to a lack of knowledge. The same evaluation criteria in first survey have been conducted after using the model, and it was noticed that the percentage of those who agree with applying the model to enhance the progress reports is more than those who disagree.

## 5 Discussions

The suggested framework of the model is a simulation of commercial applications that use 360° cameras in construction management Like OpenSpace, HoloBuilder, and StructionSite. Although the model is more difficult to use compared to the applications, it is still less expensive than them. After using this model instead of conventional pictures in progress report, it gave a high indicator of agreeing in saving time in the progress report. For example, one 360-degree picture equals around six conventional pictures, which means taking 40 shoots of 360 photos equals 240 conventional



pictures. Therefore, photo documentation becomes accurate and complete with no missing details on site because the 360° images cover the whole site or room. Moreover, the speed of decision-making in the reports and the saving of time are faster as a result of enhancing the report with 360° images that display the entire project as you are there. This also led to increase the rate of agreeing and strongly agreeing in terms of effort saving. In other words, instead of wasting numerous hours attempting to describe jobsite circumstances, the progress report with the smart 360-degree model representations that let your stakeholders to see the project site for themselves through secure, with role-based remote site access and handover. Despite the initial cost of the camera, it provides an indirect saving cost in the whole project like traveling costs, which means that this model in progress reports can reduce the need to be on-site frequently, which is important for stakeholders who overseen multiple projects at the same time. Increasing complete documentation and adding more details in progress reports leads to reduced errors and discrepancies in the future. Therefore, the rate of agreeing in using this model in progress reports to reduce the errors and resolve disputes was extreme. In the other side, the high of agreeing for using the model to enhance cooperation is due to the ease and flexibility of sharing it using many communication platforms, as the model is PDF, and it is also possible to print it on paper and then scan the QR. In terms of increasing productivity, the model provides a virtual walkthrough on-site, so all stakeholders could identify mistakes or requirements easily and make a fast decision. In addition, linking 360-degree pictures to site layouts makes it much easier to identify the location of the problem which increases the rate of productivity and saving time. Likewise, there were some obstacles and difficulties in applying the model, for example: The difficulties in the availability of the camera in the market and the overpricing of the camera have limited the widespread use of the technology. Despite the great demand on the part of the owner in applying the model, there was a kind of reservation and refusal on some stakeholders, such as contractors. On the pretext that the model increases the comments and observations by the owner or engineers that cause slowdown of work. However, the positive results from the model of in enhancing reports and the affirmative results in improving project management, training should be provided on the method of use, even if you will use of 360° photos through commercial applications.

## 6 Conclusion and Recommendations for Future Research

Enhancing progress reports by photo documents such as pictures or videos in the construction industry is essential to develop communication, transparency, and trust among stakeholders. In this paper, the model can be the first beneficial step for those who are interested in integrating 360-degree photos to enhance progress reporting or even in construction management, real estate, and virtual reality. Another potential benefit of this model is that it provides actionable insights in reports whether paper by using QR or electronic documents by using hyperlinks in multi-disciplinary projects. The positive results obtained from the survey after applying the model in

several projects give great indicators of using 360° photos instead of conventional photos in progress reports. Although the model is less expensive than commercial applications like OpenSpace, HoloBuilder, StructionSite, the applications are better in many ways, including data uploading, editing, and sharing with other programs such as BIM and Procore software. Also, we noted that as a result of the lack of knowledge of using the 360-degree camera, the applications are limited in use and academic studies are scarce. This research study can be extended in the future by being implemented in more case studies for long periods of time and that also is based on the feedback obtained in the surveys. As a result, the smart 360-degree model is the next step in creating more immersive and realistic 360-degree experiences. The good news is that a growing number of 360° camera manufacturers are incorporating 3D/volumetric capabilities into their products. Most of the project examples that we have covered in this paper are learning to create 360-degree environments as-built to enhance the reports using the model or other commercial applications. And, in general, these will unlock all kinds of potential to develop construction management. Technological advancements, such as 5G coverage, might open the way for a host of new 360/VR innovations, making the production and sharing process quicker and more economical than before. This pushes us to continue experimenting with other 360-degree camera techniques in the future.

**Acknowledgements** The cheerful assistance of Eng. Ayman El Hakea Systems Development Manager at CRC Dorra who offered access to apply the model in the New Giza University project and other residential projects is gratefully acknowledged. Financial support from the American University in Cairo and Hadramout Foundation is gratefully acknowledged.

## References

1. Afsari K, Halder S, Ensafi M, DeVito S, Serdakowski J (2021) Fundamentals and prospects of four-legged robot application in construction progress monitoring. *EPiC Ser Built Environ* 2:274–263. <https://doi.org/10.29007/cdpd>
2. Bohn JS, Teizer J (2010) Benefits and barriers of construction project monitoring using high-resolution automated cameras. *J Constr Eng Manag* 136(6):632–640. [https://doi.org/10.1061/\(asce\)co.1943-7862.0000164](https://doi.org/10.1061/(asce)co.1943-7862.0000164)
3. Cameron J, Gould G, Ma A, Chen A, Lui S (2021) 360 essentials: a beginner's guide to immersive video storytelling. Ryerson University Library Toronto. <https://pressbooks.library.ryerson.ca/360essentials/front-matter/introduction/>
4. Colligan M (2002) Canvas documentaries: panoramic entertainments in nineteenth-century. <https://books.google.com.eg/books?op=lookup&id=c5W0i25pSskC&continue=https://books.google.com.eg/books%3Fid%3Dc5W0i25pSskC%26printsec%3Dfrontcover%26hl%3Dar&hl=ar>
5. Demian P (2014) Barriers that influence the implementation of UK construction project extranets, July 2014
6. Ellis G (2021) 8 ways to make the most of construction photos—digital builder. <https://constructionblog.autodesk.com/construction-photos/>
7. Felli F, Liu C, Ullah F, Sepasgozar SME (2018) Implementation of 360 videos and mobile laser measurement technologies for immersive visualisation of real estate & properties

8. Jokela T, Ojala J, Väänänen K (2019) How people use 360-degree cameras. In: ACM international conference proceeding series, Oct 2019. <https://doi.org/10.1145/3365610.3365645>
9. Kangari R (1995) Construction documentation in arbitration. *J Constr Eng Manag* 121:201–208. [https://doi.org/10.1061/\(ASCE\)0733-9364\(1995\)121%3A2\(201\)?casa\\_token=6qrDG\\_5POIsAAAAA:8F40j3lLtvwoTarJrPW4z1F1JyzWe6FUYjovXFRoslwLPhDVYr78jQODQ3E5e31BOvEPvczFYVc](https://doi.org/10.1061/(ASCE)0733-9364(1995)121%3A2(201)?casa_token=6qrDG_5POIsAAAAA:8F40j3lLtvwoTarJrPW4z1F1JyzWe6FUYjovXFRoslwLPhDVYr78jQODQ3E5e31BOvEPvczFYVc)
10. Ozcan-Deniz G (2019) Expanding applications of virtual reality in construction industry: a multiple case study approach. *J Constr Eng Manag Innov* 2(2):48–66. <https://doi.org/10.31462/jcemi.2019.02048066>
11. San Francisco, 1851, from Rincon Hill—digital file from intermediary roll film copy | Library of Congress (n.d.). Retrieved 6 Mar 2022, from <https://www.loc.gov/resource/pan.6a01850/>
12. Smith J (2011) The life and death of building on photography and time. <https://yalebooks.yale.edu/book/9780300174359/life-and-death-buildings>
13. Sulaiman MZ, Aziz MNA, Bakar MHA, Halili NA, Azuddin MA (2020) Matterport: virtual tour as a new marketing approach in real estate business during pandemic COVID-19. In: *IMDES*, vol 502, pp 1–6. <https://doi.org/10.2991/assehr.k.201202.079>
14. Wehking F, Wolf M, Söbke H, Londong J (2019) How to record 360-degree videos of field trips for education in civil engineering. In: *Proceedings of DELFI workshops 2019*, Berlin, Germany, 16 Sept 2019, pp 177–188. <https://doi.org/10.18420/delfi2019-ws-120>
15. Yang J, Vela P, Teizer J, Shi Z (2014) Vision-based tower crane tracking for understanding construction activity. *J Comput Civ Eng* 28(1):103–112. [https://doi.org/10.1061/\(asce\)cp.1943-5487.0000242](https://doi.org/10.1061/(asce)cp.1943-5487.0000242)

# Envisioning Digital Twin-Enabled Post-occupancy Evaluations for UVic Engineering Expansion Project



Ishan Tripathi, Thomas Froese, and Shauna Mallory-Hill

**Abstract** The University of Victoria is in the process of expanding its engineering and computer science department to meet the growing demand for post-graduate programs by building two new buildings. UVic's Green Civil Engineering department is actively involved in the project and planning to use the buildings as experimental apparatuses for various building science and systems research such as energy, water, and indoor environmental quality. These buildings aspire to achieve net zero carbon certifications to promote innovations in sustainability. Post-occupancy evaluations (POE) provide scientific methods and tools to analyze how buildings function and to quantify their performance. First, this paper establishes the semantics of POE in the context of the new engineering expansion project along with project phases. Second, this paper discusses the digital twin execution plan that can guide the evolution of digital twins during each phase of the project life cycle for the purpose of POE. Third, this paper compares the proposed digital twin-based POE methodology with the conventional POE methodology. Conducting the POE on the UVic ECS expansion project will enable the researchers to determine the effectiveness of sustainable features by comparing the performance of existing and proposed facilities.

**Keywords** Post-occupancy evaluations · Digital twins · IoT sensors

## 1 Introduction

The Civil Engineering department at the University of Victoria (UVic) aspires to be 'the greenest' department in Canada by focusing on sustainable technologies for design, construction, and management of the built environment without compromising the natural environment [1]. To meet the ongoing demand in education and

---

I. Tripathi (✉) · T. Froese  
Civil Engineering, University of Victoria, Victoria, Canada  
e-mail: [itripathi@uvic.ca](mailto:itripathi@uvic.ca)

S. Mallory-Hill  
School of Architecture, University of Manitoba, Winnipeg, Canada

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_8](https://doi.org/10.1007/978-3-031-34593-7_8)

research, UVic is in the process of expanding the current Engineering and Computer Science (ECS) building and building a new ‘high-bay’ intelligent research lab [2]. Aspirations for the proposed buildings are to achieve net zero carbon certifications. Moreover, the Civil Engineering department is getting involved throughout the project life cycle’s planning, construction, and operations phases. This involvement aims to create a multi-disciplinary research platform that considers the buildings themselves as experimental apparatuses.

The design and construction of *high-performance* buildings have gained popularity during recent years [3]. However, only a few institutions worldwide are conducting building-scale research to evaluate the performance of in-use systems and capture complex interactions among buildings, people, energy systems and digital systems. UVic’s Civil Engineering department is addressing this challenge by proposing to develop a digital twin for monitoring, optimizing, and managing physical systems. Moreover, questions such as ‘*How does the research on building performance shape future design?*’ and ‘*Does constructing a sustainable facility really improve the footprint of the built environment?*’ often arise during the planning phase of the project. Post-occupancy evaluations can be used to make data-oriented and structured arguments toward finding answers to these questions. Moreover, digital twins are becoming a standard norm for the Architecture, Engineering, Construction and Operations (AECO) industry for managing design, construction, and operations [4]. The digital twin applications can also be extended for collecting, analyzing, and managing data for POE [5].

This study aims to provide a high-level methodology for a digital twin-enabled POE for UVic’s ECS expansion project. A systematic approach for integrating the semantics of the POE framework with a digital twin execution plan to create the overall methodology has been explained further in this paper (Sect. 4). Comparing this proposed methodology with a conventional methodology demonstrates the benefits of using digital twins for POE.

## 2 Points of Departure

### 2.1 Post-occupancy Evaluations

Post-occupancy evaluations (POE) are a systematic approach to determine whether decisions made by designers, constructors, and facilities managers meet the building performance and end-user requirements. They offer a wide range of benefits that mainly align with obtaining design and operation feedback for future projects, operational improvements, and benchmarking for comparing performance within the same facilities [6–8].

Typically, POE are conducted after the building is occupied for at least two years. The data is collected for parameters such as energy, water, indoor environmental quality (IEQ), and occupant behavior toward the building performance. The collected data is then used to calculate key performance indicators (KPIs) [9].

The methodology for conducting POE varies according to purpose and type of projects [6, 8]. However, challenges such as time-consuming data collection, lack of provisions for using the building sensors' data, and inefficient visualization can be observed across the majority of the projects [10]. Integrating IoT-based sensors with a GIS-based digital twin can address those challenges by providing a centralized platform for data management, analysis, and visualization for POE [5].

## ***2.2 Digital Twin Ecosystem for Smart Building Research***

The concept of a digital twin has evolved significantly during the past decade. There is no definitive definition of a digital twin. In the context of the Architecture, Engineering, Construction, and Operations (AECO) industry, a digital twin can be defined as a virtual replica of a physical asset [4]. Three key characteristics: representation of a physical building, bi-directional data exchange, and connection throughout the life cycle of the building [11] have contributed to the paradigm shift toward the use of digital twins for the AECO industry. IoT-based sensors have become a backbone of the digital twins by facilitating dynamic data gathering and data exchange [4].

One of the objectives of UVic's research program is to use the digital twin for simulation and predictive analysis that can be used to monitor, control, and optimize physical systems. Researchers can develop deeper understandings by coupling living labs with digital twins [12]. There are examples of 'smart living labs' that facilitate interdisciplinary research through experiments in actual conditions. The Smart Campus Integration and Testing (SCIT) Lab at Ryerson University, Toronto, Canada, supports pilot-scale research projects that focus on building controls, operations, and occupants' behavior toward building performance [13].

Building Information Models (BIM) provide geometric and parametric information to digital twins. A BIM execution plan is a document that envisions and documents the implementation of BIM throughout a project's lifecycle. It usually contains the guidelines for developing people and processes, mobilizing existing and new technologies, enhancing efficiency through teamwork, and collaboration through common data environments [14]. In the context of the UVic ECS expansion project, the fundamentals of the BIM execution plan can be used to develop a 'digital twin execution plan' that will provide guidelines for building and using digital twins for POE.

### 3 Methods

This section explains the approach to creating a methodology for conducting performance evaluation on the ECS expansion projects by using a digital twin as a centralized platform for Data Acquisition, Management, Analysis, and Visualization. The Civil Engineering department at UVic proposes using digital twins for various research purposes. Moreover, the use of digital twins along with IoT sensors will enable dynamic data exchange between the physical building and its digital twin. Therefore, conducting POE can be considered as an extended use case of digital twins.

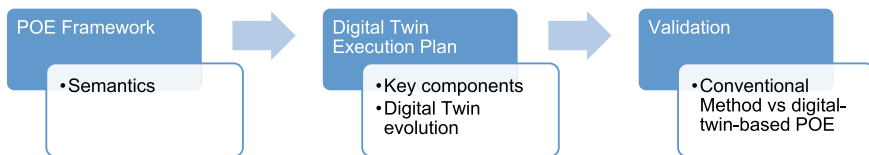
Figure 1 demonstrates the strategic approach considered for creating an overall framework that enables digital twins for POE.

#### 3.1 POE Framework

The tools and techniques for conducting POE vary with the type of project and overall purpose [8]. Therefore, it is important to establish a high-level framework for POE that explicitly defines the semantics such as purpose, objective, scope, phases, frequency, and key performance indicators (KPIs).

#### 3.2 Digital Twin Execution Plan

A digital twin execution plan uses the fundamentals of the BIM execution plan to provide a set of guidelines for developing a digital twin that accommodates the requirements of POE.



**Fig. 1** Strategic approach to creating a digital twin-based POE methodology for UVic ECS expansion project

### **3.3 Validation**

The recommended digital twin-based POE methodology will be compared to the conventional method.

## **4 Discussions**

### **4.1 Framework for Digital Twin-Based POE**

This framework helps to establish semantics for conducting both *high-level* and *detailed* POE by using digital twins. It defines the goals, objectives, levels, frequency, and KPIs of POE for the UVic ECS expansion project.

#### **4.1.1 Goals**

- Determine the 'Performance Gap' by comparing the simulation data with actual data,
- Determine the effectiveness of sustainable features by demonstrating the reduction of environmental footprint, increase in positive impact, and improvement in occupant comfort.

#### **4.1.2 Key Objectives**

- To determine gaps in expected and actual performance,
- Determine the occupant comfort level within the building,
- Provide recommendations for improvement in operations,
- Provide design feedback for future purposes,
- Monitor real-time data and use it for the realistic calibration of a simulation model periodically for predictive analysis.

#### **4.1.3 Scope**

For this project, POE should be limited to research purposes by the Civil Engineering department only. Usually, digital twins are managed by the facilities management team along with Building Management Systems (BMS). Separating POE from the scope of the Facilities Management (FM) team is necessary for ensuring safety and privacy. However, further collaboration is encouraged through structured data sharing and decision-making.



#### 4.1.4 Levels

Comparing the performance of existing facilities with the proposed new facilities is a standard way to prove the effectiveness of sustainable features. Therefore, POE can be classified into three levels.

**Level 0:** Pre-occupancy evaluations (Pre-OE) on the existing ECS building.

**Level 1:** Preliminary POE on the expansion and the new high-bay structural lab.

**Level 2:** Detailed/Advanced evaluations if required after preliminary evaluations.

Level 0 and level 1 evaluations should be conducted for the same set of KPIs for accurate comparison between existing and new facilities.

#### 4.1.5 Frequency of Evaluations

Traditionally, a preliminary POE is conducted after a minimum of 2 years of occupancy. However, with a paradigm shift from one-off to continuous, the frequency of evaluations should be decided prior to commencing as per the requirements set by the researchers and the facilities management team.

#### 4.1.6 Key Performance Indicators (KPIs)

KPIs are useful to determine the effectiveness of parameters such as energy, water, indoor environmental quality (IEQ) that affect the sustainable performance of the building. Table 1 gives a brief overview of potential parameters that can be used to evaluate the sustainability goals of the UVic ECS expansion project.

The relevant research groups at UVic can expand upon these basic KPIs to develop more rigorous indicators for detailed investigations.

### 4.2 *Digital Twin Execution Plan for POE*

#### 4.2.1 Applications of Digital Twins for UVic ECS Expansion

Determining potential applications of a digital twin during the planning phase is necessary for creating a digital twin execution plan for POE. The following are potential use cases for digital twins for the UVic ECS expansion project.

- *For research:* Simulations, predictive analysis, continuous monitoring, etc.
- *For facilities management:* Predictive maintenance and performance evaluations.
- Visualization and display of occupant comfort and satisfaction.

**Table 1** Overview of potential KPIs for UVic ECS expansion project

Sustainability goals	Relevant key performance indicators (KPIs)
<i>1. Reduce carbon emissions</i> <ul style="list-style-type: none"> <li>• Implementing sustainable transportation</li> <li>• Mass timber for reducing embodied emissions</li> <li>• Electric heat-pump for low carbon grid</li> </ul>	<ul style="list-style-type: none"> <li>• Materials and waste during construction</li> <li>• Carbon accounting for transportation</li> <li>• Carbon accounting for electric grids</li> </ul>
<i>2. High performance: energy and IEQ</i> <ul style="list-style-type: none"> <li>• LED light fixtures</li> <li>• High-performance insulation</li> <li>• Optimized window to wall ratio</li> <li>• Exterior solar shading</li> <li>• Solar panels for generating electricity</li> </ul>	<ul style="list-style-type: none"> <li>• Energy use intensity</li> <li>• IEQ: lighting, acoustics, temperature and relative humidity, indoor air quality (CO<sub>2</sub>, CO, TVOC, particulates)</li> </ul>
<i>3. Sustainable water</i> <ul style="list-style-type: none"> <li>• Low flow sanitary fixtures</li> <li>• Rain gardens</li> </ul>	<ul style="list-style-type: none"> <li>• Water use intensity</li> <li>• Water uses by source</li> <li>• Water use by end-use</li> <li>• Rainwater harvesting</li> </ul>
<i>4. Biodiversity</i> <ul style="list-style-type: none"> <li>• Bird-friendly design</li> <li>• Green roof</li> <li>• Restorative landscape</li> <li>• Indigenous plantations</li> </ul>	<ul style="list-style-type: none"> <li>• Biodiversity indicators</li> </ul>
<i>5. Occupant health and well-being</i>	<ul style="list-style-type: none"> <li>• Occupant satisfaction</li> <li>• Impact of occupant behavior on building performance</li> </ul>
<i>6. Economic factors</i>	<ul style="list-style-type: none"> <li>• Cost-feasibility of sustainable construction</li> </ul>

- Public interaction and awareness and display of positive environmental impact
  - BIM model displayed on a screen near the entrance or anticipated ‘high traffic areas’ for demonstrating design features and showing the live KPIs for occupant comfort levels.
  - For collecting feedback: at the end of the term, the occupant survey questionnaire can be distributed through UVic’s online systems for the students that used the building for a particular term.
- Gamification: for encouraging positive occupant behaviors toward building performance.

#### 4.2.2 Organizational Structure

- Based on the specialization of various research groups in the Civil Engineering department at UVic, having a separate digital twin is recommended for avoiding

confusion and complications. Those digital twins should have dynamic data exchange connections with a federated digital twin (see Fig. 4).

- Each research group should nominate a representative for managing their respective digital twin. We recommend hiring a departmental coordinator to monitor the overall quality of the digital twin, verify the integrity of connections, and communicate with the facilities management team.

### 4.2.3 IT Infrastructure

- **Hardware:** High-end graphics card for optimum performance, upgradable systems.
- **Software**
  - Autodesk BIM suite for curating BIM models and collaboration.
  - ArcGIS Pro for integrating IoT sensors with BIM models and hosting digital twins. Its spatial–temporal analysis and visualization tools will potentially enhance the capabilities of digital twins.
- **Other:** Connection with local data repositories/servers if applicable.

### 4.2.4 Components of Digital Twins

#### I. Geometry

Curated BIM models provide necessary geometric information according to the desired level of details.

#### II. Sensors—Data Types and Networking

Sensors enable digital twin for continuous monitoring, effective visualization, and informed decision-making [4]. Table 2 gives an overview of the potential type of sensors that can be used within their respective systems according to their purpose.

#### III. Data Connections

Connecting IoT-based sensors with the BIM model establishes a bi-directional communication between physical building and its digital twin. Figure 2 demonstrates integrating sensors' data with the BIM model using ArcGIS Pro (hypothetical building and data are used here for visual demonstration only).

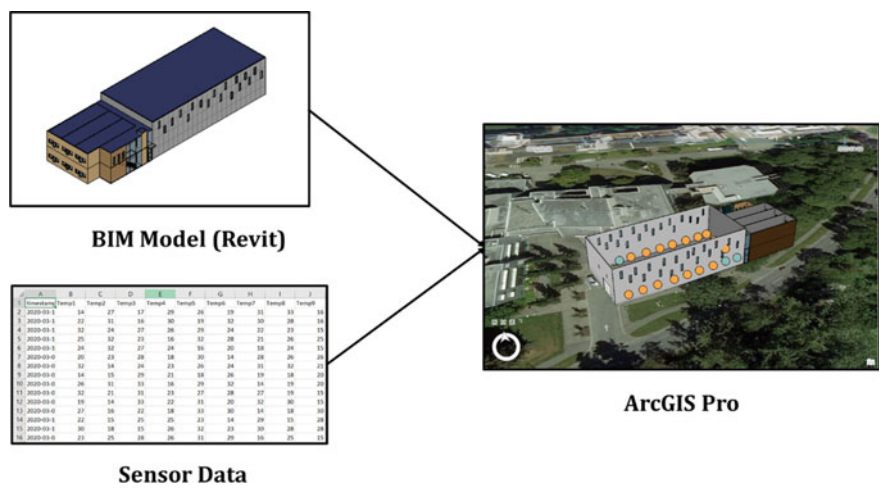
#### IV. Time: Refresh Rate

Setting refresh rate for sampling is vital for data collection and avoiding unnecessary clustering.

- **Sensors:** As per the requirements of the corresponding research group.

**Table 2** Potential type of sensors for UVic ECS expansion project

Purpose	Sensors	Related systems
Overall building performance	Thermal, lighting, HVAC, occupancy	Architectural, mechanical, electrical
Detailed area monitoring	Thermal, IEQ, occupancy	Mechanical, electrical
Water resources	Water flow, water use, soil moisture, potable water	Water
Building envelope	Heat flux, temperature and humidity, green roof	Mechanical, electrical
Structural health monitoring	Moisture, deflection, vibration, safety	Structural, building materials
Geo-technical monitoring	Settlement, strain, earth pressure, moisture	Geo-technical
Electrical	Usage, PV output	Electrical
Exposure to pollution	Mode of transport, occupancy	Outdoor air quality



**Fig. 2** Example of connecting BIM model with sensors’ data in ArcGIS Pro [5]

- **Federated Digital Twin:** Federated model should be updated periodically; for example: once every term.

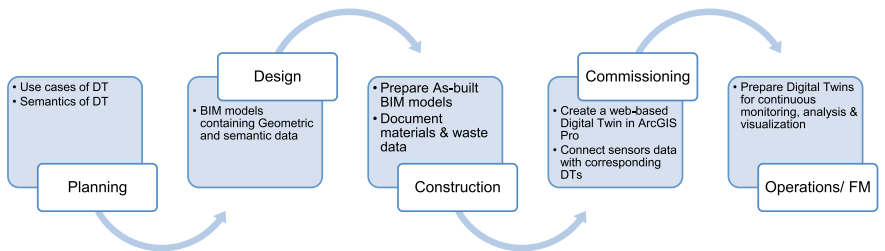
**4.2.5 Evolution of Digital Twins at Each Phase of the Project**

Figure 3 illustrates the tasks at each phase of the project life cycle that ultimately contributes to developing digital twins. For the purpose of POE, the commissioning phase is significant because it documents the initiation of different systems within

the building, and its accurate documentation can be useful for re-calibrating the simulation model.

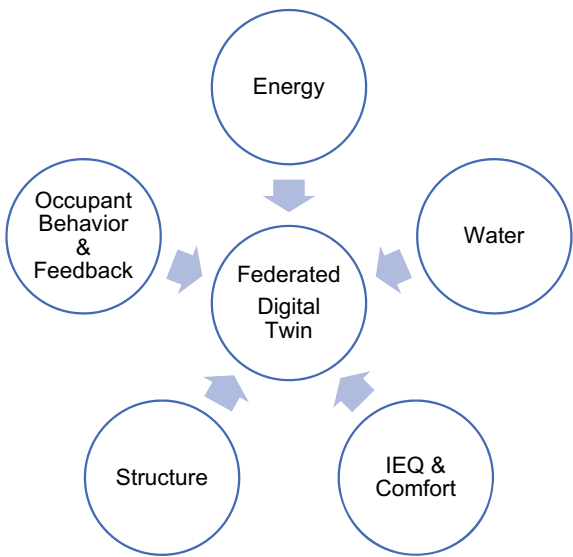
For the UVic ECS expansion project, a representative of each research group is required to create and update their version of the digital twin during the lifecycle.

As demonstrated in Fig. 4, key research areas for creating research-based digital twins are identified based on the research groups at the UVic Civil Engineering department. These research groups can manage their respective digital twins to avoid complications and privacy concerns. Moreover, a departmental representative should manage a federated digital twin that synchronizes the data from all research-based digital twins. The federated digital twin can be used to communicate with the facilities management team.



**Fig. 3** Evolution of digital twins

**Fig. 4** Conceptual illustration of a federated digital twin that is connected to various research-based digital twins



### 4.3 Challenges in Implementation

- During the planning phase, collaborating with stakeholders such as constructors is challenging, especially for communicating the concept of POE.
- Separating research-based POE from the overall scope of facilities management while establishing an operational feedback loop parallelly.
- Occupancy sensors in an educational building are challenging due to privacy and security concerns and ethical reasons.
- Uploading the building models on ArcGIS Pro cloud-based servers would be difficult because of safety and intellectual property rights concerns.

## 5 Comparison

This section compares the conventional POE process with the proposed digital twin-enabled POE process throughout the project life cycle.

Conventionally, facilities management commences during the *commissioning phase* (Table 3), where the BMS sensors' database is created after different mechanical and electrical systems have been activated. Moreover, the entire procedure for conducting POE starts during the *operations phase* [15]. Evidently, this has resulted in time-consuming data collection methods and a lack of provisions for visualization for effective communication [10].

For the digital twin-enabled POE process, relevant information can be added during each life cycle phase. Moreover, digital twins use the 3D BIM model and integrated sensors' data for effective visualization. Furthermore, ArcGIS Pro enhances spatiotemporal analytical capabilities by providing GIS-enabled advanced tools [5]. By implementing machine learning, the proposed digital twin can provide intelligent insights into building performance.

Therefore, as described in Table 3, integrating a POE framework with the digital twin evolution during the project life cycle improves the efficiency of data collection, analysis, and visualization of POE.

## 6 Conclusions

- Post-occupancy evaluation will present an opportunity to the University to lead by example by demonstrating the effectiveness and positive environmental impacts of green buildings.
- Using digital twins will enable researchers to conduct detailed analyses within their respective research groups and collaborate with other research groups.
- Integrating the POE framework with the digital twin execution plan helps create a methodology for digital twin-based POE. However, further discussion with all the stakeholders is required to optimize the methodology.

**Table 3** Comparison between conventional and digital twin-enabled POE processes

Project phases	Conventional POE process	Proposed digital twin-enabled POE process
Planning		<ul style="list-style-type: none"> <li>Propose the extended application of digital twins for POE to the stakeholders</li> <li>Establish semantics for POE and identify requirements of a digital twin platform for POE</li> </ul>
Preliminary/schematic design		<ul style="list-style-type: none"> <li>Propose the integration of IoT-based sensor networks for continuous monitoring</li> <li>Document the assumptions of occupancy</li> </ul>
Design development		<ul style="list-style-type: none"> <li>Document values provided by the simulation model for relevant KPIs such as energy use, water use, lighting levels, etc.</li> <li>Start the development of a digital twin for extended FM/POE</li> <li>Collaborate with designer for adding geo-location and built environment data</li> </ul>
Construction		<ul style="list-style-type: none"> <li>Track and collect data for materials and waste</li> <li>Track economic factors such as construction and commissioning costs for POE</li> <li>Monitor the sustainable construction procedure where applicable</li> <li>Update the digital twin using as-built model/drawing</li> </ul>
Commissioning	<ul style="list-style-type: none"> <li>Create a database for equipment and sensors for maintenance and FM</li> </ul>	<ul style="list-style-type: none"> <li>Calibrate and connect the IoT-based sensors' system and conduct preliminary monitoring tests</li> <li>Add sensors' locations to digital twin and establish a data tunnel for connecting the physical sensors to their digital representation</li> <li>Validate if the quality of the digital twin meets the requirements established during the planning phase</li> </ul>
Operations/FM	<ul style="list-style-type: none"> <li>Plan and conduct preliminary POE</li> <li>Generate KPI report and identify gaps by comparing the values with design standards</li> <li>Provide feedback for operations and design iteration</li> </ul>	<ul style="list-style-type: none"> <li>Create a plan for periodic evaluations using IoT sensors' data</li> <li>Conduct spot measurements if required and add them to relevant digital twins</li> <li>Create a digital twin-based feedback system for occupant surveys</li> <li>Generate KPI report and compare it with the simulated values to identify gaps</li> <li>Recalibrate the simulation model using sensors' data for future prediction</li> </ul>

- Digital twins will facilitate the streamlining of POE and increase the efficiency of the overall methodology.

## 7 Next Steps

- Present this high-level digital twin-enabled POE methodology to the Facilities Management team at UVic.
- Modify and detail the methodology as per the research and facilities management requirements.
- Propose pre-occupancy evaluations for the existing ECS and lab facilities used by the Civil Engineering department.

**Acknowledgements** We would like to thank the Civil Engineering department at the University of Victoria for sharing their vision and information about the UVic ECS expansion project.

## References

1. Green civil engineering—University of Victoria. UVic.ca. <https://www.uvic.ca/ecs/civil/home/green/index.php>. Accessed 21 Feb 2022
2. Engineering and computer science expansion and new lab building—University of Victoria. <https://www.uvic.ca/campusplanning/current-projects/engineering-expansion/>. Accessed 06 Mar 2022
3. O'Mara M, Bates S, Ap L. Why invest in high-performance green buildings?, p 24
4. buildingSMART International (2020) How to unlock economic, social, environmental and business value for the built asset industry. [Online]. Available: <https://www.buildingsmart.org/wp-content/uploads/2020/05/Enabling-Digital-Twins-Positioning-Paper-Final.pdf>
5. Tripathi I, Froese TM, Mallory-Hill S (2022) A BIM-IoT-GIS integrated digital twin for post-occupancy evaluations. Unpublished manuscript
6. Roberts CJ, Edwards DJ, Hosseini MR, Mateo-Garcia M, Owusu-Manu DG (2019) Post-occupancy evaluation: a review of literature. *Eng Constr Archit Manag* 26(9):2084–2106. <https://doi.org/10.1108/ecam-09-2018-0390>
7. Mallory-Hill S, Preiser WFE, Watson C. Introduction to building performance evaluation: milestones in evolution
8. Li P, Froese TM, Brager G (2018) Post-occupancy evaluation: state-of-the-art analysis and state-of-the-practice review. *Build Environ* 133:187–202. <https://doi.org/10.1016/j.buildenv.2018.02.024>
9. Mallory-Hill S, Gorgolewski M (2018) Mind the gap: studying actual versus predicted performance of green buildings in Canada. In: Preiser WFE, Hardy AE, Schramm U (eds) *Building performance evaluation*. Springer International Publishing, Cham, pp 261–274. [https://doi.org/10.1007/978-3-319-56862-1\\_20](https://doi.org/10.1007/978-3-319-56862-1_20)
10. Tripathi I, Froese TM (2020) Post occupancy evaluation in Canada: challenges, potential improvements and new frontiers. [Online]. Available: <https://dspace.library.uvic.ca/handle/1828/12338>. Accessed 30 Jan 2022
11. Trauer J, Schweigert-Recksiek S, Engel C, Spreitzer K, Zimmermann M (2020) What is a digital twin?—Definitions and insights from an industrial case study in technical product development. *Proc Des Soc Des Conf* 1:757–766. <https://doi.org/10.1017/dsd.2020.15>



12. Smart living lab. <https://www.smartlivinglab.ch/en/>. Accessed 06 Mar 2022
13. WZMH's newest project with Ryerson University will be the smart campus integration and testing hub (SCITHub). WZMH Architects, 26 Aug 2021. <https://www.wzmh.com/news/wzmhs-newest-project-with-ryerson-university-scithub>. Accessed 06 Mar 2022
14. Joblot L, Paviot T, Deneux D, Lamouri S (2019) Building information maturity model specific to the renovation sector. *Autom Constr* 101:140–159. <https://doi.org/10.1016/j.autcon.2019.01.019>
15. Pishdad-Bozorgi P, Gao X, Eastman C, Self AP (2018) Planning and developing facility management-enabled building information model (FM-enabled BIM). *Autom Constr* 87:22–38. <https://doi.org/10.1016/j.autcon.2017.12.004>

# Application of Compressed Sensing on Crowdsensing-Based Indirect Bridge Condition Monitoring



Qipei Mei

**Abstract** Bridges are a vital component of the public transit system. However, as such infrastructure systems age, they sustain various sorts of damage, decreasing their performance and service life dramatically. In this setting, effective and efficient bridge health monitoring is critical to lowering maintenance costs and extending the service life of existing bridges. Traditional monitoring techniques require sensors being installed on bridges, which is costly and time-consuming. This paper presents a novel crowdsensing-based methodology to monitor the health condition of bridges through a number of smartphones in moving vehicles, i.e., indirect monitoring. By collecting continuous data from the smartphone users and extracting features from the data while they cross the bridge, the damage can be identified through quantifying the difference of the distributions of the features. The continuous data collection and transmission with high sampling frequency pose a particular challenge to the participation of the public, because this could drain the smartphone battery and data plan quickly. In this paper, compressed sensing is introduced into this crowdsensing framework. The compressed sensing can recover the signal from much fewer samples than the ones required by Nyquist–Shannon sampling theorem through random sampling, which leads to more efficient data collection and transmission. Numerical analysis is conducted to validate the effectiveness of compressed sensing on indirect bridge condition monitoring.

**Keywords** Condition monitoring · Compressed sensing · Bridges

## 1 Introduction

Bridges, as a critical component of the public transportation system, are deteriorating due to a variety of factors. This type of degradation significantly impairs the performance and service life of bridges [4, 7, 11]. Furthermore, delayed treatment

---

Q. Mei (✉)  
University of Alberta, Edmonton, Canada  
e-mail: [qipei.mei@ualberta.ca](mailto:qipei.mei@ualberta.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_9](https://doi.org/10.1007/978-3-031-34593-7_9)

123

might result in catastrophic complications, as well as the loss of property and life. As a result, it is vital to create effective and efficient ways for the early detection of deterioration and damage [11].

Researchers have developed techniques in recent years to detect potential damage by deploying and installing different sensors on the bridges [1, 2, 4, 8, 11, 16, 18]. Despite their success in detecting damage, these technologies, however, are costly and rely on the expertise of specialists [5]. The sensors' power source is also a concern. As a result, only the major bridges can afford such monitoring systems [10, 14]. Typically, short to medium-span bridges lack sensors and so cannot be monitored in a timely manner in such ways.

To address the aforementioned challenges, a number of researchers have proposed that vehicles equipped with various types of sensors be used to monitor bridges with citizen participation in order to boost efficiency and save costs for municipal departments [10, 12] owing to the high mobility of vehicles and high accessibility of mobile sensors (e.g., smartphones). The overall framework of bridge condition monitoring using crowdsensing data is illustrated in Fig. 1. In this framework, while vehicles pass across bridges, the data from smartphones in those vehicles are collected. Analyzing the data from different vehicles at different times, the damage in bridge is expected to be identified.

In previous work [12, 14], the author and colleagues proposed a novel method that can identify the damage in bridges using data collected from smartphones in a large number of vehicles. In the work, we have shown that the accelerometers in different smartphones can provide comparable accuracy to professional sensors, and they are sufficient for bridge health monitoring purpose [12]. The method was developed based on the assumption that the distribution of the features extracted from the vehicles should stay similar if the bridge state does not change. Therefore, the damage can be identified by measuring the shift of the distribution of features.

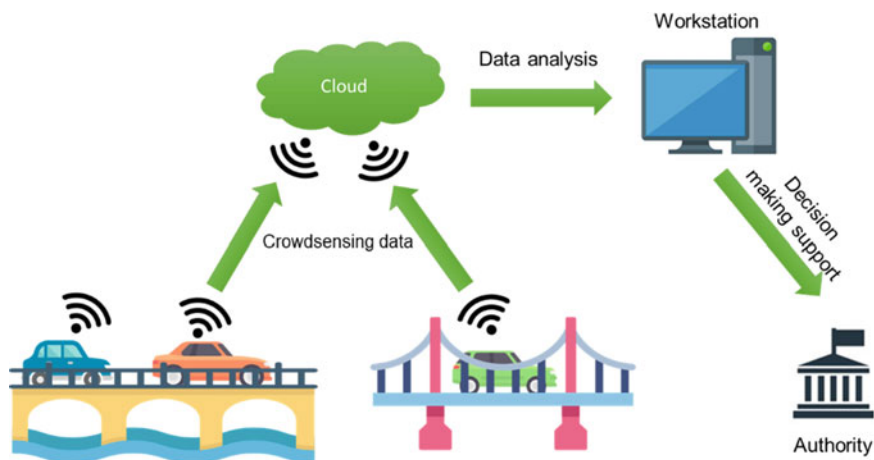


Fig. 1 Framework of crowdsensing-based bridge health monitoring

This technology has the potential to overcome issues with traditional monitoring technology with fixed sensors. First, it may drastically cut maintenance costs by eliminating the need for special monitoring measures. When vehicles cross a bridge, smartphones may automatically capture the vibration data. Second, this technology has the capacity to continuously monitor a population of bridges as long as there are vehicles passing across them. Third, because this approach is crowdsourced, it is more robust to operational impacts and human errors.

Even though the proposed framework has the potential to monitor the state of bridges efficiently, there are still challenges with it. One of the biggest ones is the dependent on the continuous collection of vibration data from smartphones. Such collection and transfer of data will consume the battery and data plan of smartphone users, which may influence the enthusiasm of the public to participate into the monitoring. Therefore, it is important to develop methods that reduce the amount of data that need to be collected and transferred.

This paper presents a method called compressed sensing that uses random sampling instead of uniform sampling for smartphone data collection. In this way, the number of data points could be fewer than the one required by the Nyquist–Shannon theorem [6]. The collected data are then applied to crowdsensing-based bridge health monitoring.

## 2 Methodology

### 2.1 Overview

In this study, compressed sensing is applied to reduce the amount of vibration data to be collected by smartphones. The original vibration signal is reconstructed from a small amount of randomly sampled data. Then, the data is applied to the method that was improved from the one developed previously by the author and his colleagues for crowdsensing-based bridge health monitoring [12]. In this method, the existence of damage is detected by analyzing the vibration data collected from smartphones in a large number of vehicles. The underlying assumption of the method is that the features extracted from the vibration data should represent the status of the bridge if a large amount of data is collected [12].

Overall, the procedure of the method presented in this paper can be summarized in three steps. First, compressed sensing is used to collect the vibration data, and reconstruction is conducted to recover the original data. Then, Mel frequency cepstral coefficients (MFCC) are extracted for each vehicle. At last, the Wasserstein distance is calculated as damage features based on the MFCCs to identify the damage in the bridge.

## 2.2 Compressed Sensing

Compressed sensing is a novel signal processing technique that was introduced in recent years, which can reconstruct the signal with much fewer samples required by Nyquist–Shannon sampling theorem [9]. In compressed sensing, we first get the signal  $x$  through random sampling, and then we need to find the  $X$  in frequency domain that can reconstruct the signal which matches  $x$  at the points of random sampling. Also, the sparsity of the acceleration signal in frequency domain is expected. Therefore, the objective function can be designed as follows for the compressed sensing.

$$\min_X \left[ (AX - x)^2 + \gamma \sum_{i=0}^n |X|_i \right], \quad (1)$$

where  $X$  is the spectrum in frequency domain,  $A$  is the multiplication of two matrices,  $\phi$  and  $\psi$ .  $\phi$  is the randomly sampling matrix and  $\psi$  the operator for inverse discrete cosine transform.  $\gamma$  is the regularization parameter. Using any optimization package, we can find the solution of  $X$ .

## 2.3 Mel Frequency Cepstral Coefficients (MFCC)

MFCC was first developed to imitate the auditory system of a human person [3]. It was introduced for indirect bridge health monitoring by the author and his colleagues [13]. One of its critical processes is to analyze the data using a collection of triangle filters with irregularly spaced intervals, which takes into account the fact that differences in lower frequencies are more important than differences in higher frequencies in bridges. The calculation of MFCC can be summarized as follows:

- (1) Calculate the Fourier transform of the acceleration data. This will convert the signal from time domain to the frequency domain, i.e., power spectrum.
- (2) Apply a series of triangular filters to the power spectrum. Each triangular filter is defined as below:

$$H_i(k) = \begin{cases} 0 & k < f_{i-1} \\ \frac{k-f_{i-1}}{f_i-f_{i-1}} & f_{i-1} \leq k < f_i \\ \frac{f_{i+1}-k}{f_{i+1}-f_i} & f_i \leq k < f_{i+1} \\ 0 & k \geq f_{i+1} \end{cases}, \quad (2)$$

$$f_i = M^{-1} \left( M(f_1) + (i-1) \times \frac{M(f_n) - M(f_1)}{n-1} \right), \quad (3)$$

where  $f_i$  denotes the  $i$ th Hertz-scale frequency with regard to the  $i$ th evenly spaced Mel-scale frequency. The functions  $M$  and  $M^{-1}$  represent the mappings between Mel-scale and Hertz-scale frequencies are shown as below.

$$M(f_i) = 5 \ln \left( 1 + \frac{f_i}{5} \right), \quad (4)$$

$$M^{-1}(m_i) = 5(e^{m_i/5} - 1). \quad (5)$$

- (3) Calculate the logarithms of the powers (sum of the multiplication of triangular filters and power spectrum) at each of the Mel frequencies using Eq. (6).

$$\log E(i) = \log \left( \sum_k H_i(k) \times X \right). \quad (6)$$

- (4) Take the Discrete Cosine Transform (DCT) of the logged powers to extract the MFCCs.

$$\text{MFCC}_j = \sum_{i=2}^{n-1} \log E(i) \cos \left[ \frac{\pi}{n-2} \left( i - \frac{3}{2} \right) j \right]. \quad (7)$$

Generally, only some selected MFCCs are used for further damage detection. More details of the MFCC calculation process can be referred to Mei et al. [13].

## 2.4 Damage Detection

There is a basic assumption in the method proposed in this paper, which is that the features extracted from the vehicles should follow some stable distributions if the bridge condition does not change. Damage can be detected by comparing the distributions of two sets of features at different times from different vehicles. Herein, the Wasserstein distance [15, 17] is used to measure the distance of distributions for each feature, and damage feature is defined as the sum of all the Wasserstein distances.

$$\text{DF} = \sum_{i=1}^q \int_{-\infty}^{\infty} |U_i - V_i|, \quad (8)$$

where  $q$  is the number of MFCCs for damage detection.  $U_i$  is the cumulative distribution function (CDF) of  $\text{MFCC}_i$  for the baseline case (intact bridge), and  $V_i$  is the CDF of  $\text{MFCC}_i$  for a damage case.

### 3 Numerical Analysis

To validate the proposed method, a number of numerical analysis are conducted in the finite element software Abaqus. In the analysis, a simply supported bridge is modeled with a spring-mass model passing across it to simulate the moving vehicle. The bridge is 25 m long and is constructed of reinforced concrete with a density of  $2400 \text{ kg/m}^3$ . The elastic modulus of the concrete is 27.5 GPa. The bridge's cross section measures  $2.0 \text{ m}^2$  in size and has a moment of inertia of  $0.12 \text{ m}^4$ . The bridge's first three frequencies are 2.08, 8.33, and 18.75 Hz.

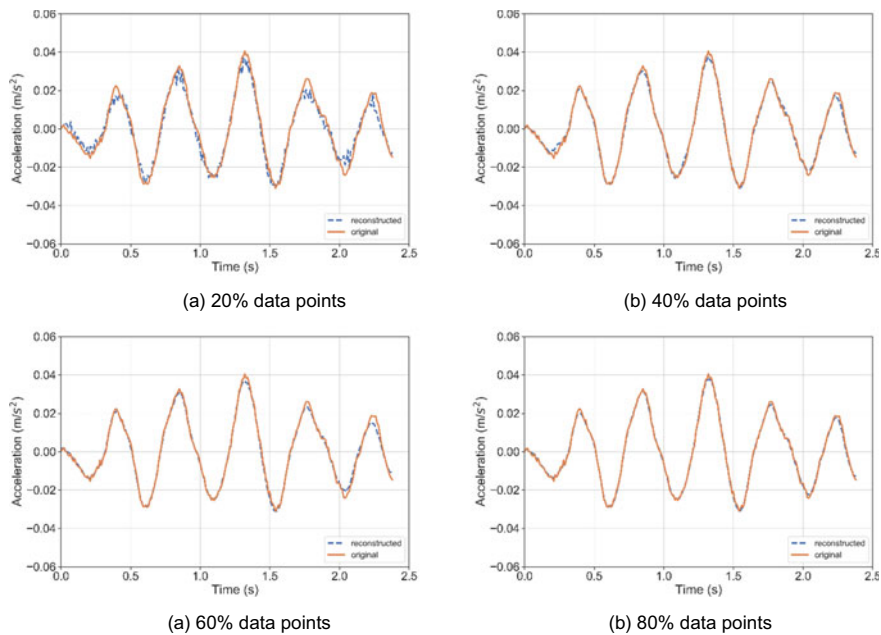
Five damage cases as well as an intact case are simulated on the bridge. In the intact case, the bridge has no damage. In DC1a and DC1b, we apply 15 and 30% stiffness reduction to the mid-span of the bridge. In DC2a and DC2b, we apply 15 and 30% stiffness reduction to the quarter span of the bridge. In DC3, the boundary conditions at both ends are changed to fix.

To simulate the situation that multiple vehicles pass across the bridge, the spring constant, mass and speed of the spring-mass model are changed at different levels. The mass of the vehicle can be 960, 1200, 1440, 1680, 1920, 2160, 2400 kg. The spring constant can be 200, 250, 300, 350, 400, 450, 500 kN/m. The speed are 28.8, 36, 43.2, 50.4, 57.6, 64.8, and 72 km/h. Combining the possible values for these three parameters, we have  $7 \times 7 \times 7 = 343$  simulations, where each of them would be corresponding to one set of vehicle configurations. Each vehicle is repeated 5 times with 5% artificial noise added. Therefore, for each damage case of the bridge, there are in total  $343 \times 5 = 1715$  vehicles passing across it. To consider the fact that different vehicles (but with similar distribution) will pass across the bridge at different times, 50% of all the data entries will be randomly sampled for each damage case. For the intact case, we randomly sample the 1715 data entries twice with 50% each time, where one of them is used as the baseline and the other one is for validation.

## 4 Results and Discussion

### 4.1 Vibration Signal Reconstruction

First, we evaluate the performance of compressed sensing for a given acceleration data entry. The data we choose herein are the vehicle with a mass of 1200 kg, a speed of 36 km/h and a spring constant of 500 kN/m. Since the finite element software cannot do randomly sampling, the acceleration data collected from the vehicle with a uniform sampling frequency of 100 Hz are first extracted. Then, random sampling is conducted on the extracted data to simulate the compressed sensing. Eventually, the full signal is reconstructed from the compressed sensing data. Therefore, for the compressed sensing with 20% data points, we are using only 20% of data points to reconstruct the original acceleration signal.



**Fig. 2** Comparison of original and reconstructed data with different level of random data points

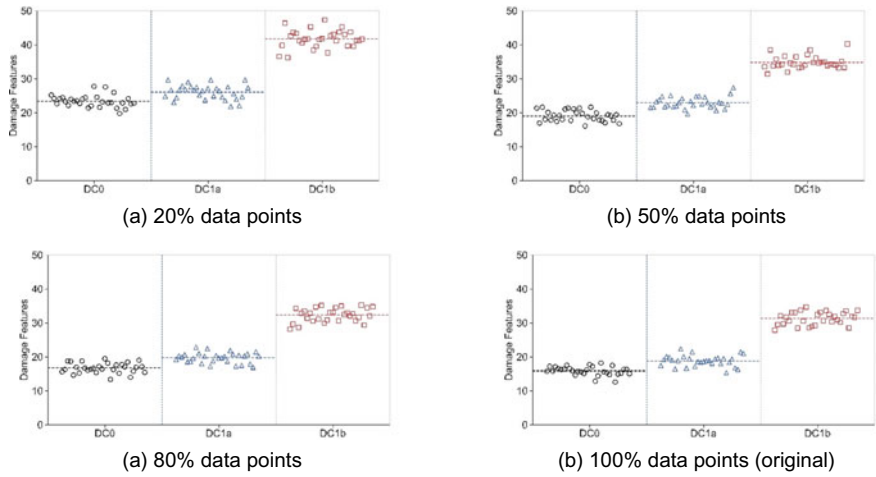
Figure 2 shows the reconstruction results from 20, 40, 60, and 80% data points. From Fig. 2, we can see that the use of even only 20% of the data points can reconstruct the original acceleration signal very well. More data points lead to better reconstruction. 40% data point can result in quite good reconstructions. The  $R^2$  of the reconstructed data and original data are 0.957, 0.989, 0.989, and 0.996 for 20, 40, 60, and 80% data points.

## 4.2 Damage Detection

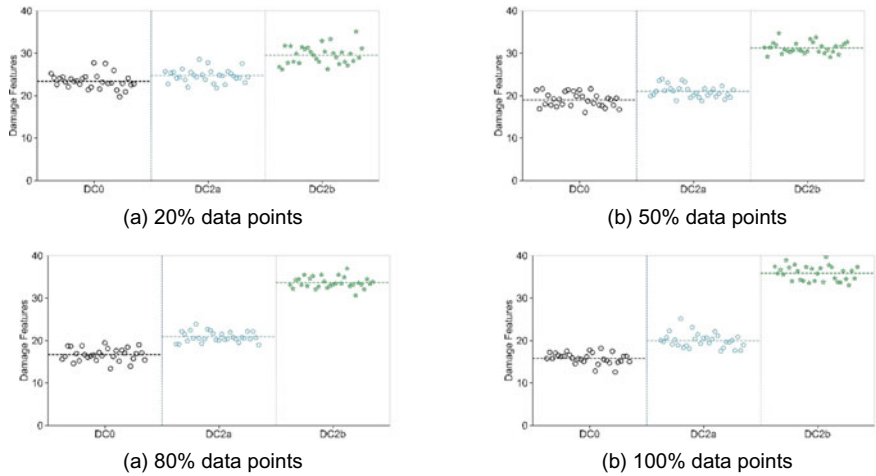
Then, the reconstructed data are applied for damage detection. To demonstrate the method's robustness, the 50% random sampling is performed 30 times, with the data for the baseline, validation, and damage instances being picked differently each time. Therefore, there are 30 DFs for each damage case. Figures 3, 4 and 5 represent the damage detection results for five damage cases with different levels of data points for compressed sensing.

Figure 3 shows the comparison of damage features (DFs) for the validation as well as DC1a and DC1b cases. Three different levels of data points (20, 50 and 80%) are used for reconstruction and damage detection. We can see from Fig. 3 that all damage has been detected successfully. When only 20% data points are used, the results are more sparsely distributed. However, the average of DFs still shows



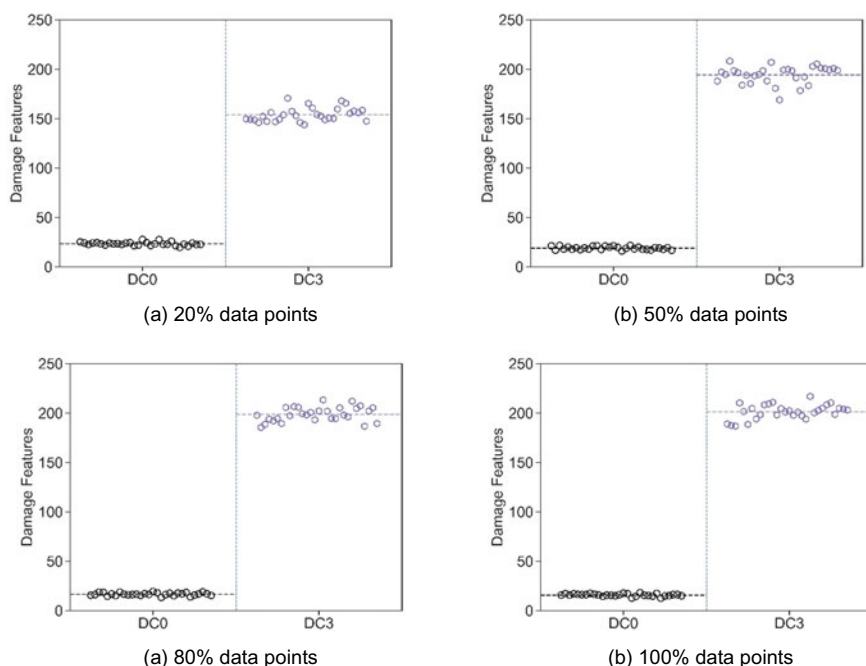


**Fig. 3** Damage cases DC1a and DC1b



**Fig. 4** Damage cases DC2a and DC2b

the existence of damage. The results are closer to the one with original data when more data points are used. Figures 4 and 5 show similar patterns which demonstrate that compressed sensing will not affect the performance of damage detection while reducing the amount of data to be collected.



**Fig. 5** Damage case DC3

## 5 Conclusions

This paper investigated the feasibility of applying compressed sensing to crowdsensing-based indirect bridge health monitoring. Numerical analysis with multiple damage cases and with data collected from moving vehicles was conducted in this study. The preliminary numerical results show that compressed sensing can significantly reduce the amount of data that have to be collected, while still maintaining the quality of damage detection. In the future, studies on laboratory experiments will be conducted to further validate the proposed method.

## References

1. Bao Y, Shi Z, Wang X, Li H (2017) Compressive sensing of wireless sensors based on group sparse optimization for structural health monitoring. *Struct Health Monit* 1475921717721457
2. Catbas FN, Gul M, Burkett JL (2008) Conceptual damage-sensitive features for structural health monitoring: laboratory and field demonstrations. *Mech Syst Signal Process* 22:1650–1669
3. Gowdy JN, Tufekci Z (2000) Mel-scaled discrete wavelet coefficients for speech recognition. In: 2000 IEEE international conference on acoustics, speech, and signal processing, 2000. ICASSP'00. Proceedings. IEEE, pp 1351–1354

4. Gul M, Catbas FN (2011) Damage assessment with ambient vibration data using a novel time series analysis methodology. *J Struct Eng* 137:1518–1526
5. Hoult NA, Fidler PRA, Hill PG, Middleton CR (2010) Long-term wireless structural health monitoring of the ferryby road bridge. *J Bridge Eng* 15:153–159
6. Kazimierczuk K, Orekhov VY (2011) Accelerated NMR spectroscopy by using compressed sensing. *Angew Chem Int Ed* 50:5556–5559
7. Kim C-W, Chang K-C, Kitauchi S, McGetrick PJ (2016) A field experiment on a steel Gerber-truss bridge for damage detection utilizing vehicle-induced vibrations. *Struct Health Monit* 15:174–192
8. Loh C-H, Hsueh W, Tu Y-C, Lin J-H, Kuo T-J (2018) Vibration-based damage assessment of structures using signal decomposition and two-dimensional visualization techniques. *Struct Health Monit* 1475921718765915
9. Mascareñas D, Cattaneo A, Theiler J, Farrar C (2013) Compressed sensing techniques for detecting damage in structures. *Struct Health Monit* 12:325–338
10. Matarazzo TJ, Santi P, Pakzad SN, Carter K, Ratti C, Moaveni B, Osgood C, Jacob N (2018) Crowdsensing framework for monitoring bridge vibrations using moving smartphones. *Proc IEEE* 106:577–593
11. Mei Q, Gül M (2016) A fixed-order time series model for damage detection and localization. *J Civ Struct Health Monit* 6:763–777
12. Mei Q, Gül M (2018) A crowdsourcing-based methodology using smartphones for bridge health monitoring. *Struct Health Monit* 1475921718815457
13. Mei Q, Gül M, Boay M (2019) Indirect health monitoring of bridges using Mel-frequency cepstral coefficients and principal component analysis. *Mech Syst Signal Process* 119:523–546
14. Mei Q, Gül M, Shirzad-Ghaleoudkhani N (2020) Towards smart cities: crowdsensing-based monitoring of transportation infrastructure using in-traffic vehicles. *J Civ Struct Health Monit* 10:653–665
15. Panaretos VM, Zemel Y (2019) Statistical aspects of Wasserstein distances. *Annu Rev Stat Appl* 6:405–431
16. Soh CK, Tseng KKH, Bhalla S, Gupta A (2000) Performance of smart piezoceramic patches in health monitoring of a RC bridge. *Smart Mater Struct* 9:533
17. Villani C (2009) The Wasserstein distances. In: *Optimal transport*. Springer
18. Zhu D, Guo J, Cho C, Wang Y, Lee K-M (2012) Wireless mobile sensor network for the system identification of a space frame bridge. *IEEE/ASME Trans Mechatron* 17:499–507

# Adoption and Implementation of Common Data Environments in the Province of Quebec: Barriers, Challenges, and Trends



Erik Andrew Poirier and Margaux Soye

**Abstract** The ISO 19650 framework for information management in the built asset industry has formalized the notion of common data environments (CDE) on a global scale. Yet, as with any new concepts and their operationalization, there remains a lot of work to be done to adopt, implement, and deploy the practices, processes, tools, and technologies underlying the CDE at both the project and asset level. The work presented in this paper discusses the results of a survey undertaken in the Province of Quebec to understand the rate of CDE adoption and implementation, the barriers and challenges, and the current trends in the province. The paper also regroups and formalizes the concept of CDE based on a broad review of the literature to understand and better frame the functions and the practices supported through CDEs. Unsurprisingly, the results show relative confusion as to what a CDE is and how it should be articulated within a project setting. Low awareness of current tools and what constitutes a CDE, its required functions and its general deployment in practice were observed. A series of recommendations are formulated to better frame and raise awareness on the concept of CDEs, namely its potential benefits and uses cases. The results indicate that there is still a lot of work to be done to support the structured adoption of CDEs to ensure their successful implementation and use.

**Keywords** Quebec · Common data environments

## 1 Introduction

The advent of “Level 2” Building Information Modeling (BIM) and its transition to “BIM according to ISO 19650” over the past few years is predicated upon *centralized* and *federated information container*-based collaboration between built asset life cycle stakeholders. The common data environment (CDE) is the core principle around which this federated, container-based working and management happens. The

---

E. A. Poirier (✉) · M. Soye  
École de Technologie Supérieure, Montréal, Canada  
e-mail: [erik.poirier@etsmtl.ca](mailto:erik.poirier@etsmtl.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_10](https://doi.org/10.1007/978-3-031-34593-7_10)

notion of CDE, acting as the “*single source of truth*” for a project and/or assets information across its life cycle, has existed in other domains for a number of decades [1]. It has been discussed as a concept in the built environment for an equally long time [13] and has been gaining traction over the past two decades [21]. The PAS 1192 series of standards, which have since become the ISO 19650 series, have popularized the concept of the CDE and its necessity within the built asset industry to successfully transition to digital ways of working supported through BIM [6, 14]. The concept of CDE is therefore gaining significant traction in the built asset industry and as with any emerging concept, a lot of vagueness still surrounds it, as was the case with the emergence of BIM in the 2000s [11]. This vagueness can significantly impact a CDEs implementation trajectory at the project, organizational, and industry levels. While there have been some studies looking into the core concepts and principles of CDEs, there is a dearth of research on their actual implementation.

The overall objective of this study was therefore to investigate the adoption and implementation of CDEs in the Quebec built asset industry. There was a need to first understand and frame the concept of CDEs. As such, a review of the literature was pertaining to the principles and concepts as well as the definitions of CDEs. Once defined, a review of features offered by CDEs and identified in past research and within industry was performed. Finally, a survey was conducted within the province of Quebec to understand adoption rates and implementation barriers.

The results are articulated in three parts. First, through the review of concepts and definitions, a framing of the concept of CDE is provided. A CDE is conceptualized as *a mutually agreed to digital collaborative ecosystem that enables the storage, exchange, use and management of information throughout an asset’s life cycle, using standardized processes and workflows*. Second, 323 discrete features distributed across 28 sub-categories and six overarching categories were identified. Lastly, the survey findings indicated confusion as to what constitutes a CDE, and 15 specific barriers across four categories: contextual, technological, procedural, and social/organizational. This work lays the foundation for a broader, multi-year research program undertaken by the *Groupe de Recherche en Intégration et Développement Durable* (GRIDD) at the *École de Technologies Supérieure* (ÉTS) on the topic of CDEs and BIM-enabled collaborative working.

## 2 Literature Review

The accelerating rate of BIM adoption and implementation, and the shift in practice that it has prompted, has cast light on the necessity to integrate project and asset information across its life cycle. This push for integration is fundamentally transforming how built asset industry actors collaborate and manage, consume, and exchange information [16]. The concept of CDE underpins this integration, supporting, on a first level, federation of discrete information containers, in a progression toward integration of data sources at the object level. Indeed, according to Comiskey et al. [8], technology is the key to evolving the traditional approach to project delivery, by

enabling the deployment of collaborative environments and 3D models embedded in a BIM process. The multi-disciplinary nature of the sector requires information that is stored, shared, managed in a more structured and standardized way, that is current and available [8]. This is where a CDE becomes indispensable. Indeed, according to Mordue [15], the CDE has two fundamental goals: (1) provide access to all stakeholders to up-to-date, reliable information in a structured and easily accessible format and (2) encourage the management, creation, assurance, sharing, disclosure, and coordination of generated information. Moreover, according to Radl and Kaiser [19], information deficiencies or shortcomings are resolvable through appropriate implementation of the CDE in projects.

## **2.1 CDE Definitions**

Several definitions of CDEs have been provided over the past few years. Indeed, CDEs have been defined within academia, standards, and industry literature. From a standards point of view, PAS 1192-2:2013 defines a CDE as “a single source of information for storing, sharing, and managing documents in graphical, structured, and unstructured forms.” ISO 19650-1:2018, defines as an “agreed-upon source of information about a given project or asset, used to collect, manage, and disseminate each container of information through a managed process.” According to DIN SPEC 91391-1, “a CDE is a web-based platform for managing processes and information throughout the building life cycle.” The standard goes on to state that “A CDE is the central storage and reference point for all project-related information. The processing of information is coordinated, redundancies are avoided, and current data is always available. Information is exchanged through structured interfaces (according to ISO 19650). The access to the CDE features and the exchange of information is done via the Internet. All project participants take care of their project-specific information exchange using the CDE. This way, project tasks and statuses are always traceable to a central location.”

In the scientific literature, Preidel et al. [17] define the CDE as “a common digital project space, which provides well-defined access areas for project stakeholders, combined with clear state definitions and a robust workflow description for sharing and approval processes.” Comiskey et al. [8] define the CDE as “an Internet-enabled cloud hosting platform, accessible to all construction team members to access shared project information.” Mordue [15] instead compares the CDE to a “technology stack,” i.e., a combination of several software products as opposed to a single product. “It’s not just a technology solution it’s a combination of software and processes” [15]. Bucher and Hall [7] see CDE as “a tool based on the idea of connecting a collaborative project space with cloud technology to form a collaborative platform. The goal is for all participants to move toward a more holistic and collaborative way of managing the project.” They see CDE as the next evolution of BIM. “The term CDE refers to

the use of cloud technology to create a central project or space to manage model-based project management” [7]. Table 1 identifies the various elements that have been developed to define CDEs in the literature.

**Table 1** Elements developed in the literature to define CDEs

	BSI PAS 1192-2:2013	ISO 19650-1:2018	DIN SPEC 91391-1 [10]	Preidel et al. [17]	Comiskey et al. [8]	Mordue [15]	Bucher and Hall [7]
Single source of information	x		x				
Store, share and manage/ information management	x	x	x				
Managed process/robust workflow description for sharing and approval processes	x	x	x	x			
Notion of information container		x					
Well-defined access areas				x			
Clear definition of states		x	x	x			
Access to information only					x		
Combination of software and processes						x	
Collaborative platform				x	x		x
Use of the cloud			x	x	x		x
Throughout the building life cycle			x				

## 2.2 *Barriers to CDE Implementation*

Alreshidi et al. [3] identify barriers to BIM adoption and collaboration practices through a survey which included 118 construction practitioners. The authors focus on issues relating to data management and governance and identify requirements for developing a BIM governance solution. The issues identified through the survey are categorized across organizational, legal, financial, and technical dimensions. As such, the most common organizational barriers are resistance to change, the BIM skills gap between generations, followed by barriers to collaboration (e.g., trust in a team) and barriers to adopting a single management process across multiple disciplines. Legal barriers include a lack of defined responsibilities for information curation, lack of intellectual property rights and fair practice standards for electronic information, a lack of clear regulations regarding the roles, responsibilities, and authorities of practitioners is a barrier and a lack of standards for collaboration. From a financial point of view, cost of training is the primary barrier, followed by the cost of the initial software configuration and the tight budget and small profit margins that exist on projects. The impact of financial barriers was seen to be correlated to the size of the organization. Finally, technical barriers include a lack of technical training, a lack of compatibility between different standards, a lack of compatibility between software, and a lack of data integration between stakeholders. Security was not identified as one of the more important obstacles. Moreover, it was noted that respondents still rely heavily on email for communication even when the company has its own CDE. Additional CDE implementation challenges include network speed, complex IT procurement challenges, varied offerings in the CDE market, cost and resources to procure a system, maintaining data security, and developing a system solution that is compatible with existing organizational systems [8, 15].

## 2.3 *CDE Implementation Process*

In terms of CDE adoption and implementation, a few studies have developed strategies and solutions. For instance, Abanda et al. [1] recommend moving toward an open standard format such as IFC, using the CDE to exchange information, ensuring usability of the solution, and that staff be trained and “educated.” According to Preidel et al. [18], when implementing a digital collaboration platform, the following aspects should be considered:

- **Appropriateness:** The objectives, effort, and benefits of the chosen measures and procedures should be proportionate.
- **Neutrality:** The selected procedures and measures should be independent of any particular software, so that the companies involved can use the software of their choice.



- **Applicability:** The selected procedures and measures should be applicable to companies and projects of different sizes and application areas.

To ensure proper use of the CDE, it is necessary to standardize processes for the production and exchange of project information, to find agreements of standards and methods to unify the form and quality of repeated and reusable information [12]. According to Preidel et al. [18], CDE users need to know the basics of data management. The structured information and exchange mechanisms must be as transparent as possible so that the user has a deeper understanding of the information exchange system. The CDE should be able to connect with any type of BIM authoring tool used by the project team, to optimize the exchange paths. According to Radl and Kaiser [19], there should be initiatives coming from the government to standardize CDE, while respecting the nature of each project and their domains (roads, residential, renovation, etc.).

According to Comiskey et al. [8], as the industry becomes more comfortable managing information with the implementation of ISO 19650, it is important that the notion and application of CDEs becomes a construction industry standard. For this to happen, barriers to adoption must be overcome. According to Mordue [15], key recommendations for CDE implementation by public agencies are to analyze initial processes and identify pain points, understand the purpose of CDE implementation, consider current infrastructure, to take the time to detail the functional requirements necessary for a CDE (security, functionalities and common requirements), to choose the right people to involve in the implementation process, to anticipate the whole cost of the CDE implementation process (technology, support, data storage, etc.), and to maintain data security. Moreover, Allen (Autodesk Construction [2]) lists important points to building buy-in (if challenging: run a pilot program to show success), choosing a CDE that is easy to use, progressively increasing the scale of implementation, creating a framework for CDE (define standards, processes, and workflows), clearly defining roles, and clearly define objectives and always analyze possible improvements.

These recommendations remain theoretical and are difficult to implement without additional detail. Above all, they target implementation within a particular company and not the broader construction industry. There lacks actionable strategies and tasks that can help improve the implementation of CDE at the industry level.

## **2.4 BIM Maturity Levels and CDEs**

The level of collaboration has been conceptualized as a progression of levels of BIM adoption also called maturity levels [4, 20]. According to DIN SPEC 91391-1, the BIM levels are the “considered development stages for the digitalization of the construction industry” [10, p. 12]. The UK government defined the main elements for the implementation of BIM in 2011 in the form of BIM levels to communicate the considered development stages of a digitization of the construction industry. They

are also part of the BSI PAS 1192-2:2013 standard and subsequent ISO 19650 series of standards.

DIN SPEC 91391-1: 2019-04 is the only standard that indicates the use of a CDE from BIM maturity level 1 [10]. It is also the only official standard given in the table, so surely the one that would be best to refer to. According to Dadmehr and Coates [9], a CDE starts to become part of the project context at level 2 due to the extent of data exchange. DIN SPEC 91391-1 confirms the use of CDE at level 2 but provides the nuance that depending on the level of BIM maturity, it is not the same level of CDE that comes into play. In fact, the DIN SPEC 91391-1 standard specifies the existence of different levels of CDE. Level 2 CDE is based on information containers and level 3 CDE is based on databases. This concept has evolved over the years, for instance, Botton and Kubicki [5] only brought the CDE into play at BIM maturity level 3 because that is when it enables centralization and accessibility of information. This source from 2014 shows that opinions have evolved regarding the place of the CDE in the BIM process.

Overall, what differentiates the different views on CDE have to do with how BIM is approached and managed within the collaborative platform. It is interesting to note that some firms that claim to practice BIM-enabled approach are in fact often only using object-oriented modeling software, i.e., BIM level 1 [9]. They do not use formalized collaboration processes as can be found at level 2 maturity. This is also the subtlety of the BIM definition: If a collaborative platform wants to fully fulfill the role of CDE, it will need to adopt an overall BIM strategy and formalized collaboration processes.

### 3 Research Methodology

The research project presented in this paper aimed to investigate the adoption and implementation of CDEs in the Quebec built asset industry. A two-staged approach was adopted to do so. First, the concept of CDE was framed to ensure a consistent basis for analysis and discussion. Indeed, as presented above, the concept of CDE is still vague, and therefore, any attempts to study it must first ensure a common understanding. To do so, a review of features offered by CDEs and identified in past research and within industry was performed to further frame the concept of CDEs. The deliverable from this first stage was a review of definitions and features of CDEs. In the second stage, a survey was conducted within the province of Quebec to understand adoption rates and implementation barriers.

Regarding the framing of the concept and features of CDEs, a table of functions of CDEs was created by aggregates criteria and features from various sources, which included:

- Scottish Futures Trust's CDE Procurement Scoping Document Template

- The functional requirements of an AIM CDE, according to the UK Government BIM Working Group
- DIN SPEC 91391 standard
- ISO 19650 standard
- Details of KROQI feature, by KROQI
- The method to choose your common data environment, by AxeoBIM
- CDE Framework by Catenda.

These sources were chosen due to the quantity of features identified and to offer a diversity of perspectives: official ISO and DIN standards, advocacy groups, and industrial bodies. The list of features is not exhaustive and can be further completed and tested. It also relies on documented features. The features were compiled and categorized into groups and sub-groups as the list was developed. These groups and sub-groups were inspired by the groupings already constituted within the references. The list was validated with one of the GRIDD's industrial partners. Of note, CDE solutions were not tested in the context of this research project.

The survey was designed iteratively by the research team and tested with domain experts. It was developed into seven parts, including (1) general information about the respondent and the company, (2) current communication and data sharing practices, (3) perception of CDEs, (4) benefits of implementing a CDE, (5) barriers to implementing a CDE, (6) areas for improvement, and (7) general comments. Question types were varied, including open, multiple choice, and Likert scales. The survey was conducted online. Invitations were open and sent in two phases. In the first phase, the questionnaire was distributed to the GRIDD's partners, which numbered about ten people. This first phase served as a test to ensure that the survey would collect usable and relevant information. In a second phase, the questionnaire was distributed through the Groupe BIM du Quebec, a BIM user group located in the province of Quebec, which has approximately 2000 people on the mailing list. The questionnaire has 35 questions in total could be completed on a voluntary and anonymous basis. A total of 46 completed and valid surveys were analyzed.

## 4 Research Findings

### 4.1 *Definitions and Features of CDEs*

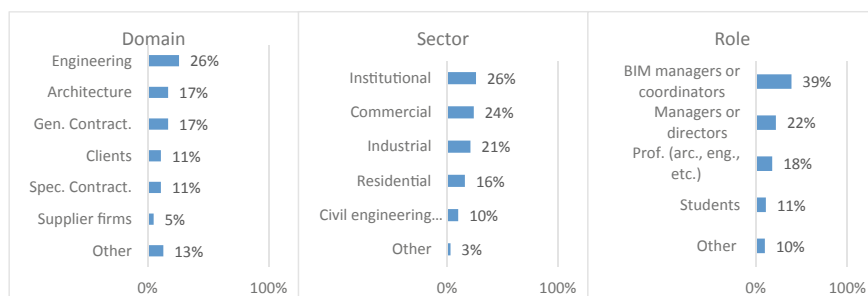
A total of 323 features of CDEs were identified through the review of the sources listed above. These features were classified into groups and sub-groups, as explained in the methodology. The groups and sub-groups identified through this review are given in Table 2. The most cited features across the different references were also identified. Table 3 shows the features that are mentioned in three references or more.

**Table 2** Classification of features listed by group and sub-groups

Function groupings and sub-groupings		Number of features
File management (FM)		75
	Model management (MM)	28
	General file management (templates and documents) (GFM)	15
	Compatibility and interoperability (C&I)	12
	Storage (St)	8
	Custom classification libraries (CCL)	7
	Document management (DM)	5
Collaboration and project management (CPM)		73
	Reports and error resolution (RER)	18
	Workflow (Wf)	13
	Communication (Comm)	10
	Tasks (T)	9
	Workspace (Wsp)	9
	Information requirements definition (IRD)	7
	Agendas and meetings (A&M)	7
Information exchange (IE)		70
	Search and filter (S&F)	18
	File repository management (FRM)	15
	Information assurance (IAss)	11
	Information acquisition (IAcq)	10
	Metadata (Md)	9
	General: information containers (G: IC)	7
Access to the CDE (Acc)		54
	Accessibility/user experience (UX)	20
	Security (Sec)	19
	Access rights (AR)	9
	Pricing (Pr)	6
Services (Serv)		48
	Linking with third-party tools (Link)	26
	Connectivity (Con)	13
	General (Ge)	5
	Tender and contract management (T&CM)	4
Compliance with standards (Comp)		3
	Total	323

**Table 3** Features identified in at least three references

Group/ sub-group	Feature	#
FM/MM	Integrated 3D viewer	6
Acc/UX	Accessibility of the CDE online via a web platform without installing third-party software	5
Acc/AR	Assignment of access rights to the CDE (different roles: manager, administrator, manager, etc.)	5
Acc/Sec	Ability to store securely (no further details)	5
FM/GFM	Assign status to documents (work in progress, shared, published, archived) according to PAS 1192 or ISO 19650	5
Acc/AR	Management of access rights/restrictions to the CDE during the project	4
Acc/AR	Creation and management of groups/teams on the CDE	4
FM/DM	Possibility to visualize in the CDE the most common extensions (images, pdf, office, etc.)	4
IE/FRM	File naming convention system (BS 1192-2007)	4
IE/S&F	Search engine (without more details)	4
Serv/Ge	Application programming interface (API)	4
Acc/UX	Video tutorials, user manual, FAQs	3
Acc/UX	Possibility to use the CDE on a mobile application	3
Acc/UX	Free trial period of the CDE	3
Acc/Pr	Inclusion of cloud storage	3
CPM/RER	View project summary and key performance indicators	3
CPM/T	Task management	3
CPM/Wf	Configuration of custom workflows	3
Comp	Conformity with the ISO 19650 standard	3
FM/C&I	Integrated BCF server	3
FM/MM	Ability to display models in IFC format	3
FM/MM	Ability to move around the model (rotate, scroll, zoom, select, measure, etc.)	3
FM/GFM	Ability to annotate/comment on models or documents directly on the CDE	3
FM/GFM	Version management (configurable number of versions + use of the most recent)	3
FM/GFM	Preservation of the history (traceability)	3
FM/St	Automatic data backup system (for recovery in case of partial or total loss)	3
IE/IAcq	Ability to download files in bulk	3
IE/FRM	Control of a deposit list	3
IE/Md	Metadata can be configured for files (project-specific classification) (e.g., phase, creator, etc.) for easier retrieval	3
Serv/Con	Open API	3

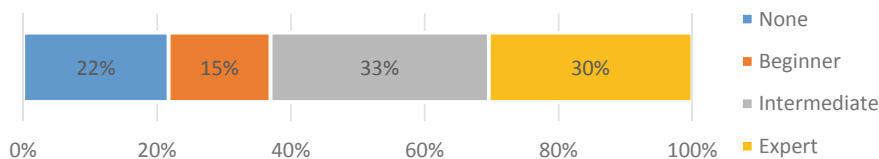


**Fig. 1** Respondent profile by domain, sector, role, experience, and size of organization

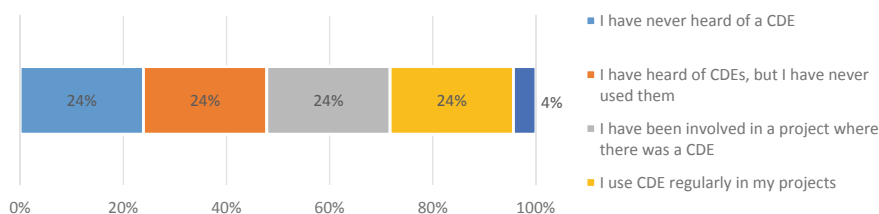
## 4.2 Survey Results

As mentioned, the survey was divided into seven parts. In terms of general information about the respondent and the company, the respondent pool was relatively diverse. Figure 1 illustrates the respondents' domain, sector, and role. Most respondents had between 11 and 29 years of experience in the industry (45%). A large proportion of respondents had between 1 and 10 years of experience (36%). 11% had more than 30 years experience in the industry. In terms of size, about 36% of businesses were from small businesses (1–99 employees), 33% were from medium sized businesses (100–499 employees) and 31% were from large-sized businesses (500+ employees). In terms of location, more than half of the respondents (51%) worked in a company that operates throughout Quebec. A large proportion of companies (34%) operate outside Quebec, while a smaller proportion operate only within specific regions of Quebec (15%).

In terms of current communication and data sharing practices, 93% of respondents identified e-mails as the dominant method to communicate and/or exchange information. A large proportion of respondents also indicated that they used specialized (63% of respondents) and standard (59%) cloud applications. Phone calls (57%) and text messages (26%) are also used, whereas 17% of respondents indicated using FTP solutions. When asked whether they were satisfied with their current communication methods, 60% indicated that they were, and were so due to ease of use and level of comfort with them. Among the dissatisfied (40%), the elements that come up often are too many different communication methods and channels, too much emphasis on email, and the need for a significant change in industry practices to embrace new communication technologies. When looking at document management systems (DMS), the number of respondents who were satisfied with their DMS (69%) was higher than those satisfied with their communication system. Overall, those satisfied mentioned their level of comfort with an existing solution that is easy to use and efficient. Those dissatisfied mentioned the difficulty of correctly and quickly establishing the necessary processes for document management.



**Fig. 2** Level of organizational BIM maturity



**Fig. 3** Level of experience with CDEs

With regard to BIM and CDEs, respondents were first asked about the maturity of their respective companies, as shown in Fig. 2, ranging from none, to expert, where BIM has been included as part of their organization's mission. In addition, 74% of respondents were familiar with the ISO 19650 series of standards, whereas 26% were not aware. Finally, Fig. 3 illustrates the level of knowledge or experience with CDEs.

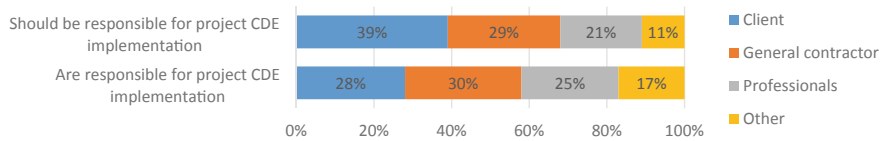
Concerning perceptions of CDEs and their use, opinions were divided regarding the possibility of implementing a "single source of information." Indeed, 60% believed it possible against 40% who believed otherwise. The survey was branched for those who indicated having at least used a CDE on one of their projects (52% of respondents). This subset was asked to identify the CDEs they had implemented through an open question. As such, a variety of answers were logged, which included responses such as "Microsoft Office Suite," "Industry Foundation Classes" (IFC) and "prolo" which is a programming language. These responses illustrate the confusion that can occur when discussing CDE. There were, however, a large number of consistent responses, including Autodesk BIM 360, including its specific products, representing 34% of the responses, followed by Procore (14%) and ProjectWise (7%). The entire list of proposed CDEs were analyzed using the definition and features developed above. 68% of the platforms cited were indeed CDEs. The subset of respondents who had indicated having used a CDE were asked about their benefits. More than 85% of these respondents believed that the CDEs can improve project performance. Open responses were then analyzed and categorized as given in Table 4. Information centralization and compliance, progress tracking, easier and faster collaboration, as well as tracking of versions/modifications, were among the top themes mentioned by the respondents. Interestingly, nearly 72% of respondents indicated that they do not use the same CDE on all their projects, whereas 24% use the same CDE on all

their projects and 4% do not know. Finally, Fig. 4 indicates who is and who should be responsible for project level CDE implementation.

When asked if a CDE should be made mandatory by the government for public projects, 95% of respondents indicated yes, while 5% had no opinion. When asked to justify their response, respondents identified several reasons supporting the importance of implementing CDEs on public projects (Table 5).

**Table 4** Respondents’ perception of CDEs impact on project performance

Theme	Instances
Information centralization and compliance	9
Progress tracking	6
Easier and faster collaboration	5
Tracking of versions/modifications	5
Save time and energy	3
Security	2
Improved communication	2
Process automation	1
Motivation of the staff	1
Avoiding errors	1



**Fig. 4** Responsibility for project CDE implementation

**Table 5** Respondents justification for mandatory CDE implementation on public projects

Theme	Instances
Standardization/compliance/uniformity	6
Accelerate digital transformation/encourage implementation	3
Reduce costs	2
Tracking and traceability	2
Increase productivity/improve project quality	2
Improved communication	2
Facilitate projects/reduce effort/optimize/avoid errors/improve process efficiency	2
Improve collaboration/break down silos	2
Solve interoperability issues	1
Centralize information	1



In terms of advantages to implementing a CDE, a list of benefits was presented to the respondents, and they were asked to rate the importance of each on a scale of 1–5 (1 being not important at all; 5 being very important). An overview of the main features that are beneficial to respondents are presented in Table 6. Overall, respondents rated most of the features very high (4 or 5). The features identified as most important were ease and speed of access to data, data security, data structure (with different states and version tracking), automation of coordination of checks and revisions. The least important were centralization of communication, open BIM standards, and automation of task coordination.

Lastly, respondents were asked to identify the barriers that might hinder the implementation of a CDE. The most important barriers for the respondents were the incompatibility of the CDEs between the project partners, the incompatibility of the CDE with certain software, the lack of desire to change methods, and the way the transition to the CDE is managed. The least important barriers were lack of technological capacity, the need for persistent Internet connection, and the specification of access rights (Table 7).

**Table 6** Ranking of benefits of CDEs according to the respondents

Feature	Avg. score
Access to all information is quick and easy from the single central platform	4.64
The CDE is secure (encryption)	4.53
The different states of the files are clear (work in progress, shared, published, archived)	4.50
Each time a modification is made, another version of the project is created (to avoid errors)	4.42
It is possible to view any type of file directly on the CDE	4.39
The coordination of controls and revisions is automated	4.31
The CDE can be used in the field (tablets, smartphones)	4.28
It is possible to manage/modify/annotate/comment any type of file	4.11
Original ownership of files is clearly identified	4.06
The CDE allows project management (task tracking, agenda sharing, performance measurements, etc.)	3.92
Task coordination is automated (reminders, etc.)	3.92
The CDE is based on openBIM standards (BCF, IFC, etc.)	3.86
The CDE ensures centralized communication (discussion threads, video conferencing, etc.)	3.69

**Table 7** Ranking of barrier to the implementation of the CDE according to the respondents

Barrier	Avg. score
Incompatibility of CDEs between project partners	4.44
Incompatibility of the CDE with certain software	4.14
Lack of desire to change usual work practices/methods	4.06
The way in which the transition to the CDE is managed	4.03
The inability of the CDE to perfectly adapt to the specific needs of the company	3.97
Lack of staff training/skills	3.97
Lack of availability of internal support staff	3.94
The cost of implementing the CDE	3.89
Lack of staff knowledge about the benefits of the CDE	3.86
Risks regarding data confidentiality and security	3.81
Difficulty of use, lack of accessibility of the CDE	3.81
More complicated use of the CDE in the field, application not adapted	3.58
The time and energy required to specify access and modification rights for each of the roles/stakeholders	3.58
The need for a persistent Internet connection	3.50
Lack of capacity of technological equipment/infrastructure	3.36

## 5 Discussion

The finding can be interpreted across three dimensions: confusion around the concept of what constitutes a CDE, perceptions and adoption of CDEs and barriers to implementation. First, the fact that 30% of the platforms considered as CDEs by the respondent are in fact not CDEs according to the features identified supports the claim around confusion on the concept. However, the main groups of features identified, captures the main areas of importance for the CDE: File Management, Collaboration and Project Management (including communication), Information Exchange, CDE Accessibility, Services, and Compliance to standards. The main themes are oriented around model management within the CDE, user experience, file management, security, reporting for project management purposes, advanced search within the CDE, access rights management, interoperability within the CDE and between CDEs and applications. The presence of the notions of metadata and information containers is also in agreement with the definition established in the literature review. That being said, it seems easier to establish what does not constitute a CDE rather than establish what is a CDE. Since the CDE is supposed to be the central reference point, it is supposed to consider all the user's needs to allow the user to use only the CDE and minimize the communication and file sharing channels. Based on this premise, every collaborative platform that lacks essential functionality or requires the use of applications external to the platform would not be a CDE. None of the CDEs identified through the survey were seen to fulfill the entirety of features identified through this

research. Indeed, it could be assumed that the higher the number of criteria met (the higher the number of features checked), the closer a CDE would be to a “perfect” solution. It is however complex to list all the features of a CDE. For instance, it was for noticed that there was no mention of electronic/digital signature for documents.

Regarding perceptions and adoption of CDEs, according to more than half of the respondents, a single source of information for projects is possible. One respondent said that it “currently works very well on many large projects with an independent BIM manager or on behalf of the owner.” A single source of information is possible if it is well managed by a designated BIM manager. Another respondent indicated that: “If one software could combine the strengths of each of these programs, it would be possible to have everything in one place.” Moreover, a single source is seen as necessary to optimize and facilitate projects, as confirmed by one respondent: “Even though we work in different firms, we all work on the same project/building. So, everything should be centralized in one place to make everyone’s work easier.” On the other hand, not all interviewees felt the same way. Many believe that no perfect or complete solution exist to meet everyone’s needs. According to one respondent, a single environment is not possible. Instead, CDEs should be comprised of several interrelated specialized environments. Another respondent stated that a single source “Doesn’t benefit everyone: It’s not relevant for all professionals to have ONE place that contains all the information. It’s mostly beneficial to the client and managers.” CDEs are seen by almost all respondents as platforms that improve project performance by enabling centralization and conformity of information, progress monitoring, simpler and faster collaboration, and version and change tracking. Less importance was afforded to the communication aspect. One of the interviewees even indicated that they used the CDE only as a duplicate of the data kept on their systems.

Finally, concerning barriers to implementation, they can be classified into four categories: contextual, technological, socio-organizational, and procedural, as shown in Fig. 5. Furthermore, respondents identified several needs and recommendations that are discussed below. One solution area is to clearly define who should lead the implementation of the CDE. According to ISO 19650, the CDE should be implemented by the appointing party (the client), provided it has the knowledge and motivation. 39% of survey respondents agreed with this. However, the client must be informed, and the appropriate practices must be in place. Regarding implementation of a CDE, a clear definition of requirements must be formulated. Moreover, a clear definition of roles and a mobilization of all stakeholders involved is necessary. Indeed, one of the major obstacles is that some parties do not have the time or inclination to participate in the CDE system, so it is essential that all parties be engaged and understand their place in the collaborative process. Managers must ensure that all stakeholders are always engaged in the CDE and throughout the life cycle by ensuring that they are informed, trained, and have clear rules.

Policy and regulation can also influence the adoption and implementation of CDEs, for instance should CDEs be mandated by public bodies on their projects. If mandated, this could incentivize organizations to invest in training and mobilization. In this sense, almost all survey respondents agreed that public bodies should make

Contextual barriers	Technological Barriers	Socio-organizational barriers	Procedural barriers
<ul style="list-style-type: none"><li>•Influence of project size</li><li>•Influence of company size</li><li>•Risks to data privacy and security</li></ul>	<ul style="list-style-type: none"><li>•Hardware and connectivity</li><li>•Interoperability</li><li>•Lack of CDE adaptation in the market</li><li>•Lack of accessibility of CDE</li></ul>	<ul style="list-style-type: none"><li>•Lack of information/ knowledge of staff</li><li>•Lack of staff training/skills</li><li>•Reluctance to change usual practices/work methods</li><li>•Lack of interest in innovation</li><li>•The cost of implementing CDE</li></ul>	<ul style="list-style-type: none"><li>•Lack of standardization/ clarity of exchange processes</li><li>•How the transition to the CDE is managed</li><li>•Need for 100% buy-in from all parties</li></ul>

Fig. 5 Consolidated barriers to CDE implementation

CDEs mandatory on their projects. Respondents felt that this would help standardize systems and speed up the transition to BIM/CDE implementation.

There is also a need to focus on clearly defining communication and data exchange processes. Indeed, should the CDE not be used correctly, for example if several applications are used for the same purpose, information is duplicated, resulting in wasted time and energy. Standardization of these processes could help streamline its use and ensure with stakeholder buy-in. Moreover, improving accessibility to the CDE and ensuring a “simple and effective” solution is seen as critical in enabling adoption. Indeed, the improvement of the user experience and the user interface can facilitate the integration and simplify the use of the platforms. To adapt to the project, parameterization should be simple. Finally, regular version changes (e.g., product nomenclature, interfaces) of the CDE should be avoided to avoid discouraging users from the constant readjustment.

6 Conclusion

This paper presented the results of a research project that aimed to investigate the adoption and implementation of CDEs in the Quebec built asset industry. A review of the definitions and features of a CDE was first performed to frame the concept. A survey was then conducted to evaluate how CDEs are perceived, their adoption rate and to understand the challenges and barriers to the implementation of common data environments in the province of Quebec. This study allowed a first approach on this subject in the province.

The results did provide an overall impression of how CDEs are perceived within the Quebec built asset industry. There is still a lot of confusion about the concept, but it is becoming more widely known. As for its adoption rate, it is difficult to establish with much precision since it depends directly on the definition one gives to CDEs. The CDE features matrix is an evolving tool and can be continuously completed with each new reference added. On the other hand, the selected references were quite diverse and can give a good overview of the functionalities (official ISO and

DIN standards, UK Government BIM Working Group, Scottish Futures Trust, and market CDE). The more a platform fulfills the functional requirements of the matrix, the closer it is to a “perfect CDE”.

The identified barriers to CDE adoption and implementation are mainly socio-organizational (lack of information about the CDE, confusion about the objective of the CDE, reluctance to change habits, etc.) and procedural (concern that the CDE will only worsen the standardization of processes, management of the transition to the CDE, need for buy-in from all parties). Some barriers are contextual (size of project, type, and size of company) or technology (cost, platforms not adapted to needs, etc.). Future developments should include practical tests of CDE to establish an exhaustive list of criteria and thus be able to define with certainty for example whether a platform is a CDE or not.

A limitation of this research is the survey response rate that is low. Results would have been more reliable and representative with a larger panel of respondents. However, the results were still usable and sufficiently significant. Indeed, the panel was diversified in terms of roles, experience, types of companies, etc. Future work includes further validation of the features matrix and broadening the scope of evaluation to the Canadian context. Individual barriers will also be addressed through targeted research projects as part of a broader research program.

## References

1. Abanda FH, Mzyece D, Oti AH, Manjia MB (2018) A study of the potential of cloud/mobile BIM for the management of construction projects. *Appl Syst Innov* 1(2):9. <https://doi.org/10.3390/asi1020009>
2. Allen J (2019) What's a common data environment (CDE)? [WWW Document]. Digital Builder
3. Alreshidi E, Mourshed M, Rezgui Y (2018) Requirements for cloud-based BIM governance solutions to facilitate team collaboration in construction projects. *Requirements Eng* 23(1):1–31. <https://doi.org/10.1007/s00766-016-0254-6>
4. Bew M, Richards M (2008) BIM maturity model. In: Construct IT autumn 2008 members' meeting, Brighton, UK
5. Boton C, Kubicki S (2014) Maturité des pratiques BIM: dimensions de modélisation, pratiques collaboratives et technologies. In: SCAN'14, 6ème Séminaire de Conception Architecturale Numérique, pp 45–56. <https://hal.archives-ouvertes.fr/hal-01025675>
6. BSI (2013) PAS 1192-2:2013 incorporating corrigendum no. 1 specification for information management for the capital/delivery phase of construction projects using building information modelling. BSI Group, London
7. Bucher DF, Hall DM (2020) Common data environment within the AEC ecosystem: moving collaborative platforms beyond the open versus closed dichotomy. [https://www.researchgate.net/publication/342719477\\_Common\\_Data\\_Environment\\_within\\_the\\_AEC\\_Ecosystem\\_moving\\_collaborative\\_platforms\\_beyond\\_the\\_open\\_versus\\_closed\\_dichotomy/references#fullTextFileContent](https://www.researchgate.net/publication/342719477_Common_Data_Environment_within_the_AEC_Ecosystem_moving_collaborative_platforms_beyond_the_open_versus_closed_dichotomy/references#fullTextFileContent)
8. Comiskey D, McKane M, Jaffrey A, Wilson P, Mordue S (2017) An analysis of data sharing platforms in multidisciplinary education. *Archit Eng Des Manag* 13(4):244–261. <https://doi.org/10.1080/17452007.2017.1306483>
9. Dadmehr N, Coates S (2020) An approach to “national annex to ISO 19650-2”. [https://www.researchgate.net/profile/Nady-Dadmehr/publication/341915122\\_AN\\_APPROACH\\_TO\\_NAT](https://www.researchgate.net/profile/Nady-Dadmehr/publication/341915122_AN_APPROACH_TO_NAT)

- IONAL\_ANNEX\_TO\_ISO\_19650-2/links/5ed96e81299b1c67d418b5f/AN-APPROACH-TO-NATIONAL-ANNEX-TO-ISO-19650-2.pdf
10. DIN Deutsches Institut für Normung (2019) DIN SPEC 91391-1: Partie 1: Composants et jeux de fonctions d'un CDE; avec adaptateur numérique
  11. Eastman C, Teicholz P, Sacks R, Liston K (2008) BIM handbook: a guide to building information modeling for owners, managers, designers, engineers and contractors. Wiley
  12. Garyaeva V (2018) Application of BIM modeling for the organization of collective work on a construction project. MATEC Web Conf 251:05025. <https://doi.org/10.1051/mateconf/201825105025>
  13. Howard HC, Levitt RE, Paulson BC, Pohl JG, Tatum CB (1989) Computer integration: reducing fragmentation in AEC industry. J Comput Civ Eng 3(1):18–32. [https://doi.org/10.1061/\(ASCE\)0887-3801\(1989\)3:1\(18\)](https://doi.org/10.1061/(ASCE)0887-3801(1989)3:1(18))
  14. ISO (2018) ISO 19650-1:2018 organization of information about construction works—information management using building information modelling—part 1: concepts and principles. <https://www.iso.org/standard/68078.html>
  15. Mordue S (2018) Implementation of a common data environment: the benefits, challenges & considerations. <https://www.scottishfuturetrust.org.uk/storage/uploads/cdeimplementationresearchaug18.pdf>
  16. Poirier E, Forgues D, Staub-French S (2017) Understanding the impact of BIM on collaboration: a Canadian case study. Build Res Inf 45(6):681–695. <https://doi.org/10.1080/09613218.2017.1324724>
  17. Preidel C, Borrmann A, Oberender C-H, Tretheway M (2016) Seamless integration of common data environment access into BIM authoring applications: the BIM integration framework. <https://doi.org/10.13140/RG.2.1.4487.4488>
  18. Preidel C, Borrmann A, Mattern H, König M, Schapke S-E, (2018) Chapitre 15: common data environment. In: Borrmann A, König M, Koch C, Beetz J (eds) Building information modeling: technology foundations and industry practice. Springer International Publishing, pp. 279–291. [https://doi.org/10.1007/978-3-319-92862-3\\_15](https://doi.org/10.1007/978-3-319-92862-3_15)
  19. Radl J, Kaiser J (2019) Benefits of implementation of common data environment (CDE) into construction projects. IOP Conf Ser Mater Sci Eng 471:022021. <https://doi.org/10.1088/1757-899X/471/2/022021>
  20. Succar B (2009) Building information modelling framework: a research and delivery foundation for industry stakeholders. Autom Constr 18(3):357–375. <https://doi.org/10.1016/j.autcon.2008.10.003>
  21. Turk Z (2000) Paradigmatic framework for construction information technology. Constr Inf Technol 948–958

# Enwave's Western Expansion Project—Key Challenges and Solutions



Mark Bruder, Ahmed Elsherif, James Scharbach, and Kris Landon

**Abstract** In 2017, Enwave Energy Corporation (Enwave) identified a need to expand their deep lake water cooling and hot water distribution network in the City of Toronto (Toronto). Enwave retained R.V. Anderson Associates Limited (RVA) to undertake project management, preliminary/detailed design, cost/schedule/risk management, stakeholder negotiation, and assistance during tendering/construction for this Western Expansion Project (WEP). Enwave's primary objectives were to connect four (4) Enwave customers with chilled/hot water supply/return along the trunk pipe route, meet the contractual online service dates for the four (4) Enwave customers, place as much of the trunk pipe route within the ROW as possible, minimize impacts to existing stakeholders, minimize the construction cost by optimizing the trunk pipe route, and future-proof the system. The final design included a deep rock tunnel, shallow soil hand-mined tunnels, and shallow open-cut trenches. The total length of the route was 1400 linear metres of new chilled/hot water supply/return trunk pipe. The pipe sizes were 600 mm inner diameter (ID) HDPE for the chilled water trunk, and 350 mm ID FRP for the hot water trunk. Pressure relief chambers and valves were included at high points along the route, and buried isolation valves were installed at strategic locations for maintenance/operation and future expansion. In late 2021, the system was energized, and it is currently in operation. With a new thermal storage hub at The Well, Enwave is now capable of further expanding their district energy system. This paper presents some of the key design challenges and solutions that were developed to address the primary objectives. This includes a discussion on information gathering and optimization of the trunk pipe route. The goal of this paper is to give owners/designers some tools to help them navigate a complex design project like the WEP in Toronto.

**Keywords** Enwave's Western Expansion Project

---

M. Bruder (✉) · A. Elsherif  
R.V. Anderson Associates Limited, Toronto, Canada  
e-mail: [mbruder@rvanderson.com](mailto:mbruder@rvanderson.com)

J. Scharbach · K. Landon  
Enwave Energy Corporation, Toronto, Canada

## 1 Problem Statement and Goal

In 2017, Enwave Energy Corporation (Enwave) identified a need to expand their deep lake water cooling and hot water distribution network in the City of Toronto (Toronto). The expansion was to be 850 linear metres west (as the bird flies) from Enwave's Pearl Street Energy Centre (PSEC) at Simcoe Street and Pearl Street to The Well development at Front Street West and Spadina Avenue. The goal of Enwave's Western Expansion Project (WEP) was to feed their new 7.6 million litre thermal storage facility located within the footprint of The Well. The Well is a mixed-use development by Allied Properties and RioCan featuring over 3 million square feet of retail, office, and residential space. Enwave designed the thermal storage facility to supply low-carbon heating/cooling to 17 million square feet of space near The Well and along the proposed trunk pipe route from Enwave's PSEC. The thermal storage facility will function as a "battery", where energy is stored overnight during off-peak times. This will ease the strain on the electricity grid and reduce overall costs for heating/cooling. See Figs. 1 and 2 below for how the system at The Well works in the winter and summer, respectively.

In 2018, Enwave retained R.V. Anderson Associates Limited (RVA) to undertake project management, preliminary/detailed design, cost/schedule/risk management, stakeholder negotiation, and assistance during tendering/construction for the WEP. RVA performed these services from 2018 to 2022.

This paper highlights some key design challenges and the solutions that were developed to successfully deliver the WEP. Construction of the WEP will be addressed in a subsequent paper. The goal of this paper is to give owners/designers some tools to help them navigate a complex design project like the WEP in Toronto.

## 2 Design Challenges and Solutions

The primary design challenge of the WEP was to find an optimal trunk pipe route in the congested downtown Toronto Municipal Right of Way (ROW) that would satisfy the following objectives (in no particular order):

- Connect four (4) Enwave customers with chilled/hot water supply/return along the trunk pipe route;
- Meet the contractual online service dates for the four (4) Enwave customers;
- Place as much of the trunk pipe route within the ROW as possible and avoid easements;
- Minimize impacts to existing utilities, owners, and stakeholders along the ROW;
- Obtain road cut permits as soon as possible and minimize disruption to traffic and the public;
- Minimize the construction cost by optimizing the trunk pipe depth to be as shallow as possible; and





Fig. 1 Winter daytime heating system at the well, per Ref. [1]



Fig. 2 Summer daytime cooling system at the well, per Ref. [1]

- Future-proof the trunk pipe route for expansion in all directions along the main and from The Well.

In general, RVA's final design consisted of the following primary components:

- Deep rock tunnel, shallow soil hand-mined tunnels, and shallow open-cut trenches;
- 1400 linear metres of new chilled/hot water supply/return trunk pipe;
- 600 mm inner diameter (ID) HDPE chilled water supply/return (CWS/CWR) trunk pipe;
- 350 mm ID FRP hot water supply/return (HWS/HWR) trunk pipe;
- Pressure relief chambers and valves at high points along the trunk pipe route;
- Buried isolation valves at strategic locations for maintenance/operation and future expansion; and
- Live tapping of existing Enwave pipes/chambers at strategic locations.

The following sections of this paper discuss some key design challenges and solutions that were developed to address the above objectives.

## ***2.1 Information Gathering***

Relevant and reliable information is critical to the success of a design project. Gathering information should be done prior to initiating conceptual design. Certain types of information (such as lists of ongoing construction projects) should be re-assessed throughout detailed design and construction. Without key information, projects are likely to encounter changes during design/construction that can impact the project schedule/cost and disrupt stakeholders. The following subsections summarize numerous resources that designers can leverage in Toronto to gain relevant/reliable information for projects within the ROW.

### **2.1.1 Road Closure Coordination Working Group and Infrastructure Coordination Unit**

Per Ref. [2], Toronto has a robust coordination and planning process with a variety of key stakeholder groups/units. This includes the Road Closure Coordination (RCC) working group and the Infrastructure Coordination Unit (ICU). The RCC is comprised of stakeholders from utility companies, the Toronto Transit Commission, and Toronto staff from both Transportation Services and Engineering and Construction Services. The RCC meets monthly to help ensure road closures within the City are coordinated in a way that minimized disruptions. The ICU acts as a coordination body for all group that perform construction work in Toronto. The ICU seeks to bundle work to avoid frequently impacting the same areas and the public.

On the WEP, early engagement with the RCC and ICU was critical to help develop a realistic design/construction schedule. It also helped RVA understand whether certain sections could be bundled with other City work, and if/how moratoriums on road work could be suspended. Ultimately, bundling was not possible in all sections of the trunk pipe route, as the timelines were not compatible for the City's planned projects adjacent to the WEP. Transportation Services approved the suspension of some moratoriums, as Enwave had agreed to enhanced road restoration details and/or utilizing lower-impact construction means/methods. These variances to moratoriums helped RVA/Enwave keep the project on schedule.

### **2.1.2 T.O.INview (Infrastructure Viewer)**

Per Ref. [3], Toronto maintains a comprehensive web-based information map called T.O.INview (see Fig. 3) of current and forecasted infrastructure projects. Projects are tracked from the following City groups or other parties: Transportation Services; Toronto Transit Commission; Toronto Water; City Planning; Economic Development and Culture; Parks Forestry and Recreation; and third-party construction or other construction/events. The map allows users to control information displayed based on the year, project status, and City group. Upon selecting a specific project, more information is presented including project location, details, duration, status, owner, contact info, website, and the councillor for the associated City ward. The purpose of this tool is to help plan capital projects and improve coordination within Toronto.

For the WEP, T.O.INview was a useful resource to help develop a realistic design/construction schedule. It helped RVA understand which sections of the proposed trunk pipe route may be delayed due to ongoing construction, moratoriums, or other events. It also facilitated more meaningful discussions with stakeholders undertaking construction within the ROW.

### **2.1.3 Application Information Centre**

Per Ref. [4], Toronto's Planning and Development group maintain a web-based Application Information Centre (AIC) for development applications (see Fig. 4). This tool allows users to see development applications to the Community Planning, Committee of Adjustment, and Toronto Local Appeal Body groups. Users can search by City district/ward, application number, address, and other methods.

For the WEP, the most useful items within the AIC were the supporting documents for applications. Projects adjacent to the proposed trunk pipe route often had useful renderings/drawings of the development and geotechnical reports. While this information was not always up to date and should not be relied upon by a third party, it helped RVA focus our stakeholder engagement and develop better alternative routes. It also helped RVA plan project-specific field investigations for the WEP. For example, a variety of recent boreholes were found across the street from the proposed deep rock shaft/tunnel at Spadina Avenue and Clarence Square (see



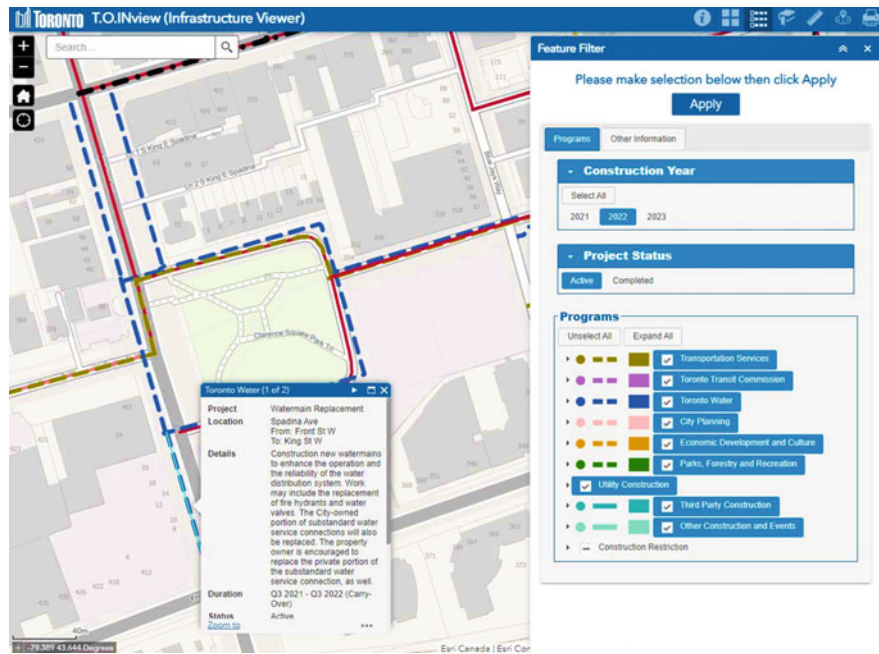


Fig. 3 Example plan from T.O.INview

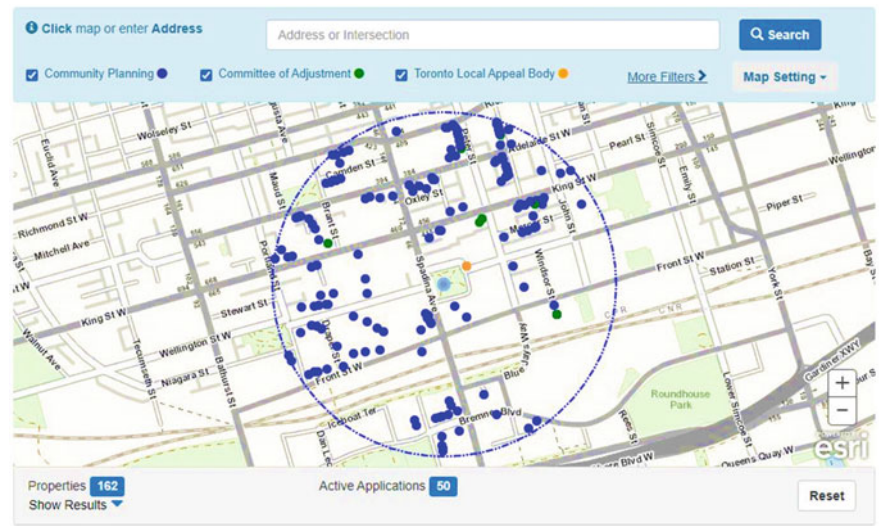


Fig. 4 Example plan from AIC

Sect. 2.2.5). This helped RVA understand whether the top of rock elevation was likely to be sloping, and what risk mitigation actions were warranted in the deep rock tunnel design.

### 2.1.4 Digital Map Owners Group

Per Ref. [5], Toronto maintains comprehensive utility mapping that is available to the public for purchase (see Fig. 5). The mapping includes the following: Toronto's Digital Map Owners Group (DMOG); Toronto's Street Furniture Programme; fibre optic networks; pedestrian street lighting; sewer/water infrastructure; ROW and curb lines; building envelopes; and municipal numbers. The DMOG consists of seven members who share in the cost of maintaining the map and database. Utilities are represented by double lines, are labelled, and are colour coded for ease of reading. Sections of the map are provided in a file format compatible with AutoCAD (DWG type) or Microdesk (DGN type).

Users of this mapping should be aware that the information is not up to date or a perfect representation of the as-built condition. The mapping primarily serves as a useful tool in conceptual design development. Designers should validate map components that are likely to be critical to the success of their design/construction. Specific items worth validating include, but are not limited to, the following: pinch points between existing utilities and newly proposed infrastructure; vertical/horizontal offsets to high-risk utilities such as hydro, gas, watermains, sewers, and private property.

For the WEP, the DMOG information was invaluable. It helped RVA develop preliminary routing, high-level pricing, and potential impacts to utilities/owners. A variety of critical areas of the DMOG were validated with subsurface utility investigations and through other means. This helped to reduce the risk of changes during design/construction.

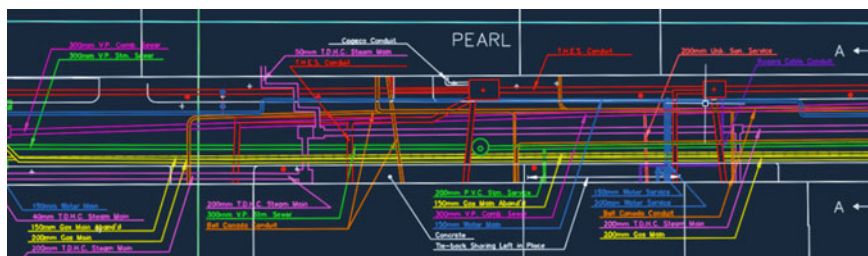


Fig. 5 Example utility map from DMOG

## 2.1.5 Project-Specific Field Investigations

Subsurface Utility Engineering (SUE) is a key exercise for nearly all underground projects. The primary purpose is to inform designs, reduce delays, manage stakeholders, and decrease disruptions. SUE is classified by quality level (QL) per ASCE 38–02 and is based on the type of activity performed: QL-D is desktop records research; QL-C is a visual field investigation; QL-B is utility locating through geophysical equipment; and QL-A is daylighting (often through vacuum excavation). Most utility projects in the ROW can immediately benefit from undertaking QL-D to QL-B at the outset of preliminary design. QL-A (also referred to as a “test pit”) is often carried out during detailed design. QL-A helps inform critical locations that require certainty on the type/size/depth of the existing utility.

For the WEP, QL-D to QL-B was undertaken for the entire route. The mapping was cross-checked with the DMOG and other information obtained. This allowed RVA to choose a suitable vertical/horizontal route for Enwave's pipes through the congested ROW. When pinch points were found, or when required vertical/horizontal clearances were encroached, QL-A was used to confirm the extent of the identified conflict. The QL-A data was often sufficient to help achieve approval from stakeholders. In some cases, stakeholders were unaware that their infrastructure was not constructed per their as-built records. See Fig. 6 for an example QL-A test pit location plan and photo from the WEP field investigation.

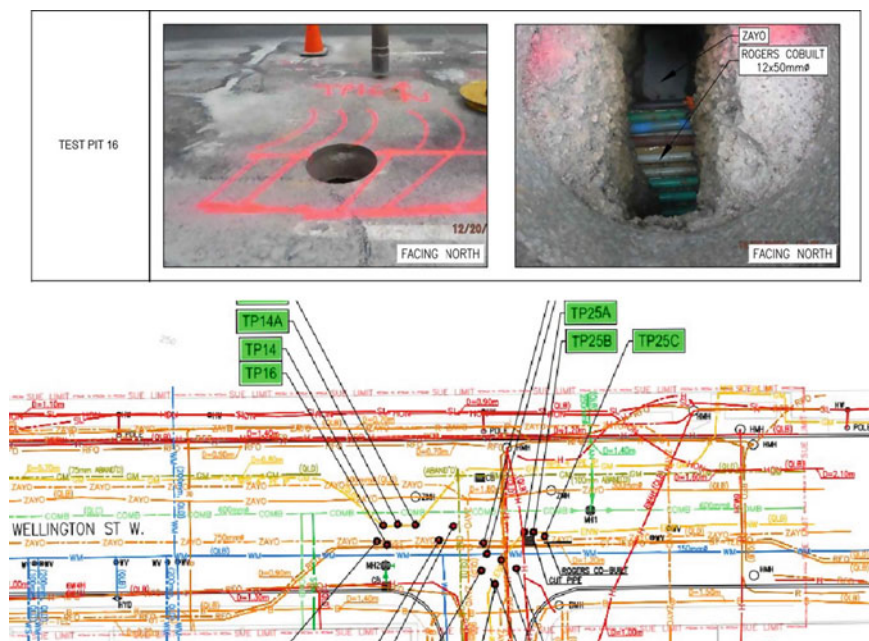


Fig. 6 Example QL-A test pit photos of field investigation and location plan

2.1.6 Toronto Public Utilities Coordinating Committee

Per Ref. [2], the Toronto Public Utilities Coordinating Committee (TPUCC) is a consortium established by the City and utility companies. TPUCC’s goals are similar to the RCC and ICU, where members present upcoming projects for discussion. Members are encouraged to promote innovative ideas/designs to reduce impacting each other and the public during construction. Members include, but are not limited to, the following: Beanfield; Bell Canada; City of Toronto; Cogeco Data Services; Enbridge Gas Distribution; Enwave Energy Corporation; Hydro One Networks Inc.; MTS Allstream; Rogers Cable Communications Inc.; Telus; Toronto Hydro; and Toronto Transit Commission.

Utility projects within Toronto’s ROW require a road cut permit (RCP) issued by Transportation Services. The design must adhere to the Municipal Consent Requirements (MCR) per Ref. [6]. The MCR is intended to ensure projects protect the City’s interests, and that applicants provide all required information for permit review. An RCP is not issued until the applicant has demonstrated that all members of the TPUCC have signed-off on the project. To gain a TPUCC sign-off, the design must satisfy the MCR, but in particular Appendix O for Vertical and Horizontal Clearance Guidelines. Each TPUCC member has defined preferred clearances from their utilities for future projects in the ROW. Preferred clearances may be reduced with the written permission of the affected utility. In some cases, clearances are defined on a case-by-case basis, as denoted by a double asterisk (see Fig. 7).

For the WEP, in advance of detailed design, RVA/Enwave directly engaged the TPUCC at their monthly meeting and presented the overall project with some conceptual trunk pipe routing. This early engagement was critical to help identify any “show-stoppers” in upcoming work or anticipated variances to preferred clearances. It also gave the TPUCC “sign-off reviewers” context prior to their review of the design drawings.

For the final WEP design, RVA/Enwave requested a variance to some preferred clearances from Toronto Hydro, Toronto Water, and Enbridge. This could have been flatly rejected by the utilities and required RVA to develop a less-preferred trunk pipe

APPENDIX O  
VERTICAL AND HORIZONTAL CLEARANCE GUIDELINES \*

All Dimensions in Millimetres (mm)	Preferred Vertical Clearance	Preferred Horizontal Clearance
WATER		
Water Supply		
Inside Diameter < 100mm	150	600
100mm <= Inside Diameter < 400mm	300	750
Inside Diameter >= 400mm	500	900
Valve Chamber	**	600

Fig. 7 Example preferred clearances from MCR appendix O



route. However, RVA/Enwave had established a collaborative relationship with the TPUC, and the teams were able to develop solutions to avoid redesigns. Some of these solutions included the following: enhanced field investigations to confirm utility offsets; robust trench backfill material adjacent to other utilities (such as 35 MPa concrete instead of unshrinkable fill); and permitting other utilities to locally encroach on Enwave's preferred clearances in the future.

Another notable solution to utility impacts was negotiating the removal of 120 linear metres of an abandoned 750 mm diameter "Metronet" pipe on Wellington Street that was owned by the City and leased to Zayo. Zayo confirmed they were not using the pipe, so the City permitted Enwave to remove the pipe and install their own CW/HW pipes. This was provided that Enwave also installed new PVC conduit in the same trench for Zayo to use at their discretion. While this was a lengthy negotiation, it proved to be a worthwhile exercise. Zayo received new conduit connected to their manholes at each end of Enwave's trench, and Enwave constructed within an unobstructed corridor for a lengthy portion of their trunk pipe route.

## ***2.2 Optimizing the Trunk Pipe Route***

Throughout preliminary design, RVA/Enwave developed many alternative trunk pipe routes. When evaluating/comparing the routes, the key considerations were cost effectiveness, schedule efficiency, and mitigating impacts to stakeholders. In some cases, a longer/skewed route was preferable to a shorter/straight route, but this was typically only when there was a significant schedule advantage despite the added cost of construction. In other cases, there were no elegant solutions, and the "best of the worst" option was chosen. For example, a deep rock tunnel crossing was required in close proximity to a high-pressure watermain, as it was not practical to open-cut across multiple active streetcar rails.

The following principles were used to develop alternative routes and select/optimize the preferred route:

- Avoid utility relocations and easements as much as possible;
- Nearly all routing must accommodate four pipes (CWS/CWR and HWS/HWR); a pipe layout in a trench of 4 Horizontal by 1 Vertical or 2 Horizontal by 2 Vertical is preferred; and a pipe layout in a trench of 1 Horizontal by 4 Vertical is the least preferred;
- Shallow open-cut trenches are preferable to shallow soil hand-mined tunnels or deep rock tunnels;
- Shallow soil hand-mined tunnels are often necessary for street crossings or near congested utilities, and 45 m is the maximum preferred length between hand-mined tunnel shafts; and

- Interior isolation valves at customers and main line isolation valves for operational control/expansion.

Regarding utilities, relocations can cost hundreds of thousands of dollars, and relocations can delay a project by months to years. For the WEP, where many utility conflicts were anticipated, RVA/Enwave first worked with TPUC members to explore options for tighter vertical/horizontal clearances. In some cases, Enwave provided enhanced monitoring/backfill details to accommodate stakeholder concerns. If permission was not granted, then preliminary routes were modified as required. Ultimately, no permanent utility relocations were necessary for the 1400 m long WEP trunk pipe route, which was a remarkable outcome.

Regarding trench sections, there are few locations within the congested downtown Toronto ROW that can accommodate four pipes laid horizontal ( $4H \times 1V$ ). Therefore, RVA developed routes that could fit two layers of two pipes ( $2H \times 2V$ ). See Fig. 8 for an example trench section. Hot water pipes are insulated and placed on top to help minimize their effect on the cold water pipes. The ideal horizontal or vertical spacing between pipes is 300 mm, but 150 mm is possible. The trenches are backfilled with unshrinkable fill and restored per City standards. Spare conduit is also included to support Enwave's future plans.

Regarding construction means/methods, depending on the length, open-cut trenches can be significantly cheaper than hand-mined tunnels or deep rock tunnels. Also, open-cut trenches are often faster to construct than alternatives. Pipes in open-cut trenches are placed as shallow as possible, where the minimum depth is the frost limit (1.2 m in Toronto). Pipes are locally deepened below the frost limit as required to provide clearance to existing utilities. For the WEP, approximately 1150 m (82%) is open-cut trenches, 200 m (14%) is hand-mined tunnels, and 50 m (4%) is deep rock tunnel.

Regarding soil hand-mind tunnels, the cap and lag construction methodology has been used frequently by Enwave in Toronto. This construction technique is

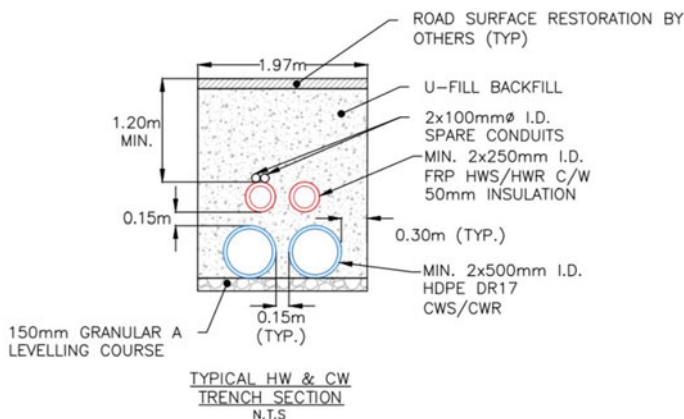


Fig. 8 Example section of trench carrying four pipes



**Fig. 9** Cap and lag tunnel under pearl street, shaft facing (left) and cut facing (right)

particularly suitable for utility crossings, as the method minimizes disturbance to surrounding soil. Although it can be expensive and slow, it is often the best alternative for highly congested areas in the ROW. To simplify health and safety requirements, access shafts are placed no further than 45 m apart. As excavation proceeds, sequential lumber frames are installed at the cut face. All voids beyond frames are filled with dry-pack grout. Existing utilities above are not required to span the tunnel unsupported. After pipes are installed inside the tunnel, the annular space is filled with unshrinkable grout leaving no air voids inside. The cap and lag framing and pipes are entombed by the grout. See Fig. 9 for photos of a cap and lag tunnel under Pearl Street.

The as-built trunk pipe route, pipe ID (in mm and inches), and valve locations are shown in Figs. 10 and 11. The primary customer that required chilled/hot water supply/return pipes was The Well, as this was the hub for Enwave's entire Western Expansion Project. Therefore, not all portions of the entire route needed to carry both chilled/hot water pipes. Trenches carry either HWS/HWR, CWS/CWR, or all four pipes as required. Chilled water is supplied to The Well directly from an existing Enwave pipe at Wellington Street and Windsor Street. Hot water is supplied to The Well directly from Enwave's PSEC. Customers at 57 Spadina, 333 King, and 19 Duncan are either along the main trunk pipe route or nominally offset.

The following subsections summarize challenges and design rationale for some of the routing segments.

### **2.2.1 From Pearl Street Energy Centre to 19 Duncan and Metro Hall**

Pearl Street was the preferred trunk pipe route from the PSEC for the HWS/HWR pipes towards The Well, and for HW/CW pipes to the customer at 19 Duncan. Pearl was the most direct and cost-effective route, as compared to Adelaide Street West (to the north) or King Street West (to the south). Adelaide would have carried the route

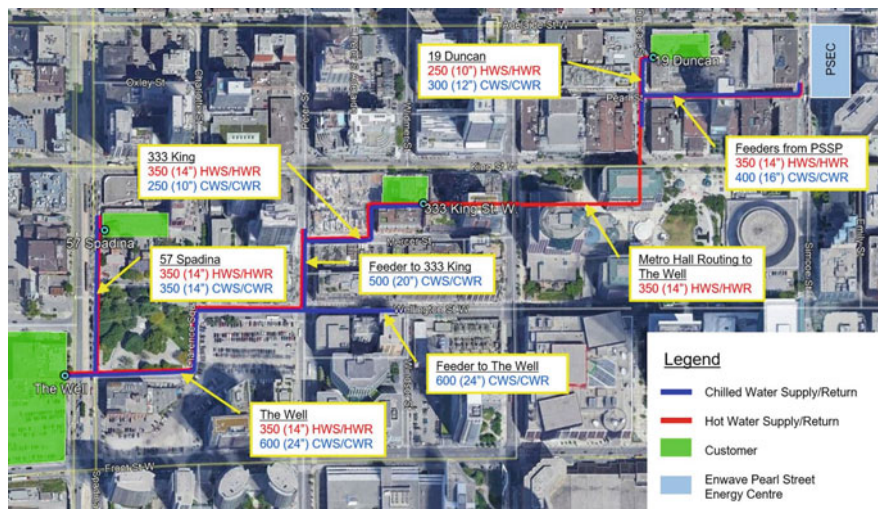


Fig. 10 Plan of WEP trunk pipe route

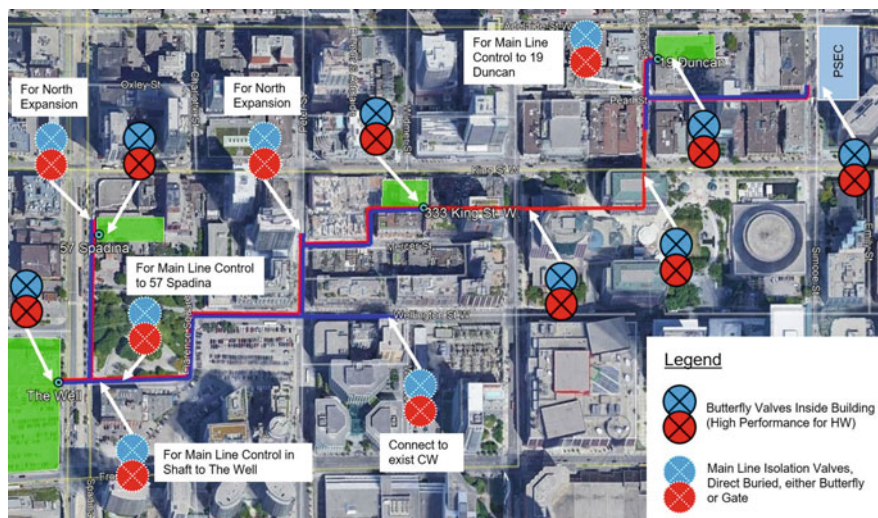


Fig. 11 Plan of valve locations along WEP route

further away from The Well, and it had numerous ongoing construction projects. King has active streetcar tracks and is heavily congested with existing utilities. Although construction along Pearl appeared feasible with an open-cut 2H × 2 V trench section, detailed SUE showed a variety of significant utility conflicts. To address these conflicts, RVA developed several alternatives that were evaluated on capital cost and schedule. Splitting the route into multiple 1H × 2 V trenches was

not cost-effective, and constructing a  $1\text{H} \times 4\text{V}$  trench directly against Enwave's existing steam corridor had challenges related to constructability, operation, and maintenance. Surprisingly, the best option proved to be a soil hand-mined tunnel for the entire street with multiple intermediate shafts. Among other benefits, this option satisfied TPUCC members by maintaining vertical/horizontal clearances, and it allowed for a single lane of traffic to be kept open throughout construction.

### **2.2.2 Through Parking Structure of Metro Hall and Metro Centre**

The preferred trunk pipe route for the HWS/HWR pipes towards The Well was through the existing Metro Hall and Metro Centre underground parking structure. The hot water pipe would be suspended from the cast-in-place concrete floor slabs. This option required a lease agreement between Enwave and the property owner. An alternative route was to place the pipes within the ROW along King Street from Ed Mirvish Way to John Street. However, as noted above, King has active streetcar tracks and is heavily congested with existing utilities. Although a hand-mined tunnel along King was technically possible to avoid utility conflicts, it was not cost-effective. Therefore, the cost of the lease was the preferred option.

### **2.2.3 From Metro Centre Along Lot 187, Lot 184, Mercer, and Blue Jays Way to Wellington**

The preferred trunk pipe route for the HWS/HWR pipes towards The Well from Metro Centre was along Lot 187, Lot 184, Mercer, and Blue Jays Way to Wellington Street. This was preferred due to the relatively low number of existing utilities in the ROW, and because an open-cut trench with  $2\text{H} \times 1\text{V}$  hot water pipe was possible without encroaching on most utility clearances. Conveniently, the Enwave customer 333 King was adjacent to this route. This customer required both CW/HW pipes, and therefore, a small chilled water feeder line from Wellington was required. This feeder was sized to accommodate further expansion north of Mercer along Blue Jays Way. Unfortunately, Lots 184/187 were very narrow which slowed construction rates and required modified means/methods. Furthermore, Lots 184/187 were being used daily to deliver products to the restaurants/businesses along King. They were also frequently used for garbage pickup. To address this concern during construction, AECON (the contractor) assisted with delivering products and facilitated garbage pickup as required.

### **2.2.4 Along Wellington and Clarence Square**

The preferred trunk pipe route to supply CW to The Well was from an existing Enwave pipe at Wellington Street and Windsor Street. This route is a straight shot with nearly no bends. This was possible due to constructing within the same corridor

as an abandoned “Metronet” pipe. See Sect. 2.1.6 for a discussion on Zayo. See Fig. 6 for a portion of the Zayo alignment near Windsor per the DMOG and SUE. Alternate routes along Wellington would have required multiple  $1\text{H} \times 2\text{V}$  trenches for CW/HW and extensive bends/fittings to maintain adequate clearance from existing utilities. A detailed SUE was completed to help assess whether the older “Metronet” corridor had adequate clearances from newer utilities. The City’s DMOG showed may encroachments on the “Metronet” corridor by either removed/abandoned chambers or newer manhole construction. Based on the SUE, all of these clearance issues were resolved to the satisfaction of TPUC members and Toronto Water. This helped gain TPUC sign-offs and a RCP from Transportation Services. No further utility clearance issues were encountered along Wellington during construction.

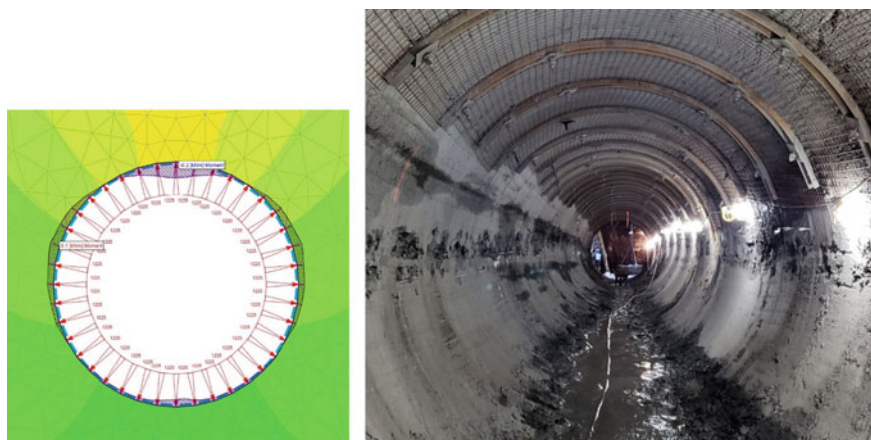
### 2.2.5 Crossing Spadina with a Deep Rock Tunnel

The connection point at The Well for the CW/HW pipes was near the east corner of the development. The pipes had to travel 50 m across Spadina Avenue from Clarence Square to enter The Well. This posed significant design and construction challenges. Spadina is a major traffic artery to/from the Gardiner Expressway, it has active streetcar tracks, and it is heavily congested with shallow utilities. An open-cut trench was not practical, and a shallow soil hand-mined tunnel could not avoid a large combined sewer located near the top of bedrock elevation. The only option was a deep rock tunnel and shaft.

The top of bedrock (shale of the Georgian Bay Formation) was 9 m below grade, and the connection point within The Well development was 14 m below grade. To avoid impacting The Well foundation wall, the tunnel obvert was located at 15 m below grade. At this depth, an existing high-pressure 1500 mm inner diameter trunk watermain was present. Therefore, the tunnel had to maintain adequate clearance from the watermain, and the design team had to demonstrate to Toronto Water that the design/construction would have no negative impacts to the watermain. RVA evaluated the interaction between the watermain and the proposed excavation through a finite element model (FEM) and subsequent analysis. Multiple loading cases were considered, and conservative assumptions were made on the adequacy of the watermain based on strength/durability. Through the stress/displacement analysis, RVA demonstrated that there were no negative impacts to the watermain from the proposed works.

To carry the CW/HW pipes and facilitate construction with a small tunnel boring machine, the excavated tunnel diameter was set at 2770 mm. After installing the CW/HW pipes, the tunnel was backfilled with 35 MPa concrete. Figure 12 shows a snapshot of the FEM of the watermain (left image) and the unlined excavated tunnel (right image). The tunnel obvert had temporary support with curved steel channels, rock bolts, and welded wire mesh. This helped to control overbreak due to stress concentrations. Figure 13 shows the shaft that was constructed on Clarence Square looking up from the bottom. To support the overburden during excavation, sequential rings of steel wide flanges were installed with wooden lagging placed in between





**Fig. 12** Spadina crossing deep rock tunnel, FEM with roscience RS2 (left) and unlined (right)

the rings. Grout was used behind the lagging as required. The exposed rock face was supported by rock bolts and welded wire mesh. For the first four (4) rings excavated from grade, the shaft diameter was smaller to avoid impacting shallow utilities. Below the fourth ring, the diameter was expanded to facilitate the tunnel boring operation. The shape of the excavated shaft looked like an upright wine bottle. An oval-shaped secant pile temporary shaft was considered during design, but it was uneconomical and provided insufficient space for construction.

### 3 Conclusions

RVA/Enwave successfully designed and delivered the WEP despite a variety of challenges. In late 2021, the system was energized, and it is currently in operation. With a new thermal storage hub at The Well, Enwave is now capable of further expanding their district energy system. During the design, some lessons were learned the hard way, while many solutions were carried forward from past projects. Future design teams of linear projects within Toronto's ROW may find their own success by (1) gathering all relevant information, (2) generating many practical alternative solutions, (3) optimizing the preferred solutions based on project-specific guiding principles, and (4) working proactively/co-operatively with other utilities to explore creative solutions.



**Fig. 13** Spadina crossing deep rock shaft, looking up from the bottom

**Acknowledgements** The authors would like to thank Aecon Group Inc. (general contractor), C&M McNally Underground Inc. (deep rock shaft/tunnel construction), Clark Construction Management (contract administration during construction), and Telecon (SUE) for taking significant leadership roles on the WEP. The authors would also like to thank the City of Toronto, Toronto Water, Transportation Services, and all TPUCC members for being strong collaborators. We look forward to working together on future projects.

## References

1. <https://www.enwave.com/case-studies/groundbreaking-expansion-project-brings-water-to-the-well/>. Last accessed 12 Feb 2020
2. <https://www.toronto.ca/services-payments/building-construction/infrastructure-city-construction/understanding-city-construction/construction-coordination-in-the-city/>. Last accessed 12 Feb 2020
3. <https://map.toronto.ca/toinview/>. Last accessed 12 Feb 2020
4. <https://secure.toronto.ca/AIC/index.do>. Last accessed 12 Feb 2020
5. <https://www.toronto.ca/city-government/data-research-maps/utility-maps-engineering-drawings/>. Last accessed 12 Feb 2020



6. <https://www.toronto.ca/services-payments/building-construction/infrastructure-city-construction/construction-standards-permits/standards-for-designing-and-constructing-city-infrastructure/>. Last accessed 13 Feb 2020

# Enhancing Bridges' Safety Training Using Augmented Reality and Virtual Reality



M. El Rifaae, S. Bader, I. Abotaleb, O. Hosny, and K. Nassar

**Abstract** The increase in urbanization rates has driven the construction of highly complex infrastructural projects across the globe, including the construction of bridges. However, such growth comes at an extremely high cost as the industry is considered to be among the most dangerous industries. Accordingly, construction safety has been raising several concerns due to the associated adverse impacts, thereby creating an urge to depart from the memorization and spoon-feeding approaches of the traditional safety training methods to novel teaching methods that incorporate new technologies such as virtual reality (VR) and augmented reality (AR). The efficacy of using VR-based training programs within the construction industry has been well established in the literature; yet there is still a lack of the simultaneous capitalization of several technologies to push the limits of the provided training programs in relation to the trainees' learning outcomes. Such a lack has been coupled with a lack in targeting the hazards associated with one of the riskiest construction project types, namely the construction of bridges. Thus, this research aims to design and test a safety training program that integrates AR and VR to address potential hazards in the construction of bridges in an attempt to introduce novel training means that would elevate the quality of training programs within the industry. It is worth noting that this research is a part of a mega-research that aims to develop a comprehensive AR and VR-based bridge construction safety training program. The results revealed a statistically significant improvement in the trainees' knowledge acquisition, safety motivation, safety awareness, and hazard identification and assessment skills. Hence, not only do the results of this research aid in enhancing the safety performances of bridge construction projects but also, act as a basis for the development of novel training approaches to elevate the learning experiences and gains of trainees.

**Keywords** Augmented reality · Virtual reality · Enhancing Bridges

---

M. El Rifaae (✉) · S. Bader · I. Abotaleb · O. Hosny · K. Nassar  
The American University in Cairo, Cairo, Egypt  
e-mail: [Mohamed.elrifaae@aucegypt.edu](mailto:Mohamed.elrifaae@aucegypt.edu)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_12](https://doi.org/10.1007/978-3-031-34593-7_12)

173

# 1 Introduction

## 1.1 Background and Overview

The construction industry is one of the most significant industries of today's world due to its crucial impact on the economy of any nation. The industry is characterized as an economic activity whose flourishing and prosperity are a direct reflection of the economic growth of a country. This is further driven by the increase in urbanization rates necessitating the construction of highly complex infrastructural projects, including the construction of bridges. The increase in population rates along with the increase in road traffic across the globe have also surged the demand for bridges, causing a rapid increase in the construction of highly sophisticated bridge structures with larger spans and capacities. In fact, the bridge construction market size was valued at \$857.7 billion in 2020 and is projected to reach \$1416.9 billion by 2030, registering a compound annual growth rate (CAGR) of 5.0% from 2021 to 2030 [24]. This growth has further been driven by increasing investments, government initiatives, and economic booms in both developed and developing countries [24].

However, such growth comes at an extremely high cost as the industry is considered to be among the most dangerous industries, thereby causing it to be concurrently recognized as an economic force yet, and one that is notoriously dangerous. This is further evidenced by the increasing number of fatal and non-fatal accidents witnessed in the industry across the globe. As stated by the International Labor Organization, "The construction industry has a disproportionately high rate of recorded accidents" [33]. To illustrate, in the USA, 40 out of 100,000 construction workers encounter fatal work injuries per year [5]. The case is quite similar in Europe where statistics reveal that about one-fifth of the work-related fatalities that took place in 2019 occurred in the construction industry.

Accordingly, construction safety has been raising several concerns throughout the years. This is not only due to its strong impact on the lives and welfare of construction workers and laborers but also, on the three constraints of construction projects, namely time, cost, and quality. Recent research has proved the inefficacy of traditional training methods in elevating the safety awareness of construction workers [3]. This is further supported by the findings of Albert et al. (2014) which revealed that construction workers were incapable of identifying more than 50% of the hazards in typical construction sites. This has been primarily attributed to the fact that trainees act as passive receivers of knowledge [3]. Thus, causing traditional training methods to be criticized for the lack of cognitive stimulation and skill development as a result of the lack of engagement and interactivity [3]. Not only does this risk the lack of knowledge due to the trainees' emerging sense of exclusion but also risks counterproductive learning. This creates an urge to depart from the memorization and spoon-feeding approaches to novel teaching methods.

In fact, the modern-day education system is witnessing a massive revolution with a dramatic shift toward experiential learning and proactive teaching approaches hence placing greater emphasis on “learning by doing” [12]. Such revolutionary approaches have necessitated the incorporation of multiple novel technologies to elevate the learning experiences and outcomes of learners. As stated by [17], “Technology-mediated learning is a relatively new phenomenon in adult learning and is rapidly becoming a vital component of the current and future workplace.” (p. 681). Within this context, VR technologies are highly being adopted as teaching platforms in higher education systems. Fromm et al. [9] further state that the opportunity to create learning experiences that were otherwise not possible in real-life is another main driver to the wide utilization of VR technology in education.

In line with this and to overcome the weaknesses and pitfalls of traditional training programs, a similar momentum toward the incorporation of VR technologies is currently being witnessed within the construction industry [15]. Effectively, the VR technology has already substantiated both its efficacy and effectiveness in construction safety training. This is primarily driven by the fact that the VR technology is capable of simulating the physical presence of trainees in an immersive virtual environment [30]. It is a cost-effective learning method that takes a proactive approach to learning, especially in fields and applications where actual practice is not feasible in terms of cost or safety [30].

From the literature, several benefits of incorporating VR technologies into safety training programs within the construction industry emerged. These include increased safety awareness, vigilance, knowledge retention, risk identification and cognitive ability [28], reduced reaction times and enhanced decision-making abilities [4], enhanced motivation and safety commitment levels [19, 30], among others. Whereas, as stated by [32], “Augmented Reality (AR) is a specialization of Mixed Reality, where virtual objects are superimposed upon the real world.” However, the literature is scarce when it comes to the use of AR within construction safety training [15, 32]. In addition, up to the researchers' knowledge, none of the previous studies have attempted to concurrently integrate AR and VR technologies in a single safety training program.

## ***1.2 Research Gaps***

From an extensive review of the literature, the following gaps emerged. Firstly, it is evident that most of the identified literature was concerned with investigating the potential of the technology in enhancing the learning outcomes of construction safety programs; thus, it could be fairly concluded that its effectiveness has been well-established in the literature. Yet, it was noticed that the majority of the previous studies were hazard-based studies meaning that they tended to focus on limited and general hazards within specific contexts relating to general construction sites, such as falling from a height, being struck by a moving vehicle, or being struck by falling objects. Thus, creating a lack in the provision of a comprehensive and

fully inclusive project-based safety training program. Not only does this reveal the limited application of the technology in general but also shows how the industry is still lagging behind when it comes to the leveraging of the full potential of the technology.

In fact, the research by Isleyen and Duzgun (2018) was the only identified research that developed a VR training program on a project level. The authors used virtual reality to train construction workers on the hazards that could arise in tunnels due to certain activities, such as tunnel-digging using explosives; they also attempted to assess the hazards associated with subsurface water. However, their training program was not inclusive of all potential hazards that could be encountered in tunneling construction projects. This creates an urge for the development of comprehensive training programs that address all project-related hazards based on the specific conditions and particularities of the given type of project at hand. Through the provision of a project-based safety training program, not only would workers be aware of and able to react to hazards within the context of their specific trades but also would be able to identify, assess, and mitigate all potential risks that are related to the type of project being studied.

Secondly, none of the reviewed literature had tackled the safety issues within the construction of bridges which shows a significant gap in addressing one of the riskiest construction project types. This is primarily attributed to the confined and crowded environments in which bridges are normally constructed including but not limited to being constructed over running roads, valleys, rivers, and railways. Up to the researchers' knowledge, the research conducted by [8] is the only identified research that aimed to provide comprehensive coverage of all the hazards encountered in bridges that are yet under construction. However, their research did not address any training needs. Not only does this gap show the disregard of a project type that is often constructed in highly complex and threatening environments but also disregards several alarming accidents that call for determined actions to minimize the extremely high rates of fatalities in the construction of roads and bridges.

Finally, and most importantly, there is a lack of research when it comes to the integration of AR and VR in construction safety training programs, thereby revealing an urge to investigate the benefits and the potential of integrating both technologies on the learning outcomes and safety awareness of construction safety trainees. Also, such integration is vital for the delivery of an effective and comprehensive safety training program. This is due to the fact that not all safety information and topics need to be modeled in a fully immersive and interactive environment such as theoretical and managerial safety aspects. Thus, through the provision of an AR training prior to the VR training, ample time could be saved as the trainees could collectively have the training at the same time using their smartphones yet still in a semi-interactive and engaging environment as compared to having such training in the traditional training form.

### ***1.3 Research Aims and Objectives***

Based on the aforementioned discussion, this research aims to bridge the existing gap in the literature by designing and testing a safety training program that integrates AR and VR to address all potential hazards in the construction of bridges in an attempt to introduce novel training means that would elevate the quality of training programs within the industry. Thus, the research objectives are as follows:

- To identify all potential hazards that could be encountered during the construction of bridges.
- To design and develop an AR and VR-based training model based on the identified hazards.
- To test the developed model to determine its efficiency and potential in relation to the learning outcomes of the trainees.

### ***1.4 Research Significance***

It is worth mentioning that this research is a part of wider-scale research which aims to develop a comprehensive training module that tackles the safety of bridge construction from a broad project-based perspective by incorporating different environments, materials, construction methods, and technologies. Thus, the significance of this research emerges from the fact that it addresses one of the riskiest construction projects that not only jeopardize the safety of construction workers but also the safety of passengers and other public assets. Accordingly, this training program would aid in the development of self-efficacious workers who are capable of identifying, assessing, and reacting to severe and life-threatening hazards in the confined and complex environments of bridge construction. Besides, this research would pioneer the incorporation of AR technology as a complementary tool to VR technology within construction safety training. Through such incorporation, the theoretical safety aspects could still be addressed in an engaging and semi-reactive environment, as compared to the traditional training methods, while saving time and costs that would be incurred if they were to be covered using VR technology. Hence, the results of this research could act as a basis for the development of novel training approaches that integrate a wide range of technologies to elevate the learning experiences of trainees, thereby augmenting their learning experiences and gains.

## 2 Theoretical Background

The significance of safety training in the construction industry has been acknowledged by multiple academic researchers throughout the past decades. However, existing safety training is conducted in the most traditional forms while using textbooks, lectures, and presentations in the construction industry; furthermore, they are mostly being conducted as mere compliance to safety regulations, thereby leading to a wide momentum toward the introduction of VR-based training. The following paragraphs provide an overview of the previous studies that incorporated VR and AR technologies for educational purposes.

### 2.1 *VR for Educational Purposes*

There has been a wide momentum toward the incorporation of VR technologies in multiple industries for various reasons including gaming, simulation, and forecasting. Nevertheless, its use for training and educational purposes has been witnessing dramatic growth across several educational fields. To illustrate, the aviation field has been heavily reliant on VR technologies in its flight simulations and aerial training [20]. Similarly, the medical field has been utilizing VR technology for diagnostic, treatment, care, and surgical simulation purposes [25, 26]. The use of VR has also witnessed an unprecedented surge in other industries including the automotive and manufacturing industries [16], emergency training and evacuations such as firefighting [4].

### 2.2 *VR in Construction Safety Training*

There have been several international attempts to incorporate VR in the construction industry; nevertheless, these attempts, although extremely beneficial, are very limited and primal in their scope, scale, technologies used, and areas covered. To demonstrate, VR has been tested in relation to training for general site safety and other specific activities such as pouring concrete [28], working from heights and fall hazards [11, 19, 28] general struck by and electrocution hazards [11], and hazard identification in general construction sites [13].

Other research papers tackled more specific learning outcomes from their VR-based training. To demonstrate, [14] employed virtual reality to improve safety during the prestressing process of concrete. Similarly, Li et al. (2012) have developed VR-based training to train students on the safety measures and procedures that should be implemented for the safe dismantling of cranes, whereas also, Iseyen and Duzgun (2019) developed a VR model to train construction workers on the safety of the tunnels in terms of the underground roof fall; while [22] developed a VR model to

train construction workers on safe postural positions to overcome muscular disorders that are often encountered from unsafe work postures. Other developed models tackled training on heavy machinery including the operation of tower cranes [28, 30], light equipment such as the safe use of a table saw [19], and off-site production training including pre-cast/prefabrication [12].

However, as stated earlier, most of the existing literature related to the incorporation of VR technologies within construction safety training was primarily focused on investigating and testing the potential and effectiveness of the technology, as compared to traditional teaching and training approaches. Thus, the comparisons were often concerned with learning outcomes of trainees [28] including but not limited to knowledge retention, risk identification, and cognitive ability [28], reaction times and decision-making abilities [4], among others. This was mostly done by dividing the research into two groups with one attending a traditional training program, based on slides, notes, and multimedia such as images, videos, and recordings, while the other attending a VR-based training program [28].

Whereas, others have compared the efficacy and effectiveness of VR technology against other training methods. To illustrate, [11] compared the training results of two main types of training, namely VR-based and PARS (Panoramas of Reality for Safety) training. Both training forms incorporated OSHA's major hazards such as falling from a height, electrocution, and being caught in between. This comparison was based on a hazard identification index (HII) that was found to be significantly higher in students who attended the VR training as compared to the ones who attended the PARS training. Similarly, [7] have compared the efficacy of VR against 360-degree panorama-based safety training. The results confirmed that the VR-based training was better than 360-panorama in terms of the trainees' hazard identification skills.

### ***2.3 AR in the Construction Industry***

As stated earlier, there is a lack of research that incorporates AR into construction safety training [15]. Rather, most of the identified literature was primarily related to the use of AR as a design and construction support tool that is further complemented with BIM [10, 27, 29]. Yet, a few beneficial research papers in relation to construction safety were identified. To demonstrate, [27] developed an integrative design framework for bringing recent advances in augmented reality (AR) and artificial intelligence (AI) to enhance the safety of highway workers in highway work zones. Yang and Miang Goh [34] developed an immersive VR/MR (mixed reality) simulation-based activity to train individuals from the architecture, engineering, and construction (AEC) industry on different safety management lessons.

While Albert et al. (2014) employed augmented reality by developing the "system for augmented virtuality environment safety" (SAVES) model to substitute the traditional and paper-based evaluation exam for safety officers and supervisors. Likewise, Kim et al. (2019) developed an AR model to replace the traditional paper-based



assessment method of construction safety officers. Their results revealed that the AR-based assessment technique was more effective in determining the exact awareness level of students and trainees. They concluded that this technique would better aid safety managers in identifying the areas of flaws and potential improvement in the safety officers' knowledge base, based on which future training programs could be designed to enhance the safety performance of their projects.

## ***2.4 Comparison Between Traditional Safety Training and AR/VR Safety Training***

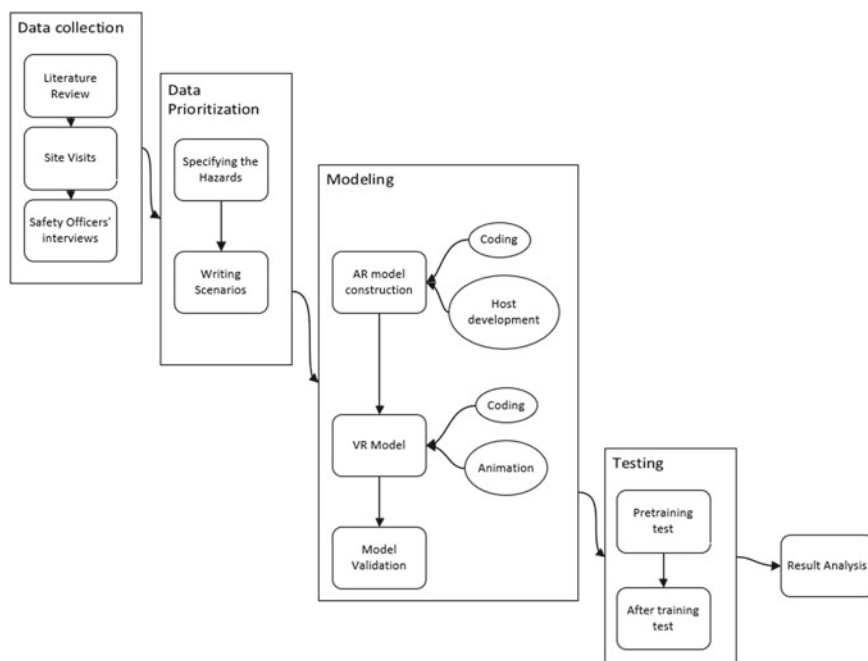
Traditional training is lecture-based training that is heavily reliant on slides, lecture notes, and multimedia, with the instructor mostly beginning to demonstrate the slide; consequently, trainees are considered passive learners [1]. In this design, there is a lack of interaction between learners and training material [23]. On the other hand, VR overcomes this gap by immersing the student in a virtual environment filled with risks to interact with or investigate [1, 4, 28]. As the trainees gain real-world building site experience, this immersion accomplishes the aim of enhancing engagement, awareness, and response time comparing to the traditional training [1, 22, 28]. Although VR fills in conventional training gaps, game creation is time-consuming and costly. AR was advised at this training to save design time and costs, and it is also accessible to everyone, since it simply takes a mobile phone and a host paper.

## **3 Methodology**

To achieve the main aims and objectives of this study, the methodology was divided into four phases, namely data collection, data prioritization, modeling, and testing, as presented in Fig. 1. The following paragraphs provide more details on each of these phases. However, it should be noted that the developed model is part of a larger and more comprehensive model that is still under development.

### ***3.1 Data Collection***

As identified in the introduction chapter, there is a lack of research when it comes to the comprehensive coverage of all potential hazards that could be encountered during the construction of bridges; thus, this stage entails the collection of data that directs the flow of the research in the design and model development stages to create a more elaborate and inclusive training model. Accordingly, data on the most encountered



**Fig. 1** Research methodology

hazards in bridge construction was gathered through two primary techniques namely, interviews and site visits.

### 3.1.1 Interviews

Interviews with safety managers who are currently working in bridge construction projects were conducted in Egypt. The type of interviews conducted was semi-structured; whereas, the purposive sampling technique was used where interviewees who play key roles in managing the safety of bridge construction were chosen. This was further accompanied by the snowball sampling technique where interviewees were asked to refer the researchers to other safety managers. In total, five interviews were conducted with safety managers that are currently supervising bridge constructions. The experience levels of the experts ranged from 14 to 20 + years of experience in managing the safety of construction projects.

3.1.2 Site Visits

Besides the interviews, site visits to bridge projects that are still under construction in Egypt were also conducted. The aim was to target projects under the main construction and finishing phases. Also, specific considerations were given to the location of the sites to cover a wide range of conditions and possible constraints. Five site visits were conducted in total for five main bridge types, each with a unique set of circumstances and hazards as illustrated in Fig. 2.

Site Visit #1

The first site visited was an elevated railway constructed amid a busy and congested area in the suburbs of Cairo and is being built over a critical road connecting two cities. The uniqueness of this bridge emerges from the fact that it is not meant for automobiles or trucks but rather serves as a rail for an elevated train.

Site Visit #2

The second bridge was being built as a crossroads so it can aid with traffic congestion problems. Besides, it is being built in a heavily populated region near buildings with 5 or more stories; thereby, the site included several constraints in terms of its method statements. Moreover, the site is dealing with major lifting issues due to the imposed limitations on crane heights as a result of the existence of electricity cables besides the neighboring structures.

Site Visit #3

The third bridge is being built across the River Nile as an extension to the width of an existing bridge. The main hazards emerging from this site were primarily related

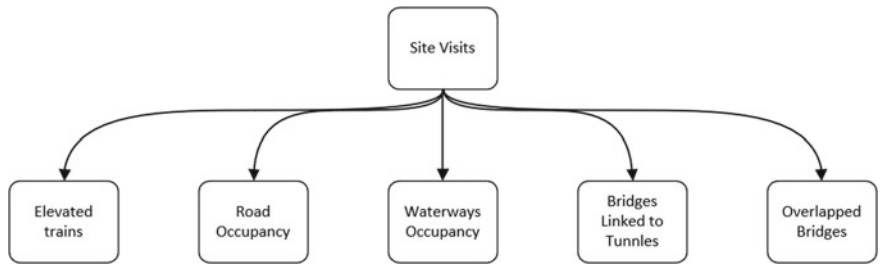


Fig. 2 Types of the site visits

to the construction over/near water. Besides, the site was also governed by other environmental and marine work regulations.

#### Site Visit #4

The fourth bridge was in an open, remote, and deserted area, thereby causing extreme weather conditions, such as high wind speeds and sand storms, to be main interfering factors and sources of hazards during construction. In fact, the bridge is to be connected to a tunnel which further magnified the hazards in this construction site.

#### Site Visit #5

Similarly, the fifth bridge was unique for the following reasons. Firstly, it is being constructed in one of the densest and most populated areas in Cairo. Secondly, it is being constructed over and above one of the busiest bridges in the city. Thirdly, it is also being built over one of the busiest bus stations that thousands of people rely on for their daily commute. Figure 3 shows a few pictures of the sites visited.

### **3.2 Data Prioritization**

Subsequently, the collected hazards from all conducted interviews and site visits were organized and prioritized. This was done to avoid repetitions in the training model and to allow for the inclusion of the most significant hazards. The prioritization was conducted using the Delphi technique with three experts who conducted a probability-impact analysis where the identified hazards were assessed based on their probability of happening and their impact on the safety of construction laborers and their impacts on the three constraints a project. Such impact was measured based on delay in time, loss of money, or loss of productivity. From this analysis, all the identified hazards along with the associated risks were categorized and prioritized to inform their design and inclusion in both the VR and the AR models. However, for the purpose of this research, hazards relating to the themes shown in Fig. 4 are the only hazards included.

### **3.3 Modeling**

As stated earlier, this research integrates both AR and VR technologies to achieve its main objectives. Thus, the modeling phase included the design and development of the AR and the VR models as illustrated hereunder.



a. Sample pictures from site visit 2.



b. Sample pictures from site visit 4.

**Fig. 3** Site pictures from the visited bridge construction sites



c. Sample pictures from site visit 5.

Fig. 3 (continued)

### 3.3.1 Augmented Reality Model

The AR model was developed using Unity Game Engine (2019.4.17f1), Blender, and Unity's animation assets and to be operated on the Android Operating System. The Adobe Photoshop Software was utilized to develop the host; whereas, Vuforia Augmented Reality SDK was used to transform the photoshopped images to targets suitable for hosting the model. As the Vuforia package was used to add an AR camera used to access the camera when the software was launched; besides, an image target was added as well that image target was the host. Model objects were added mainly from three packages (AF construction pack 1, Construction worker, and Simple vehicle pack). The model buttons and animation triggering were programmed in C#. The produced augmented reality (AR) model is a camera-based model that could be accessed using the trainees' smartphones. The model includes several hazards, each hosted on a separate image on a physical paper (Hazard sheet) to be scanned by the trainees. Once the camera detects the hosts, the model associated with the targeted images from the hazard sheet is displayed on the trainees' smartphone screens. Figure 5 shows the developed hazard sheet along with the developed targets to be hosted (marked in yellow).

The AR model consisted of three scenes with a total of twelve hazards. In the first scene, a concrete mixer was operating alongside a combustible barrel and a worker stood behind the concrete mixer. The hazards in this situation were primarily targeting coordination and workers' unsafe behavior issues. The second scene simulated a lifting procedure by having a crane lift an overly excessive load without assigning safe movement zones. Thus, this scene was primarily addressing hazards related to site layout, site boundaries, nearby structures, and moving workers and passengers. The third scene simulated road occupancy issues caused by the extension of the crane's

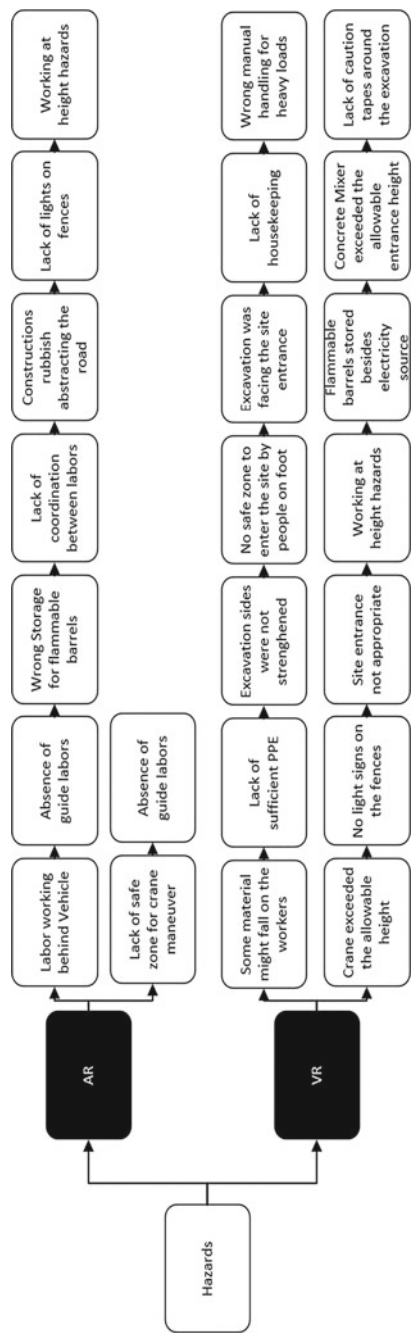


Fig. 4 Hazards included in both the AR and VR models





Fig. 5 Host's nodes

boom outside the site's boundaries while lifting an excessive load. Accordingly, this scene was primarily targeting road occupancy hazards. Figure 6 shows the AR model as viewed by the trainees using their smartphones.

### 3.3.2 Virtual Reality Model

The VR model replicated bridge construction activities including excavation, concrete pouring, pier cap installation, and rigging and lifting activities. The model was developed with two translations in English and Arabic. The trainees were instructed to navigate and inspect the site. As the trainees approach one of the hazardous locations, the right Oculus joystick (Oculus Touch) begins to vibrate, revealing the hazard and indicating the means through which such hazards are mitigated. The user interface was designed on a canvas with two buttons, one for the Arabic model and one for the English model. The Oculus Touch OVRRaycaster script was used to interface with the buttons, and the default EventSystem was changed with the one supplied in the UIHelpers. Accordingly, for each joystick (handAnchor), a canvas was added for the trainees which emerged according to the trainees' location inside a specific radius ( $R$ ). The right-hand canvas was used to clarify the hazard to the trainees as they navigate the site; whereas, the left-hand canvas was used to guide and direct the trainees toward ways to mitigate/prevent the potential consequences of the encountered hazards. These canvases were developed using the C# script Detect and Change which was added to the hazardous places.



**Fig. 6** AR model as viewed from the trainees' smartphones

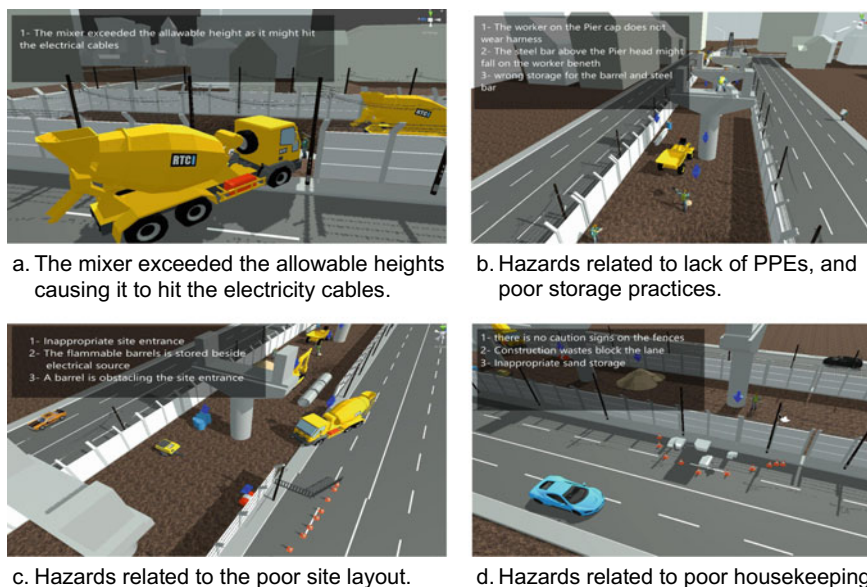


a. AR models as they emerged on the trainees' smartphones when they hovered it over the hazard sheet.



b. AR model shows further explanation to the emerging hazards.

The site Layout was imported from the LowPolyRoadPack asset, and various adjustments, such as structural piers and pier tops, were applied using Blender 2.91 and Revit 2019. This model was created with Oculus Quest in mind and was built using the Unity game engine to create the virtual environment and imitate the site layout. For various animations, such as walking for some workers, the motion of the objects was added using Maximo. Other building operations were imported from the construction worker asset. To change the location and position of the characters, an animator overwrite controller was employed. Apart from the imported animations from Maximo and the Unity asset shop, the characters were animated in Blender. The primary camera was then replaced by OVRPlayerController, which included OVRCameraRig. OVRCameraRig is developed by Meta corporation to develop VR games targeting Oculus quest and Oculus quest 2 virtual reality headsets. The model was written in C# using Visual Studio. Revit was used to model and build the structural pieces attached to the model, with the model being based on a real pier and pier cap from one of the visited bridges.



**Fig. 7** Sample of hazards in AR and VR models

### 3.3.3 Model Validation

The model was validated using three experts with 5, 6, and 10 years of experience in the field of construction safety management. The experts reviewed the model and significantly contributed to enhancing the quality of hazard representation and sequence. They also contributed toward the realism of the developed model; thereby, aiding the researchers in adding the needed amendments to create a realistic model. Figure 7 shows samples of the hazards included in the model.

## 3.4 Testing Phase

A test was administered to all the trainees to determine the effectiveness of integrating both AR and VR technologies. The test was divided into two sections: a pre-training assessment of trainees' knowledge and a post-training assessment of trainees' increment in knowledge. The same test was administered to all the trainees before and after the training. The test consisted of photographs in which the trainees were asked to analyze the hazards in each photograph. Each photograph represented one of the scenarios that the learner would encounter during the VR and AR training. Additionally, the exam was confirmed by experts in the field of construction safety,

as five experts solved it with experience ranging from five to twenty-five years in the field of construction safety. Then, prior to testing the students and trainees, their opinions about the quiz length and the clarity of the images were taken into account.

## **4 Results and Analysis**

### ***4.1 Research Participants***

In total, 24 participants, including students and engineers, undertook the safety training using both the AR and VR technology followed by the hazard identification test. Their years of experience ranged from zero to 36 years. All of the aforementioned participants confirmed that they have never attended VR training before despite being aware of the technology. However, 6 of the research participants confirmed that they have previously used VR technology for gaming applications. When it comes to AR technology, all of the participants confirmed that they have never witnessed nor used AR technology before.

### ***4.2 Pre-training Results***

The average time that the participants took to solve the pre-training exam was 13 min. An average score of 14.16 out of 37 was obtained in this test. The difficulties encountered by the trainees in solving the exam, students and experienced engineers alike were apparent during the tests. One of the most common observations witnessed during grading the tests was the trainees' failure to identify all the hazards in a given photo, despite being informed that a photo might contain more than one hazard. A possible interpretation for this phenomenon could be attributed to the trainees' lack of safety awareness in general, and in relation to bridge construction in specific. This has caused the majority of participants to disregard some of the basic, yet risky, hazards in the photos. Another possible interpretation, specifically with experienced engineers, is the fact that they might have not been able to perceive the hazards in the photos as they are used to such hazards being common and dominant within construction sites that lack effective safety management.

### ***4.3 Training Results***

Subsequently, the AR model was used as the starting point for the training. The AR software and the hosts were sent to the learners, and they were then asked to access them using their smartphones. As stated earlier in the methodology, this part

of the training was primarily concerned with theoretical and managerial aspects. The training was conducted in the form of groups where 3–4 participants undertook the training simultaneously and took an average of 10 min to be conducted. During the training session, it was observed that the participants started discussing the hazards and threats together while discussing potential reasons behind such hazards as they scanned the hazard sheet using their smartphones. Accordingly, the training took the form of an interactive session where the trainees were drawing the attention of their colleagues to their perspectives throughout the training session. This sheds light on the level of focus and attention of the trainees during the training. Moreover, it highlighted the level of interactivity and enthusiasm attained which was much higher as compared to the trainees' enthusiasm and excitement levels in traditional training forms.

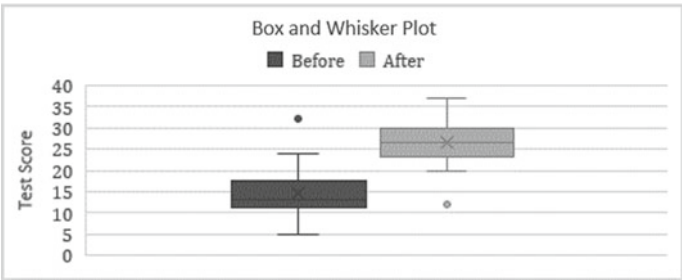
The case was quite similar as the trainees entered the VR model. In the model, the trainees promptly started witnessing the hazards that were discussed in the AR session while showing significant improvements in the assessment of the associated risks in terms of their severity (likelihood of occurrence and potential impact on workers, time, cost, and quality). These were further complemented with more practical information as the trainees were being exposed to the hazards themselves. To illustrate, some of the trainees expressed concerns over having to move under unstable loads while others were anxious about their possibility of falling into an unprotected excavation zone due to a lack of fences covering such an area. However, it was noticed that younger trainees were more comfortable with the VR experience, as it took them around 15 min on average to navigate the site and detect most of the hazards. On the other hand, the older trainees took about 22 min on average to navigate the model and complete the training. They were also not comfortable/ not used to using the Oculus touch (Joystick). Additionally, 4 of the older participants have withdrawn from the VR training in the first 1–3 min; therefore, their pre-training test results were omitted from further analysis. Again, the trainees showed extremely high levels of interactivity and enthusiasm to the extent that they started suggesting further means through which the site layout could be improved to improve the safety performance of the project.

#### **4.4 Post-training Results**

Finally, the participants took part in the post-training evaluation. Despite having a longer average time to solve the same test (15 min as compared to 13 min in the pre-training test), all of the trainees showed significant improvements in their ability to identify the hazards that were not previously identified in the pre-training test. As seen in Table 1, the mean score of the trainees has witnessed an 81.6% increase post the AR and VR training as compared to their test scores before the training (Before Training:  $M = 14.7$ ;  $SD = 5.873$ ; After Training:  $M = 26.7$ ;  $SD = 5.512$ ). The box and whiskers plot, as shown in Fig. 8, further shows the difference in mean, median, and ranges of the data from both groups.

**Table 1** Statistical results and analysis

Factors	Before training	After training
Sample size	24	24
Mean	26.7	14.7
Standard deviation	5.512	5.873
Shapiro–Wilk Normality test <i>P</i> -value	0.1886	0.08038
Distribution	Normally distributed	Normally distributed
<i>t</i> -Test (Paired two samples) to be conducted for comparing between the two groups	<i>P</i> < 0.0001, which is less than 0.05, then the variance is statically significant	



**Fig. 8** Box and whisker plot of the obtained scores

A Shapiro–Wilk Normality Test was initially conducted to further compare such data to determine the distribution of the data gathered. The results indicated that the scores of both testing groups were normally distributed (Before Training:  $p = 0.1886$ ; After Training:  $0.08038$ ). Thus, a paired sample *t*-test was conducted, as given in Table 1. The results revealed a statistically significant difference between the mean values of both tests ( $p = 4.0E-09 < 0.05$ ). Thereby, it could be concluded that there is strong evidence that the training model had contributed to enhanced learning outcomes to the trainees in terms of their knowledge acquisition which increased their hazard awareness. These improvements have positively reflected the trainees’ hazard identification and assessment skills leading to higher scores.

## 5 Discussion

From the aforementioned results, it could be concluded that the training was extremely effective in enhancing the learning outcomes of the trainees. While the aforementioned literature was primarily concerned with testing the effectiveness of

VR-based training, there have also been several recent studies that aimed to integrate AR and VR technologies. To illustrate, [32] used augmented virtuality (AV) as a safety training tool where trainees get supervised feedback. Similarly, Kim et al. (2019) developed an AR-based assessment model which aided in determining the exact awareness level of students and trainees. They concluded that this technique would better aid safety managers in identifying the areas of flaws and potential improvement in the safety officers' knowledge base, based on which future training programs could be designed to enhance the safety performance of their projects.

Despite the lack of research that uses AR technology as a complementary tool to VR-based training, there are several insights that could be obtained from existing research. To start with, it has been established that AR, in its entirety, does not significantly contribute to the learning curve of students as compared to the traditional training means [18], however, it was found to be effective in tasks that do not require complex reasoning and decision-making. Thus, it was concluded that AR-based training aids in the provision of procedural knowledge meaning that it is beneficial in showing trainees the "what" and "why" elements. To illustrate, [31] proved the efficacy of AR technologies in introducing new game rules and immediate feedback to sportspeople. This was quite the case in this research as the AR-based training mainly tackled theoretical aspects, exposing the trainees to the types of hazards and the associated risks and consequences. The high interaction between the trainees further supports this during the AR-based training session. This, in fact, has led to extremely high knowledge acquisition levels by the trainees. Accordingly, knowing the "what" (types of hazards) and the "Why" (why they are considered as hazards) along with the trainees' interaction acted as a basis for the VR-based training.

During the VR-based training, the trainees readily implemented the theoretical knowledge gained from the AR training, which was evident in their elevated hazard recognition and identification skills. Also, the VR-based training had further complemented such learning by reinforcing the "how" element, thereby elevating the hazard assessment and risk mitigation skills of the trainees. These findings conform to the findings from the literature that proved significant improvements in the trainees' hazard identification skills [28], Le et al. 2014, [2]. Furthermore, the efficacy of VR-based training in supporting the learning process through the practical acquisition of new knowledge has also been proved. As stated by [2], "VR-based training was associated with a significant increase in knowledge, operational skills, and safety behavior compared to in-person training." The authors attributed such enhancement in knowledge acquisition to the limitations of using a communication-based approach to teaching aspects that purely rely on adequate visualization and specialization, which is the case in the construction safety curriculum. While the enhanced visualization, interactivity, and immersion levels were also considered to be substantial factors that led to enhanced learning outcomes in this research, it was also interpreted that the high attention and focus levels, as a result of the high learning motivation levels of the trainees, were also leading factors that drove significantly higher scores.

These findings conform to the findings of Nykänen et al. [19], who indicate that VR-based training is capable of enhancing the trainees' safety motivation, self-efficacy, and safety-related outcome expectancies. Thus, this research provided preliminary evidence on the effectiveness of integrating AR and VR technologies in construction safety training.

## 6 Conclusion

To conclude, this research aimed to design and test a safety training program that integrates AR and VR to address potential hazards in the construction of bridges as part of a mega-research that aims to develop a comprehensive AR and VR-based bridge construction safety training program. The results revealed a statistically significant improvement in the trainees' knowledge acquisition, safety motivation, safety awareness, and hazard identification and assessment skills. Hence, this research could act as a basis for developing novel training approaches that integrate a wide range of technologies to elevate trainees' learning experiences, thereby augmenting their learning experiences and gains. However, it should be noted that future research should attempt to quantify each technology's contribution to the trainees' learning outcomes. Thus, it is suggested that a test should be conducted between the AR and VR training sessions to have a deeper insight into the specific learning outcomes that were enhanced through each technology type.

## 7 Future Research

In the future study, it is intended to compare the suggested method's outcomes to those of conventional methods and VR training. Additionally, it is intended to switch from a regular quiz to an AR exam for evaluation.

## References

1. Abotaleb I, Hosny O, Nassar K, Bader S, Elrifaae M, Ibrahim S et al. (2022) Virtual reality for enhancing safety in construction. Construction Research Congress 2022. <https://doi.org/10.1061/9780784483961.125>
2. Adami P, Rodrigues P, Woods P, Becerik-Gerber B, Soibelman L, Copur-Gencturk Y, Lucas G (2021) Effectiveness of VR-based training on improving construction workers' knowledge, skills, and safety behavior in robotic teleoperation. *Adv Eng Inform* 50:101431. <https://doi.org/10.1016/j.aei.2021.101431>
3. Al-Qahtani A, Higgins S (2012) Effects of traditional, blended and e-learning on students' achievement in higher education. *J Comput Assist Learn* 29(3):220–234. <https://doi.org/10.1111/j.1365-2729.2012.00490.x>

4. Bourhim E, Cherkaoui A (2020) Efficacy of virtual reality for studying people's pre-evacuation behavior under fire. *Int J Hum Comput Stud* 142:102484. <https://doi.org/10.1016/j.ijhcs.2020.102484>
5. BLS (2020) National Census of Fatal Occupational Injuries in 2019. US: U.S. Department of Labor. Retrieved from <http://www.bls.gov/iif/oshcfdef.htm>
6. Chuai X, Lu Q, Huang X, Gao R, Zhao R (2021) China's construction industry-linked economy-resources-environment flow in international trade. *J Clean Prod* 278:123990. <https://doi.org/10.1016/j.jclepro.2020.123990>
7. Eiris R, Gheisari M, Esmaeili B (2020) Desktop-based safety training using 360-degree panorama and static virtual reality techniques: a comparative experimental study. *Autom Constr* 109:102969. <https://doi.org/10.1016/j.autcon.2019.102969>
8. Essa R (2015) Accidents and Hazards in Construction of Bridges. Repository.sustech.edu. Retrieved 4 March 2022 from <http://repository.sustech.edu/handle/123456789/12485>
9. Fromm J, Radianti J, Wehking C, Stieglitz S, Majchrzak T, vom Brocke J (2021) More than experience? - on the unique opportunities of virtual reality to afford a holistic experiential learning cycle. *Internet High Educ* 50:100804. <https://doi.org/10.1016/j.iheduc.2021.100804>
10. Garbett J, Hartley T, Heesom D (2021) A multi-user collaborative BIM-AR system to support design and construction. *Autom Constr* 122:103487. <https://doi.org/10.1016/j.autcon.2020.103487>
11. Geisari M, Esmaeili B (2018) PARS: using augmented panoramas of reality for construction safety trainings. The Center for Construction Research and Training, The University of Florida, George Mason University
12. Goulding J, Nadim W, Petridis P, Alshawi M (2012) Construction industry offsite production: a virtual reality interactive training environment prototype. *Adv Eng Inform* 26(1):103–116. <https://doi.org/10.1016/j.aei.2011.09.004>
13. Isleyen E, Duzgun H (2019) Use of virtual reality in underground roof fall hazard assessment and risk mitigation. *Int J Min Sci Technol* 29(4):603–607. <https://doi.org/10.1016/j.ijmst.2019.06.003>
14. Joshi S, Hamilton M, Warren R, Faucett D, Tian W, Wang Y, Ma J (2021) Implementing virtual reality technology for safety training in the precast/prestressed concrete industry. *Appl Ergon* 90:103286. <https://doi.org/10.1016/j.apergo.2020.103286>
15. Li X, Yi W, Chi H, Wang X, Chan A (2018) A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Autom Constr* 86:150–162. <https://doi.org/10.1016/j.autcon.2017.11.003>
16. Lawson G, Salanitri D, Waterfield B (2016) Future directions for the development of virtual reality within an automotive manufacturer. *Appl Ergon* 53:323–330. <https://doi.org/10.1016/j.apergo.2015.06.024>
17. Mancuso D, Chlup D, McWhorter R (2010) A study of adult learning in a virtual world. *Adv Dev Hum Resour* 12(6):681–699. <https://doi.org/10.1177/1523422310395368>
18. Moghaddam M, Wilson N, Modestino A, Jona K, Marsella S (2021) Exploring augmented reality for worker assistance versus training. *Adv Eng Inform* 50:101410. <https://doi.org/10.1016/j.aei.2021.101410>
19. Nykänen M, Puro V, Tiikkaja M, Kannisto H, Lantto E, Simpura F et al. (2020) Implementing and evaluating novel safety training methods for construction sector workers: results of a randomized controlled trial. *J Saf Res* 75:205–221. Retrieved 19 April 2021 from <https://doi.org/10.1016/j.jsr.2020.09.015>
20. Oberhauser M, Dreyer D, Braunstingl R, Koglbauer I (2018) What's real about virtual reality flight simulation? *Aviat Psychol Appl Hum Factors* 8(1):22–34. <https://doi.org/10.1027/2192-0923/a000134>
21. OSHA (2019) Commonly used statistics. Retrieved from: <https://www.osha.gov/data/commonstats>
22. Ojelade A, Paige F (2020) Virtual reality postural training for construction. In: Construction research congress 2020. <https://doi.org/10.1061/9780784482872.061>



23. Ojha A, Seagers J, Shayesteh S, Habibnezhad M, Jebelli H (2020) Construction safety training methods and their evaluation approaches: a systematic literature review. In: The 8th international conference on construction engineering and project management. Retrieved form: <https://bit.ly/3OtmUx9>
24. Pawar D, Sumant O (2022) Bridge construction market trends, analysis and forecast 2030. Allied Market Research. Retrieved 1 March 2022. from <https://www.alliedmarketresearch.com/bridge-construction-market>
25. Pensieri C, Pennacchini M (2014) Overview: virtual reality in medicine. *J Virtual Worlds Res* 7(1). <https://doi.org/10.4101/jvwr.v7i1.6364>
26. Riener R, Harders M (2012) Virtual reality in medicine
27. Sabeti S, Shoghli O, Baharani M, Tabkhi H (2021) Toward AI-enabled augmented reality to enhance the safety of highway work zones: feasibility, requirements, and challenges. *Adv Eng Inform* 50:101429. <https://doi.org/10.1016/j.aei.2021.101429>
28. Sacks R, Perlman A, Barak R (2013) Construction safety training using immersive virtual reality. *Constr Manage Econ* 31(9):1005–1017. <https://doi.org/10.1080/01446193.2013.828844>
29. Schiavi B, Havard V, Beddiar K, Baudry D (2022) BIM data flow architecture with AR/VR technologies: use cases in architecture, engineering and construction. *Autom Constr* 134:104054. <https://doi.org/10.1016/j.autcon.2021.104054>
30. Song H, Kim T, Kim J, Ahn D, Kang Y (2021) Effectiveness of VR crane training with a head-mounted display: double mediation of presence and perceived usefulness. *Autom Constr* 122:1–11
31. Soltani P, Morice A (2020) Augmented reality tools for sports education and training. *Comput Educ* 155:103923. <https://doi.org/10.1016/j.compedu.2020.103923>
32. Wolf M, Teizer J, Wolf B, Bükür S, Solberg A (2022) Investigating hazard recognition in augmented virtuality for personalized feedback in construction safety education and training. *Adv Eng Inform* 51:101469. <https://doi.org/10.1016/j.aei.2021.101469>
33. World Statistic. Ilo.org (2021) Retrieved 12 September 2021. from [https://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS\\_249278/lang--en/index.htm](https://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS_249278/lang--en/index.htm)
34. Yang F, Miang Goh Y (2022) VR and MR technology for safety management education: an authentic learning approach. *Saf Sci* 148:105645. <https://doi.org/10.1016/j.ssci.2021.105645>

# Integrating Virtual Reality into IOSH Safety Training



Y. Elhakim, S. Bader, M. Elrifae, S. Ibrahim, A. Sorour, M. Soliman, M. Sherif, I. Abotaleb, O. Hosny, and K. Nassar

**Abstract** The construction industry has high rate of accidents and poor safety records around the world. In addition to the injuries and fatalities, accidents affect both time and cost of the projects, thereby leading to time delays and additional costs. Previous research studies indicate that poor training contributes to the majority of accidents in the construction sector. Traditional training methods are inefficient in inducing the desired learning outcomes due to the passivity of knowledge receivers. Accordingly, new experimental and proactive learning approaches have emerged, out of which the safest and most ethical is the virtual reality (VR) construction safety training. VR safety training enables users to simulate reality in 360 degrees at the full visual capacity and to react accordingly improving understanding and critical response to stimuli. The research stream in this area is still relatively not fully explored, and previous work mainly focuses on semi-immersive VR technologies. The goal of this research is to develop and test a fully immersive and interactive VR model for safety training in accordance with the Institution of Occupational Safety and Health (IOSH). The focus is on advancing hazard recognition and mitigation “skills” rather than just providing “information”. The VR model was developed and deployed using state-of-the-art technologies, with a focus on falls, struck-by, slips, and general site safety. A testing phase was then initiated throughout interviewing professional safety inspectors to validate the model capabilities and to give their insight into approaches for further development. After that, the model was tested on senior university construction engineering students, and their performance was compared to students who didn’t take the VR training. Interviews indicated that the developed model would be of great importance for entry level workers and for general safety inductions; in addition, it can be more developed to target professionals on training centers. Experimental testing indicated statistical significance in the effectiveness of the developed model in enhancing students’ understanding

---

Y. Elhakim (✉) · S. Bader · M. Elrifae · S. Ibrahim · A. Sorour · M. Soliman · I. Abotaleb · O. Hosny · K. Nassar  
The American University in Cairo, Cairo, Egypt  
e-mail: [yasminmohamd@aucegypt.edu](mailto:yasminmohamd@aucegypt.edu)

M. Sherif  
University of Hawai’I, Honolulu, USA

and visualization. In addition, when compared to other available VR models, the developed model performed better in terms of visualization, immersion, realism, and ability to enhance the desired hazard identification and mitigation skills.

**Keywords** IOSH safety training · Virtual Reality

## 1 Introduction

Construction industry has high accidents ratio and poor safety records around the world [15]. It produces around 25–40% of fatal accidents in developed countries, although, it constitutes only 6–10% of work force. The case is even worse in developing countries. A total of 60,000 construction fatalities is estimated to occur per year around the globe, which is equitable to one fatality in construction sector every 10 min [10]. One of the major causes of high accidents rates in construction sector is poor training, as it contributes to more than half the occupational incidents [15]. Therefore, there are different recognized safety trainings presented by various international organizations that are tailored to professionals in the construction industry. These trainings general aim is to equip personnel working in the construction industry with the crucial skills required to manage and operate sites safely in order to reduce hazards and fatalities. Among health and safety trainings, IOSH is “the chartered body and the largest membership organization.” It provides trainings for safety professionals in around 130 countries. IOSH offers different training courses based on the targeted trainees and their level of expertise. Among these trainings, the “Working Safely” and “Managing safely” offer basic and general safety trainings that equip trainees with the basic knowledge of health, safety, and environmental aspects. Working safely training provides facts, case studies, examples, and recognizable scenarios related to hazards and fatalities that the trainees may be exposed to in real life. As for the Managing safely training, it is tailored to managers to manage workplace safely and to boost safety awareness in their organizations [11]. As safety trainings are proven to be of great importance in decreasing site injuries and improving workers’ productivities, efforts are exerted to make these trainings more effective and realistic to site conditions. Conventional training methods usually deal with trainees as passive receivers of knowledge; therefore, these trainings lack cognitive simulation, engagement, and interactivity [3]. Accordingly, emphasis on experimental learning and proactive teaching approaches that boost “learning by doing” have been widely emphasized [9]. This has to be ensured in a safe environment without exposing trainees to actual hazards. Therefore, using virtual reality in enhancing safety-training programs could be considered a promising technique for better visualization and response to site conditions and potential hazards.

Virtual reality creates a virtual environment where the user is immersed and is allowed to interact actively with the surroundings. It is composed of a computer, software, and hardware to create a simulated environment that allows user interaction and tracks the user’s location and orientation [15]. Virtual reality enhances “learning

by doing” which improves critical thinking [1], in addition to acquiring the trainee with the experience of dealing with hazardous materials, restricted, and inaccessible locations without being exposed to actual risk. Therefore, VR is considered a cost-effective learning method that allows proactive learning, especially in areas where actual practice is not feasible in terms of cost or safety [20]. VR has proved to improve reaction time and decision making skills [4], in addition to boosting motivation to safety and self-efficacy [16, 20]. Several attempts to incorporate virtual reality in safety trainings have proven that VR has the potential to increase trainees’ immersion and perception of safety aspects. However, these trainings are usually focused on one site activity like crane dismantling for example or general trainings not abiding by a standard safety training format.

Therefore, the objective of this paper was to develop a VR model that could supplement IOSH trainings and to test it among professional safety inspectors and entry-level engineers. Therefore, the significance of the model for enhancing overall understanding and performance was tested on entry-level engineers. In addition, professional’s comments will be taken into account to further refine the model.

## 2 Literature Review

VR use in different disciplines has been escalating in previous years for different purposes such as education, gaming, simulation, and forecasting. It has been used in aviation industry heavily for simulation of flights and aerial trainings [5, 6, 17]. It has been also employed in the medicine industry for surgical simulation purposes, in addition to diagnosis and treatment conditions [13, 18, 19]. In the construction industry, the incorporation of VR has witnessed several attempts which are beneficial but still limited in areas covered, scale and technologies used. Nykanen et al. [16], implemented VR methods for safety trainings for construction workers and evaluated the performance based on social cognitive theory and theory of planned behavior. In total, 120 participants were engaged in this experiment, half of them trained by VR system and the other half trained based on lecture-based training. Then, there were three measurement points which are baseline, short-term follow up and one month follow up. They were trained to use personal protective equipment, identifying hazards, hazard awareness around traffic and inspecting equipment and safety for using table saw and working at height. It was concluded that VR safety training showed greater increase in self-reported safety, VR lead to higher self-efficacy based on long-term effect (one month follow up). It was reported that 81% gave positive feedback for implementing VR for better in-depth thinking and implementation of occupational safety.

Joshi et al [12] implemented virtual reality for safety training in the precast/prestressed concrete industry. The researchers developed the module of four major parts: audio/video instructions, Personal Protective Equipment, suspended heavy loads, and the stressing process. Then, they tested the module on 32 students from the Mississippi State University, where half of the students used traditional video

experiment and the other half used the VR training. The researchers analyzed the results based on efficacy and effectiveness and concluded that VR training provided better engagement and motivation than the traditional training methods. In addition, the simulation sickness in case of VR was within the acceptable range.

Gheisari and Esmaeili [7] developed a training tool based on augmented panorama of reality PARS for construction safety trainings. They compared the effectiveness of application of VR and PARS on 52 trainees, where half of the trainees were exposed to the PARS and the other half was exposed to VR and modeled the four leading causes of fatalities by OSHA. Subsequently, they assessed the performance by hazard identification index. It was concluded that VR trainees has higher scores regarding hazard identifications. However, 360 panorama gives true representation of construction sites. Therefore, PARS was better in terms of representing construction site and associated hazards.

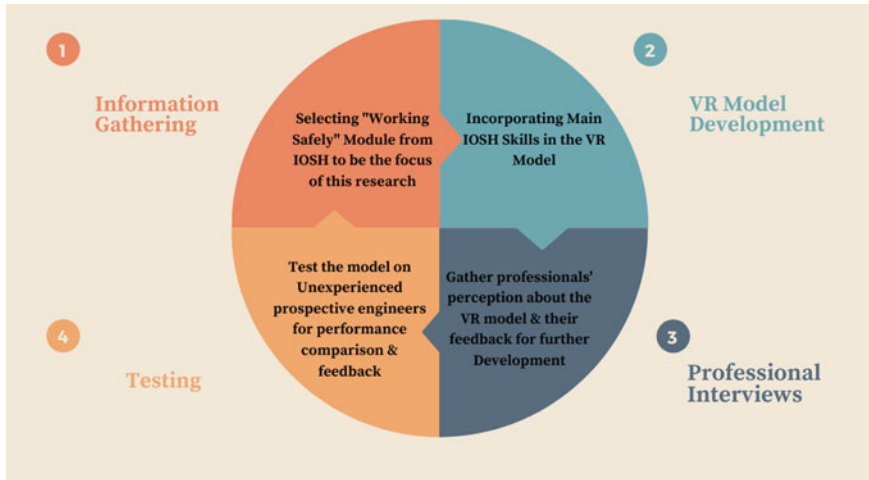
Pedro et al. (2015) developed a virtual construction safety education system to be integrated into construction methods education. The model was then tested on third year building construction class along with concrete works chapter related to construction methods and materials. It was concluded that integration of safety education into construction technology education led to an improvement in the safety knowledge and hazard identification.

Li et al. [14] developed a multi-user virtual training system for tower crane dismantling. The system had the following three main attributes: time sequence, location, and responsibility. The system was then tested on 30 construction workers who were divided into three main groups (A, B, and C) based on the level of experience. Group A did not have experience, was trained based on the proposed system, group B had experience in tower crane related operations, was trained based on the proposed system, and group C had experience but was trained traditionally in tower dismantling. It was concluded that group B performance was the best among all participants. The proposed system added valuable experience to the participants with risk free environment and affordable cost compared to mockup environment.

Therefore, based on the literature review, it is evident that VR has a promising role in enhancing overall performance, understanding and hands-on experience related to different tasks related to construction safety trainings. However, it was not yet used on a large scale in standardized training centers side by side with the conventional techniques. Therefore, in this paper, a VR model was developed to complement IOSH safety trainings to enhance visualization, hands-on experience and knowledge retention. Professionals' interviews were done to validate the model overall interface and logic. In addition, experimental testing of the model was done on senior construction engineering students to test its significance for entry-level engineers.

### 3 Research Methodology

In order to achieve the objective of incorporating VR into the IOSH trainings, first the different trainings provided by IOSH were examined. The most relevant and commonly used training was “working safely,” besides the “managing safely” training. As the “managing safely” training contains more theoretical information that needs to be learned by lecture-based techniques, the focus was on “working safely” training. Different sections of the training were studied carefully, and the main skills that should be attained at the end of the training were noticed. Subsequently, the exams provided by IOSH at the end of the trainings in order to get the certificate were obtained, i.e., information was used to prepare an overview of the main topics discussed in IOSH trainings in general, and the main skills needed to be fostered in the trainees in particular. The main skills were related to hazard identification, hazard control, hazard mitigation, and fatality processing. Therefore, the VR model was developed so that the hazard identification is manifested, choosing the right PPE that is suitable for the assigned task is solidified and avoiding sources of hazards is enhanced. Comparisons were made between the available VR headsets available in the market. These include HTC Vive, Samsung Gear, Oculus Rift, Oculus Go, Lenovo Mirage, and others. Based on the price, mobility, specifications, and quality, the Oculus Quest 2 headset were the chosen one. This headset is a standalone headset that does not require any other supporting devices and comes along with its own handheld controllers through which trainees would be able to interact with the created virtual environments. After that, the model is validated by examining it to ensure it is free from any bugs or glitches. Then, it was introduced to professional safety inspectors to gather their feedback and perception of the idea of incorporating VR into the IOSH training and whether it could replace the conventional methods of trainings or not. Their feedback would be used to further develop and manifest the developed model. Then, the model was tested on senior students on construction engineering as part of their field trips safety induction trainings—that used to be following IOSH requirements—to examine the effect of the developed model in enhancing safety knowledge for non-experienced engineers. After that, their feedback about their experience was collected. Finally, the conclusions from professionals and students will be roots for improving the proposed model and for future creation of a framework to incorporate VR into the most recognized internationally safety trainings. Figure 1 summarizes the main steps adopted in this research, it is cyclic steps as results of each step is used to better enhance the following step. Then, the conclusions would be used for further development of the VR model. Therefore, the process is iterative for better development of the model.



**Fig. 1** Research methodology steps

## 4 Model Development

The model was built in Unity®, a game engine that supports VR games. It uses C# coding language. The development of the model also considered having a realistic simulation of construction sites by incorporating construction noise, advanced lighting, and clear visual images. Full user interaction is attained by proper embodiment of the user where his actual movements are detected and imitated by the virtual avatar. This interaction will be limited to hand and head movements to be feasible for use in confined places and to limit dizziness. For building the model, the following hazards from “working safely” IOSH module were modeled in the VR model, falling from height, welding, slipping or tripping, falling in excavated area and falling objects, in addition to general site safety.

### 4.1 Model Characteristics

The model was developed such that two main characteristics are achieved which are fully immersion and realistic sense of construction sites and hazards.

Fully immersive experience is ensured based on the following main immersion techniques identified by [2]:

- (a) Tactical Immersion: which is the moment by moment immersion act of the game playing. In order to achieve that, the model is designed and validated to ensure flawless user interface with almost zero speed struggles or control disruptions.

- (b) Strategic Immersion: which allow the user to observe, make decisions on the needed actions to reach victory. This is achieved by building logic scenario with clear path and directions to ensure straightforwardness and no randomness.
- (c) Narrative Immersion: is attained based on the scenario effectiveness. This is ensured by formulating the scenario in an interesting way, so that the trainee is encouraged to invest his time and finish the game in a satisfactorily manner.

Moreover, the trainees will make decisions while being immersed in the model. Therefore, based on different possible decisions taken by each trainee, different endings will be evident which accordingly would enhance the hands-on experience for each one.

In order to attain consciousness state and realistic sense of construction sites and hazards while in the virtual reality, the following parameters were considered in developing the model:

1. Embodiment: This allows for trainees physical interactions with the virtual model in the safety-training module. Embodiment could be achieved by having an avatar virtual body whose motions align with the trainee real ones [21]. For the user input, it has been approved that hand motion could allow for sufficient degree of freedom and interaction for majority of applications [8]. Therefore, the trainee movement inside the VR model was limited to hand and head movement instead of full body movement. This would allow for using the model in any location regardless its size and ensure minimum sickness felt by the trainees.
2. Presence: Clear vivid images and no visual anomalies were assured in the model to ensure realistic sense of presence in the VR training.
3. Sensory Feedback: As mentioned by Gobbetti and Scateni [8], “Our sense of physical reality is a construction derived from the symbolic, geometric, and dynamic information directly presented to our senses.” Therefore, sound and vision perceptions were the main sensory feedback that were attained in the model by incorporating different sounds evident in construction sites, in addition to realistic and authentic site surroundings.

## **4.2 Model Formulation**

The model consists of a large construction site which has different buildings and activities. The user will need to interact with the surrounding environment, so the model will process according to the user inputs. The user will be asked to do a certain task, and his behavior will be assessed based on picking the right and suitable PPE, avoiding hazards and completing the task in a satisfactorily manner. The user may be prompt with another sudden task while doing his initial one in order to assess his reactions and behavior toward sudden situations and his ability to move from one activity to another in a safe way.

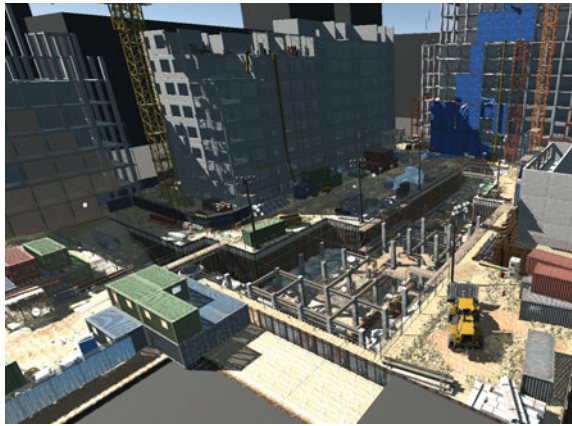


### 4.3 Model Components

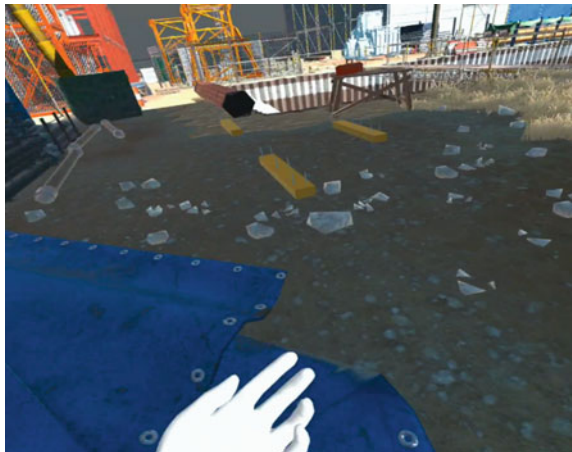
The model has the following components:

- (1) Site layout: Fig. 2 shows the site layout of the model. The site consists of different buildings and activities. The player (first person view showed in Fig. 3) will start from the personal protective equipment room (shown in Fig. 4) and will have to move to another building to perform the task assigned to him. The site has signs to guide the user to the main locations he may need to go while doing his task.
- (2) Potential hazards: The player is exposed to different kinds of hazards while moving in the site. If he picked up the right PPE at the beginning of the game, he could avoid the consequences of these hazards. Such hazards are falling in

**Fig. 2** Site overview (the developed model)



**Fig. 3** First person view of the site (the developed model)





**Fig. 4** PPE room (the developed model)

the excavated area, falling objects from any of the buildings, hurt by nails in the wooden planks, falling from heights, hurt by welding equipment and slip or tip. Figures 5, 6, and 7 show examples of these hazards from the model.

### (3) Potential Activities

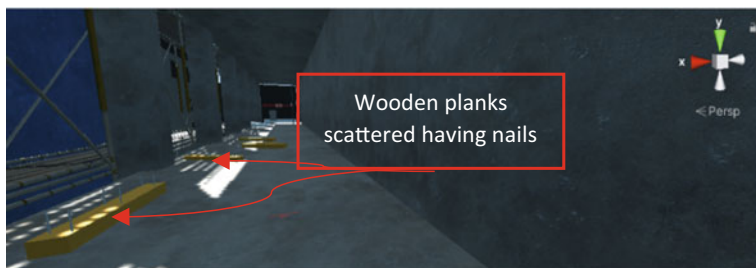
- (a) Installing bricks: The user needs to move the bricks from their storage area to the 7th floor where he will use them. He needs to ensure that they are well secured in the material conveyor that transports them upwards, so that they will not fall down and risk his or anybody's life. Figure 8 shows the location where bricks should be installed at the 7th floor.



**Fig. 5** Falling in excavated area hazard (the developed model)



**Fig. 6** Slip or tip hazard (the developed model)



**Fig. 7** Hurt by nails hazard (the developed model)

**Fig. 8** Installing bricks at 7th floor (the developed model)

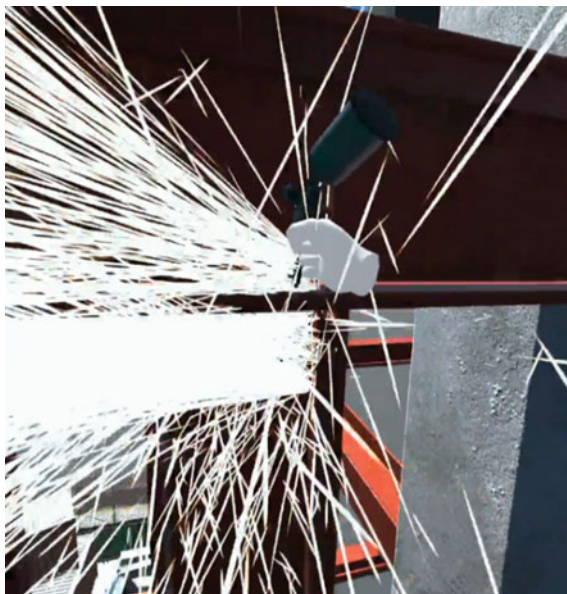


- (b) Being prompt with welding activity: While the user is installing bricks in the 7th floor, another worker will suddenly ask for his help in a welding activity in the 4th floor. The user will need to go first to the PPE room to pick up the welding face shield. If he ignores this step, he will be subjected to the harms related to welding. Figure 9 shows the welding activity.

## 5 Model Validation

In order to collect professionals' perception about incorporating the developed model into the IOSH safety trainings, interviews were done with 15 safety inspectors who have wide in site experience ranging from 5 years till 22 years of experience. A presentation was done to them to introduce the concept of the VR and the developed

**Fig. 9** Welding activity (the developed model)



model. Then, they were allowed to test the VR model using Oculus headset. After that, they were asked questions related to what they had experienced in addition to general questions, e.g.,

- (a) Their overall opinion about the developed model
- (b) Can the developed model replace the conventional training methods?
- (c) Potential areas of improvement in the model.

From the results of these interviews, professionals interviewed agreed that the concept of VR incorporation in safety trainings is beneficial and would have great impact in improving visualization and hands-on experience. They also approved that the developed model has high capabilities and is best suited for safety inductions and trainings that abide by IOSH standards for junior engineers with minimal experience because of its straightforwardness and simplicity. In addition, they said that the developed VR model would be of great value when used side by side with lecture-based trainings as a supplement source for better visualization and for testing trainees' overall understanding. However, they agreed that it cannot totally replace the conventional methods as human to human interaction is important in the educational process. After that, they proposed some ideas for further improvement of the model to address wider categories and not to be limited to a certain category of users. These aspects are: first step, the targeted trainees experience should be decided upon, if they are professional safety inspectors with more than 5 years experience, workers (Skilled or Non-skilled), or general safety induction for non-experienced personnel.

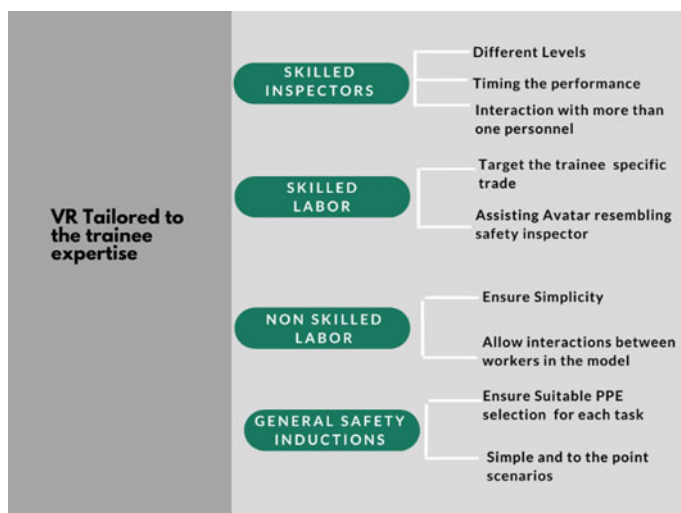
Then, according to each category, there are different aspects that should be adopted in the model to guarantee optimum performance. These parameters are explained next:

- (a) For Skilled Inspectors, the model should have levels where in each level, the degree of difficulty increases. The task for each level should be clearly identified in the beginning and with going up in levels, the inspector would have to interact with more than one personnel in the VR model. Then, with the increase in level difficulty, time of performing the task should be another aspect in assessing his performance. For each right decision the inspector makes, his score should increase. On the other hand, with wrong acts, his score decreases and in case the wrong act would cause a fatality, he should get game over.
- (b) For Skilled Workers, the model should target the specific trade that the worker would perform on site. An assisting avatar having the role of safety investigator should be evident on the VR model to give the worker guidance about the main steps to perform his task safely. Then, the worker should continue playing the game, so his performance is assessed accordingly.
- (c) For Non-skilled Workers, the model should target the tasks associated to them in a simple and straightforward manner. Interactions between workers should be evident in the model, as usually non-skilled workers have to cooperate to finalize their assigned tasks.
- (d) For General Safety Induction, the interviewed professionals see that the proposed model best suit general safety induction that is following IOSH guidelines because of the general nature of safety measurements included in the model; in addition, the model is direct, simple, and easy to move around.

Figure 10 summarizes the most crucial aspects that should be incorporated into the model according to the professionals' feedback.

## 6 Model Testing

As the interviewed professionals recommended that the developed model should be used in general safety inductions that abide by IOSH standards, the model was tested on 24 junior construction engineers who were enrolled into the course construction technology and equipment. These students used to have a field trip every week to a construction site where they usually took a safety induction first that is abiding by IOSH standards. Therefore, students were grouped into two main groups where the first group is 12 students compromising the control sample who took only conventional safety training and the other group is 12 students who took VR safety training. Then, the performance of both groups was compared by giving them the same test. The test followed the Institution of Occupational Safety and Health (IOSH) standards of working safely. The test was made up of 20 questions, each question was graded out of one point giving a maximum total of 20 points. The test presented general questions that cover the main hazard areas in IOSH working safely module.



**Fig. 10** Professionals feedback

The hazard areas would include scenarios like, excavated areas, falling objects, and falling from heights. The questions format followed IOSH testing format where the test taker is presented with a photo, then he is asked to identify the hazard he is seeing, how this hazard can cause him/her harm and what is the most suitable way of controlling the hazard. The test presented six hazard scenarios where students had to correctly choose the type of hazard that corresponds to what they are seeing in the picture. The picture would present a situation where there is a given work activity that is being done, after that the student chooses the appropriate answer from the presented answers. He/she then had to write how this hazard can cause harm to any given person in that situation and what are the best ways to control/mitigate this hazard. For example, in one of the scenarios the student is presented with a picture with someone doing welding work. The student must correctly identify that the hazard is the worker can get hurt by welding. Then, the student must mention that this activity can harm the worker's eyes or cause skin burns. Finally, the student mentions that this risk can be mitigated if the worker wears the appropriate PPE like welding goggles and gloves. The five other scenarios followed the same format. The last two questions presented the student with a work scenario with certain conditions, and he is asked what the proper PPE that is needed to safely perform this task. The results from both groups of students were then compared using the t-test to evaluate the significance of the proposed model in improving basic safety topics. The research hypothesis is that the developed VR model improves students' performance more than the traditional training methods.

After that, the students who took the VR test took a questionnaire to ask them about their perception of the VR technology versus the conventional trainings methods. The questionnaire entailed questions related to the benefits of using the VR model in



**Table 1** Summary of the results of the effectiveness of the developed VR model

Attribute	Control group: traditional training	Test group: VR training
Average score (out of 20)	16.16	18
Standard deviation	2.22	1.27
<i>P</i> -value of <i>t</i> -test	0.024 (< 0.05 indicating statistically significant difference between the group averages)	

different aspects of safety trainings, the extent to which it could substitute traditional safety trainings. Moreover, it contains questions related to the disadvantages that the trainees experienced while using the model, and their degree of satisfaction with the model, in addition whether they suffer from any sickness symptoms during the VR testing or not.

## 7 Results and Analysis

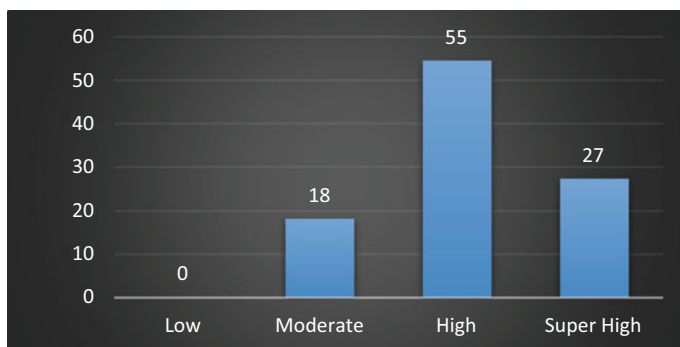
### 7.1 Testing Results and Analysis

The test results showed that there are noticeable advantages for the students who took the VR training against the students who got the traditional training. The students who received the VR training were able to achieve an average score of 18 points out of 20 with a standard deviation of 1.27, while the students who got the traditional training got an average score of 16.16 out of 20 with a standard deviation of 2.22. Comparing the safety performance of VR trained versus traditionally trained participants using the *t*-test resulted in a *p*-value = 0.024 (< 0.05), indicating a statistically significant difference between both groups. Table 1 summarizes the aforementioned results. This in fact validates the efficiency of the proposed VR model in enhancing safety and promotes its use as a standard tool for complementing IOSH safety training.

### 7.2 Questionnaire Results and Analysis

The first question was on the extent to which the trainees thought it was likely for the trainings given using the VR model to substitute traditional safety trainings including IOSH. Approximately 55% of respondents saw that it was highly possible, 27% of respondents saw that it was super highly possible, and 18% saw that it was moderately possible. Figure 11 summarizes the aforementioned results.

The trainees were also asked about the different aspects of safety trainings that they viewed as most suitable to be carried out using the VR model. For hazard



**Fig. 11** Percentages of possibility of VR substituting traditional trainings

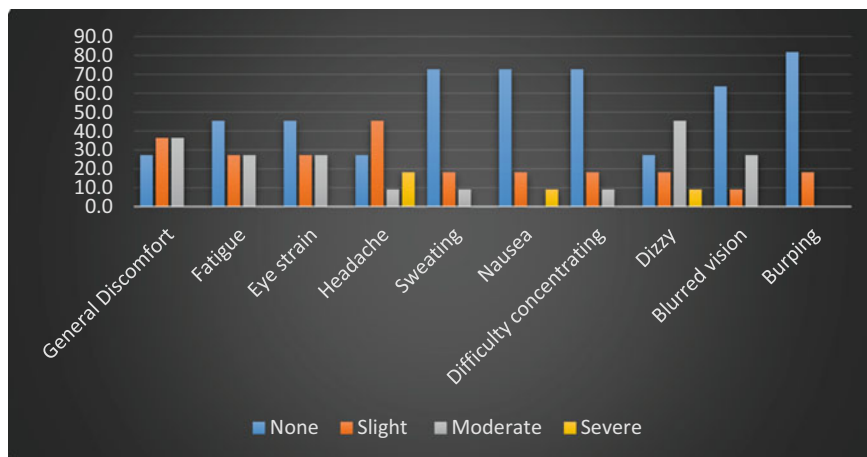
identification aspect, a majority of 73% of the respondents viewed the model as highly beneficial. For hazards control aspect, a total of approximately 55% of the respondents viewed it as highly and very highly beneficial. For dealing with hazards aspect, approximately 55% of the respondents viewed the model as highly and very highly beneficial.

When compared to traditional safety trainings in terms of the learning experience and education quality, 36% and 18% of the respondents viewed the VR model as better than and much better than traditional safety trainings, respectively. Approximately 50% of the respondents viewed the model as better than traditional safety trainings in terms of amount of information delivered, focus during content delivery and trainee participation. A majority of 64% viewed the VR experience as much better in terms of visualization.

The participants were also asked to identify the extent to which they experienced sickness symptoms related to using the VR model. Those symptoms included general discomfort, fatigue, eye strain, headache, sweating, nausea, difficulty concentrating, dizziness, blurred vision, and burping. A majority of 73% declared that they experienced no sweating, nausea, or difficulty concentrating. Approximately 82% and 64% declared that they have not experienced blurred vision and burping, respectively. Only three of the symptoms were experienced in a severe manner by some of the participants which were headache (18%), nausea (9%) and dizziness (9%). Figure 12 summarizes percentages for different sickness symptoms.

When asked if the participants would use the VR model frequently, 73% agreed that they would. When asked about simplicity and clarity of the model, about 91% of the participants responded that the model was not complex, and 55% responded that the model was easy to use. Approximately 82% of the participants viewed the functions of the prototype as well-integrated, consistent, and not cumbersome to use, while 73% viewed the VR model as easy to learn. Finally, approximately 46% of the participants were confident using the model; however, 64% of the participants thought that they would need assistance in order to use it. Finally, the overall satisfaction rating of the participants was 4.6 (out of 5).





**Fig. 12** Percentages of different sickness symptoms while using VR

## 8 Conclusions

In conclusion, virtual reality has proved its ability to foster visualization and hands-on experience, as seen in previous research. Therefore, the objective of this paper is to incorporate VR into IOSH standard working safely module training. The VR model was built according to IOSH main topics, then interviews with professionals were conducted to gather their feedback about the model. From these interviews, it was concluded that it is important to know the level of expertise of the targeted trainees in order to customize the model to the needed skills that should be fostered at the end of the training. In addition, they recommended that the developed model would be of great benefit if used for general safety inductions following IOSH standards and for non-experienced personnel, mostly due to its simplicity and clarity. In addition, the model contains the main skills needed for general site safety. Subsequently, the model was tested on non-experienced engineering students where it was proved significantly effective in enhancing their overall performance. The model can be further developed to deal with hazards after they occurred. In addition, further development could include interaction between different trainees in the virtual reality to observe how their different reactions affect each other.

**Acknowledgements** The authors are grateful to the support provided by the American University in Cairo in terms of funding the research under Grant Number SSE-CENG-I.A.-FY21-FY22-RG(2-20)-2020-Mar-01-17-09-27, provided by the office of the associate provost for research and creativity.

## References

1. Abulrub AHG, Attridge AN, Williams MA (2011, April) Virtual reality in engineering education: the future of creative learning. In: 2011 IEEE global engineering education conference (EDUCON). IEEE, pp 751–757
2. Adams E (2021) Postmodernism and the three types of immersion. Designersnotebook.com. Retrieved 30 Sep 2021 from [http://designersnotebook.com/Columns/063\\_Postmodernism/063\\_postmodernism.htm](http://designersnotebook.com/Columns/063_Postmodernism/063_postmodernism.htm)
3. Al-Qahtani A, Higgins S (2012) Effects of traditional, blended and e-learning on students' achievement in higher education. *J Comput Assist Learn* 29(3):220–234. <https://doi.org/10.1111/j.1365-2729.2012.00490.x>
4. Bourhim E, Cherkaoui A (2020) Efficacy of virtual reality for studying people's pre-evacuation behavior under fire. *Int J Hum Comput Stud* 142:102484. <https://doi.org/10.1016/j.ijhcs.2020.102484>
5. Chittaro L, Corbett C, McLean G, Zangrando N (2018) Safety knowledge transfer through mobile virtual reality: a study of aviation life preserver donning. *Saf Sci* 102:159–168. <https://doi.org/10.1016/j.ssci.2017.10.012>
6. Clifford R, Engelbrecht H, Jung S, Oliver H, Billingham M, Lindeman R, Hoermann S (2020) Aerial firefighter radio communication performance in a virtual training system: radio communication disruptions simulated in VR for air attack supervision. *Vis Comput* 37(1):63–76. <https://doi.org/10.1007/s00371-020-01816-6>
7. Geisari M, Esmaili B (2018) PARS: Using augmented panoramas of reality for construction safety trainings. The Center for Construction Research and Training, University of Florida, George Mason University
8. Gobetti E, Scateni R (2020) Virtual Reality: Past, Present, and Future. Sardinia Cagliari, Center for Advanced Studies, Research and Development, Italy
9. Goulding J, Nadim W, Petridis P, Alshawi M (2012) Construction industry offsite production: a virtual reality interactive training environment prototype. *Adv Eng Inform* 26(1):103–116. <https://doi.org/10.1016/j.aei.2011.09.004>
10. Lingard H (2013) Occupational health and safety in the construction industry. *Constr Manag Econ* 31(6):505–514
11. IOSH Website. Retrieved Feb 2022. <https://iosh.com/>
12. Joshi S et al. (2020) Implementing virtual reality technology for safety training in the precast/prestressed concrete industry. Elsevier Ltd
13. Le D, Le C, Tromp J, Nguyen N (2018) Emerging technologies for health and medicine. Wiley
14. Li H, Chan G, Skitmore M (2012) Multiuser virtual safety training system for tower crane dismantlement. *J Comput Civil Eng ASCE*. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000170](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000170)
15. Li MR (2018) Virtual reality and construction safety. Hong Kong Shu Yan University
16. Nykanen M et al. (2020) Implementing and evaluating novel safety training methods for construction sector workers: results of a randomized controlled trial. *J Saf Res*. Elsevier Ltd
17. Oberhauser M, Dreyer D, Braunstingl R, Koglbauer I (2018) What's real about virtual reality flight simulation? *Aviat Psychol Appl Hum Fact* 8(1):22–34. <https://doi.org/10.1027/2192-0923/a000134>
18. Pensieri C, Pennacchini M (2014) Overview: virtual reality in medicine. *J Virtual Worlds Res* 7(1). <https://doi.org/10.4101/jvwr.v7i1.6364>
19. Riener R, Harders M (2012) Virtual reality in medicine. Springer, London
20. Song H, Kim T, Kim J, Ahn D, Kang Y (2021) Effectiveness of VR crane training with a head-mounted display: double mediation of presence and perceived usefulness. *Autom Constr* 122:1–11
21. VR Training Technology. RTE global. Retrieved 30 Sep 2021. <https://rte.global/vr-training/>

# Semantic Segmentation of Synthetic Images into Building Components for Automated Quality Assurance



H. X. Zhang, L. Huang, W. Cai, and Z. Zou

**Abstract** Quality assurance (QA) plays an essential role in the construction project life cycle. During the construction phase, discrepancies between as-built structures and as-designed models can lead to schedule delays and cost overruns. Currently, QA for buildings is primarily conducted by inspectors physically touring the building to visually inspect and manually measure discrepancies between the design model and the finished structure. This manual approach is time-consuming, costly, and error prone. In this study, we proposed a vision-based approach toward automated QA using images collected via virtual cameras in a game engine. Specifically, our approach aimed to address the problem of the lack of large-scale labeled open datasets for training reliable machine learning models for the task of semantic segmentation of building components (i.e., labeling of each pixel to a specific class of building component). The approach leveraged Building Information Modeling (BIM) authoring tools and a game engine to automatically generate images of virtual buildings with pixel-wise labels. Given the labeled images, a convolutional neural network (CNN)-based model can be implemented for accurate segmentation of the images. To validate the approach, we used one building information model as the testbed. In total, 20,700 images (18,000 for training, 2200 for validation and 500 for testing) were generated from the BIM. Performance of the CNN segmentation model was measured by the mean Intersection over Union (MIoU), which achieved 0.89. The result is significant since it rivals the current state-of-the-art from the Architecture Engineering and Construction (AEC) domain. The approach proposed in this study lays a concrete step toward automated QA, where inspectors can leverage the trained CNN model to automatically label images collected onsite during or after construction to avoid labor-intensive manual inspections.

**Keywords** Quality assurance · Building components · Semantic segmentation · Synthetic images

---

H. X. Zhang · L. Huang · W. Cai · Z. Zou (✉)

Department of Civil Engineering, Faculty of Applied Science, The University of British Columbia, Vancouver, Canada

e-mail: [zhengbo@civil.ubc.ca](mailto:zhengbo@civil.ubc.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_14](https://doi.org/10.1007/978-3-031-34593-7_14)

# 1 Introduction

Quality assurance (QA) is essential throughout the life cycle of a construction project to minimize the discrepancies between as-built structures and as-designed models. Such discrepancies can lead to schedule delays and cost overruns caused by rework [9, 10], which can constitute more than 5% of the total construction cost [6]. Currently, QA is labor intensive, requiring inspectors to compare design drawings with built structures from manual measurements [21]. The use of remote-sensing instruments such as total stations introduced possibilities for automation. However, the current practice is still time-consuming and error prone since inspectors need to manually compare the collected measurements with 2D shop drawings or BIM.

The development of vision-based sensors (e.g., laser scanners, RGB-D sensors) and advanced computer vision algorithms (e.g., convolutional neural networks) created opportunities for quick and automated QA. Research studies in the Architecture Engineering and Construction (AEC) domain explored possibilities of using point clouds and RGB images as input to help QA through as-built model generation [1], construction progress monitoring [12], discrepancy detection [20] and surface defect detection [14]. The specific inspection task varied in these studies, but one foundational step remained common—the semantic segmentation of the building components (e.g., conduct pixel-by-pixel labeling of a construction site image into construction equipment, worker, etc.), because it constitutes the first step toward allowing an automated algorithm to understand the types of building objects in an input image or point cloud. This understanding is crucial for next steps of identifying discrepancies between the as-built structure and the design model. Therefore, in this study, we focused on improving the performance of automated semantic segmentation of building components from imagery inputs.

We currently face two challenges in achieving automated semantic segmentation of images and point clouds for construction projects. First, it is challenging to collect large-scale real-world data of images or point clouds in buildings and/or on construction sites, which is essential for training high-performing algorithms for semantic segmentation. The more pressing challenge, on the other hand, lies in the labeling of the collected datasets. Currently, human labelers draw contours of objects and categorize them into a variety of classes from the images, which can be time-consuming and expensive. To address these challenges, researchers in the AEC domain have been keen on leveraging synthetic datasets of images and point clouds [16, 23]. The advantage of synthetic data is the flexibility and efficiency in generating images or point clouds. The generation process can be executed rapidly while allowing for diversity in the dataset by using a large number of building information models with different design purposes [16].

In this study, we combined Building Information Modeling (BIM) and a game engine to simultaneously generate large datasets of synthetic images and their corresponding labels. The proposed approach started with a BIM authoring tool (i.e., Revit), which provided the model geometry and building information. Next, we imported the BIM model into a game engine (i.e., Unreal Engine 4), which was

connected to Simulink in MATLAB to run simulations for collecting virtual images and corresponding labels from different locations and angles. Finally, we applied a convolutional neural network (CNN) model (i.e., U-Net) to conduct semantic segmentation of the images generated by the game engine. The performance of the CNN model was measured by the mean Intersection over Union (MIoU). We tested the effectiveness of this method using the BIM of a residential building. The main contribution of the study is the proposed approach which allows for automated generation and labeling of synthetic images of building components.

## 2 Background

Computer vision-based QA has been applied to surface defect detection, geometric discrepancy detection and relationship reasoning [17, 18]. Defect detection has mainly been implemented in assessing the condition of horizontal civil infrastructures, including reinforced concrete bridges, underground concrete pipes, asphalt pavements and vertical structures such as steel frames, reinforced concrete columns, and exterior masonry walls [11]. For example, an integrated vision-based system was developed to automatically inspect the quality of slate slabs for defects and ensure the adherence to construction guidelines [7]. Other than defect detection, studies had also experimented with computer vision algorithms for detecting dimensional, geometrical, and positional variations of building components for QA. For example, a vision-based approach was proposed to inspect the alignment of tile installation by analyzing the geometric characteristics of the tile finishes [11]. Despite existing applications of vision-based QA, a recent survey [17, 18] demonstrated that research and applications of this area are still limited.

A significant bottleneck for developing vision-based QA for AEC, cited by recent reviews [17, 18, 3, 27], is the lack of open training and testing data. Noticeable efforts had been made in creating open data in the community. For example, the first workshop of computer vision in the built environment set out to provide an open dataset for the task of scan-to-BIM, where point clouds were given as inputs and the output was an automatically created building information model [4, 25]. However, as pointed out in previous studies [3, 27], collecting large-scale real-world data is costly and time-consuming, impeding the validation progress of vision-based algorithms in AEC.

One approach to solving the lack of open data is to use synthetic data. Synthetic data generation has seen success in several areas, including autonomous driving, object segmentation, and recognition [2, 8, 26]. For instance, [22] created an open dataset of synthetic urban environments, aiming to train deep neural network (DNN) algorithms for segmenting objects from a collection of diverse urban images for the purpose of autonomous driving. Eleven classes of objects were included in the virtual environments created using a game engine (i.e., Unity 3D). Advancement has also been made in synthetic data generation for indoor environments. Notable efforts from [5] showed the possibility of creating synthetic images for indoor scenes by

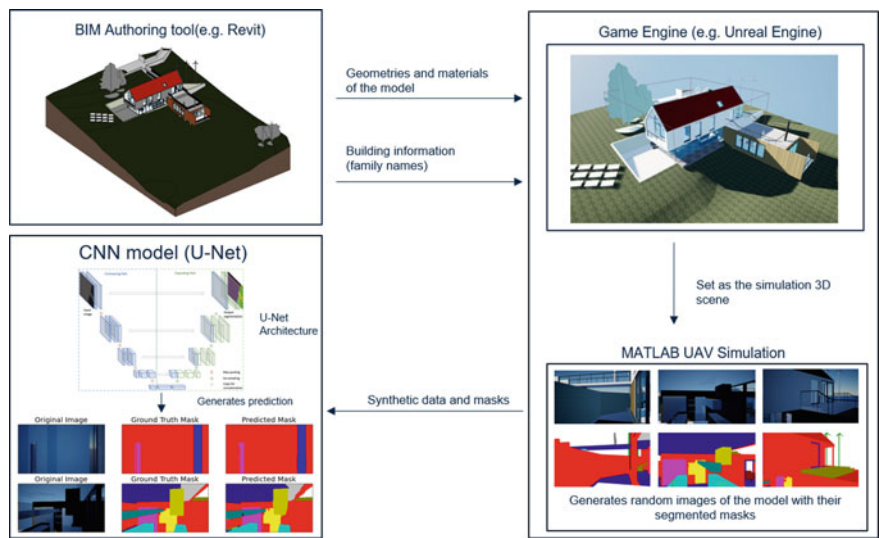
randomly placing a virtual camera in 3D models. Despite significant research strides, a main research gap exists, since the connection between BIM and synthetic data has rarely been examined. Even when BIM was used as the base model to generate synthetic data [16], it only provided 3D geometric input, leaving out other valuable information from BIM. Motivated by previous research and the research gap, in this study, we proposed an approach to generate diverse synthetic images from BIM for QA by combining BIM with a game engine and computer vision algorithms.

3 Methodology

In this study, we incorporated a BIM authoring tool for providing the geometric and building information, a game engine connected to MATLAB for generating synthetic images, and a CNN-based semantic segmentation model trained from the generated images to create pixel-level semantic segmentation of building components. The main components of the proposed approach are shown in Fig. 1.

3.1 Exporting BIM into a Game Engine

In the first step of our approach, we exported the building information model from a BIM authoring tool (i.e., Revit), which served as the 3D geometry and building



**Fig. 1** Main components of the approach for semantic segmentation of images into building components

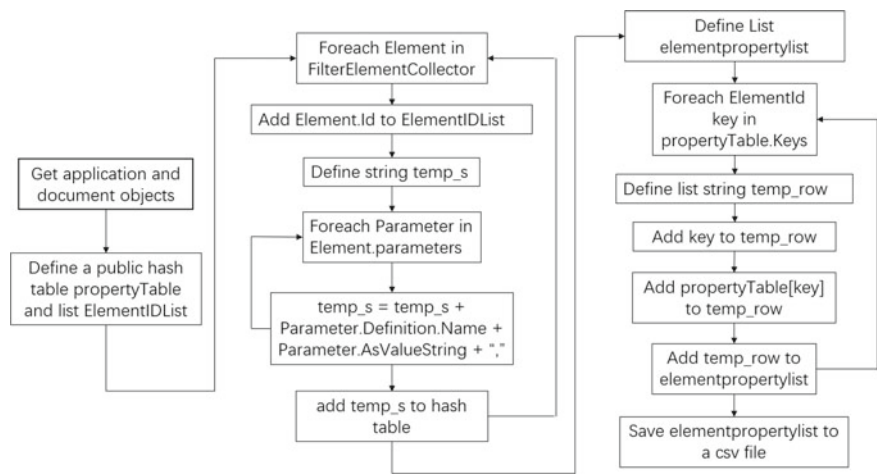
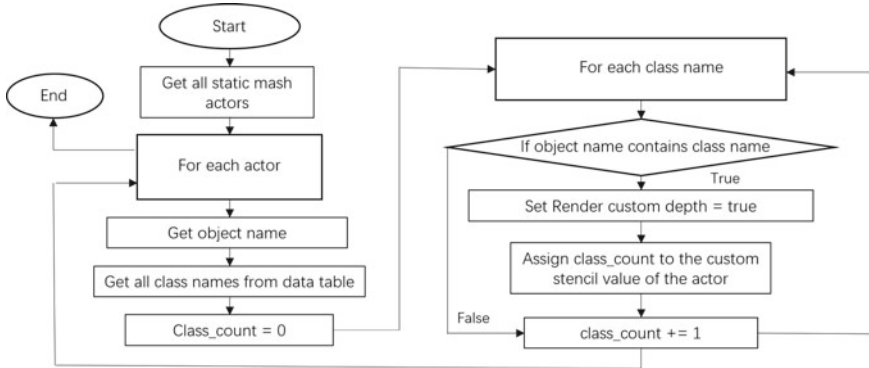


Fig. 2 Flowchart of the custom plugin in Revit for retrieving building information

information contributor. Once the BIM file has been exported, we used the Datasmith plugin to import the model into Unreal Engine. Next, we created a custom plugin in Revit to find all building components in the model. The details can be referred to in Fig. 2. Essentially, the custom plugin looped through all objects in a Revit model and retrieved all objects’ family names and their properties to store in a HashTable (i.e., a data structure storing information in key–value pairs), and then stored them in a csv file for later use.

3.2 Labeling Building Components in Game Engine

An important step in this implementation was to assign a unique depth value to building components of the same class in Unreal Engine. To achieve this, we created a level blueprint, which was a custom script in Unreal Engine that controlled objects in the scene. To link the building components’ information in Unreal Engine to BIM, we imported the csv file generated in Revit into the Unreal engine as a data table (i.e., a data structure that contains the building component family type). Next, in the level blueprint, we created two loops to match names of every object in the Unreal Engine scene (i.e., static mesh actors in the flowchart) with the class name in the csv file and assign objects in the same class a unique depth value (as seen in Fig. 3). This step ensured the virtual camera generated in the simulation would label objects with the same depth value using the same color.



**Fig. 3** Flowchart of the level blueprint assigning unique depth values to building components

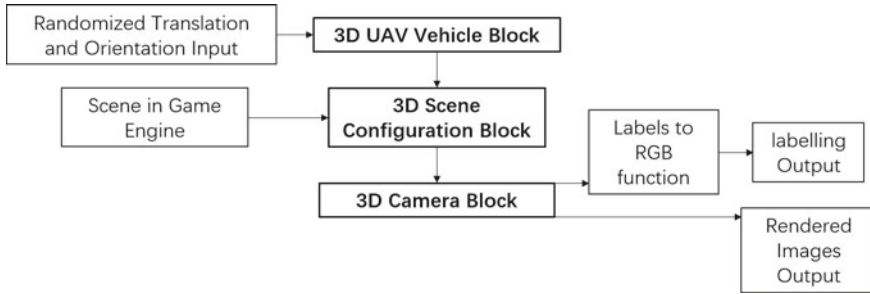
### 3.3 Generating Synthetic Images in Game Engine

To generate synthetic images of building components, we connected a simulation in MATLAB Simulink to the Unreal Engine. The simulation was designed to generate randomly placed cameras in the game engine and take a “photo” of the building components. We built the simulation with three blocks: 3D Scene Configuration, 3D Camera, and 3D Unmanned Aerial Vehicle (UAV) (as shown in Fig. 4). In the 3D Scene Configuration block, we established the connection between Simulink in MATLAB and Unreal Engine by setting the scene in Unreal as a rendering environment. The 3D UAV block was then used to set the camera location. Thereafter, we used the Random Number block in Simulink to randomize translation and orientation of the camera to get images from different locations and angles. The Random Number block provided a normally (Gaussian) distributed random signal, and we set the sample time to 0.10 s, which meant the camera changed position and angle every 0.10 s. The goal of this randomization was to generate enough variations in the dataset so that the real world can be represented as one of the variations [24]. Finally, by using the 3D Camera block, we can export the images of building components and their labels in the RGB format.

### 3.4 Semantic Segmentation of Synthetic Images into Building Components Using U-Net

We used a CNN-based algorithm (i.e., U-Net) to segment the synthetic images into building components, which was proven effective by previous studies [21]. A typical CNN architecture includes an input layer, alternating convolutional layers, pooling layers, fully connected layers, and an output layer. Simply put, a convolutional layer





**Fig. 4** Setup of the simulink simulation

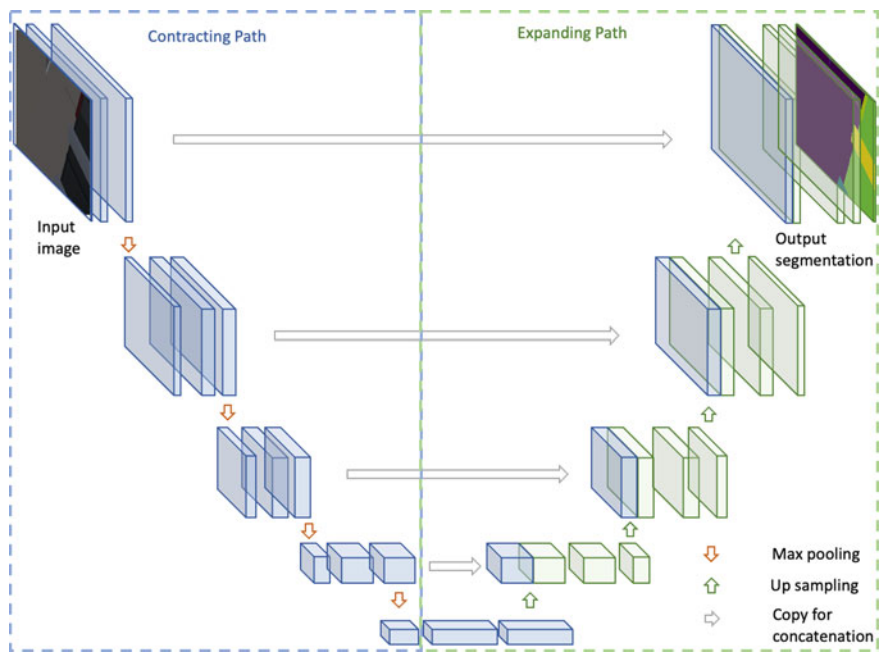
applies a set of filters to the input image to create a set of feature maps that summarizes all the pixels into a single value as input for the next layer. In this study, we adopted U-Net [21] as the underlying CNN architecture. U-Net is based on the fully convolutional network [15] and required fewer data to achieve a high level of performance. The architecture of U-Net is shown in Fig. 5 below. The main idea of the fully convolutional network is to supplement a contracting network by successive layers, where pooling operators are replaced by upsampling operators [21]. In U-Net, the upsampling part is modified to have a large number of feature channels, which allows the network to propagate context information to higher-resolution layers. The resulting architecture is almost symmetric consisted of a contracting path and an expansive path.

The contracting path consists of four successive blocks. Each block contains two  $3 \times 3$  convolutions (unpadded convolutions), followed by a rectified linear unit (ReLU) and a  $2 \times 2$  max pooling operation with stride 2 for downsampling. During the contraction, the spatial information is reduced while feature information is increased. On the other hand, the expansive path decreases the number of layers and creates a high-resolution segmentation map as the output.

### 3.5 Loss Function and Metrics Used for Semantic Segmentation of the Images

During training of the U-Net model, we used multi-class (categorical) cross-entropy as the loss function, and mean IoU as the performance metric. Cross-entropy is a measure of the difference between two probability distributions for two random variables. Because we have more than two classes of building components, we used multi-class cross-entropy loss, which is a combination of the softmax activation and the cross-entropy loss, as seen in the equations below.

$$l(x, y) = \sum_{n=1}^N \frac{1}{\sum_{n=1}^N w_{y_n} * 1} * \left( -w_{y_n} \log \frac{e^{x_{n,y_n}}}{\sum_{c=1}^C e^{x_{n,c}}} \right) \quad (1)$$



**Fig. 5** U-Net architecture

In Eq. (1),  $x$  is the input (prediction),  $y$  is the target (true label),  $w$  is the weight,  $C$  is the number of classes and  $N$  is the number of samples in a batch. In our implementation, we set weights equal to 1 for all classes, assuming all classes are equally important. By using cross-entropy loss, the model would train a CNN to output a probability distribution over the  $C$  classes for each image.

For the performance metric, we used the mean intersection over union (MIoU), which is an effective measurement for the performance of the segmentation model. MIoU can be defined as the overlapped section of the prediction and the true label divided by the union of the prediction and the true label (as seen in Fig. 6). This calculation was applied across all classes, and the mean of the IoU scores was be taken as the final output.

**Fig. 6** An intuitive representation of the IoU score

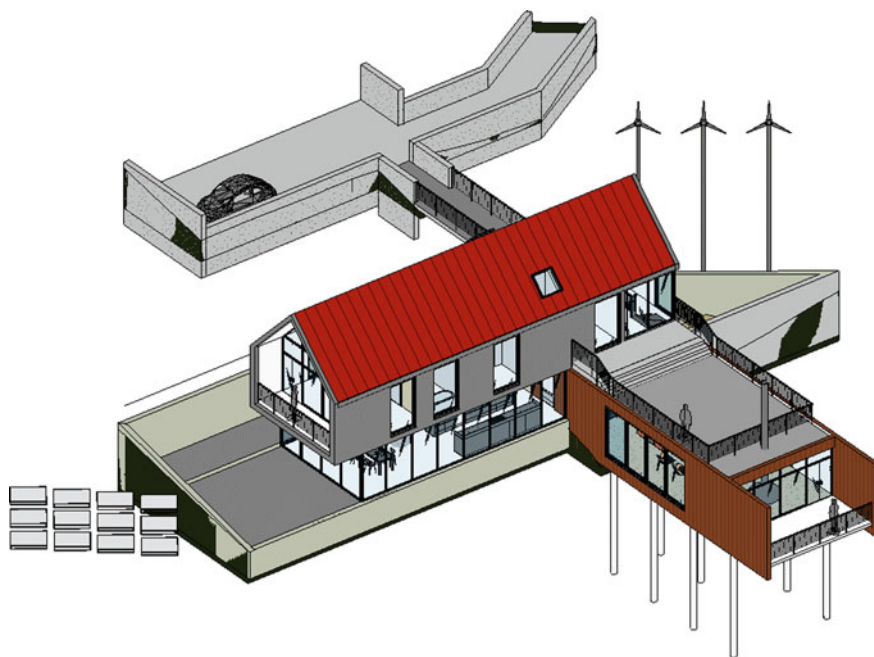
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

## 4 Experiment and Results

### 4.1 Test Model Used for Experiment

To validate the effectiveness of using synthetic data in image segmentation of building components, we used a residential building information as the testbed model (as shown in Fig. 7). It is a single-family house with 21 building component types (i.e., Revit families). Objects other than the 21 component families was labeled “Void”. The resulted 22 classes of building components are shown in Table 1.

In total, we generated 20,700 images from the residential building model with corresponding pixel-level labels. We used 18,000 images for training, 2200 for validation and 500 for testing. It should be noted that some component families would dominate in quantity and size in the model (e.g., walls), causing issues with small data samples if certain families are under-represented in the collected images. This issue can be addressed by generating a large synthetic dataset with sufficient number of images or resampling of the collected images to rebalance the class distribution for building components.



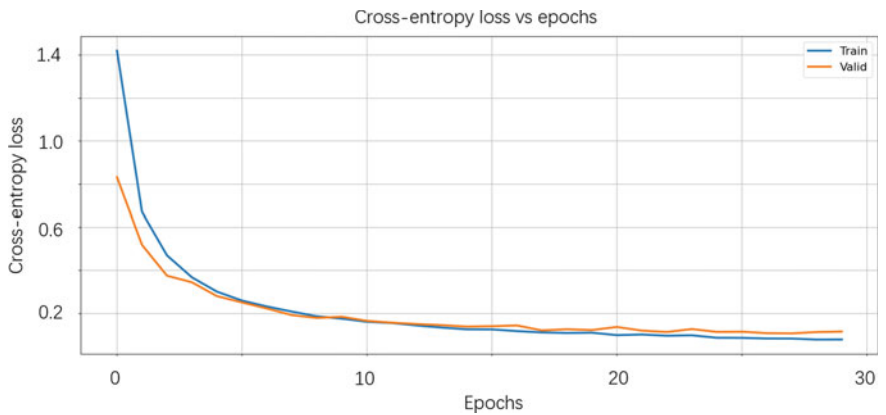
**Fig. 7** BIM used for synthetic data generation

**Table 1** Classes of building components retrieved from Revit based on Revit family categorization

Architectural	Structural	Mechanical electrical and plumbing	Others
Walls	Structural foundation	Electrical equipment	Specialty equipment
Doors	Structural columns	Lighting fixtures	Generic models
Curtain panels		Plumbing fixtures	Furniture
Floors			Furniture systems
Stairs			Casework (e.g., bookshelf)
Windows			Site (e.g., Wind power generator)
Roofs			Entourage (e.g., RPC male)
Ceilings			Void

4.2 Performance of the Semantic Segmentation Model

For the testbed model, we modified the output of the U-Net to 22 classes to match the number of building component classes generated from simulation. Cross-entropy was used as the loss function, and the learning rate of the U-Net model was set at 8.00 e-4. We used a batch size of 16, and trained for 30 epochs. As seen in Fig. 8, the best cross-entropy loss was reached at epoch 26 with 0.08. The best MIoU score was 0.89. Figures 8 and 9 show the categorical cross-entropy and MIoU score over 30 epochs. It can be seen that both graphs reached a plateau after epoch 25. This means the optimization of the model had reached its local minimum, providing the best prediction it could generate given the dataset. To visualize the segmentation results, a subset of the prediction result is shown in Fig. 10.



**Fig. 8** Cross-entropy loss over 30 epochs

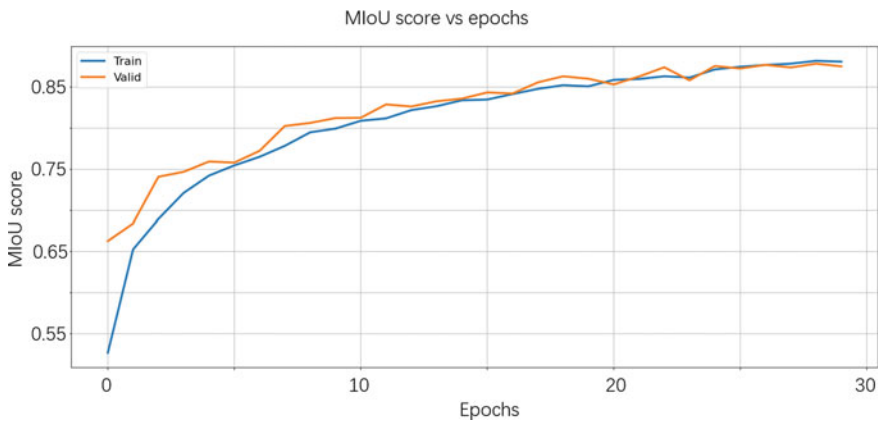


Fig. 9 MIoU over 30 epochs

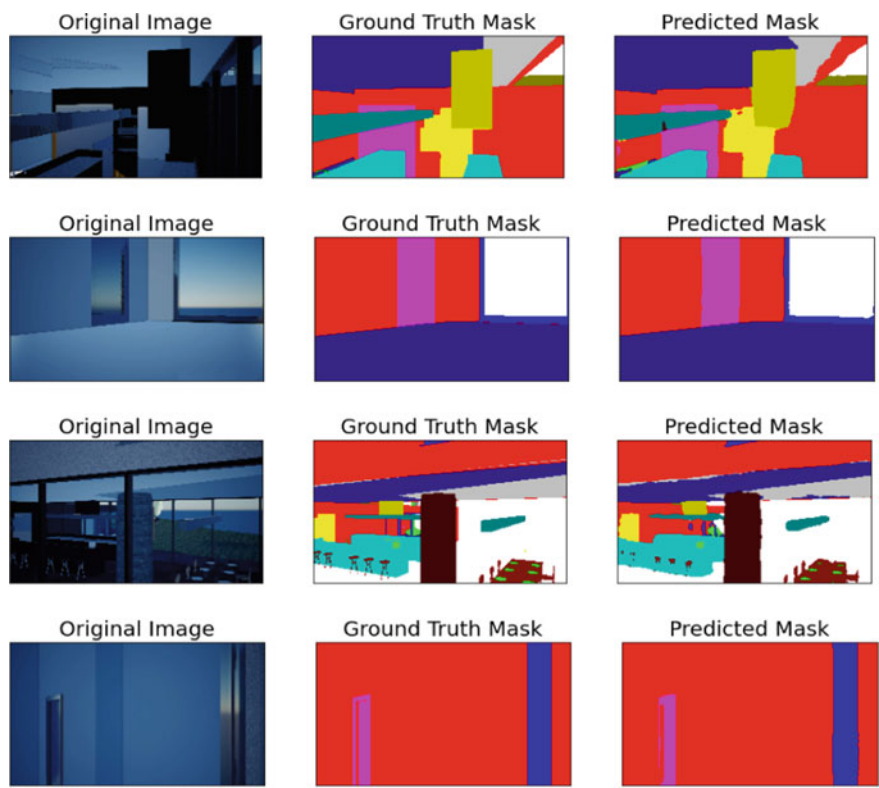


Fig. 10 Segmentations generated by the trained U-Net model

## 5 Conclusions and Future Work

This study proposed an automated method for generating diverse images of building components in a game engine to provide a wealth of training and testing images for semantic segmentation of building components. In addition, we introduced an automated method for labeling the synthetically generated images from a game engine, helping address the problem of lacking labeled training data in the AEC domain for semantic segmentation of building components, which can be seen as the first step toward automated QA using computer vision. A residential building was used as a testbed to show the viability and effectiveness of the proposed approach. From the result of our segmentation model, which reached a mean IoU of 0.89, it can be concluded that the synthetic data generation approach was effective. As the performance of segmentation models continues to improve, we can deploy the model in real-world vision-based QA tasks, which may help the inspectors to automatically label images collected onsite for later comparisons between as-built structures with as-designed models.

For future work, we aim to apply the trained model using only synthetic data to real-world data to test its performance. Past studies indicate that segmentation models trained on synthetic point clouds of building components [16] demonstrate higher performance when synthetic data was used together with real-world data. However, the detailed correlation between mixing synthetic and real data and the performance of the model has not yet been discovered, which opens opportunities to further improvement of the model performance. Another thread of future work lies in defining more appropriate metrics for QA specific computer vision tasks. This is because there are still gaps between the current computer vision metrics and civil engineering applications because evaluations like MIOU score cannot be directly applied to QA. To bridge the gaps, one improvement that can be made is to incorporate metrics that are more relevant in practice in the AEC community, such as geometrical and topological connections between different spaces and openings.

## References

1. Czerniawski T, Leite F (2020) Automated segmentation of RGB-D images into a comprehensive set of building components using deep learning. *Adv Eng Inform* 45:101131
2. Danielczuk M, Matl M, Gupta S, Li A, Lee A, Mahler J, Goldberg K (2019, May) Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. In: 2019 International conference on robotics and automation (ICRA). IEEE, pp 7283–7290
3. Guo BH, Zou Y, Fang Y, Goh YM, Zou PX (2021) Computer vision technologies for safety science and management in construction: a critical review and future research directions. *Saf Sci* 135:105130
4. Han J, Rong M, Jiang H, Liu H, Shen S (2021) Vectorized indoor surface reconstruction from 3D point cloud with multistep 2D optimization. *ISPRS J Photogramm Remote Sens* 177:57–74
5. Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R (2015) SceneNet: understanding real world indoor scenes with synthetic data. arXiv preprint 2015. arXiv preprint [arXiv:1511.07041](https://arxiv.org/abs/1511.07041)

6. Hwang BG, Thomas SR, Haas CT, Caldas CH (2009) Measuring the impact of rework on construction cost performance. *J Constr Eng Manag* 135(3):187–198
7. Iglesias C, Martínez J, Taboada J (2018) Automated vision system for quality inspection of slate slabs. *Comput Ind* 99:119–129
8. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014) Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint [arXiv:1406.2227](https://arxiv.org/abs/1406.2227)*
9. Josephson PE, Hammarlund Y (1999) The causes and costs of defects in construction: a study of seven building projects. *Autom Constr* 8(6):681–687
10. Kim MK, Wang Q, Li H (2019) Non-contact sensing based geometric quality assessment of buildings and civil structures: a review. *Autom Constr* 100:163–179
11. Koch C, Georgieva K, Kasireddy V, Akinci B, Fieguth P (2015) A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Adv Eng Inform* 29(2):196–210
12. Kopsida M, Brilakis I (2020) Real-time volume-to-plane comparison for mixed reality-based progress monitoring. *J Comput Civ Eng* 34(4):04020016
13. Lin KL, Fang JL (2013) Applications of computer vision on tile alignment inspection. *Autom Constr* 35:562–567
14. Liu Z, Cao Y, Wang Y, Wang W (2019) Computer vision-based concrete crack detection using U-net fully convolutional networks. *Autom Constr* 104:129–139
15. Long J, Shelhamer E, Darrell T (2014) Fully convolutional networks for semantic segmentation. *arXiv:1411.4038 [cs.CV]*
16. Ma JW, Czerniawski T, Leite F (2020) Semantic segmentation of point clouds of building interiors with deep learning: augmenting training datasets with synthetic BIM-based point clouds. *Autom Constr* 113:103144
17. Martinez P, Ahmad R, Al-Hussein M (2019) A vision-based system for pre-inspection of steel frame manufacturing. *Autom Constr* 97:151–163
18. Martinez P, Al-Hussein M, Ahmad R (2019) A scientometric analysis and critical review of computer vision applications for construction. *Autom Constr* 107:102947
19. Phares BM, Washer GA, Rolander DD, Graybeal BA, Moore M (2004) Routine highway bridge inspection condition documentation accuracy and reliability. *J Bridg Eng* 9(4):403–413
20. Rahimian FP, Seyedzadeh S, Oliver S, Rodriguez S, Dawood N (2020) On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Autom Constr* 110:103012
21. Ronneberger O, Fischer P, Brox T (2015, October) U-net: convolutional networks for biomedical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Springer, Cham, pp 234–241
22. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 3234–3243
23. Soltani MM, Zhu Z, Hammad A (2016) Automated annotation for visual recognition of construction resources using synthetic images. *Autom Constr* 62:14–23
24. Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P (2017, September) Domain randomization for transferring deep neural networks from simulation to the real world. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp 23–30
25. Wu Y, Xue F (2021) FloorPP-Net: Reconstructing floor plans using point pillars for scan-to-BIM. *arXiv preprint [arXiv:2106.10635](https://arxiv.org/abs/2106.10635)*

26. Yue X, Wu B, Seshia SA, Keutzer K, Sangiovanni-Vincentelli AL (2018, June) A lidar point cloud generator: from a virtual world to autonomous driving. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval, pp 458–464
27. Zhong B, Wu H, Ding L, Love PE, Li H, Luo H, Jiao L (2019) Mapping computer vision research in construction: developments, knowledge gaps and implications for research. *Autom Constr* 107:102919



# Lessons Learned from Developing and Testing an Augmented Reality Application for Just-in-Time Information Delivery to Improve Construction Safety



Krithikashree Lakshminarayanan, Mayank Arvindbhai Patel, Zia Din, and Lingguang Song

**Abstract** The use of Building Information Modeling (BIM) is rapidly gaining popularity among designers, owners, and construction project managers thanks to its ability to generate and communicate construction-related information. However, the use of BIM to provide job-related details to construction workers is limited so far. There is a need for tools such as augmented reality (AR) applications to enable the BIM information flow to field workers to benefit fully from it. Due to a lack of guidelines, there is a need to provide AR development resources to potential developers, such as construction professionals, construction educators, and new AR developers. This article describes the development of a marker-based AR application for handheld devices to deliver just-in-time information to construction workers. The resulting application aims to help workers to learn task-specific construction methods and safety information before and during task performance by interacting with the models of the task at hand. The authors collected the proposed AR application requirements through a literature survey. The application was developed in the Unity game engine using the Vuforia AR software development kit. Three-dimensional BIM models and simulations of building elements, such as slab-on-grade reinforcement mesh and formwork models, were developed in Autodesk Revit and Autodesk Navisworks. Models and simulations can be visualized in the marker-based AR application. The application development process and resources documented in this paper will help new developers, especially non-programmers, avoid the pitfalls that the authors experienced during the development of this application.

**Keywords** Construction safety · Augmented reality application

---

K. Lakshminarayanan

Department of Computer Science, University of Houston, Houston, TX 77004, USA

M. A. Patel · Z. Din (✉) · L. Song

Department of Construction Management, University of Houston, Houston, TX 77004, USA

e-mail: [uziauddi@central.uh.edu](mailto:uziauddi@central.uh.edu)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_15](https://doi.org/10.1007/978-3-031-34593-7_15)

# 1 Introduction

The design, development, and use of augmented reality (AR) applications are resource-intensive and require technical skills, subject knowledge, and high-performance hardware, such as computers and handheld devices. Most non-programmers interested in developing AR applications face challenges due to a lack of learning resources and guidance regarding AR application development. There is a need to provide AR development resources to new AR developers, such as construction professionals and researchers.

Despite architectural and engineering designs being commonly produced using computer programs [17], the resulting digital models created by designers are only available to management-level project staff, for example, clients, architects, project managers, and superintendents. As front-line workers on construction sites, construction workers rarely take advantage of these digital models to improve construction performance [10]. There is a need to develop tools such as AR applications to present digital models to end-users, i.e., construction workers. Apart from the numerous advantages digital designs offer for construction, they can minimize the need for physical prototypes or construction mock-ups to communicate design concepts to project stakeholders, including construction workers [9]. Physical mock-ups are expensive, time-consuming, and unsafe to build and use [13]. The digitalized mock-up model can be cost-effective, quicker, and easy to reproduce [7] for construction workers without the high price tag associated with the physical version. AR can further improve the visualization of digital models to prepare workers for just-in-time information delivery, such as construction methods and safety [3]. Digital mock-ups have limitations in that they cannot give information, such as directly feeling the texture and polish of installation materials or conducting tests such as water leak tests. However, it can provide critical information on the construction process and is not constrained by space or time.

By importing a BIM model into an AR application installed on a mobile device, construction workers can learn about their tasks in the virtual space before performing the actual construction task. By visualizing construction task models in AR on a mobile device, workers can identify safety and constructability issues in less time than with the traditional approach by reading blueprints [21]. The authors developed an AR software application called a just-in-time (JIT) information presentation system that can present 3D data from a BIM model and other useful information in AR, such as simulations, stepwise execution processes, and construction task-related 2D drawings. This paper presents the lessons learned while developing an AR application to present the BIM model in a marker-based AR environment.

## **2 Application Development**

This section describes the steps involved in developing an AR application called the just-in-time (JIT) AR app. The hardware and software requirements for the application, the development process, and the post-development protocols are given below.

### ***2.1 Pre-development Process***

This section discusses hardware requirements for application development and testing, including computer and portable device specifications and the computer programs necessary for the application development process, such as a game development engine and a software development kit. Additionally, hardware and software selection and their features are also described.

#### **2.1.1 Development Team**

The development team is diverse, consisting of researchers from different disciplines, including computer science and construction management.

#### **2.1.2 Gathering AR Application Development Requirements**

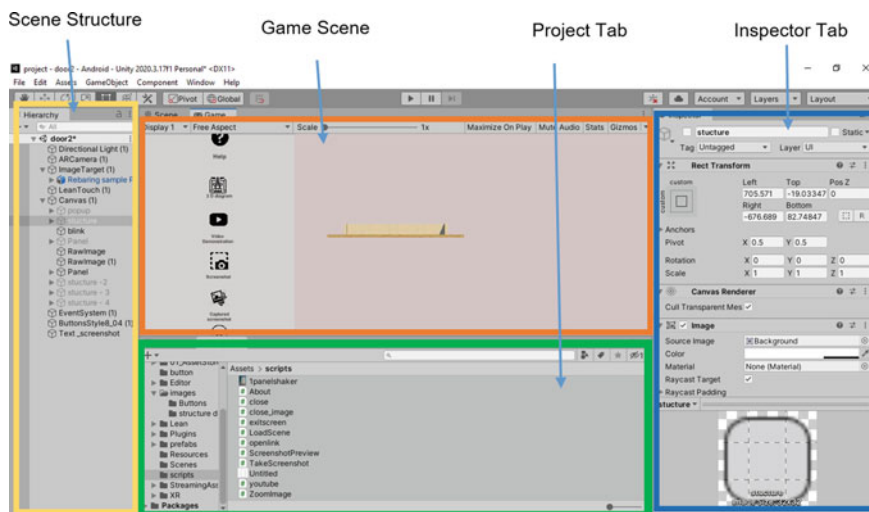
The list of commonly used game engines provided by Din and Gibson [4] was used to choose the most suitable development platforms. Unity, Unreal, and Gamemaker Studio are the three most popular gaming engines. These are commercial game development platforms that are currently available on the market. The following factors were considered when selecting a game engine: the learning curve, the availability of learning resources, the licensing fee, the availability of assets, the scripting requirement, and the ability to deploy applications across multiple platforms such as iOS, PC, and Android.

Based on the above features, the authors created a list of pros and cons to select the best tools to develop the JIT information presentation system, an AR application.

#### **2.1.3 Hardware for Development and Testing**

The authors used Samsung's Galaxy S7 Plus [16] to test and deploy the application. The reason to choose this handheld device for testing is its processing power. It has a Qualcomm Snapdragon 865 processor, and this intelligent chipset is fast and robust, allowing it to run intense games or efficiently multitask with minimal lag.





**Fig. 2** AR interface development in unity game engine

game designers to create objects and environments quickly and efficiently. With the help of a wide variety of shaders, materials, processing, and intense lighting, Unity can deliver powerful graphics [8].

Unity was written in C++ and has been fine-tuned for performance over time. Unity excels at cross-platform deployment, which is a key attraction for today's developers. Unity supports almost all operating systems, making applications developed using Unity deployable on any platform.

Figure 2 shows a screenshot of the Unity development engine. The left panel shows the scene hierarchy. The middle tab is the output of the screen sample. The right corner is the inspector tab. This tab has the selected component property, such as the size and position of the object. The bottom-most tab is the project tab. It consists of all the assets required for the application, including scripts and images.

In this research, a slab-on-grade model was created in Autodesk Revit and then imported into Unity. The scene structure shows the active scene. It shows the presence of canvas, panels, and buttons that will be visible in the application user interface. The game scene is the sample view of how the application's screen will look when deployed to a handheld device. The project tabs show the list of custom scripts the authors have developed for the JIT application. Finally, the inspector tab shows the details about the structure panel. It shows the positions, colors, and other features of the panels.

### **2.1.5 Vuforia AR Software Development Kit (SDK)**

Vuforia is an augmented reality (AR) software development kit (SDK) for mobile devices [12]. Computer vision technology recognizes and tracks planar graphics and 3D objects in real time [6]. This image registration functionality allows developers to position and orient virtual objects when viewed through a mobile device's camera. The virtual object tracks the image's location and orientation in real time, ensuring that the viewer's perspective on the object is identical to that of the target. A virtual object is a part of the scene that exists in the actual world [12].

Vuforia provides APIs in C++ , Java, Objective-C++ , and .NET. As a result, the SDK supports development on the iOS and Android platforms. This makes the development of AR applications in Unity easy and portable.

## **2.2 JIT AR Application Development Process**

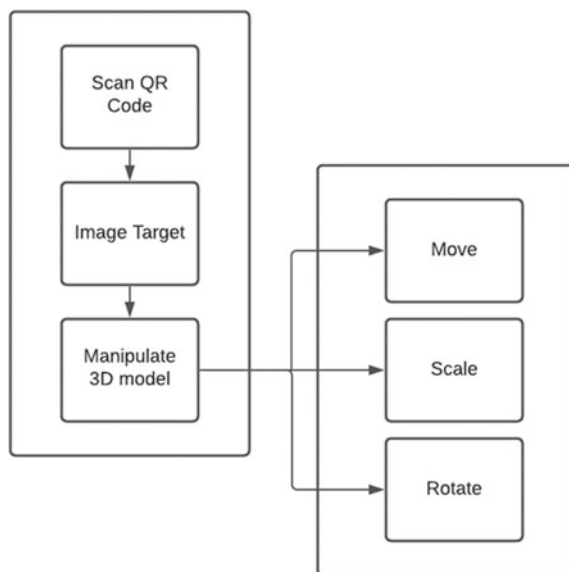
This section discusses the AR application's development phase, which includes the following steps: developing a software architecture, selecting modeling and simulation tools, developing a quick response code scanning function using an image target, interacting with the resulting model, and deploying it for visualization.

Software architecture is a conceptual model or blueprint that defines the building elements of a software application [15]. Figure 3 shows the software architecture of the proposed JIT application. An Android mobile application was created to achieve a JIT information presentation function for construction workers. The application was built in Unity. The Unity platform includes different packages and libraries such as AR foundation, AR XR core, and Vuforia Engine. The Vuforia Engine helps create a database for various quick-access (QR) code collections. This database (DB) was imported into the Unity asset to link through a unique key to the 3D models in the Unity assets. After scanning the QR code, the 3D model related to the marker can be retrieved from the local database in Unity assets. A detailed description of the process and tools is as follows.

### **2.2.1 3D BIM Model and Simulation Development**

BIM is a smart 3D model-based approach that includes a digital representation of a facility's physical and functional properties [14]. BIM is a shared knowledge resource for facility information that serves as a strong foundation for decision-making throughout a facility's life cycle, defined as the period from conception through disposal. Revit is a BIM-based program where architects, engineers, contractors, and designers collaborate and generate a unified model using real-world data [14]. The authors used Autodesk Revit to create 3D models, drawings, and documents enabled by BIM technology.

**Fig. 3** Key features of the AR application

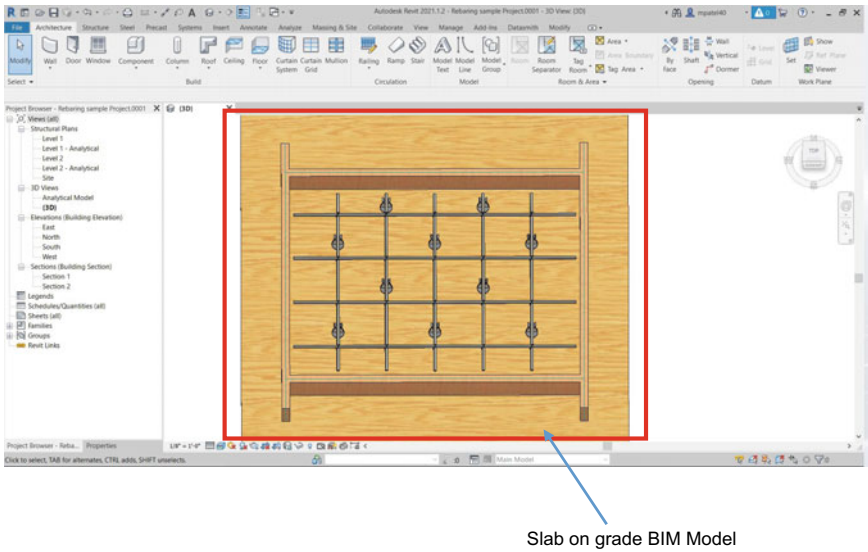


The authors developed a 3D BIM model of reinforcement mesh and formwork for slab-on-grade construction. A screenshot of the BIM model development phase is shown in Fig. 4. All 3D model components, such as plywood base, wood planks, reinforcing bars, plastic spacers, and tie bars, are created in Autodesk Revit 2020. Additionally, Autodesk Revit 2020 is used to draw two-dimensional drawings. Other components of the information, such as simulation and step-by-step task execution, are built using Autodesk Navisworks 2020. The 3D model must be exported to Navisworks using the Navisworks plugin for Autodesk Revit to create simulations. The challenging aspect was to convert a 3D BIM model to an AR model without affecting its materials, textures, and dimensions. After multiple attempts at various approaches, the authors used Revit's twin motion plugin to convert a 3D BIM model to a Fbx. file format. Finally, the Unity engine helped visualize the processed Fbx file as an AR model.

### 2.2.2 QR Code Scan

The authors used the AR camera feature of the Vuforia plugin to scan a QR code. It is one of the most user-friendly plugins with much power and adaptability. A QR code is used to uniquely identify the content presented in the AR environment. The QR code is created and stored in the Vuforia database. A Vuforia account is required to store image targets [18].

Furthermore, with a Vuforia developer account, developers can create and manage their target database using the Vuforia developer webpage's target manager. The target manager supports target images of all types [12].



**Fig. 4** Slab-on-grade model in Autodesk Revit

### 2.2.3 Image Target

A license key must be assigned to each database in the target manager. Users can add numerous targets to a device database, such as photographs and objects. It has both multi-targets and image targets.

Each target must be given a name, a measurement in meters, and the source image from which we want to create an image target. The image's real aspect ratio should correspond to the target size in meters. For example, a 50-cm-wide printed image target will have a width of 0.50. To define the image-based target, developers may need to fill in extra dimensions depending on the target type.

### 2.2.4 Manipulation of the 3D Object

Lean Touch is an asset available in the Unity Asset Store that makes scaling and rotating objects in AR possible. It allows players to simulate multiple finger gestures (e.g., pinch, twist) using a computer and a mouse to test the developer without deploying it on a mobile device. Lean Touch provides an API that provides access to many flexible example components, allowing customization of everyday interactions with minimal coding.



### 2.2.5 BIM Model Deployment in AR Application

The BIM AR model built using Revit is imported into the Unity projects asset folder and is added as a prefab. The Vuforia is added to the project by the Package Manager. The license key found in the developer portal is added to the image target to integrate the target manager database. Slight asset touch (lean pinch scale, lean rotate axis, and lean drag translate) is added to the prefabricated properties. The C# script for additional functionality is integrated where necessary.

### 2.2.6 Technical Steps and Pseudo Codes

In this section, the authors discuss the steps involved in the development of the JIT application. They also discuss the logic behind each application's functionality. The authors have also given sample starter code snippets for the reader to build upon.

#### Basic Setup

A detailed description of the basic setup and requirements to build an AR application with a marker is discussed below. The procedure is as follows.

To create a marker for the application, a Vuforia developer account must be created, and the image target must be added to it. A Unity project of type 3D must be created. To link the application to the Vuforia image target, the Vuforia SDK must be added to the asset folder, and the downloaded plugin must be installed. The AR camera and image target component must be added to the project. Figure 5 shows a screenshot of this process.

Then the Vuforia license key needs to be added to the advanced settings of the image target to add the Vuforia database to the project, as shown in Fig. 6. Then the developed BIM model from the asset folder is added to the image target component of the project.

By following the steps mentioned above, a simple working marker-based AR application can be developed. The next section will give a description of how the authors have added custom features to improve the efficiency of the application.

#### Application's Major Features

The authors discussed with a couple of construction experts and conducted a literature survey to understand which form of communication is effective for construction workers. From the literature study and the expert opinion, the authors concluded that the combination of traditional and modern approaches gives the best results. The authors verified the necessity and impact of these features with other successful applications such as IKEA [5], ARKI [1], and Augment [2]. We formulated the

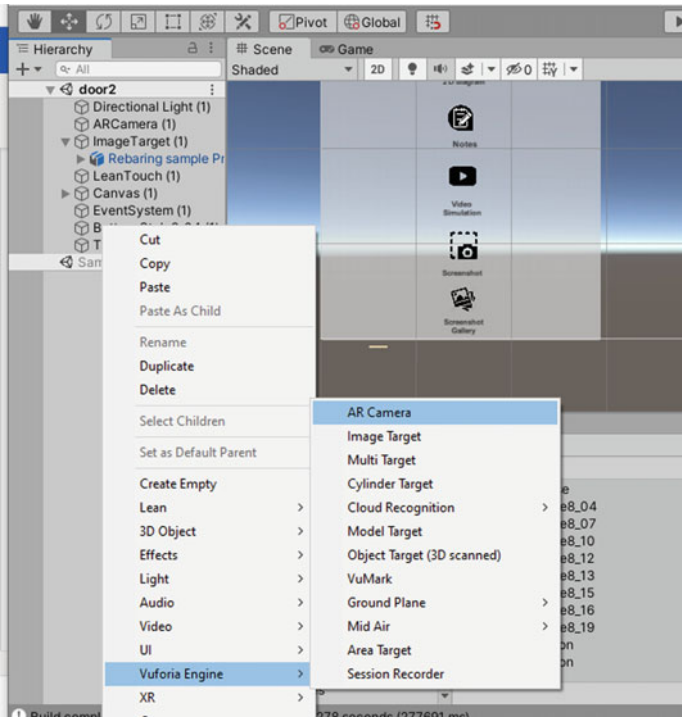
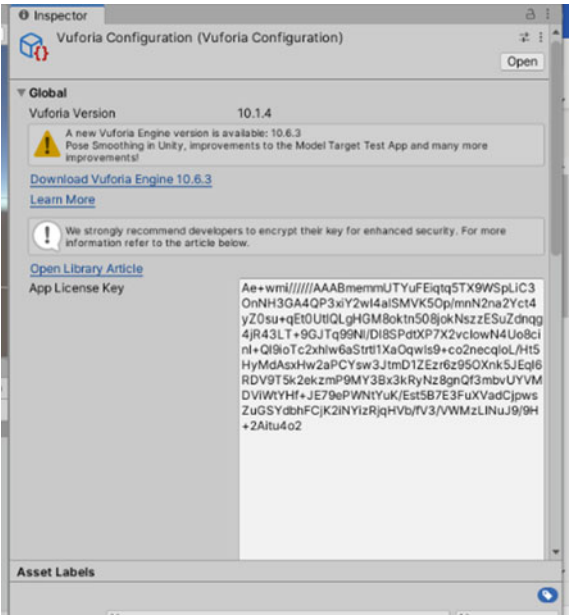


Fig. 5 Adding Vuforia components to user interface (UI)

Fig. 6 License key for Vuforia



necessary UI features from these applications to provide the JIT information to the user.

The authors decided to feed the end-user with the 2D drawings (traditional approach) combined with visual aids like simulations, digital 3D models, etc., to give all the information required to learn before actually performing the task.

**Feature Development.** After the basic setup to enable the touch gestures for the imported 3D model, the authors used the “Lean Touch” plugin from the Unity Asset Store.

The 3D model under the target image was selected, and components such as lean drag translate for movement, lean pinch scale for resizing, and lean twist rotate axis for model rotation were added. These features help the end-user to gain more perspective about the task.

The authors added the panel in the canvas for the user interface, and functionalities were added for the immersive experience button. Figure 1 shows a screenshot of the user interface of the application. The development screen and individual button functionalities are shown in Fig. 1.

1. Help button: This button gives a description of the application, the task description, and detailed help about navigating inside the application.
2. 2D diagram: This button opens the 2D drawings for the BIM model. The image was created using Autodesk Revit. We provide the zoom-in and zoom-out features for this panel. The logic behind the zooming is as follows in Code Block 1.

```
if (Input.touchCount == 2)
{
    Touch touchZero = Input.GetTouch(0);
    Touch touchOne = Input.GetTouch(1);

    Vector2 touchZeroPrevPos = touchZero.position - touchZero.deltaPosition;
    Vector2 touchOnePrevPos = touchOne.position - touchOne.deltaPosition;

    float prevMagnitude = (touchZeroPrevPos - touchOnePrevPos).magnitude;
    float currentMagnitude = (touchZero.position - touchOne.position).magnitude;

    float difference = currentMagnitude - prevMagnitude;

    zoom(difference * 0.01f);
}
```

Code Block 1: Zoom 2D drawing

The logic here is that we checked if there is a touch count of 2 (2-finger pinch zoom) if yes, we identified the position of both fingers and changed the image size accordingly.

3. Video Simulation: We have created a simulation video for our slab-building experiment using Navisworks. We have uploaded our video to popular streaming platforms such as YouTube and Microsoft Stream. By clicking the button,

the application navigates to the streaming channel. We did this using the **application.openURL** function present in C#.

4. Screenshot Capture: To keep track of the progress and to communicate in the near future, the authors have enabled users to take a screenshot of their progress/doubts during the process. The authors developed this feature using custom codes shown in Code Block 2. The logic behind this process is as follows.

```
IEnumerator CaptureIt()
{
    // Texture2D texture = null;
    string timeStamp = System.DateTime.Now.ToString("dd-MM-yyyy-HH-mm-ss");
    string fileName = "Screenshot" + timeStamp + ".png";
    string pathToSave = fileName;
    ScreenCapture.CaptureScreenshot(pathToSave);

    yield return new WaitForEndOfFrame();

    Instantiate(blink, new Vector2(0f, 0f), Quaternion.identity);}
```

Code Block 2: Screenshot Capture

The above logic uses the CaptureScreenshot method to click on the picture. To name each picture uniquely, the authors used the timestamp of the exact moment when the action was taken and used it as the name of the png. The authors have also added a blink feature to let the end-user know that the screenshot was taken.

5. Gallery: To view the captured screenshots, we redirected to the file's path and checked if there exists any image. The authors then added the previous and next buttons to navigate among the screenshots in the gallery. The skeleton for this process is shown in Code Block 3.

```
Texture2D GetScreenshotImage(string filePath)
{
    Texture2D texture = null;
    byte[] fileBytes;
    if (File.Exists(filePath))
    {
        fileBytes = File.ReadAllBytes(filePath);
        texture = new Texture2D(2, 2, TextureFormat.RGB24, false);
        texture.LoadImage(fileBytes);
    }
    return texture;
}
```

Code Block 3: View Screenshot

6. Notes: Similarly to the screenshot, the authors have provided users with a place to take notes during the process so that they can visit later to these comments. This feature is achieved by using the input field component in the UI feature of

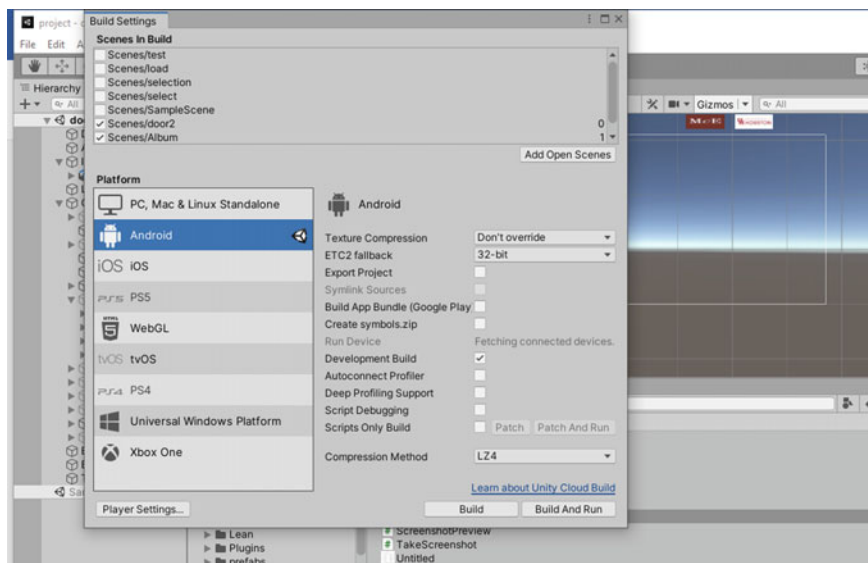


Fig. 7 Build environment

the Unity project. The logic behind saving the notes is described in Code Block 4.

```
inputField.text = PlayerPrefs.GetString("NoteContents");
string path = Application.dataPath + "/notes.txt";
File.WriteAllText(path, "Note test \n\n");
```

Code Block 4: Notes

## Building and Deploying the Application on Android Devices

Once the development was done to test the application, the authors deployed the JIT app on a Samsung Galaxy tablet. To achieve this, go to Files > Build settings, Select Android. The top panel shows the list of scenes present in the project. The authors selected the scenes necessary for the build. We could configure the application in the player settings. A screenshot of building an apk and deploying the application on a handheld device is shown in Fig. 7.

### Transfer from BIM to Unity

To transfer the BIM model directly into Unity, the authors used software called Tridify [19]. It imports processed 3D BIM and CAD models with BIM data attached straight to Unity software using Tridify BIM. The steps involved in establishing this pipeline are as follows:

1. Download the Tridify BIM software.
2. Go to Assets > Import-Package > Custom package in Unity and locate the package file to import.
3. The Tridify menu should appear under the Tools menu. If not, make sure you are using .NET 4.X as the Scripting Runtime Version (Edit > Project Settings > Player > Other Settings > Configuration) and restart Unity.
4. Log in to your Tridify account (Tools > Tridify > Conversion Projects) and locate and import your processed files into Unity.

This process requires that the application be rebuilt and redeployed for every new model used inside our JIT system.

### 2.3 *Post-development and Testing*

The authors plan to use objective and subjective measurements to evaluate the results used by Pranoto et al. [11] in a usability measure study. Quantitative measurements and qualitative expert opinions will be collected and analyzed to assess the effectiveness of the AR application.

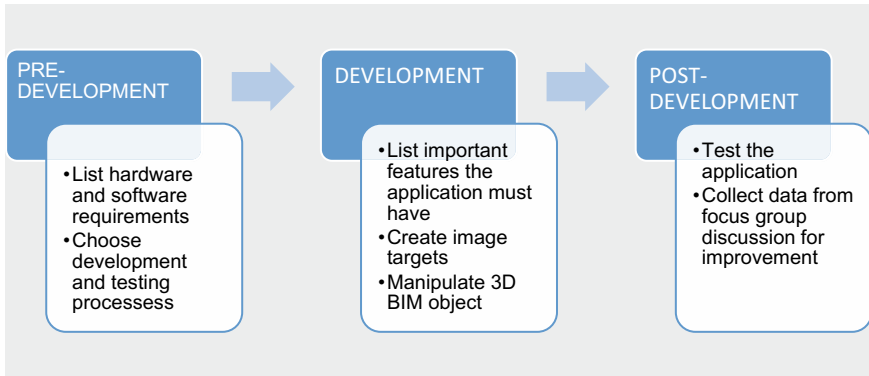
All AR application testing participants will be asked to complete a survey. The survey will ask about the participant's academic and construction professional background and previous experience with any AR-related application. Participants will also sign a consent form prior to the experiment. Participants will receive a briefing on how to use the JIT AR application.

After the experiment, participants will be asked to complete a post-experiment survey. This survey will act as an objective measure that provides quantitative observations through rating and judgment based on participants' subjective opinions.

Finally, all participants will be invited to a focus group discussion where they could give the authors suggestions for enhancements.

## 3 Summary of Lesson Learned

This section summarizes the development process and highlights the issues faced during implementation. The development process has three phases, as illustrated in Fig. 8.



**Fig. 8** Main steps in the AR application development

1. **Choosing an appropriate development platform:** A list of all available development platforms and their pros and cons was identified. This study selected the Unity game engine due to its component-based development procedure, allowing developers to use existing prefabricated components to create powerful applications.
2. **Choosing the right plugins:** There are a variety of plugins available on the market, such as AR kit, and AR core. Before choosing the right plugin, the authors developed a requirement specification document based on the water flow model. The authors have selected Vuforia because it supports multi-target features and facilitates cloud-based image target storage based on the requirements checklist.
3. **Choosing prefabs:** A wide range of prefabricated materials is available for each function. The authors selected appropriate prefabs like Lean Touch based on the number of downloads and user comments in Unity's asset store.
4. **Writing custom code:** The authors wrote custom codes for various features of applications such as open video, simulation in the stream, and open and closed panels. The authors suggest using Unity's manual to obtain a reliable and proper syntax for coding.

## 4 Future Work

The current application offers useful features that make the tool ready for construction information presentation. AR applications for construction-related tasks should consider adding advanced features such as task completion, Chabot, video call, issue report, and walk-throughs in the future application developments. These features are not in the scope of this work, but the authors plan to include them in a future enhancement. Many of the features mentioned can be achieved by integrating existing tools available on the market, such as Slack for texting.

**Acknowledgements** This study was supported by the McElhattan Foundation. Its contents and opinions are solely those of the authors and do not represent opinions and views of the McElhattan Foundation.

## References

1. ARKI (2022) Augmented reality platform for architecture, engineering and construction. <https://www.darfdesign.com/arki.html>
2. Augment app (n.d.) Our built-in solutions get your AR-ready 3D products fast and at an unbeatable price. Increase Sales Engage e-Commer 3–4
3. Chi HL, Kang SC, Wang X (2013) Research trends and opportunities of augmented reality applications in architecture, engineering, and construction. *Autom Constr* 33:116–122. <https://doi.org/10.1016/j.autcon.2012.12.017>
4. Din ZU, Gibson GE (2018) Leveraging pedagogical innovation for prevention through design education: lessons learned from serious game development. *ASCE* 706–716
5. IKEA (2017) Say Hej to IKEA Place. YouTube 3–4
6. Kim K, Lepetit V, Woo W (2010) Scalable real-time planar targets tracking for digilog books. *Vis Comput* 26(6):1145–1154
7. Kim S, Park H, Lee K, Jung C (2006) Development of a digital mock-up system for selecting a decommissioning scenario. *Ann Nucl Energy* 33(14–15):1227–1235. <https://doi.org/10.1016/j.anucene.2006.07.009>
8. Linowes J (2021) Augmented reality with unity AR foundation: a practical guide to cross-platform AR development with Unity 2020 and later versions. Packt Publishing. <https://books.google.com/books?id=iBk-EAAAQBAJ>
9. Memon AH, Rahman IA, Memon I, Iffah N, Azman A (2014) BIM in Malaysian construction industry : status , advantages , barriers and strategies to enhance the implementation level. 8(5):606–614
10. Mirarchi C, Pavan A, de Marco F, Wang X, Song Y (2018) Supporting facility management processes through end-users' Integration and coordinated BIM-GIS technologies. *ISPRS Int J Geo-Inf* 7(5). <https://doi.org/10.3390/IJGI7050191>
11. Pranoto H, Tho C, Warnars HLHS, Abdurachman E, Gaol FL, Soewito B (2018) Usability testing method in augmented reality application. In: Proceedings of 2017 international conference on information management and technology. ICIMTech 2017, 2018-Jan (November 2020), pp 181–186. <https://doi.org/10.1109/ICIMTech.2017.8273534>
12. PTC Inc (n.d.) Vuforia library. Retrieved 21 March 2022, from <https://library.vuforia.com/>
13. Reiners D, Stricker D, Klinker G, Stefan M (2020, Nov) Augmented reality for construction tasks: doorlock assembly. *Augmented Reality* 51–66. <https://doi.org/10.1201/9781439863992-10>
14. Revit B (2022) Run projects more efficiently 1–10
15. Richards M, Ford N (2020) Fundamentals of software architecture: an engineering approach. O'Reilly Media
16. Samsung (n.d.) Samsung.com Advantage 7(349):1–10
17. Schön DA (1992) Designing as reflective conversation with the materials of a design situation. *Knowl-Based Syst* 5(1):3–14. [https://doi.org/10.1016/0950-7051\(92\)90020-G](https://doi.org/10.1016/0950-7051(92)90020-G)
18. Simonetti Ibanez A, Paredes Figueras J (2013) The QR code is created and stored in the Vuforia database. A Vuforia account is required to store image targets. Universitat Politècnica de Catalunya



19. Tridify BIM (n.d.) Tridify BIM tools for unity installing tridify bim tools for unity how to use tridify BIM tools for unity. 2–3
20. Unity (2021) The platform of choice for multiplayer hits. 1–13
21. Wang X, Truijens M, Hou L, Wang Y, Zhou Y (2014) Integrating augmented reality with building information modeling: onsite construction process controlling for liquefied natural gas industry. *Autom Constr* 40:96–105. <https://doi.org/10.1016/j.autcon.2013.12.003>

# The Effectiveness of Data Augmentation in Construction Site-Related Image Classification



Mansoor Asif, Shuai Liu, Ghulam Muhammad Ali, Ahmed Bouferguene, and Mohamed Al-Hussein

**Abstract** In modern construction, the construction sites are congested, busy, and full of obstacles. The environment in the construction site is dynamic. In recent times, deep learning-based models remain the main tools used for image classification. However, the performance of the deep learning models in construction site-related image classification is not convincing due to the dynamic nature and busyness of the construction sites. The availability of construction site-related image datasets also remains an obstacle in achieving the best performance from the deep learning models. Data augmentation is a technique used to apply random but realistic transformation to the images. Data augmentation will not only help to diversify the image dataset but also assist in increasing the size of the dataset. This study used a state-of-the-art YOLOv4 deep learning model and implemented data augmentation techniques like gamma transformation to control the intensity of light and mimic sunny, cloudy, day, and night situations in the construction site images. The other data augmentation technique used is Gaussian blur to minimize the details in the images, and salt-and-pepper noise to degrade the quality of construction site images. The model is trained and tested on Alberta Construction Image Dataset (ACID) and construction workers hand signal image datasets, with and without the implementation of data augmentation. The performance of the model is evaluated based on the test dataset while keeping all the parameters of the model the same. It is observed that the model trained on the augmented dataset performed better than the model trained on the non-augmented dataset by 4%.

**Keywords** Data augmentation · Construction site-related · Image classification

---

M. Asif (✉) · S. Liu · G. M. Ali · M. Al-Hussein  
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada  
e-mail: [amansoor@ualberta.ca](mailto:amansoor@ualberta.ca)

A. Bouferguene  
Campus Saint-Jean, University of Alberta, Edmonton, Canada

# 1 Introduction

In recent times as the construction industry is moving from traditional construction operations to modular construction techniques, the construction sites are getting more overwhelmed with on-site resources like construction machinery, workers, etc., that make the construction operation less productive and unsafe for workers. The implementation of recent technological advancements can be useful to make the construction operation's working conditions more smooth, productive, and safe.

The adoption of technological advancement in the construction industry is moving at a very slow pace compared to other industries like automotive, health care, manufacturing industry, etc. However, the researchers in the construction industry implemented some of the technological advancements like deep learning techniques, IOTs, Big Data, etc., to solve the construction site-related problems. Deep learning has been studied and implemented by researchers in the construction industry to assess construction worker's ergonomics [22], on-site construction machinery, and their activity detection [19], detections and classification of multiple objects like construction hardhat [5] and waste materials [4] etc.

Meanwhile, the construction industry is facing multiple challenges in adopting the deep learning techniques in day-to-day construction operations for example issues on data privacy, high cost of computational machines to process the data, requiring expert engineers with the knowledge of technology, etc. The biggest challenge is the lack of ability of construction site-related image datasets. Where deep learning models requires significant amount of data for training purposes from multiple backgrounds, varying weather and lightning condition, etc., to improve the accuracy of the deep learning model. The case of implementing deep learning in construction sites has multiple challenges. Firstly, the amount of open-source construction site-related data is very limited in the industry, secondly construction site is dynamic in nature where the workers, machinery and other background work conditions are changing every minute, and finally, the construction sites are congested which makes deep learning model less effective to detect and classify objects. The weather and lighting conditions of the construction site also affect the accuracy of deep learning models.

To solve the issue of small dataset and different lightning conditions data augmentation can be used. Data augmentation is a technique that is used to apply realistic transformation to the dataset [8]. This technique can assist to reduce the chances of model overfitting, solve data imbalance and data biasness issues [14, 17]. The data augmentation will also help to increase the size of the image dataset by duplicating the samples in the dataset. In this paper, we will be using You Only Look Once (YOLOv4) state-of-the-art deep learning model to classify construction site-related images of construction equipment and construction site workers hand signals. YOLOv4 deep learning model is capable to look at an image and find the subset of object class, enclose it with the bounding boxes and identify its class when trained [1]. The reason of using state-of-the-art YOLOv4 model is because of its superior performance

over other state-of-the-art deep learning models like YOLOv3, Faster Recurrent-Convolutional neural network (Faster *R*-CNN), Single Shot Detector (SSD) etc., [1].

The present work is organized as follows: Sect. 2 describes the related work in the context of implementing data augmentation in image classification. Section 3 defines the methodology and image dataset selection. Sections 4 and 5 introduce the different augmentation techniques used in research and the architecture of the state-of-the-art YOLOv4 model. Section 6 presents the experimental setup for the research. Sections 7 and 8 contain the results and the conclusion of the study.

## 2 Related Work

The implementation of data augmentation techniques is very important to improve the performance of deep learning models. For example, [18] introduced an attentive CutMix data augmentation method to enhance the performance of deep learning image classification. The experiment is conducted on CIFAR-10/100 image dataset. The author used different convolutional neural network (CNN) architectures like ResNet, DenseNet, and EfficientNet. The results indicated that the developed approach increase the classification accuracy by 1.5% from the traditional data augmentation technique and a 3.04% increase in accuracy from baseline non-augmented techniques. Perez and Wang [16] used traditional transformation like zoom out, zoom in, shaded, etc., and Generative Adversarial Network (GAN) like Cezanne, Enhance, etc., augmentation techniques on ImageNet and MNIST datasets. The smallNet CNN model is trained on the augmented and non-augmented datasets. The model trained on the augmented dataset showed a 6% increase in performance from the non-augmented dataset. Mikołajczyk and Grochowski [14] trained VGG 16 model on skin melanomas diagnosis, histopathological images and breast magnetic resonance imaging (MRI) datasets. The author used GAN augmentation techniques and shear, reflection, rotation, etc., augmentation techniques and notice that implementation of data augmentation techniques improved the performance of the VGG 16 model. Gu et al. [6] used VGG-16 architecture with the custom CNN model on CIFAR-10 image datasets. Rotation, width and height shift, flip, etc., augmentations are used to train the model. The result showed a 2.1% improvement in the accuracy of the model when trained with the augmented dataset. Lei et al. [12] implemented rotation, solarize, invert, shear, color balance, etc., augmentation techniques on CIFAR-10, MNIST, and Fashion-MNIST datasets. The ResNet and LeNet-5 models are trained on augmented and non-augmented image datasets. It is noted that the model trained on the augmented image dataset showed better accuracy in the classification of the test dataset.

The augmentation techniques and deep learning models discussed above showed convincing results, but the construction site has its challenges. For example, the low visibility due to dusty conditions in the construction site, congestion of construction sites, full of the obstacle, changes in outdoor lighting condition, dynamic nature

and business of construction site make the construction site-related image classification a challenging task. Therefore, we propose three different augmentation techniques (gamma transformation, Gaussian blur, and salt-and-pepper noise) to mimic the construction site changing scenarios and as well as to improve the classification performance of deep learning models in construction site-related images classification.

### 3 Methodology

We propose three different data augmentation techniques in which five different experiments are conducted. Initially, the YOLOv4 model is trained on the original image dataset without any augmentation techniques implemented. Secondly, we applied gamma transformation to the original image dataset where lightning conditions in the images are adjusted. Thirdly, we applied Gaussian blur to the original image dataset. Fourthly, we tested the model by adding salt-and-pepper noise augmentation technique to the original image dataset; and finally, we combined all the above augmentation techniques to test the performance of the model.

#### 3.1 Dataset Selection

For this work, we use Alberta Construction Image Dataset (ACID) image dataset [20] and the construction site worker's hand signal image dataset. ACID is a construction machine detection dataset developed at the University of Alberta [20]. ACID was developed as a source to assist the use and development of deep learning applications in the construction automation field. The dataset contains images collected from construction sites all over the world [20]. The construction site worker's hand signals image dataset is collected manually in  $720 \times 720$  pixels in Indoor and outdoor environments. The dataset has static and dynamic backgrounds as well as sunny and cloudy weather conditions. For the original image dataset, we used 1000 images of construction vehicles from the ACID dataset and 1000 images of construction site workers' hand signal image datasets. Further data augmentation techniques are applied to the original image dataset.

### 4 Data Augmentation Techniques

Data augmentation is a method to increase the size of the collected dataset. Using data augmentation techniques, the dataset can be diversified by applying a random but realistic transformation. The implementation of data augmentation can improve the performance of the deep learning models [7], reduce the likelihood of model

overfitting, as well as better, generalize the dataset for model training purposes [14]. The following data augmentation techniques are used for the present work.

### ***4.1 Gamma Transformation***

Gamma transformation is a technique to control the intensity of light in the images. With gamma transformation, we can adjust the contrast and brightness of the original image [3]. This technique can bring more originality to the dataset and help the model to perform better in any lighting condition. For the current dataset, the range of gamma values in which to adjust the brightness and contrast is set to between  $-35$  and  $+35$ , where a set of  $-35$  will result in darker images and  $+35$  will result in lighter images. This range is chosen to make the conditions more realistic and representative of weather conditions ranging from sunny/clear to cloudy/overcast.

### ***4.2 Gaussian Blur***

Gaussian blur is used to blur the images in the dataset; it has been widely used for multiple purposes such as reducing the detail in the images. This technique smooths out the randomness in an image based on a chosen blur radius. Each pixel in the frame will adopt a new value based on the weighted average of its surrounding pixels [15], where more weight is given to pixels in closer proximity. The amount of blur in the frame is measured in pixels (px), where a higher value of px means more blur and a lower value means less blur will appear in the frame. To add blur to the dataset, a blur value between 1.15 and 3.25 px is chosen for the present work. Adding blur to the images in the dataset ensures that the model will be trained for a scenario where the camera may lose focus, thereby allowing the model to achieve better accuracy.

### ***4.3 Salt-and-Pepper Noise***

The “salt-and-pepper” noise data augmentation technique is used for image degradation. In this technique, some pixels of the frame are kept very noisy. The effect is likened to sprinkling salt and pepper on the frame [2], hence the name. The intensity of salt-and-pepper noise is measured in percentage, where 0% means there is no noise in the frame. For the current work, 10–40% salt-and-pepper noise is applied. This helps the model to be trained for unintentional and unwanted changes in the scenes. It also helps the model to correctly classify hand signals captured from low-quality cameras.

The reason for using these augmentation techniques is to better generalize the training dataset and to mimic the construction site environment, which is typically



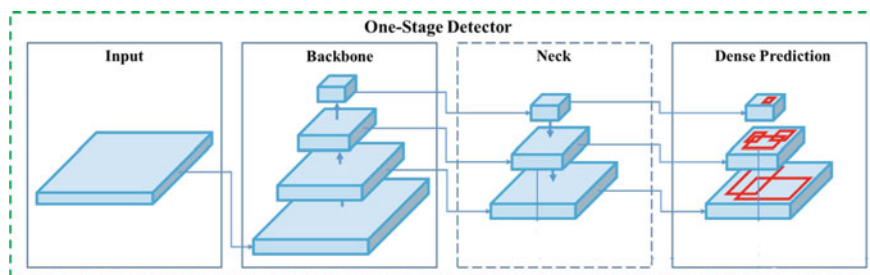
**Fig. 1** Example of different augmentation techniques

subject to changing weather, dusty conditions, among other dynamic and unpredictable conditions. After applying these transformations, the data augmentation techniques randomly choose each image, apply the transformation, and store the image as a different image. The example of discussed augmentation is shown in Fig. 1.

## 5 State-of-the-Art Yolov4 Model Architecture

YOLOv4 is a deep learning-based object detection and classification model capable of inspecting an image to find the subset of object class, enclosing it within bounding boxes, and identifying its class when trained [1]. YOLOv4 is the updated version of the YOLO algorithm series, which has been used in many applications in various fields, such as in the classification of safety helmets [10], construction vehicles [9], defects in sewer pipes [21], etc. The YOLOv4 model is selected for the present work based on its superior accuracy and speed compared to other deep learning models, such as YOLOv3, Faster R-CNN, SSD, and RetinaNet [1]. The YOLOv4 model architecture consists of three parts—the backbone, neck, and head—as shown in Fig. 2.

The backbone of YOLOv4 is composed of cross-stage partial connection CSPDarknet53. CSPDarknet53 is a convolutional neural network with residual connections that are responsible for extracting features from the image dataset [1]. The neck of YOLOv4 architecture is composed of spatial pyramid pooling (SPP)



**Fig. 2** Architecture of state-of-the-art YOLOv4 model

block [11] and path aggregation network (PANet) [13] is responsible to collect the feature maps from different stages in the network. The YOLOv4 network architecture used the same YOLOv3 head, aiming to predict objects in multiscale. The details about the architecture of YOLOv4 are given in [1]

## 6 Experimental Setup

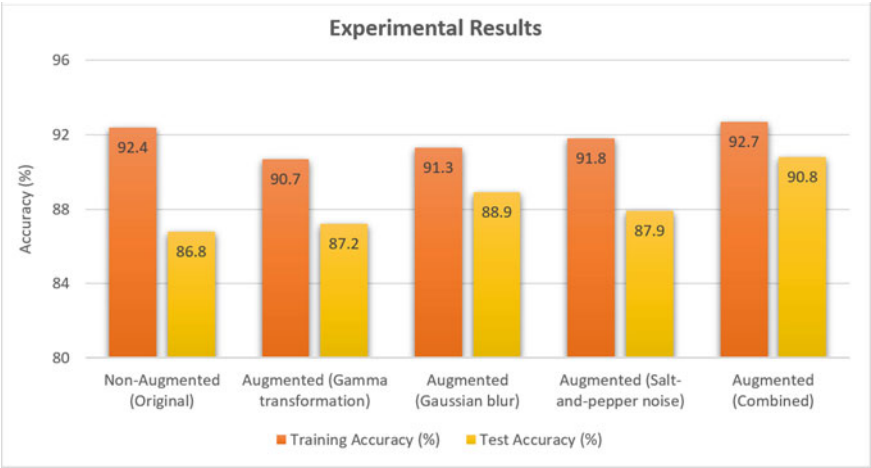
The YOLOv4 model has trained on ACID and construction site workers hand signals the image datasets. The images in the datasets are of varying resolutions. To feed the images in the YOLOv4 model, the resolution of the images is scaled down to  $416 \times 416$  pixels. The reason for scaling down the resolution of the images is to reduce the computational expense of the system. Initially, the model is trained on the original image datasets. In the original dataset, the images were used without any enhancement or augmentation to train the YOLOv4 model. After training the model, the training accuracy of the model is recorded. Further, the model is evaluated on the test dataset. The test dataset is the images that are not used while training the YOLOv4 model. Secondly, the model is trained on the images with gamma transformation. The detail about gamma transformation is discussed in Sect. 4.1. Here, the original images and some images with gamma transformation are used to train the model. The accuracy of the model is evaluated based on training and test dataset. Thirdly, the original dataset is used with Gaussian blur data augmentation. The detail about Gaussian blur is explained in Sect. 4.2. The YOLOv4 model is trained on the original images along with Gaussian blur images, and the performance of the model is evaluated on training and test datasets. Further salt-and-pepper noise data augmentation technique is implemented to the original image dataset. The original images and some images with salt-and-pepper noise are used to train the YOLOv4 model. The accuracy of the model is recorded on the training and test dataset. Finally, all the augmented techniques are combined and implemented on original images. Here, each augmented images have gamma transformation, Gaussian blur, and salt-and-pepper noise. The model is trained on the combined augmented and original dataset. The model is evaluated on training and test dataset accuracies. Note that



the same test dataset images are used to evaluate all the models while the models are trained on different datasets. As well as the same architecture of the YOLOv4 state-of-the-art model is used in this work without any changes or improvement.

### 7 Results and Discussion

A total of five experiments are conducted where the YOLOv4 model is trained on the non-augmented dataset, augmented with gamma transformation, augmented with Gaussian blur, augmented with salt-and-pepper noise, and finally, the model is trained with combined augmented images. The model is trained on 6500 epochs/iteration. Based on the experimental results as shown in Fig. 3, the state-of-the-art YOLOv4 model achieved the training accuracy of 92.4% on a non-augmented dataset where the model achieved an accuracy of 86.8% in the test dataset. When random transformation is implemented on the original image dataset like gamma transformation the training accuracy of 90.7% is recorded, while the model is used to evaluate the test dataset the model achieved an accuracy of 87.2%. When the YOLOv4 model is trained on, the images with Gaussian blur the model achieved an accuracy of 91.3 and 88.9% in training and test datasets. The YOLOv4 model is also trained on the images with salt-and-pepper noisy images; the model achieved a training accuracy of 91.8% and test accuracy of 87.9%. Further, by combining all the augmentation techniques and the YOLOv4 model is trained on the augmented combined dataset, the model achieved an accuracy of 92.7% in training. When the model is deployed for the test dataset, the model achieved an accuracy of 90.8% in the test dataset.



**Fig. 3** Comparison of YOLOv4 model accuracy between original and different augmented image datasets

The results showed that the YOLOv4 model achieved the highest accuracy in the augmented combined dataset in training as well as in the test dataset. For this case, we will be evaluating the performance of the models on the test dataset. The test dataset used in this work contains the same images while the models are trained on different datasets. The augmented combined dataset trained model achieved an accuracy of 90.8% in the test dataset, which is 4% better than the model trained with the non-augmented dataset. The combined augmented dataset trained model also performs better than the model trained on augmented with gamma transformation, Gaussian blur, and salt-and-pepper noise by 3.6%, 1.9%, and 2.9%, respectively. It is also noted that all models trained with an augmented dataset perform better than the model trained on a non-augmented dataset when evaluated on the test dataset.

These results clearly indicate that the data augmentation techniques implemented lead to an improvement in the classification results of the state-of-the-art YOLOv4 deep learning model. Data augmentation is an effective technique to increase the size of the dataset, better generalize the dataset, and help improve the classification accuracy of deep learning models. It also seems that the augmentation techniques chosen in the study to train the model perform satisfactorily to mimic the construction site changing scenarios. However, the implementation of data augmentation techniques to the image dataset is a time-consuming task. The applicability of data augmentation for video analysis of construction sites is under study for future work.

## 8 Conclusions

In this work, three different data augmentation techniques (gamma transformation, Gaussian blur, and salt-and-pepper noise) are used to evaluate the performance of the state-of-the-art YOLOv4 model with augmented and non-augmented construction site image datasets (ACID and construction site worker's hand signals). A total of five experiments are conducted. Initially, the state-of-the-art YOLOv4 model is trained on the original image dataset without any augmentation and evaluated based on the accuracy of the training and test dataset. Secondly, data augmentation is applied to the dataset one by one using gamma transformation, Gaussian blur, and salt-and-pepper noise. Finally, all the data augmentation techniques are combined, and the model is trained on a combined dataset and the performance of the model is evaluated. Based on the experiments conducted the highest test dataset accuracy is achieved by the state-of-the-art YOLOv4 model when trained on the combined dataset with 90.8%, which is 4% more than the model trained on the non-augmented dataset. In addition, the model performed better when trained on the augmented dataset, from the model trained on the non-augmented dataset. The result is convincing that the data augmentation technique is an effective way to mimic the construction site changing scenarios. Data augmentation also helped to better generalize the dataset as well as increased the size of the image dataset. The future work will include applying data augmentation techniques in videos and real-time classification in dynamic construction sites.

## References

1. Bochkovskiy A, Wang C, Liao HM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
2. Bonceleat C (2009) Image noise models. In: *The essential guide to image processing*. Academic Press, pp 143–167
3. Bull D (2014) Digital picture formats and representations. *Communicating Pictures* 99–132
4. Davis P, Aziz F, Newaz MT, Sher W, Simon L (2021) The classification of construction waste material using a deep convolutional neural network. *Autom Constr*, Elsevier 122:103481
5. Fang Q, Li H, Luo X, Ding L, Luo H, Rose TM, An W (2018) Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom Constr*, Elsevier 85:1–9
6. Gu S, Pednekar M, Slater R (2019) Improve image classification using data augmentation and neural networks. *SMU Data Sci Rev* 2(2):1–43
7. Hernandez-Garcia A, König P (2018) Further advantages of data augmentation on convolutional neural networks. *Springer Nature Switzerland*, pp 95–103
8. Hernández-García A, König P (2018) Further advantages of data augmentation on convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp 95–103.
9. Hou X, Zhang Y, Hou J (2021) Application of YOLO V2 in construction vehicle detection. Springer International Publishing, *Lecture Notes on Data Engineering and Communications Technologies*
10. Hu J, Geo X, Wu H, Gao S (2019) Detection of workers without the helmets in videos based on YOLO V3. In: *12th International congress on image and signal processing, biomedical engineering and informatics*, pp 1553–1560
11. Huang Z, Wang J, Fu X, Yu T, Guo Y, Wang R (2020) DC-SPP-YOLO: dense connection and spatial pyramid pooling based YOLO for object detection. *Inf Sci* 522:241–258
12. Lei C, Hu B, Wang D, Zhang S, Chen Z (2019, October) A preliminary study on data augmentation of deep learning for image classification. In: *Proceedings of the 11th Asia-pacific symposium on internetware*, pp 1–6
13. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path Aggregation network for instance segmentation. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp 8759–8768
14. Mikołajczyk A, Grochowski M (2018) Data augmentation for improving deep learning in image classification problem. In: *International interdisciplinary PhD workshop (IIPhDW)*. IEEE, pp 117–122
15. Misra S, Wu Y (2020) Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. *Mach Learn Subsurf Charact* 289
16. Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint [arXiv:1712.04621](https://arxiv.org/abs/1712.04621)
17. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *J Big Data*, Springer Int Publishing 6(1):1–48
18. Walawalkar D, Shen Z, Liu Z, Savvides M (2020) Attentive cutmix: an enhanced data augmentation approach for deep learning based image classification. In: *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings*, pp 3642–3646
19. Xiao B, Kang S-C (2021a) Vision-based method integrating deep learning detection for tracking multiple construction machines. *J Comput Civ Eng* 35(2):04020071
20. Xiao B, Kang S-C (2021b) Development of an image data set of construction machines for deep learning object detection. *J Comput Civ Eng* 35(2):05020005
21. Yin X, Chen Y, Bouferguene A, Zaman H, Al-Hussein M, Kurach L (2020) A deep learning-based framework for an automated defect detection system for sewer pipes. *Autom Constr* 109:102967

22. Yu Y, Li H, Umer W, Dong C, Yang X, Skitmore M, Wong AYL (2019) Automatic biomechanical workload estimation for construction workers by computer vision and smart insoles. *J Comput Civ Eng* 33(3):1–13

# An Integrated Approach Combining Virtual Environments and Reinforcement Learning to Train Construction Robots for Conducting Tasks Under Uncertainties



Weijia Cai, Lei Huang, and Zhengbo Zou

**Abstract** Robots can support onsite workers with repetitive and physically demanding tasks (e.g., bricklaying) to reduce workers' risk of injuries. Central to the wide application of construction robots is solving the task of motion planning (i.e., moving objects optimally from one location to another under constraints such as joint angle limits). Currently, robots are mostly deployed in the manufacturing phase of a construction project for off-site production of building components. Motions of these robots are pre-programmed and follow strictly designed trajectories and actions. However, the motions of robots on construction sites require considerations of uncertainties, including the onsite movement of material and equipment, as well as changes to workpieces and target locations of the work piece. Therefore, it is essential to enable construction robots to handle these uncertainties while executing construction tasks to extend their applicability onsite. In this study, we proposed an integrated approach combining virtual environments and reinforcement learning (RL) to train robot control algorithms for construction tasks. We first created a virtual construction site using a game engine, which allows for the realistic simulation of robot movements. Next, the physical characteristics of the workpiece (e.g., location) were randomized in the virtual environment to simulate onsite uncertainties. An RL-based robot control algorithm (i.e., Proximal Policy Optimization) was implemented to train the robot for completing a construction task. We tested the robustness and effectiveness of the approach using a testbed construction site for window installation. Results showed that the proposed approach is effective in training the construction robot arm to handle window installation under the uncertainties of window location, with a success rate of 75% for picking up (i.e., grasping) the window and a success rate of 68% for placing the window to its target placement without crashing into other objects onsite. Researchers and practitioners can use the proposed approach to train control algorithms for their specific construction tasks to allow for flexible robot actions considering onsite uncertainties.

---

W. Cai · L. Huang · Z. Zou (✉)  
University of British Columbia, BC, Canada  
e-mail: [zhengbo@civil.ubc.ca](mailto:zhengbo@civil.ubc.ca)

**Keywords** Construction robot · Reinforcement learning · Virtual environment

## 1 Introduction

The construction industry is known for its physically demanding, repetitive, and potentially dangerous tasks, including heavy lifting, drilling, and welding, all with potential exposure to dust, heights, and other unhealthy onsite conditions [1]. Such complicated workplaces have a high probability of occupational injuries. According to the US Bureau of Labor Statistics [2], the construction industry claimed around 20% of occupational fatalities among all industries in 2019. The unsafe work environment also directly contributed to the skilled labor shortage in the construction industry since the 1980s [3]. It is estimated that the British Columbia's construction industry will be at a deficit of 11,700 workers over the next two years and about 23,000 workers by 2029 [4]. As one of the potential solutions to these issues, construction robotics promises to bring higher productivity, safer working environments, lower costs, and superior quality, which mitigates the risk of onsite occupational injuries and can potentially relieve human workers from high-risk construction tasks [3, 5].

Many construction tasks can be broken down into smaller sub-tasks that include the fundamental step of motion planning [6], which aims to safely move construction materials or equipment to a target location. For instance, in the case of window installation, the motion planning task can be broken down as a human worker or a robot first identifies the location of the prefabricated window, holds up the window; and then moves and fits it to an opening. Currently, construction robots are mostly pre-programmed to work in an isolated environment (e.g., robotic wall panel assembly line in modular construction factories) [7]. However, pre-programmed robots are unable to adapt to onsite environments which can contain uncertainties such as the constant movements of construction material, equipment, and personnel [8]. Therefore, robotic motion planning on construction sites has higher requirements for sensing the surrounding environments and can often include on-the-fly decision-making for the robot, which increases the difficulty of robotic applications onsite [9].

To improve the flexibility of robotics operation in dynamic environments such as construction sites, reinforcement learning (RL) has been explored to enable optimal control of the robot actions [10]. In RL, an agent's goal is to learn actions according to its observed states and reward from the environment. Using window installation as an example, a robot arm can be seen as the agent; the states of the agent are the positions of joints and gripper on the robot arm; the actions of the agent are rotations of the joints. The goal of the RL agent is to obtain an optimal policy that maps the states to the actions (e.g., a smooth operation of picking up the window and then placing it at the target placement for window installation). RL agents learn the control policy through a series of trial and error, where the agent explores a large number of state-action pairs to experience different levels of reward. The final performance of the learned policy significantly relies on the degree to which the agents explore

[10]. However, it is costly, time-consuming, and potentially damaging to the robot to explore actions onsite, because the number of actions and states is exponentially proportional to the number of combinations of the robot arm joints [11]. With the advancement of 3D simulation, motions of objects can be efficiently simulated in physics-based game engines [12]. Therefore, the exploration of robots' actions and states can be accelerated in virtual environments built using game engines, without actual executions in the real world, which eliminates the possibility of damaging the robot or causing harm to onsite workers [11].

In this study, we proposed an approach that integrates RL and a virtual construction environment (VCE) to enable robots to complete tasks under uncertainties. The RL agent acts as the brain of the robots that make optimized actions, and the VCE provides simulations of scenarios for the agent to explore. Our approach includes three main steps. First, a realistic virtual construction site and a 6 degree-of-freedom (DoF) robot arm were modeled in a game engine (i.e., Unity3D). Next, the locations and physical characteristics of the objects were randomized to simulate the onsite uncertainties. Finally, an RL agent was trained in the simulated virtual environment to learn an optimized policy for construction tasks. To evaluate the effectiveness of the proposed approach, we used window installation as a test case.

## 2 Related Works

Construction robots have long been sought as a viable solution for automation in construction [13]. Previous works on construction robot applications involving motion planning can be summarized into two main categories. The first category revolves around implementing classic robotic motion planning methods. Classic motion planning methods, such as probabilistic roadmap motion planner [14] and exact cell decomposition [15], try to build up a graph consisting of sampled nodes, which represent reachable locations in the space close to the robot arm, and then search for the optimal path through these nodes from a start location to a target location. In a similar manner, recent studies [16] extended the idea to improve the efficiency of the search algorithms. These algorithms rely on improved sensing capacity of the robot to capture accurate geometric information of the environment (e.g., starting location, target location, and obstacles' shape and locations). For example, [17] used two laser rangefinders for an autonomous excavator robot to detect the topology of soil for completing the truck loading task. However, the resolution (i.e., 15 cm) of their detections was deemed unsafe for onsite personnel [18]. Later advancement of affordable non-contact vision sensors inspired researchers [8] to use computer vision algorithms for estimating the dynamic poses and locations of a robot arm and cube-shaped targets. However, their sensing technique is not applicable for construction tasks that involve small objects or require precise control of equipment, such as drilling, welding, and rebar installation.

Other than improving the sensing capacities of the robots, another direction of using classic robotic motion planning methods is to design task-specific robots with

the goal of simplifying the motion planning problems. For example, [19] designed a robotic wall construction system “JA-WA” that aims for executing concrete laying tasks for composite walls. The system utilized a supporting rail, assembled alongside the unfinished wall onsite. A mobile robot mounted on the rail moves alongside the wall to apply thermal insulation and cement. The supporting rail simplified the motion planning strategies because the workspace of the robots had been well structured. However, the supporting rail may have to be redesigned and assembled for new wall types, due to the changing shape and form of walls for different construction projects [8]. Similarly, [20] designed a cable-driven parallel robot (CDPR) working together with a robot arm for curtain wall module installation. However, the collaboration between the robot arm and the CDPR had to be programmed by experts for every construction project, which hinders its wide application.

To improve the flexibility of the robot, the second category of motion planning for construction robot aims to improve the decision-making process using RL. RL has been applied to a variety of robotic motion planning tasks, including pick-and-place and push-and-grasp [11, 21]. In general, RL-based motion planning produces control policies (a series of actions by the robot) by using states of the robots (e.g., joint angles and gripper position) and states of the environment (e.g., location of the object to be picked up) as input and receive rewards based on a pre-defined reward function by executing the learned policy. Through well-designed reward functions, RL improves robots’ ability to adapt to a complex environment even only partial information of the target and environment is provided [22]. Because of RL’s adaptability, there is an emerging number of RL applications for robots in manufacturing [23–25]. However, the popularity of RL has not caught up in the construction domain [5]. A significant reason lies in the costly setup of robots on construction sites, and the inefficiency of the training process for RL given the need for exploration by the RL agent [5]. The problem is exacerbated when considering the unstructured nature of construction sites [8]. To address current limitations for training RL agents onsite, in this study, we proposed an integrated approach that combines virtual construction sites built using a game engine and RL algorithms to provide flexible and safe training environments for robots (Fig. 1).

### 3 Methodology and Experiment Design

The proposed approach includes two components: a VCE and an RL agent, controlling actions of a robot arm. An overview of our approach is shown in Fig. 2, which describes how the two components interact with each other to produce an optimal control policy for a window installation task using the robot arm. The VCE (i.e., Fig. 2a) includes a robot arm with a gripper as the end effector, a window as the Target, and an opening as the TargetPlacement (i.e., desired final locations for the Target). The RL agent (i.e., Fig. 2b) takes the states  $S_t$  and reward  $R_t$  as input and produces the action  $A_t$ , which is then sent back to the VCE for execution. The following sections will introduce the VCE and the RL agent components in detail.



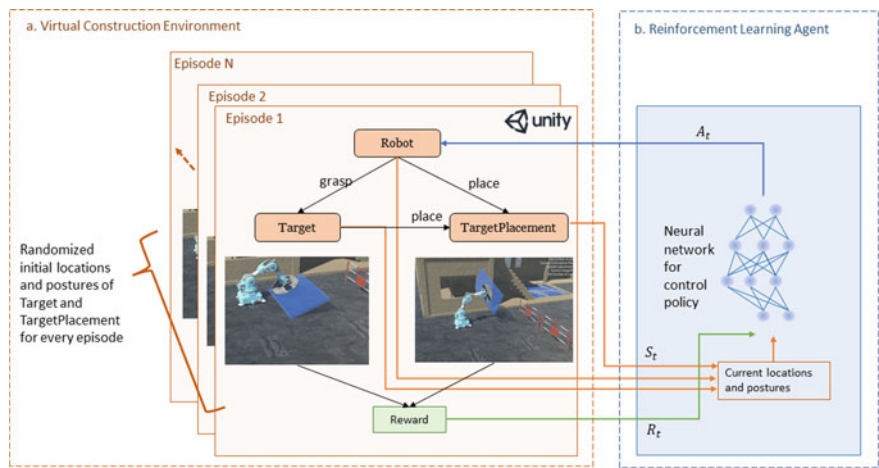


Fig. 1 Overview of the integrated approach

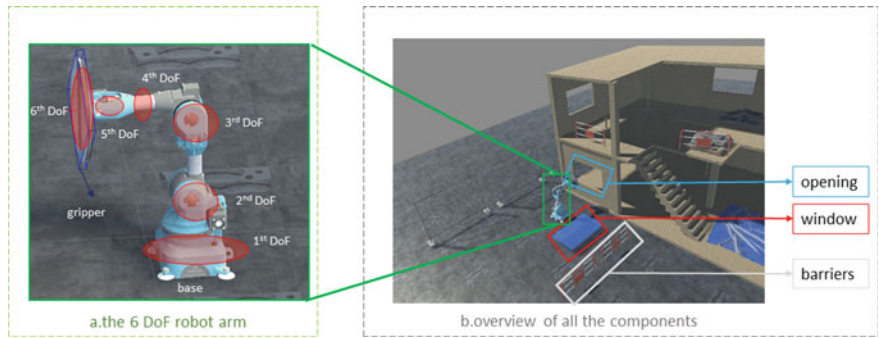


Fig. 2 Overview of the VCE

3.1 Virtual Construction Environment

In this section, we introduce the process of building the VCE using Unity3D, which includes three steps as: (1) set up physical characteristics and geometric constraints of the robot and other objects in VCE; (2) equip virtual sensors and apply detection techniques for the robot arm, and (3) link the RL agent controlling the robot to the VCE. In the first step, we designed the VCE (as seen in Fig. 2) to include a two-story building with openings for window installations, a cube-shaped window, several barriers for defining the reaching limit of the robot, and a 6 DoF robot arm (see Fig. 2a) placed in front of the opening. The configurations of the robot arm used in this study are detailed in Table 1. We assumed that the base of the robot arm is fixed so that the training processes will not fail due to unstable situations for the robot arm. We also assumed the payload of the robot is sufficient for lifting the window which

**Table 1** Mechanics-related parameters of the robot arm

Damping (Ns/m)	Stiffness (N/m)	Force limit (N)	Speed (m/s)	Torque (N)	Acceleration (m/s <sup>2</sup> )
100	10,000	100	10	100	2

has a mass of 12 kg. Next, the location constraint for the window was determined by placing down two barriers.

To detect whether the robot has achieved its goals, virtual sensors were installed on the end effector (i.e., gripper) of the robot arm. In this study, the window installation task can be broken down into two separate sub-tasks, as grasping the window and placing the window to the desired opening location. The virtual sensor used for these two sub-tasks is a ray-cast sensor, which is similar to an ultrasonic sensor in the real world. The ray-cast sensor was placed on the gripper, facing outward of the robot arm. It was used for determining whether the gripper had successfully grasped the window. Similarly, we also developed scripts in the game engine attached to the wall and window objects to detect if the wall had collided with other objects.

Finally, we linked the RL agent to the VCE by utilizing ML-Agents toolkits [26], which enables the RL agent to control the robot arm in the VCE. The ML-Agents has two main components for RL training as the Agent and the Brain. The Agent component allows us to receive observations from the environment and customize the reward function, whereas the Brain component defines the state and action space and provides decisions about the robot actions. During training, we randomized the rotation and position of the window and the rotation of the opening at the beginning of each training episode to make the learned policy more robust.

## 3.2 Reinforcement Learning Agent

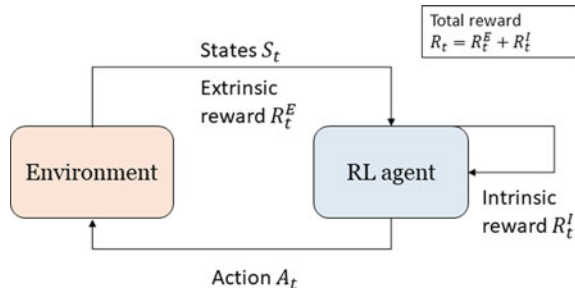
### 3.2.1 Reinforcement Learning Problem Formulation

The second component of the proposed approach is the RL agent, which learns to execute construction tasks through interacting with the VCE. In general, an RL agent learns the control actions by interacting with the environment that offers observation of states and reward (as seen in Fig. 3). Essentially, the agent aims to find an optimal policy  $\pi(A|S)$  that maximizes the expectation of the future reward from time  $t$  [10]:

$$\max E_{\pi(A|S)} \left( \sum_{k=0}^{\infty} R_{t+k+1} \times \lambda^k \right) \quad (1)$$

In Eq. (1),  $R_{t+k+1}$  is the reward at time  $t+k$ ;  $\lambda$  is the discount factor controlling the reduced reward of the current action in the future;  $A$  is the action, which in this study includes a discrete vector indicating the specific joint to rotate and a continuous

**Fig. 3** Interactions between an RL agent and the environment



vector controlling the values of the rotations for the selected joint; and  $S$  refers to states, which include joint position  $p_t^{\text{joint}}$ , gripper posture  $q_t^{\text{gripper}}$ , gripper velocity  $v_t^{\text{gripper}}$ , the target posture  $q_t^{\text{target}}$ , and the target placement posture  $q_t^{\text{place}}$ . To elaborate, a position vector  $p$  represents the 3D coordinate  $(x, y, z)$  of an object, while a posture vector  $q$  contains an object's 3D coordinate  $(x, y, z)$  and its orientation  $(r_x, r_y, r_z)$ .

To find the optimal policy, classic RL algorithms, such as value iteration, rely on the value function that essentially calculates the importance of each state [27]. After converging from value iteration, the value function is then used to determine an optimized sequence of actions [28]. Despite its wide application, value iteration suffered from slow convergence to the final policy [10]. To address this issues, Policy Gradient (PG) methods were developed to guarantee local optima by using neural network approximation to represent and learn the policy without consulting the value function [28]. However traditional PG methods are unstable, which means they do not guarantee steady policy improvements [29]. To combat this issue, later PG methods like Trust Region Policy Optimization (TRPO) [29] and Proximal Policy Optimization (PPO) [30] were developed to ensure monotonic policy improvements by maximizing a lower bound estimation of the neural network that generates the control policy, hence, we implemented PPO in this study.

### 3.2.2 Reward Function Design

One of the challenges for RL algorithms is the design of reward functions [31]. There exist two fundamental types of reward in RL algorithms, as extrinsic reward  $R_t^E$  gained from the environment and intrinsic reward  $R_t^I$  retrieved from the agent [32]. In this study, we designed the extrinsic and intrinsic rewards in the following manner. For extrinsic reward design, the basic logic is to give a positive reward when the agent completes a task (e.g., installed the window at the target placement) and to give a negative reward when the agent fails at the task or takes too long to complete the task [33]. It is also important to balance the extreme values for positive and negative rewards for more stable learning. Based on these considerations, we designed our extrinsic reward function in the form of Algorithm 1, as shown in Fig. 4.

As seen in Algorithm 1, the extrinsic reward is the cumulated reward  $R_t^E$  from every episode during training. The components of the extrinsic reward function

---

**Algorithm 1:** Extrinsic reward function

---

```

1 for every episode  $t$  do
    Initiate extrinsic reward  $R_t^E$  for current episode
    for  $i = 1$  to  $MaxStep$  do
         $R_t^E = R_t^E - \frac{1}{MaxStep}$ 
        if Grasped then
             $R_t^E = 1$  (only reward for the first time)
            if Placed then
                 $R_t^E = 2$ 
            End current episode
        else if Crashed then
             $R_t^E = -0.5$ 
            End current episode
        else
            continue
    
```

---

**Fig. 4** Designed extrinsic reward function

include: (1) a time penalty of  $\frac{1}{MaxStep}$  for every action step taken, where  $MaxStep$  refers to the maximum number of actions the agent can take in one episode; (2) a reward of 1 for successfully grasping the target; (3) a reward of 2 for successfully placing the target; and (4) a penalty of 0.5 for allowing the robot or the target to crash with other objects.

In addition to the extrinsic reward, we also adopted a curiosity-based intrinsic reward  $R_t^I$  to enable the robot agent to explore unfamiliar states [32], which was inspired by the human nature of curiosity. The authors proposed the Intrinsic Curiosity Module (ICM) including a feature extractor, an inverse model, and a forward model. The feature extractor encodes current state  $S_t$  and the next state  $S_{t+1}$  into features  $\omega(S_t)$  and  $\omega(S_{t+1})$ . The inverse model takes learned features  $\omega(S_t)$  and  $\omega(S_{t+1})$  as input and produces  $\hat{A}_t$  as a prediction of the current action  $A_t$  from the current policy. Then the forward model takes  $A_t$  and  $\omega(S_t)$  and outputs  $\hat{\omega}(S_{t+1})$  as a prediction of  $\omega(S_{t+1})$ . Finally, the intrinsic reward is defined as the difference between  $\hat{\omega}(S_{t+1})$  and  $\omega(S_{t+1})$ , as shown in Eq. 2.

$$R_t^I = \frac{\delta}{2} \times \|\omega(S_{t+1}) - \hat{\omega}(S_{t+1})\|^2 \quad (2)$$

In Eq. 2,  $\delta$  is the strength factor that scales the intrinsic reward to prevent it from being overwhelmed by the extrinsic reward. For a specific state-action pair  $\{s, a\}$  and the next state  $s'$  from executing  $a_t$ , the difference between  $\hat{\omega}(s')$  and  $\omega(s')$  (i.e., the intrinsic reward) would be larger, if the agent takes explorative actions.

**Table 2** Curriculum learning (CL) training paradigm parameters

Training criteria	Tolerance of reaching target (m)	Tolerance of reaching TargetPlacement (m)	Consider crashing
cp1*	0.15	0.30	No
cp2	0.15	0.30	No
cp3	0.15	0.15	Yes
cp4	0.10	0.15	Yes

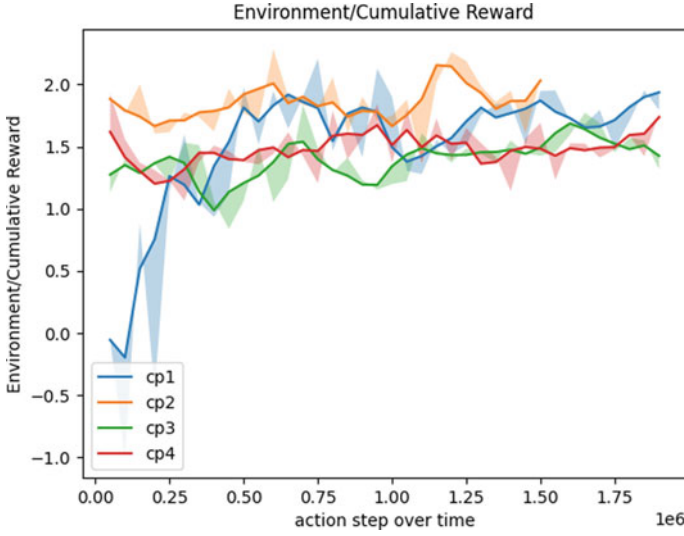
\* cp stands for curriculum plan

**3.2.3 Reinforcement Learning Agent Training**

The training paradigm is vital for RL agents, since well-structured training can ensure smooth reward improvements and a quick convergence. In this study, we used curriculum learning (CL) [34], which applies a sequence of training criteria (e.g., tolerance of reaching the target) that ascend in difficulty. The general idea of CL is that after perfecting a low-level criterion, the agent enters the next level and continues training until meeting all criteria. In this study, we focused on three adjustable factors in our training paradigm, as shown in Table 2. We first decreased the tolerance of reaching the TargetPlacement (i.e., the opening for the window) in the first two levels of criteria (i.e., cp1 and cp2 in Table 2) from 0.30 to 0.15 m, so that the agent can learn reaching the TargetPlacement more precisely. Then, we added the detection of crashing between the Target (i.e., the window) and the wall as the third-level criterion (i.e., cp3). Finally, we decreased the tolerance of reaching the Target from 0.15 to 0.10 m as the final criterion (i.e., cp4). By following this training paradigm, the agent can gradually improve its performance during training rather than trying to perfect the task at once.

**4 Results and Discussions**

In this study, we used window installation as a demonstrating construction task to show the flexibility of the RL agent and the effectiveness of our approach. Similar to a pick-and-place task, the window installation task can be simplified into two sub-tasks as: (1) the gripper reaching and grasping the window and (2) the window reaching the desired opening. The agent was instructed to learn the two sub-tasks within a single episode. Four RL agents were trained according to the training paradigm in Table 2. Results of the training are shown in Fig. 5. Each line in Fig. 5 represents the cumulative reward for one agent over 2 million action steps taken. The graph shows a trend of convergence for each line. The shading represents the raw cumulative reward for each agent, while the solid line represents the moving average with a step size of 5000. The final rewards of the first two agents (i.e., cp1, cp2) are around 1.7, with the upper bound of reward being 3, consisting of 2 for the placing reward plus



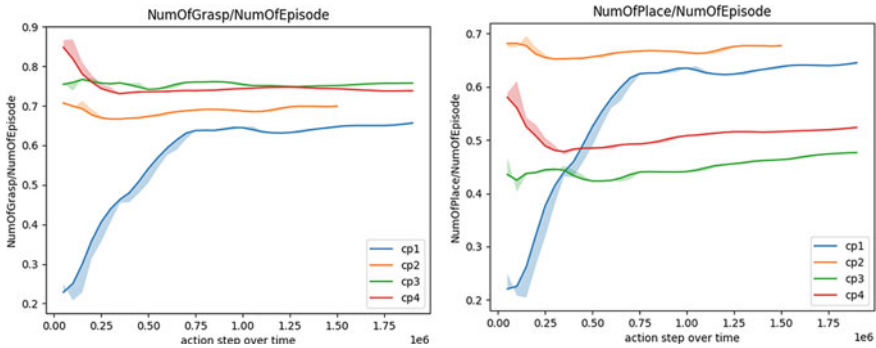
**Fig. 5** Cumulative reward over 2 million action steps taken

1 for the grasping reward. The last two agents (i.e., cp3, cp4) have a slightly lower reward at 1.5, but have a lower probability for crashing.

To illustrate the effectiveness of the robot arm at achieving the two sub-tasks, we also calculated the success rates of grasping and placing using: (1) the proportion of successful grasping, i.e.,  $\frac{N_{\text{grasp}}}{N_{\text{episode}}}$  and (2) the proportion of successful placing, i.e.,  $\frac{N_{\text{place}}}{N_{\text{grasp}}}$ . Figure 6 shows that the results of the grasping success rate increase from 65% for cp1 to 75% for cp4. The placing success rate increased from 63% for cp1 to 67% for cp2 and increased from 46% for cp3 to 51% for cp4. The placing success rate decreased between cp2 and cp3 because of the constraint of adding crashing to the agent, which added difficulty to the task. Overall, the results showed a high success rate for both grasping and placing while avoiding crashing, which indicates the effectiveness of the trained RL agent and its flexibility for dealing with randomized locations of target and avoiding obstacles, without the explicit instructions from human experts.

## 5 Conclusion and Future Works

In this study, we proposed an approach that integrates an RL agent with the VCE for training a robot arm to conduct construction tasks. A realistic virtual construction site was built using Unity3D. An RL agent was assigned to control the robot arm. We used the PPO algorithm to train the RL agent and used curriculum learning as the training paradigm. Results showed that our approach enabled the robot to become flexible and robust under uncertain conditions (e.g., uncertain locations of the target)



**Fig. 6** Cumulative ratio  $\frac{N_{grasp}}{N_{episode}}$  and  $\frac{N_{place}}{N_{episode}}$  along action step during the training

in the VCE. The success rate of grasping reached around 75%, and the success rate of placing reached around 68% without crashing.

The main contribution of this study is the proposed approach. This approach is scalable because the time and capital investment for building a virtual construction site are low as compared to implementing the robotic solution on a real construction site. The virtual construction site built in a physics-based game engine can also provide realistic feedback to the robots [35, 36] and prevent unsafe incidents in which workers come into contact with the robots. This approach is also flexible because the physical characteristics of the virtual construction sites (e.g., position, dimension, the weight of the target, and the friction effect of the construction product) can be modified, catering to the need of specific construction tasks such as ceiling installation, framing, and paneling. The fact that the robot arm learned to achieve a window installation task in the test environment without any explicit instructions from human experts opens doors for future applications of the proposed approach to train robots for more complex tasks.

For future work, we aim to explore the possibility of limiting information flow from the environment to the robot control agent (e.g., location of the target and wall), hence requiring the robot to infer the environment states from non-contact sensors, which will create an even more realistic training environment for the robot. Sensors such as cameras and laser scanners will be investigated to enable information provision for the robot arm.

## References

1. Arndt V, Rothenbacher D, Daniel U, Zschenderlein B, Schuberth S, Brenner H (2005) Construction work and risk of occupational disability: a ten year follow up of 14,474 male workers. *Occup Environ Med* 62(8):559–566. <https://doi.org/10.1136/oem.2004.018135>
2. Bureau of Labor Statistics (2019) Fatal occupational injuries by industry and event or exposure. US Department of Labor. <https://www.bls.gov/iif/oshwc/cfoi/cftb0331.htm>

3. Karimi H, Taylor TR, Dadi GB, Goodrum PM, Srinivasan C (2018) Impact of skilled labor availability on construction project cost performance. *J Constr Eng Manag* 144(7):04018057
4. Vancouver Regional Construction Association (2019) Skilled labor shortage project. <https://vrca.ca/advocacy/skilled-labour-shortage/>
5. Liang CJ, Kamat VR, Menassa CC (2020) Teaching robots to perform quasi-repetitive construction tasks through human demonstration. *Autom Constr* 120:103370
6. Lundeen KM, Kamat VR, Menassa CC, McGee W (2019) Autonomous motion planning and task execution in geometrically adaptive robotized construction work. *Autom Constr* 100:24–45
7. Yang CH, Kang SC (2021) Collision avoidance method for robotic modular home prefabrication. *Autom Constr* 130:103853
8. Feng C, Xiao Y, Willette A, McGee W, Kamat VR (2014) Towards autonomous robotic in-situ assembly on unstructured construction sites using monocular vision. In: *Proceedings of the 31th international symposium on automation and robotics in construction*, pp 163–170
9. Everett JG, Slocum AH (1994) Automation and robotics opportunities: construction versus manufacturing. *J Constr Eng Manag* 120(2):443–452
10. Sutton RS, Barto AG (2018) *Reinforcement learning: an introduction*. MIT press, New York
11. James S, Johns E (2016) 3D simulation for robot arm control with deep q-learning. *arXiv preprint arXiv:1609.03759*
12. You S, Kim JH, Lee S, Kamat V, Robert LP Jr (2018) Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments. *Autom Constr* 96:161–170
13. Paulson BC Jr (1985) Automation and robotics for construction. *J Constr Eng Manag* 111(3):190–207
14. Kavradi LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans Robot Autom* 12(4):566–580
15. Latombe JC (1991) *Exact cell decomposition. Robot motion planning*. Springer, Boston, MA, pp 200–247
16. Elbanhawi M, Simic M (2014) Sampling-based robot motion planning: a review. *IEEE Access* 2:56–77
17. Stentz A, Bares J, Singh S, Rowe P (1999) A robotic excavator for autonomous truck loading. *Auton Robots* 7(2):175–186
18. Telecommunications Industry Association and Electronic Industry Association (TIA/EIA), *Standard for Physical Location and Protection of Below-Ground Fiber Optic Cable Plant*. ANSI/TIA/EIA-590-A-1996
19. Więckowski A (2017) “JA-WA”-A wall construction system using unilateral material application with a mobile robot. *Autom Constr* 83:19–28
20. Iturralde K, Feucht M, Hu R, Pan W, Schlandt M, Linner T, Bock T, Izard JB, Eskudero I, Rodriguez M, Gorrotxategi J (2020) A cable driven parallel robot with a modular end effector for the installation of curtain wall modules. In: *ISARC Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 37. IAARC Publications, Munich, pp 1472–1479
21. Zeng A, Song S, Welker S, Lee J, Rodriguez A, Funkhouser T (2018) Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In: *Proceedings of the 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, New York, pp 4238–4245
22. Jaakkola T, Singh S, Jordan M (1994) Reinforcement learning algorithm for partially observable Markov decision problems. *Adv Neural Inform Process Syst* 7:173
23. Thomas G, Chien M, Tamar A, Ojea JA, Abbeel P (2018) Learning robotic assembly from cad. In: *Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, New York, pp 3524–3531
24. Luo J, Solowjow E, Wen C, Ojea JA, Agogino AM, Tamar A, Abbeel P (2019) Reinforcement learning on variable impedance controller for high-precision robotic assembly. In: *Proceedings of the 2019 international conference on robotics and automation (ICRA)*. IEEE, New York, pp 3080–3087



25. Apolinarska AA, Pacher M, Li H, Cote N, Pastrana R, Gramazio F, Kohler M (2021) Robotic assembly of timber joints using reinforcement learning. *Autom Constr* 125:103569
26. Juliani A, Berges VP, Teng E, Cohen A, Harper J, Elion C, Goy C, Gao Y, Henry H, Mattar M, Lange D (2018) Unity: a general platform for intelligent agents. *arXiv preprint [arXiv:1809.02627](https://arxiv.org/abs/1809.02627)*
27. Pashenkova E, Rish I, Dechter R (1996) Value iteration and policy iteration algorithms for Markov decision problem. In: *Proceedings of the AAAI'96: workshop on structural issues in planning and temporal reasoning*
28. Sutton RS, McAllester D, Singh S, Mansour Y (1999) Policy gradient methods for reinforcement learning with function approximation. *Adv Neural Inform Process Syst* 12:173
29. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: *International conference on machine learning*. PMLR, Singapore, pp 1889–1897
30. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. *arXiv preprint [arXiv:1707.06347](https://arxiv.org/abs/1707.06347)*
31. Ng AY, Russell SJ (2000) Algorithms for inverse reinforcement learning. In: *ICML*, vol 1, p 2
32. Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-driven exploration by self-supervised prediction. In: *International conference on machine learning*. PMLR, Singapore, pp 2778–2787
33. Xie L, Wang S, Rosa S, Markham A, Trigoni N (2018) Learning with training wheels: speeding up training with a simple controller for deep reinforcement learning. In: *Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, New York, pp 6276–6283
34. Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: *Proceedings of the 26th annual international conference on machine learning*, pp 41–48
35. Kurien M, Kim MK, Kopsida M, Brilakis I (2018) Real-time simulation of construction workers using combined human body and hand tracking for robotic construction worker system. *Autom Constr* 86:125–137
36. Rahimian FP, Seyedzadeh S, Oliver S, Rodriguez S, Dawood N (2020) On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Autom Constr* 110:103012

# Investigating the Feasibility of Using Virtual Reality Devices to Present Construction Information in Both Mixed Reality and Virtual Reality Environments



Hardith Suvarna Murari, Zia Din, and Christiane Spitzmueller

**Abstract** Mixed reality (MR) and virtual reality (VR) technologies are frequently recommended to present design models of construction projects to perform various visualization tasks, such as improving hazard identification training and constructability analysis. Significant financial investment is commonly highlighted as a major impediment to using these technologies, particularly MR. The cost of purchasing head-mounted display (HMD) hardware, such as Microsoft HoloLens for MR, Oculus Quest, or HTC Vive for VR, is substantial. However, the Oculus Quest manufacturer states that, due to recent software improvements, users can now deploy and run both MR and VR applications on the same Oculus Quest HMD, which was initially built only for the VR environment. This development appears to provide significant financial savings to those in the Architecture, Engineering and Construction (AEC) industry who are interested in gaining access to MR and VR. This article explores the possibility of using the Oculus Quest device to deploy MR and VR environments in construction. First, 3D models required for the experiment were created using Autodesk Revit. Then two applications were created using the Unity game engine, the Mixed Reality Toolkit for MR, and the Oculus Quest software development kit for VR. To ensure that the results were comparable, only one type of device and one construction task model were used in MR and VR settings. The MR and VR functions of Oculus Quest HMD were tested. Data on performance and user feedback were collected and analyzed. MR was easier to use than VR for this particular experiment because MR allowed the user to walk freely and view the room in real time, better understand the task, and get better details, accuracy, and scale. This way, one device can provide both environments, MR and VR.

---

H. S. Murari

Department of Computer Science, University of Houston, Houston, TX 77004, USA

Z. Din (✉)

Department of Construction Management, University of Houston, Houston, TX 77004, USA

e-mail: [uziauddi@central.uh.edu](mailto:uziauddi@central.uh.edu)

C. Spitzmueller

Department of Psychology, University of Houston, Houston, TX 77004, USA

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_18](https://doi.org/10.1007/978-3-031-34593-7_18)

**Keywords** Construction safety · Augmented reality · Virtual reality · Building information modeling · Construction project

## 1 Introduction

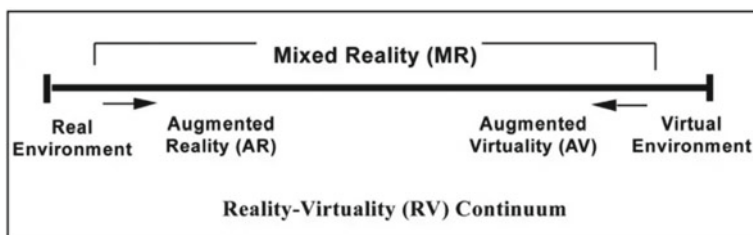
To improve hazard detection training and constructability analysis, mixed reality (MR) and virtual reality (VR) technologies are widely advocated for displaying design models of construction projects [1]. Virtual Design and Construction models (e.g., building information models) can present the steps involved in the construction process of a particular task that can enable users to visually identify and mitigate safety risks before facing the risk on the site [2]. However, one of the most typical hurdles to employing these technologies, particularly MR, is a significant financial commitment [1]. Head-mounted display (HMD) devices, such as Microsoft HoloLens for MR, are expensive.

According to the Oculus Quest manufacturer [3], users can now deploy and operate MR and VR applications on the same Oculus Quest HMD, initially designed exclusively for the VR environment [4]. This development appears to bring significant cost savings to those in the AEC business who want to use MR and VR. This article explores the possibility of using the Oculus Quest device to deploy MR and VR environments in construction. Oculus Passthrough is a new MR technology that allows users to see their hands and other objects in the virtual world [4]. This new technology offers a more natural experience with virtual reality by adding more depth. HoloLens, on the other hand, is an augmented reality device that provides information through holograms rather than virtual images. It also offers spatial sound and voice recognition capabilities [5]. This article will compare the two technologies and document the experience of using the same device for VR and MR purposes for AEC.

## 2 Literature Review

### 2.1 *Virtual Reality (VR), Mixed Reality (MR), and Augmented Reality (AR)*

VR, MR, and AR are new technologies that allow users to engage in a range of digital (artificial) immersions. A head-mounted display for virtual reality provides an immersive environment [6]. The VR material is virtual and created by a computer. A new 3D digital world replaces the current reality, isolating the user from reality. AR is a technique that overlays computer-generated content on top of the real environment.



**Fig. 1** Milgram's virtuality continuum [7]

AR is typically accessed through a smartphone application or a wearable glass device [6]. MR is a wearable gadget that blends many technologies into one. MR glasses or headsets display a digital overlay that interacts in real time with items in the natural environment [6]. Although most of the things are still in the research and development stage, MR can be viewed through transparent glasses. Milgram and Fumio [7] presented a virtuality continuum that is used to explain various types of real and digital worlds, as shown in Fig. 1.

On the one hand, there is the circumstance where nothing is virtual, but the real world is being portrayed, which is on the left. The opposite extreme, on the right, is the user's view of a completely simulated environment. This is essentially a virtual reality realm where the user is completely immersed. The spectrum was designed to identify three concepts: enhanced virtual reality, MR, and AR, which are sandwiched between the "real environment" and "virtual environment."

## 2.2 MR Devices

Mixed reality technology scans the real world in order to integrate computer-generated objects into it [6]. Compared to virtual reality and AR headsets, MR headsets are significantly less common. Table 1 lists different MR devices and their features [8].

### 2.2.1 Microsoft HoloLens

Microsoft HoloLens is a standalone device that runs without additional power [5]. It comprises three different sensors and five cameras to provide MR. HoloLens has a 50-degree field of view. Figure 2 is an image of Microsoft HoloLens.

**Table 1** Comparison of MR devices (as of February 2022)

Name	Resolution (per eye)	Price	Pros	Cons
Microsoft HoloLens	$1268 \times 720$	\$3000 \$5000 (commercial suite)	Microsoft platform support, controlled with hand movements and voice command	Price, small field of view
Microsoft HoloLens 2	$2560 \times 1440$	\$3500 \$125 (per month enterprise) \$99 (per month developer)	High resolution, Microsoft platform support, comfortable	Price, battery life
Magic leap 1	$1280 \times 960$	\$2295	Field of view is more, and controller is used for movements	Price, lack of applications available

**Fig. 2** Microsoft HoloLens

### 2.2.2 Magic Leap

Magic Leap [9] is a mixed reality headset. It uses a digital light field supported by Google and Qualcomm to guide light directly into the player's eye. When it is published, it is expected to be a game changer. Figure 3 is an image of a Magic Leap device.

### 2.2.3 Other MR Devices

Apple and Google have many MR and VR headsets patents and are working with Samsung to create MR devices.



**Fig. 3** Magic leap

### **2.3 VR Devices**

Varying degrees of options are available in virtual reality HMDs that are used to display virtual reality worlds. Virtual reality devices can be classified as standalone or tethered. Following a survey of several articles, the authors decided to build an application for a standalone device, Oculus Quest 2. The system requirements for Oculus Quest 2 are a 2.0 + GHz central processing unit (CPU) and 2 GB of system RAM. Standalone VR has various benefits, such as low-cost full mobility (Yevhenii Ratushnyi [10]. Table 2 compares the different standalone VR devices, and it is adopted from [11].

### **2.4 MR and VR in Construction**

In the AEC industry, effective training in identifying hazards and analyzing constructability is essential to instill a culture of safety that reduces risks and hazards in construction projects. However, investigations have revealed that the lack of fundamental safety information and procedures is one of the leading causes of construction accidents [12]. These investigations identified serious weaknesses in the current safety management system, education, and training, resulting in unprepared workers for on-the-job training [13]. In a traditional classroom setting, safety management concepts are difficult to convey using 2D media, such as text, videos, and photographs [14].

In recent years, technological advancements have increased the practicality and use of virtual reality and mixed reality in the AEC business. The use of virtual reality provides an immersive experience of task-level constructability analysis. For example, constructing a reinforced concrete column can be modeled at the detail level, where the digital model shows the formwork, reinforcement bars, and the temporary structure. This detailed three-dimensional model can be imported into a program such as Autodesk Navisworks or Fuzor [15] to create a construction sequence. Simulation of concrete column construction will show the formwork assembly and placement

**Table 2** Comparison of standalone VR devices (as of February 2022)

Name	Display	Resolution (per eye)	Price	Pros	Cons
Pico G2 4 K	LCD	1920 × 2160	\$300 (32 GB) \$350 (128 GB)	High-resolution display, price	Lack of applications available, 3 degrees of freedom
Oculus quest	OLED	1440 × 1600	\$399 (64 GB) \$499 (128 GB)	6 degrees of freedom, positional tracking, price, Passthrough API	Not comfortable for long periods, battery life, and controllers are not rechargeable
Oculus quest 2	LCD	1832 × 1920	\$299 (128 GB) \$399 (256 GB)	6 degrees of freedom, positional tracking, price, high-resolution display, Passthrough API	Battery life, controllers are not rechargeable, and not comfortable for long periods
HTC vive focus	AMOLED	1440 × 1600	\$599	Comfortable, and movement tracking is good	3 degrees of freedom, low storage of data
HTC vive focus 3	LCD	2448 × 2448	\$1300	Field of view is wide, comfortable, high resolution	Price, and performance is not good
Pico Neo CV	LCD	1280 × 1440	\$239 (lite) \$540 (standard)	6 degrees of freedom, price, comfortable	Lack of applications
Pico VR goblin	LCD	1920 × 2160	\$269	Price, high resolution, comfortable	Lack of applications, 3 degrees of freedom
Pico Neo 3 Pro	LCD	1832 × 1920	\$677	Field of view is wide, high resolution, comfortable	Battery life, lack of applications
Lenovo mirage solo	LCD	1280 × 1440	\$450	Battery life, field of view is wide, and good positional tracking	Display not good, heavy, 3 degrees of freedom

of the reinforcement bars and then the concrete pouring into the formwork. Typical hazards in reinforced concrete column construction are concrete that can cause bodily injuries, falls on the deck or off the edge of the deck, cuts from reinforcement bar ends or tie wire, concrete splashes in the eye, skin rashes, and/or allergies [16]. Visualization tools let you identify such issues in advance using a virtual model.

Workplace safety management is a prevalent topic researched using virtual reality and MR educational programs [17]. VR and MR technology have many advantages over traditional educational media for safety management, AEC skills, and training. Unlike VR and MR tools, traditional two-dimensional (2D) media cannot deliver virtual 3D simulations, enhanced presentations of architectural information, or virtual environments. Because most data in the AEC industry are inherently 3D, 2D visualizations are more conceptual and more complicated for clients to understand [18].

In education, 3D representations can engage students more than 2D material. Users can interact with virtual objects and scenarios in VR and MR, providing a more proactive educational experience than traditional 2D media [17].

## 2.5 *Oculus Quest 2 for MR and VR*

Mixed reality headsets, such as HoloLens, come at a high price. On the other hand, Oculus Quest 2, shown in Fig. 4, is much less expensive. Oculus Quest 2 was originally designed for virtual reality experiences, but the Oculus team has recently released a Passthrough API that can be used to build MR apps. In Oculus headsets, Passthrough gives a perceptually pleasant, real-time 3D representation of the actual environment, which is not present in other devices other than Oculus Quest devices. Passthrough is a virtual reality feature that lets users see a live perspective of their environment beyond their eyesight. Passthrough uses the sensors in your headset to mimic what one would see if you could see directly through the front of your headset and into the real world. Passthrough displays right away while creating or modifying your Guardian. The apps can also display Passthrough to blend your physical and virtual surroundings. Developers can use the Passthrough API to incorporate Passthrough imagery into their virtual experiences.

The Oculus team has also provided some samples on the development of MR apps using Passthrough. Passthrough samples [3] are found in the SampleFramework folder of the Unity Oculus Integration SDK. A Passthrough scene may be found

**Fig. 4** Oculus quest 2





in the Assets/Oculus/SampleFramework/Usage folder, showing the most significant functions and ideas in one scene. Furthermore, example scenes for various features or concepts may be found in the Assets/Oculus/SampleFramework/Usage/Passthrough/Scenes folder. The first introduces the Passthrough API's capabilities, while the second and third are helpful places to learn about individual features.

### **3 Development of MR and VR Applications for Experimentation**

#### ***3.1 Development of the 3D Model***

The authors used Autodesk Revit [19] to create the model, including all materials and textures. The models were then saved as 3D files. The authors discovered that if the model were exported in fbx format and then imported into Unity, the model's materials and textures would be lost. In 3ds format, the model could be imported into Unity without losing its materials and textures.

#### ***3.2 Development of VR Application***

The authors selected Unity as the game engine for the virtual reality application to be built for Oculus Quest 2. Unity 2020.3.14f1 (LTS) [20] was used to develop the VR application. The Unity game engine requires the authors to be familiar with C# (c-sharp), the programming language supported by Unity. The developer may use Unity Hub to manage the project's progress and team cooperation. The authors followed Justin [21] initial setup of the application development outlined in a YouTube video. The video covers building a VR environment using OpenXR [22], managing extended reality (XR) plugins, and using XR interaction tools.

#### ***3.3 Development of MR Application***

The authors again selected Unity as the game engine for the MR application to be built for Oculus Quest 2. The authors used the same initial setup for VR, with Unity 2020.3.14f1 (LTS) game engine [20], C# as the programming language for developing the application, and Unity Hub. The authors followed Dilmer Valecillos's video of "Passthrough API Hand Tracking with OpenXR and Oculus Quest 2" [23] to do the initial setup of the Unity environment for Passthrough API. The video provides structured steps to create a Unity project and add all the XR components, including Oculus SDK, XR Management Plugin, Oculus XR Plugin, etc. The video

also explains the various features of the Passthrough API, such as the OVRCameraRig experimental features and additional settings.

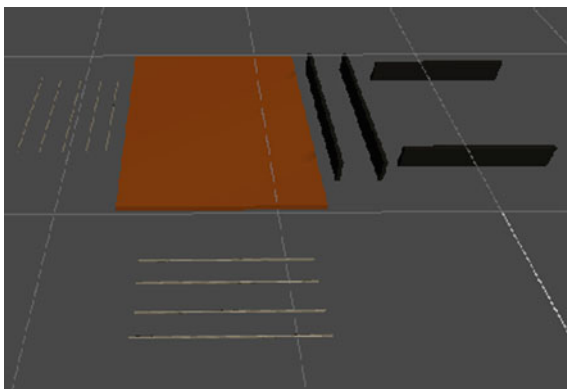
The Oculus team has also provided a detailed explanation of the Passthrough API [4] on how it works, the prerequisites needed, how to configure the Unity project, how to create the Passthrough layer, and how to customize Passthrough. The documentation is also frequently updated. They have also provided Passthrough samples [3] of various scenes, which is helpful for developers to learn more about the Passthrough API.

## 4 Testing

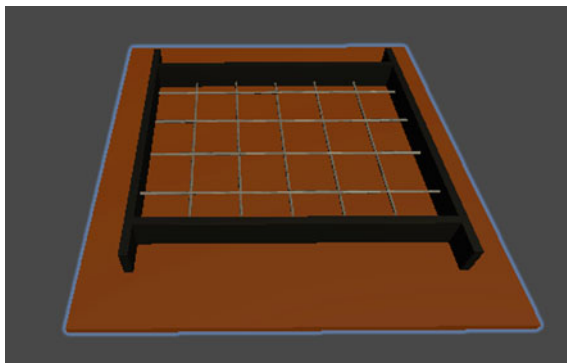
The goal of the experiment was to create a reinforcing mesh by placing individual rebars in a mesh form. The 3D model in Fig. 5 was before the user interacted with digital objects such as rebars and formwork. The 3D model in Fig. 5 was built in Autodesk Revit and imported into Unity in 3D format. The users used the controllers of Oculus Quest 2 and, using the ray cast (laser) interactor, were able to grab the individual components and move them around; using these controllers, the users grabbed the individual components to interact with the various components to complete the task. The 3D model in Fig. 6 shows after the user completed the reinforcement mesh fabrication and the placement of the formwork.

The experiment was carried out by a student who is pursuing a master's degree in construction. The user must sign a consent form before participating in the study. As part of the experimental protocol, the participant was first told what the purpose of the experiment was. Subsequently, the user received a brief overview of how Oculus Quest 2 works with MR and VR. The user was given 10 min in both MR and VR to learn about the various controls and features and to become acquainted with the appropriate environment. The experiment was initially carried out in VR and then

**Fig. 5** Material before the user interacts



**Fig. 6** After the user completes the experiment



in MR. The user was invited to complete a post-experiment questionnaire form to receive feedback after completing the experiment.

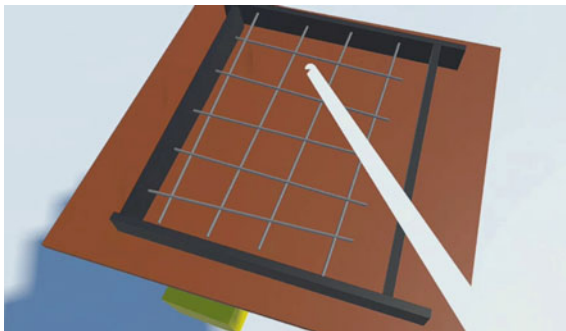
#### **4.1 VR Results**

The user spent 7 min in total performing the reconstruction activity in virtual reality. The user stated in his feedback that the identification of elements and materials of the object in VR was achievable and that the VR device allowed the participant to play with the 3D models in the VR environment, which allowed the participant to construct the tasks of the experiment faster than the user could in the real world. According to the user, the VR environment also helped gain a better conceptual understanding of its operation. The user also mentioned how challenging it was to obtain 100% accuracy when performing the assignment. The user commented that conducting a precision-focused activity would be difficult and that better haptics and user interaction in VR would be required to complete such activities. Figure 7 shows the result of the VR experiment. The placement of formwork sides and rebars in VR was not accurate, as well as the cover space of the rebars compared to Fig. 6 visually; when we compare Figs. 6 and 7, the completion of tasks in VR was not highly accurate. We can improve the accuracy by improving the VR app with better user interaction and better haptic functionalities, improving the accuracy of reconstruction in VR.

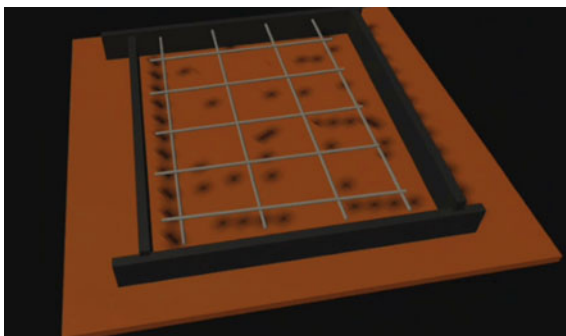
#### **4.2 MR Results**

The user spent 5 min in mixed reality executing the reconstruction activity. Due to the context and feeling of working in the real world, the user stated that the MR experience was more realistic than VR. The user also claims that, unlike VR, MR allows them to walk around the real world and view the room, allowing them to

**Fig. 7** Completion of the task in VR



**Fig. 8** Completion of the task in MR



evaluate accuracy, scale, and other factors more efficiently. After becoming familiar with the controls, the user claims that MR was easy to use and completed the tasks. The user says that achieving precision-focused MR activity was easier than in the VR environment. Figure 8 shows the result of the experiment in MR. The placement of the boundary planks and bar rebars in the MR and the cover space of the bar rebars was almost similar to Fig. 7.

## 5 Summary

After developing the MR and VR applications, conducting the tests, and reviewing the experiment results, the authors believe that MR would be easier to use than VR for this particular experiment objective because the user mentioned that MR allowed the user to walk freely and view the room in real time and get a better understanding of the task and get better details, accuracy, scale, etc. In contrast, when a user performs a job in VR, the user does it in a gamified fashion. While the user understands the ideas, they cannot match the object precisely with the real-world scenario for this objective. The authors also argue that understanding when VR is better and when MR is better depends on the objective of the work to be completed. However, another result of

the experiment is that Oculus Quest 2 can be used as an MR-related application and is less expensive than HoloLens. The Mixed Reality Toolkit (MRTK) library, which is used to build advanced features in HoloLens, can also be used in Oculus Quest 2 and the Passthrough API to develop even better MR applications.

**Acknowledgements** This publication (conference paper) was supported by CPWR through NIOSH cooperative agreement U60 OH009762, CFDA #93.262. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CPWR or NIOSH.

## References

1. DaValle A, Azhar S (2020) An investigation of mixed reality technology for onsite construction assembly. In: MATEC web of conferences, EDP Sciences, Les Ulis
2. Zhang S, Boukamp F, Teizer J (2015) Ontology-based semantic modeling of construction safety knowledge: towards automated safety planning for job hazard analysis (JHA). *Autom Constr* 52:29–41
3. Oculus Developers (2021a) Passthrough samples. Oculus Developers, Irvine
4. Oculus Developers (2021b) Passthrough API overview. Oculus Developers, Irvine
5. Microsoft (2019) Microsoft HoloLens/mixed reality technology for business. Microsoft
6. Aniwaa (2021) The ultimate VR, AR, and MR guide. Aniwaa, Singapore
7. Milgram P, Fumio K (1994) A taxonomy of mixed reality visual displays. *IEICE Trans Inform Syst* 77(12):1–15
8. Wikipedia (2021) Windows mixed reality. Wikipedia
9. Magic Leap (2021) Magic leap: augmented reality platform for enterprise
10. Yevhenii Ratushnyi (2020) Standalone VR versus PC VR: key differences. Visartech Blog
11. Wikipedia (2022) Comparison of virtual reality headsets. Wikipedia
12. Mazwin Mazlan E, Hanim Osman M, Sukri Saud M (2019) Investigating the safety cognition of construction personnel based on safety education. *IOP Conf Ser Mater Sci Eng* 513:012033
13. Yang F, Miang Goh Y (2022) VR and MR technology for safety management education: an authentic learning approach. *Saf Sci* 148:105645
14. Dhalmahapatra K, Maiti J, Das S (2020) On accident causation models, safety training and virtual reality. In: On accident causation models, safety training and virtual reality
15. Kalloc (2021) Fuzor. <https://www.kalloctech.com/index.jsp>. Accessed 11 Mar 2021
16. WorkSafe (2014) Hazard identification tool: concrete placement. <https://www.commerce.wa.gov.au/publications/hazard-identification-tool-concrete-placement>. 23 Mar 2021
17. Wang P, Wu P, Wang J, Chi HL, Wang X (2018) A critical review of the use of virtual reality in construction engineering education and training. *Int J Environ Res Public Health* 15(6):1204
18. Dadi GB, Goodrum PM, Taylor TRB, Carswell CM (2014) Cognitive workload demands using 2D and 3D spatial engineering information formats. *J Constr Eng Manag* 140(5):04014001
19. Autodesk (2021) Revit software/get prices and buy official revit 2022. Autodesk, Singapore
20. Unity Technologies (2020) What's new in Unity 2020.3.14. Unity, Gujarat
21. Justin PB (2021) How to start a VR game using unity 2021: YouTube. Youtube
22. Khronos Group (2021) OpenXR overview. The Khronos Group Inc., Oregon
23. Valecillos D (2021) Passthrough API hand tracking with OpenXR and oculus quest 2. YouTube

# The Incorporation of Learning Theories in VR-Based Safety Training Programs Within the Construction Industry



S. Bader, I. Abotaleb, O. Hosny, and K. Nassar

**Abstract** Virtual reality (VR) technology has gained wide momentum over the past few years. Accordingly, it is rapidly becoming a solid foundation for existing and future educational programs in medical and engineering fields. VR aids in the provision of a student-centered learning approach; creating learning experiences that were not otherwise possible in real life. In the construction industry, the use of VR showed significant potential in enabling the development of a preventative culture in safety training programs through enhanced hazard identification, assessment, response, and mitigation skills. However, none of the few existing attempts—albeit beneficial—has considered major learning theories in the design of their VR-based training. In fact, the understanding of adult learning processes is a matter of serious concern for scholars and practitioners advocating any educational reforms. This concern has been exacerbated with the increasing advocacy for incorporating learning theory foundations in game-based learning. In line with this, this research aims to develop a conceptual framework for the integration of social learning theories in the design and development of VR-based safety training programs within the construction industry. Accordingly, six main theories were analyzed to identify the main principles that are most beneficial to VR-based safety training models. In doing so, major adult learning attributes, such as self-direction, problem/life-centeredness, learning orientations, reflective learning, and others, were taken into account. The results revealed that combining behaviorism, constructivism, and andragogy would tap into all the learning needs of adult learners, thereby aiding in the development of self-actualized individuals who are capable of acquiring declarative and procedural knowledge while adopting new problem-solving skills. They also reveal the importance of stimulating emotions and linking them to hazards. This research will help in significantly improving future VR training models and amplifying their desired outcomes, which in turn will enhance safety in construction sites.

**Keywords** Safety training programs · VR · Construction industry

---

S. Bader (✉) · I. Abotaleb · O. Hosny · K. Nassar  
The American University in Cairo, Cairo, Egypt  
e-mail: [Sahar-bader@aucegypt.edu](mailto:Sahar-bader@aucegypt.edu)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_19](https://doi.org/10.1007/978-3-031-34593-7_19)

285

# 1 Introduction

## 1.1 Background and Overview

The construction industry is often characterized by its unique, dynamic, and fragmented nature where multiple stakeholders are involved while several complex activities are being conducted simultaneously [1]. Therefore, the safety of construction sites has been an issue that raised several concerns over the past few decades. Recent statistics show that the number of fatalities, accidents, and non-fatal injuries is still high relative to other industries, even in developed countries such as the USA and the UK. This is the case despite the increasing level of awareness and the precautionary measures that are being implemented in construction projects [2]. In recognition of this, it has been acknowledged that effectively designed safety training programs are crucial for the elimination of both fatal and non-fatal injuries in construction sites [2].

Currently, there is a wide momentum toward the adoption of new technologies, specifically VR-based technologies, to enhance the outcomes of several educational programs. As stated by [3], “Gamification has been one of the effective approaches to student-centered learning, it allows students to build skills, acquire knowledge and develop an attitude in a game world specifically created for educational purposes” (p. 325). In fact, VR technology has matured, receiving increasing popularity as it became relatively cheap and affordable [4]. In addition, the growing compatibility of the available software and tools along with the massive reduction in the development time has also aided in the widespread adoption of the technology [5].

Similarly, attention to VR technologies has also increased significantly within the construction industry over the past few years with multiple research papers incorporating VR in safety training [1, 5–7]. The aforementioned research papers have all reflected the potential of VR in enabling the development of a preventative culture through enhanced hazard identification, assessment, response, and mitigation skills, thus enhancing the learning outcomes of trainees and improving the overall safety performances of construction projects. Yet, none of the aforementioned efforts is based on solid educational foundations, which would have better informed their methodological designs to maximize the learning outcomes to students.

The understanding of adult learning processes is a matter of serious concern for scholars and practitioners advocating any educational reforms [8]. Merriam and Bierema [9] stated that the five main and most commonly referred to learning theories with direct application to adult learning are behaviorism, humanism, cognitivism, socialism, and constructivism. These theories are further complemented by the adult learning principles, also known as the andragogy theory, which address the specific factors to be taken into account while targeting adult learners [10]. In fact, researchers have been trying to implement learning theories in traditional training for decades in an attempt to enhance education quality and students’ learning experiences [11]. With the advancements in technology, there have also been similar advocacy for such incorporation in game-based learning [4]. Nevertheless, only a few researchers

have shed light on the association of social learning theories and the development of game-based educational programs [4, 12–14].

To illustrate, [14] provided an explanation of how the Game Rules, Game Play, and Game Narratives could be addressed in game-based learning from the perspective of four different learning theories, namely behaviorism, cognitivism, humanism, and constructivism, while [13] investigated the applicability of VR technology to the experiential learning theory. Also, [12] provided a framework for the application of the ten phases of the transformative learning theory in simulation-based learning (SBL). Similarly, [3] have acknowledged the significance of incorporating the behaviorist learning theory along with the reward/punishment system in gamification for educational purposes. However, as stated by [4], research around VR in education is primarily concerned with the constructivism theory that focuses on the provision of an experiential learning experience to students.

## ***1.2 Research Gaps***

The aforementioned discussion reveals several gaps in the literature. To start with, there is a general lack of discussion when it comes to the incorporation of learning theories into VR applications within educational contexts. Rather, most of the papers discussing potential improvements to the outcomes of VR-based learning were concerned with the quality of the development of the prototype itself; these include the rendering quality, user/learner immersion, and interaction levels [4, 13]. In addition, the few research papers that addressed the issue were primarily focused on one general theory, thereby disregarding the possibility of integrating several theories together. Not only does this diminish the value that could be obtained from the simultaneous capitalization on the strengths of different theories but also ignores the fact that each theory independently taps into the different needs of adult learners. Moreover, none of the previous attempts had provided a conceptual framework that could be utilized as a basis for the development of VR-based educational programs. Hence, there is a general lack of practical guidance to researchers willing to develop their training programs on solid and scientifically proven foundations.

The case is quite similar when it comes to the development of construction safety training using VR technology. In fact, the research by [15] was the only identified research that incorporated the behaviorist learning theory in VR-based training within the construction industry. However, it is apparent that the social learning behaviors of construction workers in hazardous construction sites were the research's primary concern. This sheds light on how other crucial influential factors, that are directly addressed in other learning theories, are disregarded. Examples of such factors include motivation, self-direction, problem/life-centeredness, learning orientations, and reflective learning, among others.



### ***1.3 Aims and Objectives***

Within this context, this research aims to develop a conceptual framework for the integration of learning theories in the design and development of VR-based safety training programs within the construction industry. In line with this, the research objectives are as follows:

- To identify the main learning outcomes targeted from VR-based safety training programs.
- To analyze existing learning theories and choose the ones that would act as a basis for the methodological design of VR-based safety training models in the construction industry.
- To develop a conceptual framework that would aid future researchers in developing a sound methodology for their VR-based safety training programs.

### ***1.4 Research Significance***

The significance of this research emerges from the fact that it bridges the existing gap in the literature by drawing the attention of those who are interested in developing VR-based training programs to vital aspects that would further complement the strengths of VR technology. Accordingly, it sheds light on crucial factors that should be considered above and beyond the technical aspects of developing VR training models. This would aid researchers in developing a thorough and all-encompassing VR-based safety training program that would tap into the different needs of adult learners through the consideration of major adult learning attributes. Moreover, by following this conceptual framework, researchers could push the limits of their VR training programs and elevate the learning outcomes without incurring additional costs, thereby aiding in the development of self-actualized individuals who are capable of acquiring declarative and procedural knowledge while adopting new problem-solving skills. All of the aforementioned factors would positively contribute to the overall safety of construction sites.

## **2 Background and Learning Theories**

### ***2.1 Introduction***

The modernization of education has been associated with a growth in the use of novel technologies including VR technology. However, there is a lack of research when it comes to the consideration of learning theories in the development of VR-based training programs. This chapter provides an overview of the six main learning

theories that are most commonly referred to when it comes to adult learning. The following paragraphs provide deeper insights into the foundations of major learning theories.

### **2.1.1 Behaviorism**

To start with, behaviorism is a school of psychology that was founded by John B. Watson in 1913. The theory implies that learning involves a change in behavior that directly results from a particular stimulus in the environment [16]. Accordingly, the theory is based on the assumption that if a behavior is to be reinforced or rewarded, then it is more likely to be adopted by adult learners, otherwise, the behavior disappears [14]. Behaviorism involves two main learning types, namely respondent and operant conditioning [9]. Respondent conditioning is derived from natural selection and evolution theories and occurs when conditional reflexes take place as a result of learning. These reflexes act as responses that are elicited by stimuli in a one-to-one causal relationship. On the other hand, operant conditioning/behaviors are triggered by a series of consequences that tend to shape behaviors, therefore, they occur as a result of the interaction of a stimulus and an activity. These consequences are primarily manifested through reinforcements and punishments [9].

Behaviorism is based on quantifiable, systematic, and observable outcomes as markers of learning [14], which is among its main strengths. This is since it focuses on the present and facilitates the collection and analysis of data to rectify maladaptive behaviors [17]. However, this theory is often criticized for being too deterministic as it focuses on physical aspects while neglecting mental aspects of learning that control human behavior, these include intelligence, talents, feelings, and interests of individuals [16]. Also, it disregards the genetic and internal influences on human behavior, such as moods and feelings, thereby neglecting the uniqueness of individuals [14].

### **2.1.2 Cognitivism**

Cognitivism, on the other hand, is a general psychology school that is based on “meta-cognition” to understand how thought processes influence learning [18]. It conflicts with the behaviorist theory as it acknowledges that people mentally process the information they receive before responding to a stimulus [19]. Hence, it is primarily derived from the way humans gain knowledge, interpret emotions, and dig in memories [16]. The theory emphasizes the thinking aspect of humans based on two main assumptions: (1) the crucial role of the memory as an active and organized processor of information and (2) the crucial role of prior knowledge in shaping the behaviors and reactions [14]. Among the main strengths of this theory is that it provides a more comprehensive understanding of how the human brain works and processes information, therefore, it is of great importance to practical medical applications [14]. However, it has been criticized for being purely focused on cognitive processes which cannot be observed and are hard to assess and evaluate [20]. This necessitates

the conduction of complex experiments to determine cause and effect relationships [19]. Also, it disregards biological factors, the environment, and the upbringing of an individual [20].

### **2.1.3 Humanism**

Moving on to humanism, this theory is derived from humanistic psychology that was fathered by Abraham Maslow [16]. Humanism is primarily concerned with the development of a person based on the assumption that human beings have the potential for growth and development [9]. Therefore, the theory implies that individuals act with intentionality and values with the main goal of developing themselves as self-actualized individuals [14]. In its entirety, humanism promotes the notion that human beings should be treated as whole beings while focusing on their subjective awareness and sociative and productive human capacities [16]. Accordingly, humanism emphasizes a student-centered approach to learning where teachers facilitate creativity, self-actualization, and self-directed learning [21]. Also, the theory teaches students how to learn, adapt, and change to become lifelong learners [9]. Among the main strengths of this theory is that it focuses on the entire person along with their values and self-fulfillment needs [20]. It also emphasizes human choices and responsibility which allows for the provision of a person-centered teaching approach [21]. Nonetheless, this theory has been criticized for being the most abstract and obscure of all existing learning theories [16]. This places a huge burden on teachers to understand different learning styles. Also, the theory assumes that individuals are intrinsically good and will choose positive paths [21].

### **2.1.4 Socialism**

Social cognitive theory draws from both behaviorism and cognitive theories and is mainly considered as a political ideology [22]. The theory emphasizes the role of humans as part of the larger society and highlights the idea that human learning occurs within the social context by observing other people. Through observations and imitations, people acquire knowledge, rules, skills, strategies, beliefs, and attitudes [9]. Thus, it focuses on social influences and considers the dynamic interactions between environment, behavior, and individuals. The theory is based on four basic concepts which are the behavioral potential, expectance, reinforcement value, and psychological situations of individuals [23]. Hence, the combination of behavioral, cognitive, and environmental aspects is among the main strengths of this theory [22]. However, again, it is often criticized for being loosely structured and its disregard of biological differences, hormones, emotions, and motivations [9]. Moreover, it focuses on what happens in the surrounding environment rather than what the learner actually does [23].

### 2.1.5 Constructivism

Finally, constructivism is primarily concerned with creating meanings from personal experiences as a learning process. The theory departs from the belief that knowledge is a process of formation that continues to develop and evolve [16]. Thus, according to this theory, knowledge is constructed rather than received [24]. This, therefore, necessitates the active role of students during the learning process through active engagement in activities, discussions, reflections, and hands-on experiences. The theory relates to aspects such as self-directed learning, transformational learning, experiential learning, reflective practice, and problem-solving [9]. One of the main strengths of this theory is that it places mutual emphasis on the skills learned through the learning process as well as the learning outcomes. This is since it intends to stimulate students to be critical thinkers when confronted with new facts [24]. Also, it emphasizes sensory inputs, eliminates grade-centered approach, puts more emphasis on value, stimulates self-confidence, critical thinking, and problem-solving by treating students as active participants rather than passive receivers of knowledge [25]. However, it is criticized for requiring expensive set-ups, extended preparation times, and varied assessment strategies [25]. The transformative and experiential learning theories are the main learning theories under the constructivist theory.

#### Transformational Learning

As stated by [26], the transformative learning theory is “based on constructivist assumptions, and the roots of the theory lie in humanism and critical social theory.” (p. 5). It is based on the assumption that adults change their perspectives based on the new information they receive by evaluating and critically reflecting on their past experiences and reconciling new information with what they already know in life [26]. Therefore, it involves a change in the learners’ frame of reference which consists of associations, concepts, values, and feelings [27]. Thus, critical reflections act as a main and integral component of initiating transformational learning [28] followed by critical discourse where learners validate the best judgment.

#### Experiential Learning

Experiential learning emphasizes the role of connecting current experiences to previous experiences; based on such connections, possible future implications are derived [28]. Accordingly, experiential learning simply means constructing knowledge and meaning from real-life experiences [8]. The theory is based on the following assumptions: (1) Learning is a process that entails feedback, (2) the process often includes the unlearning of previous knowledge; (3) learning is driven by conflicts, disagreements, and differences; (4) learning is a form of adapting to the environment;

(5) learning results from the assimilation of new experiences with existing ones and vice versa; and (6) based on such assimilations, learners are capable of creating new knowledge [13], whereas [8] stated that experiential learning should include both cognitive and emotional processes, action-reflection cycles, and ideals of personal transformation.

For this reason, Kolb's experiential learning theory presents the learning cycle which is based on the difference in preferences in learning styles [29]. Kolb's learning model consists of four primary stages, namely Concrete Experience (CE), Reflective Observation (RO), Abstract Conceptualization (AC), and Active Experimentation (AC) [29]. According to the learning model, the cycle begins with the learners' personal involvement in an experience [11], which acts as the basis for observations and reflections [29]. By drawing logical conclusions, these reflections are then assimilated and distilled into abstract concepts which later act as a basis for future action [29].

## ***2.2 Andragogy/Adult Learning Principles***

Lindeman (1926) had developed four key principles about adult learners. First, adult learners are motivated to learn as long as they perceive the needs and interests that the learning will satisfy. Secondly, adults tend to have a life-centered approach to learning [10]. Thirdly, adult learners' richest resource to learning is experience. Fourthly, they tend to have a tendency toward self-directed learning [10], whereas [10] state that there are six main principles to the adult learning theory. To start with, adult learners need to know about the benefits of learning along with the negative consequences of not learning. With regards to self-concept, adult learners need to be treated by others as capable of self-direction, therefore, facilitators need to assist adult learners in transitioning from being dependent to self-directing learners [10]. Concerning learners' experience, there is a need to emphasize experiential learning instead of transmittal techniques, these include group discussions, simulation exercises, and problem-solving activities. Moving on to the adults' readiness to learn, adults develop higher levels of readiness when they perceive the benefits that such learning would induce in allowing them to cope with real-life situations. Regarding adults' orientation, it has been acknowledged that adult learners have a problem/task/life-centered approach to learning. Finally, adults need to have both external and internal motivations to learn such as better jobs, promotions, job satisfaction, and self-esteem [10].

## 3 Methodology

### 3.1 *Research Philosophy and Approach*

To start with, the philosophy of research dictates the methods through which data could be gathered, interpreted, and generalized. Two main research philosophies exist, namely positivism and interpretivism. The former holds an objective stance and necessitates the collection of measurable and quantifiable data that could be statistically analyzed to create factual knowledge [30]. The deductive reasoning approach is often associated with the positivist philosophy. It entails starting with a research hypothesis that could, later on, be tested to draw conclusions and generalize findings [31].

On the other hand, interpretivism is a subjective and quality-based approach that is primarily concerned with the gathering of qualitative data that could be analyzed by integrating the researcher's values, beliefs, explanations, and constructions [30]. It is associated with the inductive reasoning approach where data is firstly gathered and analyzed to identify patterns and trends, based on which new knowledge could be constructed [31]. Since this research aims to develop a conceptual framework that requires the incorporation of qualitative attributes, such as perspectives, opinions, and judgments of experts, an interpretivist approach is adopted. This is coupled with the inductive reasoning approach where the conceptual framework would be developed based on the subjective analysis of the gathered data.

### 3.2 *Research Strategy*

Two main types of research exist, namely primary and secondary research. Primary research entails the gathering of first-hand data that meets the specific aims of the research, whereas secondary research entails the gathering of data that was originally published for other purposes [32]. To have a deeper understanding of the learning process and the factors that should be considered for the effective delivery of information to students, this research opted for a mixed-strategy approach where both primary and secondary data were used for the development of the conceptual framework. To start with, secondary data was gathered from an extensive review of books and peer-reviewed journal articles related to the different learning theories along with their integration into VR-based learning models.

Whereas, case studies, interviews, and focus groups could be used to gather qualitative data [31]. However, due to the lack of practical application of learning theories in VR-based educational contexts, the case study method is not applicable. Interviews, on the other hand, are regarded as an excellent one-to-one means for the gathering of in-depth qualitative data [31]. Similarly, focus group discussions allow for the gathering of rich information based on a moderated interaction among a group of experts [32]. Since this research aims to integrate the opinions of experts from two different fields, namely educational studies and construction engineering, the focus group method was found to be the most suitable for the gathering of differing perspectives. Accordingly, the purposive sampling technique was deployed in this research, where the research's experts were chosen based on their characteristics and the research's objectives [32]. In total, 5 experts with educational studies backgrounds and 2 experts with construction management backgrounds were selected and recruited. Years of experience ranged from 10 to 25+ years within their respective fields. It is worth noting that all educational studies experts have research interests related to the adoption of new technologies, including VR, in higher education (Table 1).

### **3.3 Focus Group Design**

The focus group research design approach developed by Nyumba et al. (2018) was adapted and used in this research. It entails four main steps, namely research design, data collection, data analysis, and data reporting. The steps followed are presented in Fig. .1

#### **3.3.1 Targeted Learning Outcomes**

The research started with the identification of the targeted learning outcomes from the VR-based safety training program. Wohlgenannt et al. [4] revealed that four primary competencies are targeted in higher education, namely declarative knowledge, procedural knowledge, problem-solving skills, and communication skills. Declarative knowledge is based on the learning of facts, abstract concepts, and scientific principles. Procedural knowledge includes the learning of tasks that foster the conduction of processes. Problem-solving skills include complex decision-making such as risk assessment. Finally, communication skills include interactivity and collaboration [4]. Based on the aforementioned competencies, five main learning outcomes from the VR-based safety training model were identified as follows: (1) to provide a better heuristic approach to site inspection, (2) to enhance hazard recognition and identification skills; (3) to elevate awareness of the consequences of hazards; (4) to Improve risk assessment skills; and (5) to enhance the selection of the right course of action

**Table 1** Demographics of focus group experts

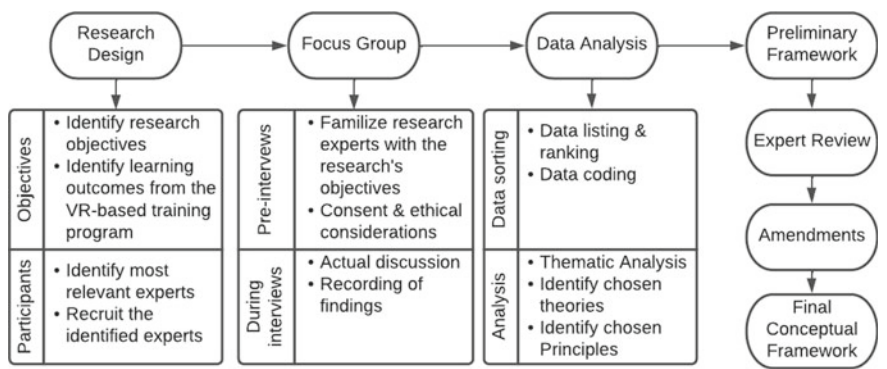
Experts	Background	Position	Years of experience	Research interests
1	Construction management	Professor	25+	A wide range of interests in the use of technologies such as artificial intelligence and digital technologies in the project management field with a specific focus on the use of novel technologies for safety training in construction sites
2	Construction management	Assistant professor	8	A wide range of interests in construction engineering and management with a current focus on the use of virtual reality (VR) and augmented reality (AR) in safety of construction sites
3	Educational studies/science education	Professor	25+	The provision of quality education within the STEM/STEAM areas, with a focus on the use of novel educational technologies
4	Educational studies	Professor	13	Educational reforms and learners' development theories with particular emphasis on modern teaching approaches
5	Educational studies	Associate professor	15	International and comparative higher education with specific interest in the use of digital technologies in teaching and learning

(continued)



**Table 1** (continued)

Experts	Background	Position	Years of experience	Research interests
6	Educational studies	Assistant professor	10	Curriculum and instruction through innovative learning solutions with a particular focus on educational technology and teaching methodology
7	Educational studies/ science education	Assistant professor	7	Educational administration, supervision, planning, and economics; sustainability of educational reforms with specific focus on organizational behaviors



**Fig. 1** Research methodology

based on practicality, affordability, and time consumption. Further details on the gaps and pitfalls in existing traditional safety training programs along with how such targets were developed would be published in the future research.

### **3.3.2 Focus Group Discussion**

The focus group discussion was divided into two phases. In the former phase, the research experts were introduced to the research topic along with its main aims and objectives. This was followed by some ethical considerations where experts consented to participate. The second phase entailed the actual conduction of the moderated discussion in a semi-structured format. The discussion was divided into three phases; the first phase entailed discussing the different existing theories, as identified from the literature, along with their applicability to the research's aims and objectives. The second phase entailed discussing all crucial influential factors that would have a positive impact on both the learner and the learning process. The third phase entailed discussing the most productive combination of learning theories that would aid in maximizing the learning outcomes of VR-based learning.

### **3.3.3 Data Analysis**

The content analysis technique was used for the analysis of the transcribed data gathered from the focus group discussion. This technique allowed for the identification of patterns and trends that acted as the main basis for the development of the conceptual framework. During the analysis, the following factors were considered to support the qualitative analysis of the gathered data: the order of the discussed topics, the presence and absence of certain topics from the discussion, the time spent discussing each topic, the intensity of expressions in relation to the topics being discussed, the reasons and reactions of experts to each topic, and the consensus over the topic being discussed. The following paragraphs reveal the results and discussion of the data gathered and analyzed from the group discussion.

## **4 Results and Discussion**

### ***4.1 Chosen Learning Theories***

Based on the analysis, a general consensus on the elimination of two main theories, humanism, and cognitivism, was observed. To start with, humanism was eliminated for being a loosely structured theory that conforms to the findings from the literature [16]. The experts asserted that following this theory would necessitate the development of different scenarios for each learning topic/module to satisfy the learning needs of different individuals, which is not practical in terms of the use of VR technology. Furthermore, there was a general consensus that this theory is centered on a western culture which assumes that individuals are intrinsically good and will naturally choose to follow the safety procedures in construction sites. However, this might not be the case in the cultures of undeveloped countries where construction workers

need to be trained on how to responsibly act and react in the extremely hazardous conditions of construction sites.

Similarly, all experts agreed on the elimination of the cognitivist theory for the following reasons. First and foremost, it was considered to be based on deep clinical foundations that necessitate the understanding of several complex brain processes. This conforms to the findings from the literature that critiqued this theory for its dependency on complex experimental processes [19], thus, it was agreed that this would be beyond the scope of the aims and objectives of this research. Furthermore, there was a general consensus from experts that this theory ignores crucial influential factors, specifically in undeveloped countries, such as the culture, social class, and upbringing of individuals. It was, therefore, ascertained that these factors play a crucial role in determining the learning potential and observed behaviors of construction workers who typically tend to be from lower social classes.

Moving on to socialism, the experts were divided equally as to whether this theory should be regarded as a basis for the development of the VR-based safety training program. The experts who were in favor of this theory based their arguments on the naturally occurring phenomenon of imitation of misbehaviors in social groups. Their perspective is supported by the findings of [33] which reveal that construction workers are often confronted with severe conformity pressures, thereby causing them to easily conform to the unsafe behaviors of the group. On the other hand, the opposing experts viewed this as insufficient for the inclusion of the theory. The experts perceived this view as over-generalization which would lead to the disregard of other factors such as emotions, motivations, and personality types. Accordingly, the researchers opted to eliminate this theory based on the contentment with the views of the opposing experts. This is further supported by the following reasons. Firstly, as stated in the literature, this theory focuses on what happens in the surrounding environment rather than what the learner actually does or learns. Considering the VR environment, it would be impractical to demand trainees to imitate positive behaviors to learn [16]. The barriers to this include but are not limited to the development time of the model, technical difficulties, and the limitations of the technology itself where not all behaviors/actions could be easily replicated and imitated in the virtual environment.

Finally, the experts were left with two main theories which are behaviorism and constructivism. All of the experts were in favor of the behaviorism theory for the following reasons. Firstly, the behavioristic approach is based on the natural objectives of the provision of training programs which are primarily concerned with competence and skill development. This is further strengthened with the main objectives of this research which aims to introduce behavioral changes to enhance the safety of construction sites. Secondly, the theory provides a solid foundation that would act as a guide for the development of the VR safety training program. This is going to be achieved by exposing the trainees to stimuli in the virtual environment and designing such stimuli to elicit desired responses, thereby inducing behavioral changes in the desired direction toward enhancing safety performance. Thirdly, the theory allows for the development of new skills and competencies based on both desired and undesired stimulus which would bridge the existing gaps in existing

traditional safety programs. Finally, the theory provides a strong basis for observing, gathering, and analyzing behavioral changes based on concrete justifications.

Likewise, all of the experts showed their support for the constructivism theory based on the following reasons. The main rationale is related to the nature of the VR technology itself which meets and supports all of the requirements of the theory. This includes the provision of a personal learning experience where the trainees actively participate to solve real-life problems in an immersive and interactive environment. This conforms to the findings from the literature which indicate that learning occurs through the integration of real-life current and past experiences [28]. Also, the experts were in favor of this theory as it provides the basis for the shift from traditional learning and assessment techniques. Thus, it would complement the main drivers for the wide-scale adoption of VR technologies in modern-day education systems. Accordingly, the following research adopts a behavioristic/constructivist approach to the development of the conceptual framework for the design of VR-based safety training programs.

## **4.2 Factors Considered**

### **4.2.1 Behaviorism**

The factors to be considered in relation to this theory are primarily associated with the stimulus provided to the trainees to elicit the desired behaviors. As stated by [17], a stimulating environment could be in the form of reinforcements/rewards or punishments. Each is further divided into two subcategories indicating whether they are positive or negative reinforcements/punishments. Positive reinforcement entails adding reinforcers to increase the likelihood of the adoption and practicing of a certain behavior, whereas negative reinforcement entails the removal of undesirable stimulus to increase the likelihood of a behavior. On the other hand, positive punishment entails the addition of undesirable stimulus to decrease the likelihood of the adoption and practicing of a certain behavior, whereas negative punishment entails the removal of desirable stimulus to decrease the likelihood of a behavior [17].

From the focus group discussion, the need to include both reinforcements and punishments in each learning topic/module was ascertained by the majority of the experts. This was chosen based on a general belief that this would further reinforce and strengthen the learning process of trainees and maximize their learning outcomes. While several research papers have advocated against the inclusion of punishments in the learning process claiming that the inclusion of reinforcements is associated with better learning outcomes [34], the experts asserted that this does not apply to the virtual environment. This is since no actual harm would be caused to the trainees as a result of the punishment provided, rather, the experts believed that this is a crucial factor to demonstrate to the trainees the intensity and seriousness of construction accidents along with the severe consequences that could occur as a result. In relation to whether reinforcement and punishments should be positive or negative, again the

majority believed that this should be left to the judgment of the developer of the program. This primarily emerged from the wide scope of hazards and accidents that could be covered in construction sites. Thus, it was concluded that developers are more capable of determining the reinforcement/punishment types that would better serve the scenario at hand.

#### 4.2.2 Constructivism

With regards to constructivism, the following factors emerged: real-life experiences, interactivity, engagement, problem-oriented training, and reflections. To start with, there was a huge emphasis on the provision of a training environment that resembles real-life experiences to the trainees. This was primarily derived from the urge to maximize the trainees' sense of immersion and inclusion in the virtual environment. As ascertained by the experts, this would need consideration of the technical aspects of the VR model, whereas the planning of the scenarios along with the associated reinforcements/punishments used are the means through which interactivity and engagement should be considered. These findings are supported by [13] who stated that concrete experiences require realistic surroundings, character movement, basic interaction with objects, users, and intelligent agents, and realistic scenarios for the design of an effective VR experience.

Secondly, the experts agreed that the problem-oriented approach should be adopted in this design of the VR-based safety training program. This is done primarily to ensure the mental engagement of the trainees, besides their physical engagement and immersion in the model. The experts asserted that the use of a problem-oriented approach to the introduction of the learning modules in VR training would stimulate critical thinking and critical reflections, thereby enhancing the problem-solving capabilities of the trainees. This is a vital skill in construction sites as hazards/accidents often require prompt and responsible reactions to be solved while mitigating all potential consequences [2].

Finally, the majority of experts indicated the significance of reflections on the learning process. They confirmed that the trainees' active reflections on their actions would allow them to reconstruct knowledge through an enhanced knowledge assimilation process. As stated by Hilgard (1964), the cognitive feedback of learners where learners reflect on their decisions based on the consequences is crucial for adult learning. Also, immediate feedback that clearly reveals the weaknesses in their thought processes should be provided; this should also be accompanied by the provision of models of superior performance which would aid in introducing a shift in the trainees' perspectives. These findings are supported by Kolb's learning cycle that emphasizes action-reflection cycles for constructing new meanings and logical conclusions which would act as the basis for the acquisition of new knowledge [29].

### 4.2.3 Adult Learning Principles

From the adult learning principles, the experts agreed on the inclusion of the following factors. Firstly, it is crucial to increase the learners' needs to acquire knowledge. This would be achieved by showing them the benefits of learning and the consequences of not learning. This conforms to the findings of Lindeman (1926) who state that adult learners are motivated to learn as long as they perceive the needs and interests that the learning will satisfy. Secondly, the experts agreed that the training model should integrate external and internal motivators that would further drive the learners' readiness to learn. This could also be accompanied by "goal-setting" practices where trainees are allowed to set goals that would be achieved upon their effective grasp and understanding of the learning content. As stated by [26], the process of goal-setting by learners is a significant motivational tool for learning and personal development. This factor is supported by [10] who stated that external and internal motivations to learning significantly influence the learning process. Thirdly, the experts agreed that the learners' frustrations, as a result of failure/potential punishments, should be accommodated. This is done to prevent a drop in their enthusiasm or excitement levels as they encounter failures throughout the learning process. This emerges from the fact that adults have lower learning potential and tolerance as compared to infants. Finally, the experts agreed that the trainees should be allowed to implement the newly learned knowledge in new experiences.

## 4.3 Conceptual Framework

The developed conceptual framework is presented using a fictitious example with reference to a "fire accident" problem that is to be explained under a "safety of workplaces" module for the sake of illustration. Firstly, before starting a new module/learning topic (Safety of workplaces), the learners' need to learn should be established. This is done by showing trainees the benefits of learning this module and the potential consequences of not learning. To illustrate, the benefits shown could include saving construction time and costs, whereas the consequences of not learning could include life-threatening accidents that would jeopardize the trainees' lives and health. This need to learn is then reinforced by the provision of both external and internal motivators that would reveal to the trainees the expected gains that could be achieved upon the effective learning of the content of the module. These include getting back home safely to their families and ensuring that their colleagues are safe until the end of the project. Subsequently, the trainees are allowed to set goals to be achieved throughout the learning module. These could include minimizing their punishments and maximizing their rewards during the training.

Eventually, the learning process starts with the confrontation of the main problem (Fire accident) to which the trainees are required to inspect the site of the accident (learning objective 1). Based on the scenario, the trainees will be allowed to react and identify the causes of the accident (The Fire) (learning objective 2). If the trainees did not choose all the potential causes of the accidents correctly, they will be faced with punishment. Examples of such punishments include protruding fire, emerging black smoke, and workers suffocating and losing consciousness (learning objective 3). The trainees are then provided a voice-over explaining how tricky the scenario was to accommodate their frustrations. They are then given some time to reflect on their choices before giving them feedback on the errors conducted.

Subsequently, the trainees are redirected to the scene and are allowed to re-choose the answers based on the feedback provided (learning objective 4). Upon choosing the right answers, a reward/reinforcement is presented. This could be distinguishing the fire without causing any harm to the workers. Then, a voice-over presenting superior performance and all the safety measures that should be taken in similar circumstances (learning objective 5) is presented. In the following modules, the trainees will be presented with small quizzes where they can actively implement and practice their newly gained knowledge.

## 5 Conclusion

To conclude, this paper provides a comprehensive conceptual framework for the development of VR-based safety training programs within the construction industry. The framework was based on a behaviorist/constructivist approach with reference to a few of the main adult learning principles. Accordingly, this framework taps into the different needs of learners, thereby aiding in the development of self-actualized individuals who are capable of acquiring declarative and procedural knowledge while adopting new problem-solving skills. This research will help in significantly improving future VR training models and amplifying their desired outcomes, which in turn will enhance safety in construction sites. It is worth mentioning that the developed framework is to be implemented and tested on a safety training program that is concerned with all safety measures related to high-rise building construction (Fig. 2).

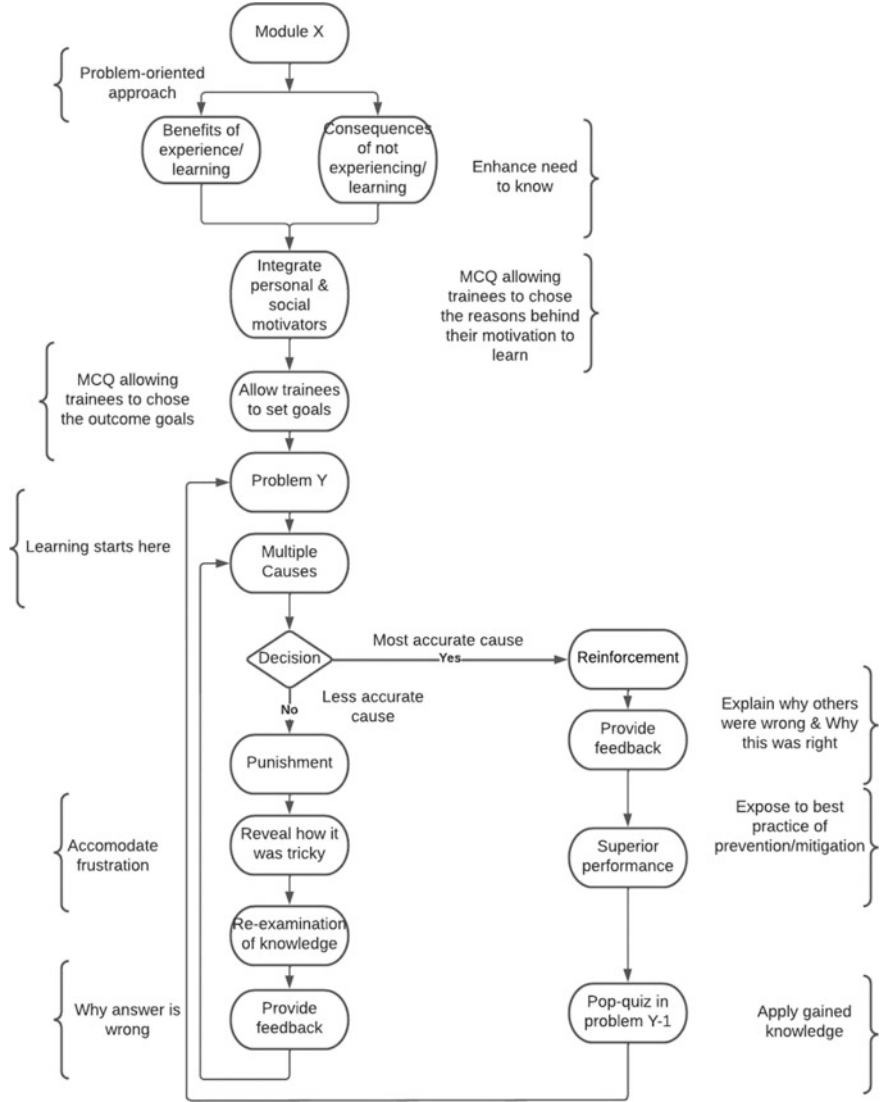


Fig. 2 Conceptual framework for the development of VR safety training programs



## References

1. Dhalmahapatra K, Maiti J, Krishna O (2021) Assessment of virtual reality-based safety training simulator for electric overhead crane operations. *Saf Sci* 139:105241. <https://doi.org/10.1016/J.Ssci.2021.105241>
2. Li F, Zeng J, Huang J, Zhang J, Chen Y, Yan H et al (2019) Work-related and non-work-related accident fatal falls in Shanghai and Wuhan, China. *Saf Sci* 117:43–48. <https://doi.org/10.1016/J.Ssci.2019.04.001>
3. Zakaria N, Saripan M, Subarimaniyam N, Ismail A (2020) Assessing ethoshunt as a gamification-based mobile app in ethics education: pilot mixed-methods study. *JMIR Serious Games* 8(3):e18247. <https://doi.org/10.2196/18247>
4. Wohlgenannt I, Fromm J, Stieglitz S, Radianti J, Majchrzak T (2019) Virtual reality in higher education: preliminary results from a design-science-research project. In: International conference on information systems development. Toulon, France. Accessed 29 Nov 2021
5. Mora-Serrano J, Muñoz-La Rivera F, Valero I (2021) Factors for the automation of the creation of virtual reality experiences to raise awareness of occupational hazards on construction sites. *Electronics* 10(11):1355. <https://doi.org/10.3390/Electronics10111355>
6. Goulding J, Nadim W, Petridis P, Alshawi M (2012) Construction industry offsite production: a virtual reality interactive training environment prototype. *Adv Eng Inform* 26(1):103–116. <https://doi.org/10.1016/J.Aei.2011.09.004>
7. Sacks R, Perlman A, Barak R (2013) Construction safety training using immersive virtual reality. *Constr Manag Econ* 31(9):1005–1017. <https://doi.org/10.1080/01446193.2013.828844>
8. Seaman J, Brown M, Quay J (2017) The evolution of experiential learning theory: tracing lines of research in the JEE. *J Exp Educ* 40(4):NP1–NP21. <https://doi.org/10.1177/1053825916689268>
9. Merriam S, Bierema L (2013) *Adult learning: linking theory and practice*, 1st edn. John Wiley & Sons Incorporated, Hoboken
10. Knowles M, Holton Iii E, Swanson R, Robinson P (2005) *The adult learner: the definitive classic in adult education and human resource development*, 6th edn. Elsevier, Amsterdam
11. Svinicki M, Dixon N (1987) The Kolb model modified for classroom activities. *Coll Teach* 35(4):141–146. <https://doi.org/10.1080/87567555.1987.9925469>
12. Briese P, Evanson T, Hanson D (2020) Application of Mezirow's transformative learning theory to simulation in healthcare education. *Clin Simul Nurs* 48:64–67. <https://doi.org/10.1016/J.Ecns.2020.08.006>
13. Fromm J, Radianti J, Wehking C, Stieglitz S, Majchrzak T, Vom Brocke J (2021) More than experience? On the unique opportunities of virtual reality to afford a holistic experiential learning cycle. *Internet Higher Educ* 50:100804. <https://doi.org/10.1016/J.Iheduc.2021.100804>
14. Wu W, Hsiao H, Wu P, Lin C, Huang S (2011) Investigating the learning-theory foundations of game-based learning: a meta-analysis. *J Comput Assist Learn* 28(3):265–279. <https://doi.org/10.1111/j.1365-2729.2011.00437.x>
15. Shi Y, Du J, Ahn C, Ragan E (2019) Impact assessment of reinforced learning methods on construction workers' fall risk behavior using virtual reality. *Autom Constr* 104:197–214. <https://doi.org/10.1016/J.Autcon.2019.04.015>
16. Alauddin M (2020) Basic of learning theory (behaviorism, cognitivism, constructivism, and humanism). *Int J Asian Educ* 1(1):37–42
17. Baum W (2017) *Table of content section understanding behaviorism, behavior, culture, and evolution*, 3rd edn. John Wiley & Sons Incorporated, Hoboken
18. Tennyson R, Rasch M (1988) Linking cognitive learning theory to instructional prescriptions. *Instr Sci* 17(4):369–385. <https://doi.org/10.1007/Bf00056222>
19. Prestine N, Legrand B (1991) Cognitive learning theory and the preparation of educational administrators: implications for practice and policy. *Educ Admin Quart* 27(1):61–89. <https://doi.org/10.1177/0013161x91027001004>
20. Daniels H, Lauder H, Porter J (2013) *Educational theories, cultures, and learning*. Routledge, New York

21. Huitt W (2009) Humanism and open education, educational psychology interactive. Valdosta State University, Valdosta, GA
22. Rotter JB, Chance JE, Phares EJ (1972) Applications of a social learning theory of personality. Holt, Rinehart & Winston, New York
23. Feather N (1982) Expectations and actions: expectancy-value models in psychology, 1st edn. Routledge, New York
24. Colliver J (2002) Constructivism: the view of knowledge that ended philosophy or a theory of learning and instruction? *Teach Learn Med* 14(1):49–51. [https://doi.org/10.1207/S15328015t1m1401\\_11](https://doi.org/10.1207/S15328015t1m1401_11)
25. Olusegun S (2015) Constructivism learning theory: a paradigm for teaching and learning. *IOSR J Res Method Educ* 5(6):66–70
26. Taylor E, Cranton P (2013) The handbook of transformative learning. Jossey-Bass, Hoboken
27. Gerster-Bentaya M, Knierim A, Herrera Sabillón B (2021) Translating the transformative learning approach into practice: the case of a training of trainers' pilot in client-centred extension approach. *J Agric Educ Extens* 19:1–21. <https://doi.org/10.1080/1389224x.2021.1953549>
28. Bass C (2012) Learning theories and their application to science instruction for adults. *Am Biol Teach* 74(6):387–390. <https://doi.org/10.1525/abt.2012.74.6.6>
29. Healey M, Jenkins A (2000) Kolb's experiential learning theory and its application in geography in higher education. *J Geogr* 99(5):185–195. <https://doi.org/10.1080/00221340008978967>
30. Crossan F (2003) Research philosophy: towards an understanding. *Nurse Res* 11(1):46–55. <https://doi.org/10.7748/nr2003.10.11.1.46.c5914>
31. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, Burroughs H, Jinks C (2017) Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant* 52(4):1893–1907
32. Gideon L (2016) Handbook of survey methodology for the social sciences, 1st edn. Springer, New York
33. Li Z, Bao X, Sheng Y, Xia Y (2021) Research on unsafe behavior of construction workers under the bidirectional effect of formal rule awareness and conformity mentality. *Front Psychol* 12:394. <https://doi.org/10.3389/Fpsyg.2021.794394>
34. Dad H, Ali R, Janjua M, Shahzad S, Khan M (2010) Comparison of the frequency and effectiveness of positive and negative reinforcement practices in schools. *Contemp Issues Educ Res* 3(1):127. <https://doi.org/10.19030/Cier.V3i1.169>
35. Mcleod S (2016) Albert bandura's social learning theory. [Simplypsychology.Org. https://www.simplypsychology.org/Bandura.html#:~:Text=Social%20learning%20theory%2c%20proposed%20by,Influence%20human%20learning%20and%20behavior](https://www.simplypsychology.org/Bandura.html#:~:Text=Social%20learning%20theory%2c%20proposed%20by,Influence%20human%20learning%20and%20behavior). Accessed 25 Feb 2022
36. Timokhin A, Khoronko L (2021) Virtual reality as a main basis for forming modern educational technologies. *E3s Web Conf* 273:12076. <https://doi.org/10.1051/E3sconf/202127312076>
37. Yardley S, Teunissen P, Dornan T (2012) Experiential learning: transforming theory into practice. *Med Teach* 34(2):161–164. <https://doi.org/10.3109/0142159x.2012.643264>

# Resources Deployment Optimization in Scattered Repetitive Projects



Heba Kh. Gad, Mostafa H. Ali, Khaled Nassar, Yasmeen A. S. Essawy, and Abdelhamid Abdullah

**Abstract** The uncertain nature of construction processes and the variability of utilized resources make resource deployment planning a complex process, specially on an organizational level. Multiple operations modes of resources add even more dimensions to the planning and decision-making process of resources deployment. In scattered short-term repetitive projects, the frequency of decisions related to resources allocation and deployment constitutes a logistics burden on any organization. Conventional Planning tools are time consuming and fall short of meeting the periodical need to update resources deployment plans required in scattered repetitive projects. This paper presents a new outlook on construction operations in scattered repetitive projects and introduces the basis of an optimization model to minimize the time and cost of these operations on an organizational level. The model utilizes Genetic Algorithms (GA) and stochastic simulations as an adaptation of two famous problems in literature: the Multiple Traveling Salesman problem (MTSP) and the Multimode Resource-Constrained Time–Cost Tradeoff Problem of Repetitive Projects (MRCTCTP-RP). The proposed model considers both aspects of resources allocation and mobility between projects in optimizing the minimum time and cost of projects on an organizational level.

---

H. Kh. Gad (✉) · M. H. Ali · K. Nassar · Y. A. S. Essawy · A. Abdullah  
Department of Construction Engineering, The American University in Cairo (AUC), Cairo, Egypt  
e-mail: [heba.kh@aucegypt.edu](mailto:heba.kh@aucegypt.edu)

K. Nassar  
e-mail: [knassar@aucegypt.edu](mailto:knassar@aucegypt.edu)

Y. A. S. Essawy  
e-mail: [yasmeen.sherif@eng.asu.edu.eg](mailto:yasmeen.sherif@eng.asu.edu.eg); [y\\_essawy@aucegypt.edu](mailto:y_essawy@aucegypt.edu)

A. Abdullah  
e-mail: [abdelhamid.abdullah@m-eng.helwan.edu.eg](mailto:abdelhamid.abdullah@m-eng.helwan.edu.eg); [abdelhamid.abdullah@aucegypt.edu](mailto:abdelhamid.abdullah@aucegypt.edu)

Y. A. S. Essawy  
Department of Structural Engineering, Ain Shams University (ASU), Cairo, Egypt

A. Abdullah  
Department of Architectural Engineering, Faculty of Engineering at Mataria, Helwan University, Helwan, Egypt

**Keywords** Construction operations · Multimode resources · Time–cost tradeoff · Repetitive scattered projects · Multiple traveling salesman · Genetic algorithms · Simulation

## 1 Introduction

Simulation in the construction industry has been a growing field of research the past couple of years, especially in the area of construction planning and project progress tracking. Simulation in general has a very wide scope of research with enormous achievements and application across all industries. Yet, the construction industry is lagging behind in terms of application of simulation in its processes. Although simulation in construction is adequately covered in the literature, it remains in the design phase and has not been implemented sufficiently in the industry for two main limitations. The first limitation is that research in construction simulations is normally conducted for a specific project; consequently, the research cannot be adapted nor generalized to be applied in other projects. In the construction industry, each project is unique in its nature, which justifies the need for customized tools, yet, excessive customization prevents the market from benefiting from these tools. The second limitation is the neglect of developing user-friendly tools that do not require an immense amount of effort to be comprehended and used by professionals in the construction industry.

The world is currently witnessing a wide movement of infrastructure investment in road networks, water and drainage systems and power transmission lines. All of the aforementioned infrastructure projects need continuous maintenance operations. These operations are repetitive in nature (similar work done for the different segments) and are scattered (by the fact of huge spatial extension of the networks), and hence having a developed tool to plan resources deployment for the maintenance work, scattered repetitive projects, is of great value. Organizational Planning for Scattered Repetitive Projects (OPSRP) constitutes a great logistical burden on organizations. Planning and utilizing the resources efficiently enhances the productivity of the work and results in savings in terms of time and money.

Modeling and optimizing operations for OPSRP is complex and requires adding an element of probability to be representative of the natural variation occurring in such operations. Agent-Based (AB) modeling presents the suitable modeling technique for the OPSRP problem. AB modeling is representing the modeling elements as instances where they interact with each other in a stochastic horizon allowing a high level of simulation to real-life operations complexity [1]. The aim of this research is to introduce a simulation element in creating a model concept for optimizing resource deployment in repetitive scattered projects to minimize the time and cost associated with them. The model concept is derived from AB modeling techniques utilizing Anylogic software.

## 2 Literature Review

This paper showcases a model that addresses resource utilization in scattered repetitive projects. The model tackles two main elements. First is the resource deployment, which is very similar to the multiple traveling salesman problem in the literature. Second is the resource optimization of repetitive projects, which is widely discussed in the literature under the name “Multimode Resource-Constrained Time–Cost Tradeoff Problem of Repetitive Projects (MRCTCTP-RP)”. This section focuses on the contribution of the literature in solving the MRCTCTP-RP and examines other related literature that addresses the commercialization of such tools and models.

### 2.1 *MRCTCTP-RP*

In “Space–time repetitive project scheduling considering location and congestion”, the authors developed a model to solve the MRCTCTP problem. They tackled the important point of multiple crew usage, where the same resource can have different functions and coexistence of multiple modes in an activity, where one activity can be done with various methods and each needs different resources. Many researchers looked at solving the MRCTCTP-RP, while taking into consideration the important point of multiple crew usage, where the same resource can have different functions in different activity types. Tao developed a model that solves the problem using soft logic and Genetic Algorithms (GA) in 2018, while Tang used Constraint Programming (CP). These models however were not very efficient in terms of computational time, not to mention their reliance on fixed discrete values to estimate the productivities and the time taken to perform required tasks.

### 2.2 *Developed Commercial Models/Tools for MRCTCTP-RP*

This section discusses some of the fully developed models covered in literature. The unique aspect of the two models discussed here is that the researchers developing these wanted to have the models available as a commercial tool for professionals in the industry to use. The first model is called MACROS, short for Multiobjective Automated Construction Resource Optimization System (Kandil 2006). The MACROS model’s objective is to optimize the utilization of resources to minimize project cost and duration, while maximizing quality. This is achieved by integrating Microsoft Scheduler, a well-known scheduling tool in the industry, to import the project schedules and perform the optimization. Although this integration made the model more commercial, it was not as user friendly as it should be. This is due to the fact that it asked the user to set the optimization-run parameters without having a default settings preset in the tool. Each user had to input the population size, crossover rate

and mutation rate for the genetic algorithm to operate the model. The value of each of the three parameters needs research and experience to set, especially that it requires a high level of academic knowledge to know which values would be used to achieve a certain goal.

The second commercial model developed was introduced in a recently published paper named “Time–cost optimization in repetitive project scheduling with limited resources”, (2021). In this paper, Zou developed an efficient user-friendly model that covers time–cost optimization of repetitive projects with limited resources using constraint programming. The model is inclusive in several aspects, yet it can be further improved by including a stochastic element to it and not having it run on discrete well defined and fixed values of time and cost since such accurate information is not readily available in the construction industry until the project is finished.

### **2.3 Literature Gaps**

After deep analysis of the work accomplished in the literature on MRCTCTP-RP, few research gaps were identified for the proposed model to tackle. First, the developed model needs to adapt an organizational level view on the scattered repetitive projects’ resource deployment. Scattered projects are often short-term projects or minor tasks that an organization would have plenty of to handle at a time, scheduling them is a tedious and complex task. The developed model concept in this paper addresses the specific nature of scattered repetitive projects and takes into account the mobility time and associated costs of moving from one project to another. Nevertheless, introducing a stochastic simulation element in the model is a competitive edge to this model as it enables it to accurately mimic the varying working conditions and the uncertainty of the time taken to complete activities due to the human-element variation.

## **3 Model Logic Development**

### **3.1 Agent-Based Analysis**

AB modeling is a growing stream in literature where researchers try to utilize its potential in mimicking real life conditions in every work field. The main concept behind AB modeling is representing the elements of simulations in terms of instances that are left to interact with each other in a space defined by parameters and probabilities [1]. This technique of modeling is suitable to represent the complexity of the OPSRP problem where resources and projects interact with each other within the defined environment guided by certain routes. AB can be utilized to form resource deployment planning tools in which resources and projects are considered agents that get matched to perform a task until all project agents have been matched. In this

sense, OPSRP is considered the product of merging MTSP and the MRCTCTP-RP. Anylogic software is used in the developing the OPSRP model.

### **3.2 Model Data**

It must be noted that project agents here refer to the different locations requiring tasks to be done by the resources; these locations can fall within the same network or construction project but are here referred to as projects for ease of understanding. Resource agents is a general term describing the type of resource that requires a deployment plan; it can be human resources (labors, engineers, crews) or machinery (trucks, cranes, riggers, etc.). However if the resource is a machinery, then human factor in operating the machine is assumed to be included within the resource agent. To plan resource deployment for scattered repetitive projects, some important decision-making data for the projects and resources must be available. These general data are obtained from the user, and the fixed values (such as project locations) are stored in a database, while variable data are required for each plan development process. Some of the important data for both resources and project agents are mentioned below. These parameters are defined through simple data entry in an excel sheet structured in a user-friendly format. The model is then linked with the Excel sheet by identifying the sheet's location in the model.

#### **Resources agents' parameters:**

- Number of resources
- Resources starting location (latitude and longitude)
- Skill level defined for the resource
  - Skill level can be capacity for equipment or experience for human resources
- Standard deviation of resource
  - Indication of the resources ability to complete a task within the expected estimated time

#### **Projects agents' parameters:**

- Projects location (latitude and longitude)
- Priority ranking
  - The priority of the task needed for a project compared to the remaining projects tasks
- Estimated task duration
  - The estimated duration that a project task should take as determined by experience
- Skill level required to perform the task.

## 4 Methodology

### 4.1 Selection and Allocation Criteria

The model consists of two sets of independent agents, one carrying the projects' parameters, while the other carrying the resources' parameters. The two agents (projects and resources) interact with each other in an environment defined by a map covering the span of the projects. Spatial movement within the map is confined to the routes available. The resources agent is matched to a project agent where it performs two operations: mobility to the location of the project and performing the task required for the project agent. Resource agents get matched to projects agents one after another until all project agents have been matched one time.

The resources line up randomly with no specific order, and all the sorting and shuffling is done to the project agents. Project agents are sorted first according to the priority rank that they are assigned to by the user. The top 10 project agents move to the second phase where they get sorted based on distance to the resource in line. The top 5 projects closest to the resource in line are then matched to the skill level of the resources. If the closest project is compatible in skill level with the resource in line, they get matched and processed to the mobility time and task time of the project. If not, the second closest project is considered and so on. If, however, the closest 5 projects to the resource in line are not compatible in skill level, the following 5 s closest projects are considered.

After a resource agent is matched to a project agent, all remaining projects go back to the start of the sorting and matching process. The resources that finish the task come back in line to be re-matched to another project where the resource's location parameter is changed to be the last project location (new starting point). Figure 1 explains the sorting and matching hierarchy.

### 4.2 Simulation Time Calculations

Simulation time is the time taken to perform the required tasks for all projects considering operation time and mobility time for the resources to move from one location to another. Hence, the simulation time can be defined as mentioned in Eq. (1).

$$T_S = T_r + T_T \quad \text{Simulation Time Equation} \quad (1)$$

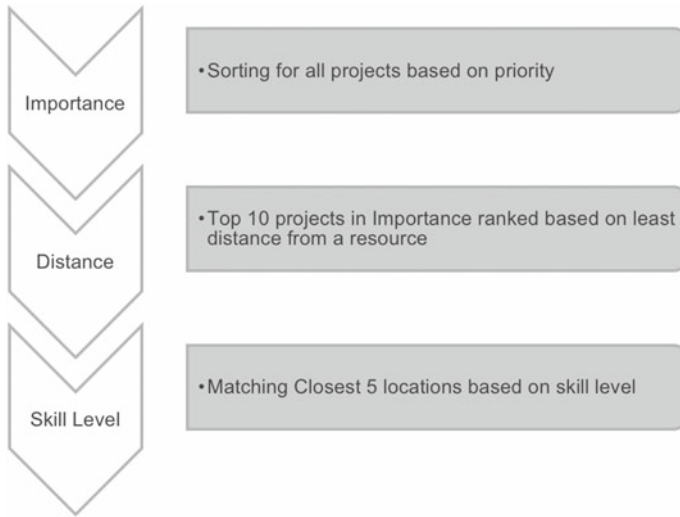
In which

$T_S$ : (Simulation time) time needed to perform all required tasks in all projects.

$T_r$ : (Travel Time) time taken in the mobility of resources between projects.

$T_T$ : (Task Time) time taken to perform required task at the projects.





**Fig. 1** Sorting and matching hierarchy

#### 4.2.1 Travel Time

Travel time is defined as the shortest distance ( $D$ ) between two projects divided by the velocity ( $V$ ) [2]. The distance ( $D$ ) is defined as the routing distance indicated by a GIS map imported on Anylogic using their built-in server which retrieves geographical, roads and traffic data. The coordinates of any points are read from the linked Excel file to indicate starting and ending points for the route. The model, then, utilizes the GIS map to navigate the crews from one point to another using the actual road network. The velocity follows a normal distribution [3] where the mean is the Traffic Average Velocity (AVT) and the standard deviation is the crew's standard deviation, to take into account the lost time and efficiency of each crew in following specified routes. AVT is determined based on a distribution extracted from Google Maps plotting the average velocity of road networks versus time to take into account the traffic congestion.

$$T_r = \frac{D}{V} \quad \text{Travel Time Equation} \quad (2)$$

In which

$D$  : shortest distance between two project locations following GIS maps routing.

$V$  : average velocity as defined in Eq. (3).

$$V = \text{Normal Distribution}(\mu, \sigma) = \text{Normal Distribution}(\text{AVT}, \sigma) \quad \text{Average velocity Equation (3)}$$

In which

$\mu$  = AVT : traffic average velocity value based on traffic distribution versus time.

$\sigma$ : standard deviation of the crew's performance time.

#### 4.2.2 Task Time

Task time [4] is simulated as a normal distribution that takes into account the nature of the task and crew's efficiency variability. The mean value of the task time normal distribution is taken as the estimated task time, while the standard deviation is the crew's standard deviation. Both parameters are defined by the user.

$$T_T = \text{Normal Distribution}(\mu, \sigma) \quad \text{Task Time Equation} \quad (4)$$

In which

$\mu$ : estimated task time, defined by the user.

$\sigma$ : standard deviation of the crew's performance time, defined by the user.

### 5 Financial Analysis

After investigating the operation costs, several elements were identified to be the main contributors. A bulk of the operation cost is attributed to depreciation and maintenance expenses. Depreciation expense for resources including but not limited to vehicles and equipment, where depreciation is the expense incurred on a company under operational cost. As depreciation here is used just only to spread the expense of owning and operating those vehicles and equipment through their useful and safe operation life, therefore using the straight line method for distributing the cost uniformly would be convenient.

$$\text{Depreciation expense} = \frac{\text{Initial Cost} - \text{Salvage Value}}{\text{Useful life}} \quad \text{Straight Line Depreciation Equation} \quad (5)$$

The maintenance cost is calculated by obtaining the average regular maintenance cost divided by the interval between two successive maintenance. After obtaining the following data, average fuel consumption (Liters/100 km), local fuel price (\$/l) and average travel speed (km/h), the fuel cost could be calculated using the below equation

$$\text{Fuel Cost}(\$/h) = \frac{\text{avg fuel consump.} * \text{fuel price}}{100} * \text{avg travel speed} \quad \text{Fuel Cost per Hour Equation} \quad (6)$$

Operation costs must incorporate the labor wages as they are a key resource in conducting the activities. Wages are influenced by multiple factors; based on the

duration of the shift, time of the day that the shift starts and ends, number of tasks to be completed and much more. Therefore, it was assumed that the only variance in wages came from the difference in skill levels and years of experience.

## 6 Model Commercial Development

Development of a commercial tool is one of the most important research gaps that was found. Not only that but it will also expose the research to a much wider audience. Therefore, more audience would most probably yield more feedback that would be beneficial for future development and improvement. The tool developed to date still needs some refinements before it could be introduced to the commercial market, but it has gone a long way.

### 6.1 User Interface

The model is developed to incorporate a user-friendly interface that eases its utilization for construction professionals. First a prestructured Excel sheet is filled by the user including all resources and project parameters as well as the financial values (fuel price, resources depreciation data and wages of labor). The user then proceeds with the welcoming window of the model (Fig. 2) where he placed the general data and the location of the structured excel sheet.

The user then proceeded to run the model and wait for the report stating the allocation of resources to which projects in order. The report includes planning and financial information as listed below.



Fig. 2 Model interface starting window

### Report Outputs:

- Sequence of projects for each resource
- Travel time and task time for all operations
- Associated operational costs.

## 6.2 Model Validation

The model was validated on a case study of 30 locations (each representing a project) within Cairo governorate region, Egypt. The resources were assumed to be maintenance crews each in a truck with different skill levels. The region GIS map was imported to the Anylogic model. Traffic data were extracted from Google Maps, where hourly congestion for one day on one of Cairo's major roads (the ring road) was taken as a representative of the average speed on the road network. Four resources (trucks with maintenance crews) with different starting locations and skill levels were included in the validation case. A 12-h shift was assumed for the validation case. Fuel price was taken as (0.49\$), and depreciation costs were leading to a depreciation cost of (0.012) \$/km of service. The model was utilized, and the sequence of projects to each resource was obtained in Table 1 as well as the estimates of the traveling time and task time listed in Table 2. Table 1 shows that for truck 1, the sequence of work is to cover projects 8, 14, 22, 27, 19, 9, and 29 in that order. The 30 projects were to be completed in two-day operations (two shifts). The operating cost of the validation case was nearly 650.94 \$, 69% of which is attributed to the depreciation of the resource (Trucks). Table 2 presents the total time spent to complete the work needed in the 30 locations and details the task time component compared to the traveling time component.

**Table 1** Sequence of projects for each resource (model output)

Resource (trucks)	Projects (locations)								
	P.1	P.2	P.3	P.4	P.5	P.6	P.7	P.8	P.9
Truck 1	8	14	22	27	19	9	29	–	–
Truck 2	6	13	18	30	20	3	24	5	7
Truck 3	16	1	23	26	10	15	12	23	–
Truck 4	2	11	4	28	25	21	17	–	–

\*Deployment schedule of each resource (truck) to cover which projects (locations) by order of operations

**Table 2** Simulation time results (validation case)

Parameter	Time (min)
Total simulation time	3902
Total task time	3003
Total travel time	899

## 7 Limitations and Recommendations

The developed model concept has two main limitations. First, the traffic data used is not accurate enough as it does not consider: days' variations, unusual traffic and weather conditions. More accurate representation of traffic data should be included. The second limitation is actively simulating the limits of the working shift. Currently, the developed model concept deals with the plan as a continuous operation. While in fact wasted time exists because a resource has time in the working shift, but not enough to perform any task. This influences the accuracy of the financial results.

The developed model concepts must be tested against implemented resource deployment plans to evaluate the magnitude of its contribution, which the authors of this paper are currently working on. Industry professionals need to be surveyed about other decision-making parameters that should be included in the sorting and matching process.

## 8 Conclusion

In conclusion, this research proposes a simulation model for developing resources deployment plans for scattered repetitive projects. The model incorporated GIS maps and traffic data to simulate the traveling time of resources between the different projects. The human element is also factored in by taking into consideration crew variability and stochastic values for the time. The model is effective in forming a fully functioning deployment plan and calculating the associated costs. The model utilizes a user-friendly interface allowing the user to set the parameters and easily view the simulation results.

## References

1. Borshchev A, Brailsford S, Churilov L, Dangerfield B (2014) Multi-method modelling: AnyLogic. In: Discrete-event simulation and system dynamics for management decision making, pp 248–279
2. Li H, Chan N, Huang T, Guo HL, Lu W, Skitmore M (2009) Optimizing construction planning schedules by virtual prototyping enabled resource analysis. *Autom Constr* 18(7):912–918. <https://doi.org/10.1016/j.autcon.2009.04.002>

3. Heravi G, Moridi S (2019) Resource-constrained time-cost tradeoff for repetitive construction projects. *KSCE J Civil Eng* 23(8):3265–3274
4. Tao S, Wu C, Sheng Z, Wang X (2018) Space-time repetitive project scheduling considering location and congestion. *J Comput Civil Eng* 32(3):04018017

# Pull-Based Simulation Modelling for Modular Construction Supply Chain Analysis: A Case Study in Northern Canada



Keagan Hudson Rankin, Zhuo Cheng, Zhen Lei, Samira Rizaee, Brandon Searle, Cynthia Ene, and Solomon Amuno

**Abstract** Modular construction, as an alternative to traditional stick-built construction, allows for schedule overlapping and efficient utilization of resources. In Canada, modular construction techniques are applicable to northern sites where labour and materials are scarce, the shipping season is short, and the method minimizes exposure to harsh winter conditions. However, when a modular construction supply chain gets longer and increasingly complex, analysing and managing it becomes a challenge. Additionally, there is a lack of research in modular scheduling compared to research in production. This paper introduces a case study where a supply chain model with manufacturing coordination was created and applied to analyse the shipment of housing modules from a factory in Southern Canada to communities in Nunavut. The case study was completed in partnership with Nunafab, a company trying to help solve Nunavut's housing crisis using modular construction. The model uses a pull-based approach, where the inputs are desired completion dates and information about the supply chain and the output is a schedule for module shipment. A discrete event simulation is also used to estimate manufacturing output at the beginning of the supply chain. The results show that the model can replicate a single season of housing module shipment, and work improvement performed on the factory reveals the opportunity to deliver more modules in less time. While having limitations, the proposed model is simple and can be generically applied to analyse similar supply chains. The case study highlights the importance of coordinating modular manufacturing with supply chain research when performing work improvement, and it has led to further collaboration between the researchers and Nunafab.

---

K. H. Rankin (✉) · Z. Cheng · Z. Lei · S. Rizaee · B. Searle  
Offsite Construction Research Centre, Department of Civil Engineering, University of New Brunswick, Fredericton, NB, Canada  
e-mail: [keagan.rankin@mail.utoronto.ca](mailto:keagan.rankin@mail.utoronto.ca)

K. H. Rankin  
Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON, Canada  
C. Ene · S. Amuno  
Nunafab Corp, Cambridge Bay, NU, Canada

**Keywords** Pull-based simulation modelling · Modular construction supply chain analysis

## 1 Introduction

Both academia and industry have seen growing interest in researching and applying modular construction principles since the 1990s. Modular construction (often used interchangeably with offsite or prefabricated construction) is an alternative to traditional construction methods where elements of the final build are prefabricated offsite and shipped to their final location for assembly [1]. Modular construction has promised improved construction productivity (both overall and onsite), higher quality, lower costs, and the potential for improved environmental performance [2–4]. Modular construction has also been suggested as a solution in regions where the site may be difficult or labour scarce [5]. Construction companies have started to look towards remote sites in Canada’s northern communities as potential markets for modular construction. In addition to being a viable market, an increase in residential modular construction may help alleviate the severe, ongoing housing shortage in northern communities [6].

Though research in offsite construction has increased in absolute terms (especially in Australia and Canada), most research has focused on modular production rather than project management, optimization, and scheduling [1, 7]. At the same time, recent research in supply chain simulation and successful implementation of modular construction has shown that good supply chain management and coordination with production scheduling is necessary for meeting deadlines and eliminating leftover stock [8, 9]. These so-called lean supply chains leverage construction technologies like digital twins and discrete event simulations (DES) to theoretically deliver more modules in less time [10, 11].

This paper attempts to address the lack of research in modular construction scheduling by providing a case study completed with an industry partner on a simple supply chain model that coordinates with manufacturing and is applicable to construction in Northern Canada. The case study is derived from the University of New Brunswick Offsite Construction Research Centre’s (OCRC) collaboration with Illu Incorporated/Nunafab (hereinafter Nunafab) during the summer of 2021. Nunafab is a company based out of Cambridge Bay, Nunavut that delivers prefabricated housing modules to parts of Northern Canada, primarily for residential purposes. Since completing a pilot project in 2019, they have been in the process of scaling up production. They are building a precast factory in Southern Canada which will ship housing modules to an assembly centre in Nunavut for processing. The modules will then go to sites around Northern Canada for final construction. Ultimately, Nunafab would like to help meet Nunavut’s housing needs using precast concrete modular construction. To achieve this goal, they needed a way to track, improve, and coordinate their remote supply chain. To address this need, researchers at the OCRC developed a pull-based model that took information about the supply



chain as inputs and returned a schedule as the output. The model consisted of site nodes connected by routes, and it was coded as a preliminary application in Microsoft Excel using Visual Basic Application (VBA). In addition to the supply chain model, a DES of Nunafab's planned precast factory was created using a time and motion study provided by the company. The DES was used to predict activity time at the precast factory node in the supply chain, and it was leveraged to perform work improvement on the given layout. This paper presents the background, methods, and results/deliverables of the work performed by the OCRC. It also discusses the model's limitations and future steps for collaboration between Nunafab and the OCRC.

### ***1.1 Supply Chain Obstacles***

Nunafab and the OCRC worked together on improving the engineering elements of their business, including the logistics behind their supply chain. Nunafab faced unique obstacles in delivering their modules to remote arctic destinations. These obstacles included:

- The shipping window from Quebec to Nunavut being limited to a period of a few months due to the freezing of shipping lanes.
- The construction time on site being limited due to winter weather.
- The predetermined schedules of shipping companies dictating the movement of modules.
- The need to track and assemble modules in remote regions of the supply chain.
- The need to produce enough modules in a suitable location offsite to meet demand.

Nunafab needed a way to overcome these obstacles and determine a schedule with beginning and end dates for the delivery of their modules.

## **2 Methodology**

The model developed in three distinct phases: the conceptual phase, the logic definition phase, and the application phase. During each phase, the OCRC presented outcomes to Nunafab and got feedback on the direction of research. The precast factory DES was created iteratively using an AI extension for Microsoft Excel called Axcel.io and information provided by Nunafab.

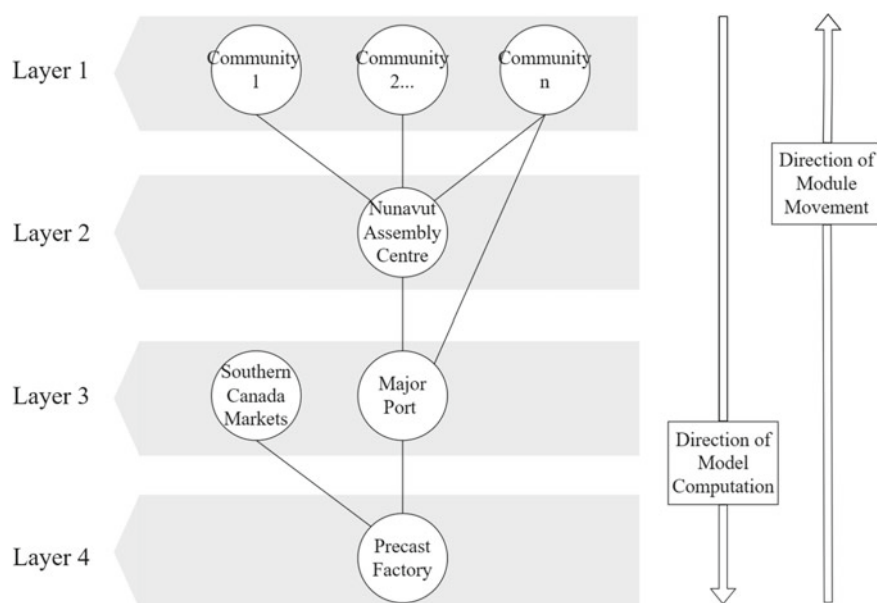
### ***2.1 Supply Chain Model***

During the conceptual phase of development, researchers at the OCRC and collaborators at Nunafab decided on a pulled-based approach for the model. The model

was conceptualized as two elements: sites and routes. Sites (or nodes) were locations where modules were stationary and undergoing an activity (manufacturing, assembly, construction) for some time. Sites were connected by routes. Routes contained shipping vessels or trucks with some volumetric capacity that transported modules from one site to another. The primary inputs of the model were the completion dates of construction at the northern communities for each individual module. Other preliminary inputs included information about ships along the routes, the lengths of routes, information about the modules, and the length of the shipping/building seasons in northern Canada. The model pulled the modules back through multiple nodes in the supply chain layer by layer, with certain modules skipping layers if conditional checks were satisfied. Once modules had been pulled all the way through, the model outputted a zero-float schedule.

After studying the planned supply chain, three “layers” of sites were defined: the multi-site Layer 1 containing northern communities, and Layers 2 and 4 which each contained a single site—the assembly centre in Nunavut and the precast factory. Layer 3, containing the major port from which modules were shipped, was included after consultation with Nunafab. Figure 1 includes a diagram of the sites, routes, and layers in the model.

After the concept of the model was defined, the OCRC set out to create a preliminary app which would demonstrate the application of the model to Nunafab. First, the logic was established. The logic used object-oriented programming and blocks



**Fig. 1** A high-level visualization of the supply chain model. Layers are numbered in the direction of calculation or “pulling”. Some modules may bypass the assembly centre through a conditional check

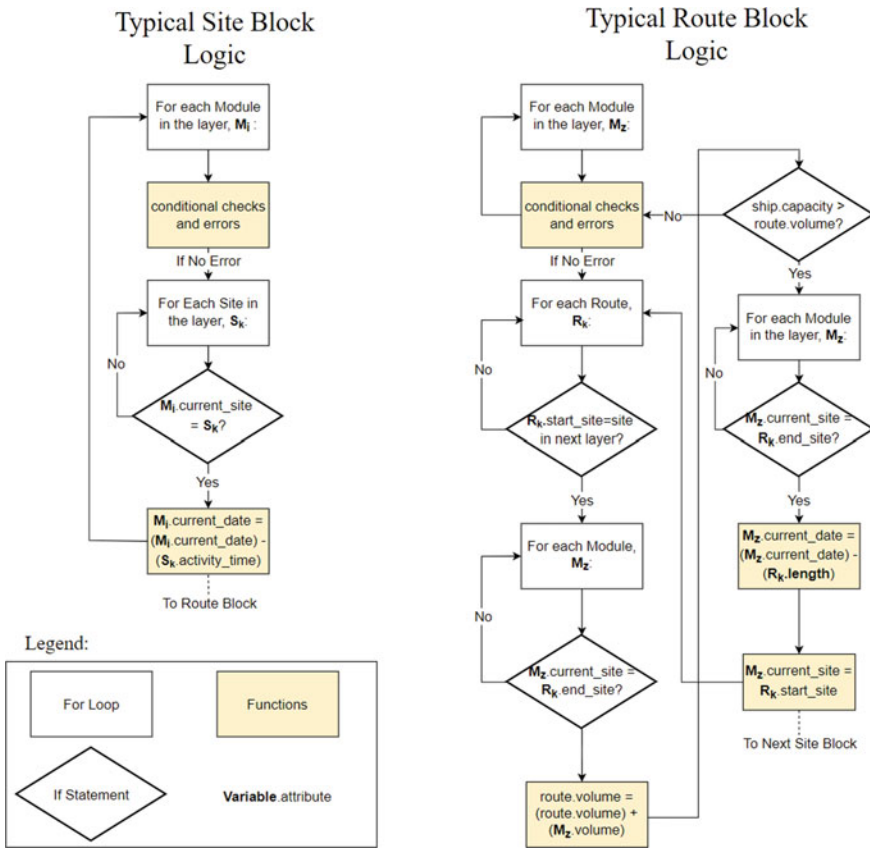
of code that followed the sites-and-routes concept of the model. Inputs were stored as class instances. These instances were all stored in lists to be accessed during the execution of the code. The blocks of code for each site and each route followed a similar format. To start, a site block was executed. The code looped through all sites in a layer to see if modules were currently at that site. If they were, then the site's activity time (stored from the inputs) was subtracted from the modules current time. The code also checked whether the seasonal building limit had been exceeded. Once all activity had been performed in a layer, the code would execute the next route block. The code looped through all routes which started at sites in the next layer (the layer closest to/pulling towards the precast factory). Modules in the current layer would move to the next layer and have the travel time subtracted from their current time unless a Boolean exception was specified in the inputs. Modules with exceptions could skip certain layers and blocks of code after a conditional check (e.g. if modules moved directly from the precast factory to a community). This process was repeated for the next layer of sites and routes until the precast factory in the final layer was reached.

When executing the routes blocks, the code checked to make sure each route did not exceed its capacity. The capacity of a route was the maximum volume of modules it could handle in one trip, and it was given by the ships and trucks that travel along it. The outcome when the capacity of a route was exceeded depended on whether it was a sea or land route. If capacity was exceeded on a sea route travelling North, the code would halt as the model assumes a single trips per shipping season. If the capacity was exceeded on a land route, multiple trips would be simulated, and times would be adjusted accordingly. The limitation of the model to a single season reflected Nunafab's production cycle and the realities of commercial shipping and sea ice formation; shipping to Northern Canada can only take place during the few months where there is no sea ice in the Arctic Archipelago, so most shipments occur in one bulk delivery. Figure 2 contains a logic diagram of the two code blocks described above.

Finally, the logic for the code was applied in a preliminary application built using Excel and VBA. The OCRC chose these programs as they were familiar to Nunafab and did not require investment into creating a new GUI. The application was organized into different sheets for different inputs, intermediate steps, and outputs. The code was structured to match the predefined logic. Many elements of the inputs were automated and organized to ensure no errors were made when running the code. Instructions were added to the application with links to shipping schedules. The application was verified against a previous project performed by Nunafab.

## ***2.2 Precast Factory Simulation***

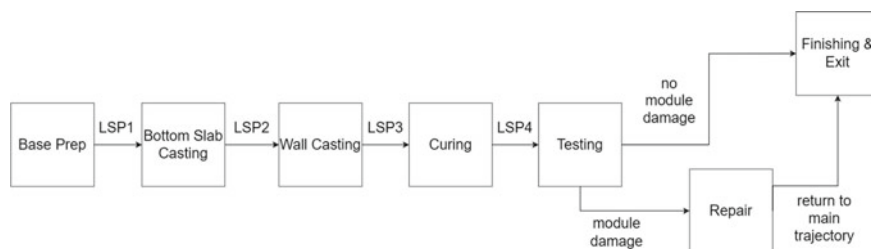
The precast factory DES was created using a factory layout and time and motion study provided by Nunafab. It provided the activity time for the first node in the supply chain model. When creating the factory simulation, events were first sorted



**Fig. 2** Typical site and route block. Dots denote properties of the objects (i.e.  $M_i$ .current\_site; current\_site is a property of module  $M_i$ )

and given durations based on the time and motion study. The number of stations and their buffers were inferred from the factory layout as well as through communications with Nunafab. The structure of the simulation was then considered. Modules were transferred from station to station via four “LSP” conveyor systems (referred to as LSP1–LSP4). The time that the LSPs took to transfer the modules between stations was estimated by multiplying their speed (given by Nunafab) by twice the average distance to the next set of stations. The entire system flowed in one direction as shown in Fig. 3. The only exception to the one-way flow was the repair station branch. If modules failed during testing, they were sent through the branch before returning to the main trajectory. Failure of modules was assumed to follow a Poisson process with a mean failure rate  $\lambda$  of one per day.

Once stations and buffer sizes were assigned, the DES was built in Excel using an artificial intelligence plugin called Axcel.io. Axcel’s could simulate production through its “machines” function. The DES held modules at a station or buffer area



**Fig. 3** Discrete event simulation structure implemented in Excel using the Axcel.io AI plugin

until the next station was free. To complete the DES, a constant time interval between the release of new modules was assumed based on Nunafab's targeted production output of 4 modules a day. Axcel returned information about each module's activity time as well as graphs showing resource utilization. The set-up allowed locations of queue build up to be identified. The simulation ran for a year of production assuming 12 h of daily operation.

The given layout of the factory had a massive queue build up and could not achieve Nunafab's desired output. Work improvement was performed to eliminate the build-up and even out utilization. Work improvement involved conducting a sensitivity analysis. Since the activity time at stations and floor area of the factory could not be changed, improvements consisted of rearranging of stations and buffer areas as well as altering the time interval between the release of new modules. Stations were rearranged within the limits of the building's geometry. Each iteration of work improvement aimed to lower the resource utilization of stations below 50% while also achieving Nunafab's output goal.

### 3 Case Study Results

This section details the results and deliverables from the case study. The supply chain model was verified using a real modular construction project completed by Nunfab 2019. Outcomes of the DES and work improvement are also considered.

#### 3.1 Supply Chain Replication

The final application of the supply chain model was an Excel program with 7 sheets that organized user inputs, ran the pull-based supply chain model using VBA, and outputted a zero-float schedule with dates for each module at each site. The first sheet of the file contained basic information on what the model was and how to use it. The next four sheets contained tables of a similar format for users to input information about shipment of modules for the year. Some tables had pre-inputted

information that was not altered by the user. Figure 4 contains an example of an input table in Excel. Final inputs included information about the shipping season, module dimensions, listing of sites and connecting routes (see Fig. 4), ship capacity, productivity information, module aggregation/building definition, and the desired construction completion date.

Repeated, unique keys/IDs were given to objects in the model in case Nunafab desired future integration with a database. They were also used to coordinate object interactions. Some inputs were found using third party sources. For example, route lengths were inputted with units of days and were found using an online shipping schedule [12].

The final two sheets of the program outputted the model’s intermediate processes and a zero-float schedule.

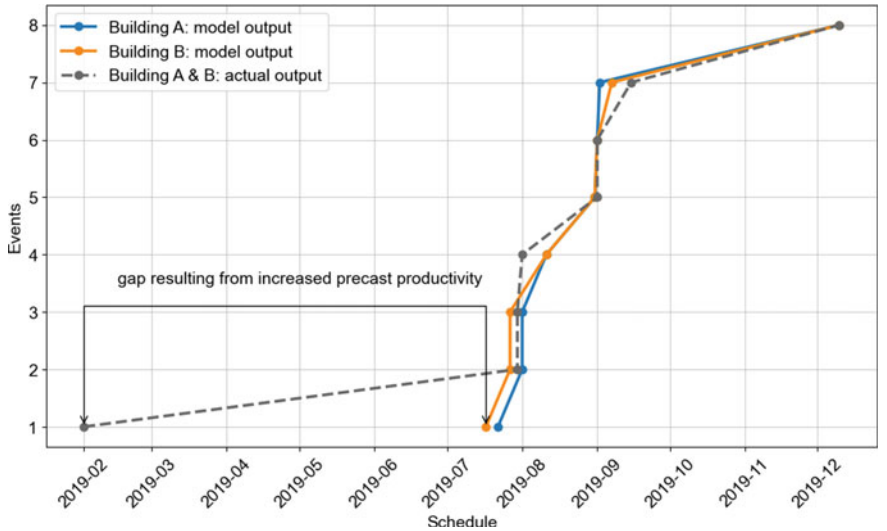
Information from a previous project completed by the Nunafab team in 2019 was used to verify the model. The project involved building two eightplexes (each consisting of 26 modules) in the Nunavut communities of Kugluktuk and Taloyoak. The targeted completion time for construction on site was the 10th of December. The model took these inputs and ran using a shipping route from the arctic sealift fleet. The outputs were compared to the actual schedule of the project created using dates given by Nunafab. Table 1 contains a side-by-side summary of the model’s outputs and the actual dates for the project, and Fig. 5 compares the dates graphically. “Building A” denotes the building erected in Kugluktuk and “Building B” the one in Taloyoak. The dates given for the model outputs were the earliest dates for individual modules in their respective buildings. The model fit the actual outputs well other than some variation in shipping lengths (due to uncertainty in actual model delivery) and the large gap in the production of modules at event 1. The gap resulted from the model’s assumption of increased productivity at the precast factory due to work improvement. This result shows the importance of coordinating production with the supply chain in modular construction; an improvement in module output will affect Nunafab’s time to deliver buildings and therefore their scheduling decisions within the narrow shipping season in Northern Canada.

SHIPPING ROUTES					
RouteID	$R_{ij}$	Route Name	Start Site	End Site	Length (days)
12		FactorytoStC	1	2	0.1
23		StCtoCBay	2	3	20
34		CbaytoKug	3	4	1
35		CbaytoGjoa	3	5	4
36		CbaytoTalo	3	6	6

**Fig. 4** Example of an input table implemented in Microsoft Excel. Grey cells show predetermined inputs that were not changed by the user

**Table 1** Summary of the supply chain replication

Events	Model outputs (date)		Actual outputs (date)	
	Building A	Building B	Building A	Building B
1. Started at precast factory	2019/07/22	2019/07/17	2019/02/01	2019/02/01
2. Shipped from factory	2019/08/01	2019/07/27	2019/07/30	2019/07/30
3. Arrived at shipyard	2019/08/01	2019/07/27	2019/07/30	2019/07/30
4. Shipped from shipyard	2019/08/11	2019/08/11	2019/08/01	2019/08/01
5. Arrived at assembly	2019/08/31	2019/08/31	2019/09/01	2019/09/01
6. Shipped from assembly	2019/09/01	2019/09/01	2019/09/01	2019/09/01
7. Arrived in community	2019/09/02	2019/09/07	2019/09/15	2019/09/15
8. Construction completed	2019/12/10	2019/12/10	2019/12/10	2019/12/10



**Fig. 5** Visualization of supply chain replication

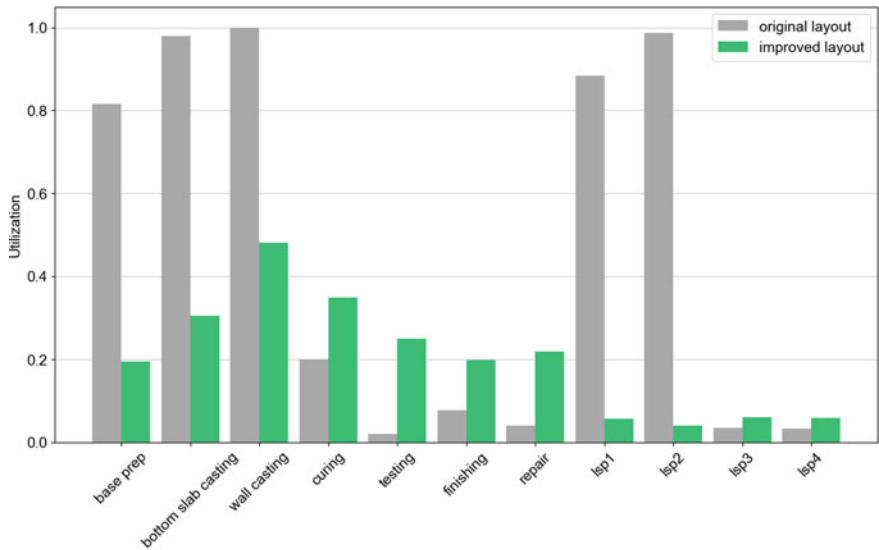
### 3.2 Precast Factory Work Improvement

The DES of the initial factory layout revealed a large queue build up at the first station (base prepping). This was because the wall casting station acted as a bottleneck to the system. It had a low number of stations and 8-h processing time. The bottleneck was eliminated through work improvement. Improvements made to the system are given in Table 2. These improvements generally consisted of reallocating space from finishing and testing stations to bottom slab casting and wall casting stations. The number of base prepping stations was also reduced to 1 as it was a quick activity and prepped bases could be queued in free bottom slab casting stations. The time

interval between the release of new modules was set at 160 min as it allowed for the completion of 4 modules a day over a year of simulation while minimizing percent utilization. The improvements to utilization are shown in Fig. 6. All final utilizations were under 50%. There was still some significant variation between the utilization of stations and the LSPs used to transport modules between stations; nevertheless, the final layout was superior to the original layout given the constraints.

**Table 2** Summary of precast factory work improvements

Resource	Original layout		Improved layout	
	# Of stations	Buffer size	# Of stations	Buffer size
Base prep	4	Infinite	1	Infinite
Bottom slab casting	2	2	4	0
Wall casting	2	0	7	0
Curing	9	0	9	0
Testing	10	0	3	0
Finishing	19	4	17	2
Repairing	4	0	5	0



**Fig. 6** Changes in utilization of factory stations after work improvement



### 3.3 *Discussion of Results*

Most of the error between dates in the supply chain model verification was likely caused by systematic differences between module production and shipping in the actual project compared to the model. The 2019 project did not use the new precast factory, hence the large difference in time required to produce the modules. The difference provided confidence to Nunafab in achieving their goal of scaling up yearly production. Another source of error was the shipping company. The shipping company used in the actual project was not the same as the one used in the model. This could account for some variations from the arrival at the shipyard and onward, though most companies ship on similar schedules within the short open-water season.

The preliminary application had several limitations. First, some of the planned features were missing. This included lack of ability to add float, running over multiple seasons, and the option to export to CSV. Nunafab also expressed a desire to host the model online to make team coordination easier. Finally, the preliminary application was coded without runtime optimization in mind due to the low number of modules in the case study. Rebuilding the app in Python and vectorizing operations using NumPy would reduce the runtime of a more complex simulation and be easier to host online. These changes could be made while retaining the core model. Exporting to csv would also only involve a small addition of VBA code. Applying the model over multiple seasons would involve looping modules which could not be completed in one season into another pull through the supply chain. Additionally, a more in-depth time and motion study of Nunafab's factory could yield better inputs for the DES and distributions for station times rather than constants, and further work improvement will be done in collaboration with the company as they finalize their precast factory layout. A final consideration moving forward with work improvement is the plugin that was used to create the DES. Axcel.io was recently purchased by Salesforce and is no longer available to the public. Final deployment of the DES for further work improvement or for use with the supply chain model will require the simulation to be rebuilt in a different program or using a custom simulation developed by the OCRC.

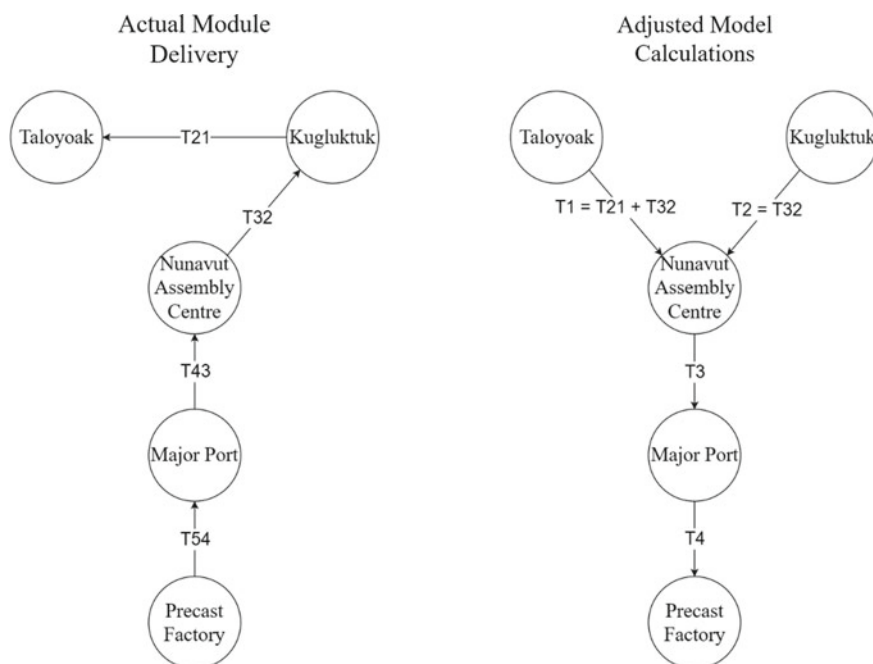
The final limitation was the difference between the model structure and the way that real modules were shipped to northern Canada. The model included all communities in single layer, so modules at each site were pulled back to the Cambridge Bay simultaneously along a single route. Contrarily, when modules were shipped to Northern communities in real life, they were brought by a shipping vessel that visited multiple communities on a set, linear timeline. Departure dates and instructions on modifying route lengths were added to overcome the model's inability to simulate this process. Departure dates determined when modules were taken to the next site on a route. The model had modules wait at a site for the departure date if they were early, or it would display an error message if the modules could not make it to a site before the departure date. Also, users would modify inputted route lengths according to the extra time taken and sites visited between departing the assembly centre and arriving at the final build site. For example, if a module was being shipped from Cambridge Bay to Taloyoak, but the ship stopped at Kugluktuk along the way, then

the length of the route between Cambridge Bay and Taloyoak would be the length from Cambridge Bay to Kugluktuk plus the length from Kugluktuk to Taloyoak. Figure 7 and Eq. 1 show a generalized form of this modification.

For site  $i$  in the final layer with  $n$  preceding sites on the ships path starting at  $m$  (the assembly centre):

$$T_i = T_{i1} + T_{i2} + \cdots + T_{im} \quad (1)$$

The preliminary application resulting from the case study was created as a basic tool and a proof of concept. Future collaboration between the OCRC and Nunafab will include creating an improved application based on the same model. The improved application will be hosted online, have a custom front end, and feature more tools to help Nunafab manage their supply chain. The additional features would consist of the improvements discussed above. Current model outputs are already being used to produce supply chain and precast factory animations for Nunafab. Stills from these animations are included in Fig. 8. These animations are important for helping the company understand the outcomes of the supply chain research and look further into the spatial constraints of the precast factory.



**Fig. 7** Actual delivery of modules and the adjusted method used in the application. The simple principles of the model allow it to adjust to the shipping route constraints of the case study



**Fig. 8** Stills from a preliminary animation of the supply chain

## 4 Conclusions

Researchers at the OCRC completed a case study in modular construction scheduling and production coordination in collaboration with industry partner Nunafab. The case study demonstrated the information advantage of production coordination and the potential for using modular construction to expedite residential construction in Northern Canada. The outcome of the case study was a simple but flexible pull-based model and preliminary application for Nunafab's remote supply chain. The model used physical inputs and desired finish dates to determine a schedule for precast module's delivery. Sites and nodes were the two key components of the model. Development was completed over three stages: conceptualization of the model, definition of the logic, and coding of the preliminary application. The model was coordinated with a production simulation used to perform work improvement and quantify potential efficiency gains by comparing outputs against a previous project. The rest of the verification showed that the model generally agreed with an actual northern supply chain. Finally, given the limitations associated with the application, future work was suggested to improve and expand the features in collaboration with Nunafab as well as continue work improvement on their precast factory.

**Acknowledgements** The authors would like to thank Nunafab and Axcel.io for their help in completing this project. They would also like to thank the funding from their Industrial Research Assistance Program (IRAP), National Research Council (NRC) of Canada, and Discovery Grant (Grant Number: RGPIN-2020-04126), Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

1. Hosseini MR et al (2018) Critical evaluation of off-site construction research: a scientometric analysis. *Autom Constr* 87:235–247. <https://doi.org/10.1016/j.autcon.2017.12.002>
2. Blismas N, Pasquire C, Gibb A (2006) Benefit evaluation for off-site production in construction. *Constr Manag Econ* 24(2):121–130. <https://doi.org/10.1080/01446190500184444>
3. Heravi G, Kebria MF, Rostami M (2019) Integrating the production and the erection processes of pre-fabricated steel frames in building projects using phased lean management. *Eng Constr Arch Manag* 28(1):174–195. <https://doi.org/10.1108/ECAM-03-2019-0133>
4. Jin R, Hong J, Zuo J (2020) Environmental performance of off-site constructed facilities: a critical review. *Energy Build* 207:109567. <https://doi.org/10.1016/j.enbuild.2019.109567>
5. Gibb AGF, Isack F (2003) Re-engineering through pre-assembly: client expectations and drivers. *Build Res Inform* 31(2):146–160. <https://doi.org/10.1080/09613210302000>
6. Nunavut Housing Corporation (2016) Nunavut is facing a severe housing crisis. [https://assembly.nu.ca/sites/default/files/TD%201584\(3\)%20EN%20Nunavut%20is%20Facing%20a%20Severe%20Housing%20Crisis.pdf](https://assembly.nu.ca/sites/default/files/TD%201584(3)%20EN%20Nunavut%20is%20Facing%20a%20Severe%20Housing%20Crisis.pdf). Accessed 8 Feb 2022
7. Hussein M et al (2021) Modelling in off-site construction supply chain management: a review and future directions for sustainable modular integrated construction. *J Clean Prod* 310:127503. <https://doi.org/10.1016/j.jclepro.2021.127503>
8. Hsu PY, Aurisicchio M, Angeloudis P (2019) Risk-averse supply chain for modular construction projects. *Autom Constr* 106:102. <https://doi.org/10.1016/j.autcon.2019.102898>
9. Hussein M, Zayed T (2021) Critical factors for successful implementation of just-in-time concept in modular integrated construction: a systematic review and meta-analysis. *J Clean Prod* 284:124716. <https://doi.org/10.1016/j.jclepro.2020.124716>
10. Innella F, Arashpour M, Bai Y (2019) Lean methodologies and techniques for modular construction: chronological and critical review. *J Constr Eng Manag* 145(12):04019076. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001712](https://doi.org/10.1061/(asce)co.1943-7862.0001712)
11. Pan Y, Zhang L (2021) A BIM-data mining integrated digital twin framework for advanced project management. *Autom Constr* 124:103564. <https://doi.org/10.1016/j.autcon.2021.103564>
12. Arctic Sealift (2021) Sealift schedule. <https://www.arcticsealift.com/en/schedule.php>

# Construction Payment Problems: A Critical Review of Payment Problems, Prompt Payment Legislation, and Challenges



Dalia H. Dorrah and Brenda Y. McCabe

**Abstract** The increasing complexity of construction projects combined with the inefficiencies in conventional construction management techniques has led to unnecessarily elongated payment cycles and other payment problems which can adversely affect the performance of projects and involved stakeholders. The rising concerns around construction payment problems have led to the introduction of prompt payment legislation in various jurisdictions worldwide with the common objective of increasing cashflow down the construction pyramid in a fair and reasonable manner. While prompt payment legislation aims to alleviate payment problems, their impacts and challenges remain unclear. In light of this, the main objective of this paper is to investigate construction payment problems, their common root causes, impacts, and potential mitigative options. Moreover, the paper also provides a critical review of prompt payment legislation introduced and implemented in different international jurisdictions for addressing the timeliness of payment in the construction industry. This review was followed by an analysis of the challenges in the implementation of such legislation as well as recommendations for effectively addressing them. Finally, this research can be utilized as a valuable resource for various jurisdictions who wish to address construction payment problems by introducing and effectively implementing prompt payment legislation.

**Keywords** Construction payment problems · Legislation and challenges

## 1 Introduction

Money is often referred to as one of the 4 M's of construction along with manpower, material, and machinery. Without adequate funds, owners can neither initiate new projects nor pay their contractors. At the same time, contractors need continuous

---

D. H. Dorrah (✉) · B. Y. McCabe  
Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON, Canada  
e-mail: [dalia.dorrah@mail.utoronto.ca](mailto:dalia.dorrah@mail.utoronto.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_22](https://doi.org/10.1007/978-3-031-34593-7_22)

333

cashflow to maintain their work progress and pay subcontractors who in turn pay for materials, equipment, and labour; the cycle progresses down the construction pyramid. This critical role of cash and cashflow makes payments among the main negotiated provisions in construction contracts and subcontracts to ensure a smooth cashflow for project performance and work progress.

Unfortunately, the dynamic nature of construction projects combined with the large number of stakeholders complicates and lengthens the payment process. Payment problems are often regarded as a main cause of dispute in the industry worldwide [1]. One common example is the industry-wide practice of late payments. Payment delays may be attributed to conflicts over work performance, incomplete invoices, and financing issues [2]. Other problems may arise from the unequal balance of power among stakeholders, such as unfair contractual agreements and payment provisions that transfer financial risks to others and put them in involuntary financing roles [3]. However, not all stakeholders have the authority to negotiate such conditions nor the financial capacity to withstand them. Such poor payment practices may result in adverse impacts that cascade down the construction pyramid and affect project performance, stakeholder financial stability, employment, productivity, and investment [2, 4].

With the persistence and critical impacts of payment problems, prompt payment legislation has been implemented in various global jurisdictions to alleviate the financial risks encountered in the construction industry. The aim of the legislation is to address the timeliness of payment and enforce payment for work performed [3]. Nevertheless, their implementation, challenges, and impacts on projects and stakeholders remain unclear. Based on the above, this paper will provide a critical review of payment problems in the construction industry, their causes, impacts, and solutions proposed in the literature. In addition, it will explore the prompt payment legislation adopted in a few international jurisdictions, their challenges, and effective implementation.

## **2 Construction Payment Problems and Related Legislation**

This section provides an overview of the construction payment process, its problems, impacts, and mitigative options. In addition, it investigates prompt payment initiatives and legislations proposed to address construction payment problems in various global jurisdictions.

### ***2.1 Causes of Construction Payment Problems***

The typical payment process starts with the contractor submitting a payment application to the owner for review and certification for payment. This application should include supporting documents for work performed under the respective contract

agreement such as invoices, monthly schedule update, schedule of values, and proof of payment to subcontractors and suppliers. Once an application for payment is submitted to the owner, a process of certification begins with the consultant's review, followed by the owner's approval, and then payment processing and transfer within the deadlines specified in the contract. The involvement of multiple stakeholders and their tasks make it time consuming and error prone. Accordingly, any flawed or incomplete documents, inefficiencies, or delays at any stage would further elongate the process and affect the performance of projects and stakeholders. Other factors contributing to payment problems include bureaucratic delays in approving payments, financing problems or insolvencies of owners or general contractors, unexplained delays of the general contractor, disputes over alleged deficiencies, and delays of by the payment certifier.

This paper reviews the literature of construction payment problems. The review started by investigating the root causes of payment delays, reduced payment, or nonpayment. The common root causes were identified and organized in seven categories: (1) contracts and regulations; (2) finances and cashflow; (3) claims and disputes; (4) payment process; (5) poor practices; (6) variations; and (7) other. These categories are further detailed in Table 1. It is shown in the table that payment problems are symptoms of more persistent underlying issues in construction projects. Some of these issues can be attributed to stakeholders across the construction pyramid and not only upper tier. For instance, contractors and/or subcontractors may have poor documentation procedures that disrupt the smooth flow of payments. Similarly, owners may have financing difficulties. Some factors can contribute to payment problems at the design, contracting, construction, or closeout phases.

## ***2.2 Impacts of Payment Problems***

Regardless of the cause, payment problems can adversely affect the performance of projects and stakeholders. The common effects of payment problems in the literature were reviewed and organized in project and stakeholder-related impacts as shown in Table 2. Affected project performance includes additional costs, disrupted or elongated durations, and poor-quality work. As for stakeholders, the impacts may cascade down the payment pyramid and affect small-size entities who have little bargaining strength to negotiate payment terms [3]. For instance, owners may delay payments by imposing unfair payment conditions through the tendering process. Similarly, general contractors may impose conditional provisions as pay-when-paid or pay-if-paid on subcontractors so as not to release payments to them until after receipt of payments from owners. Such practices emerge from the focus of stakeholders on transferring financial risks to other stakeholders regardless of the impacts on others and on projects.

**Table 1** Common root causes of construction payment problems in the literature

Aspect	Cause	Sources
Contracts and regulations	Failure of parties to understand and/or satisfy the contract agreement and preset procedures for payment	[1, 5, 6, 7, 8, 9, 10, 11]
	Improper choice of the standard form of contract or the use of contracts that may not be explicit on payment provisions	[1, 12]
	Adoption of contingent and/or unfair payment clauses that are burdensome and transfer liabilities to other parties	[1, 13–15]
	Lack of statutory legislation governing payment timing	[1]
Finances and cashflow	Owner's poor financial management and ineffective use of funds	[1, 7, 13, 14, 16–18]
	Cashflow difficulties or scarcity of capital to finance projects	[1, 7, 13, 14, 19]
	Poor estimation of the project cost and setting unrealistic cashflow plan with work done exceeding allocated budget	[1, 20]
	Financial failure due to bankruptcy	[1]
Claims and disputes	Failure in managing claims and/or settling disputes	[1, 14, 20]
	Disputes over major defects or poor quality of works	[1, 13, 14, 21, 22]
	Delays or errors in supporting documents for payment claims	[1, 13]
Payment process	Many parties in payment approval and certification process	[13]
	Lack of coordination and adequate flow of information between project teams	[5, 6, 14]
	Bureaucracy in payment process	[1, 13]
Poor practices	Owner's deliberate withholding of payments for work done for their own financial advantage and economic standing	[1, 13, 14, 18, 19]
	Tolerance of contractors and subcontractors in accepting late or nonpayment to preserve the relationship with upper tiers	[1, 13, 14, 18, 23, 24]
	Failure to agree on the valuation of work done	[1, 5, 7, 11, 13, 14, 15, 25]
	Poor documentation procedures and/or delay in preparing and submitting payment documents	[1, 8, 12, 14, 15, 26]
Variations	Variation order and unclear requirement for variation work	[1]
	Slow process of approving variations	[1]

(continued)



**Table 1** (continued)

Aspect	Cause	Sources
	Shortage of funds when variations increase contract sum	[13, 19]
	Inability of parties to agree on value of variation work done	[13]
Other	Lack of personnel knowledge and experience	[7, 21]
	Ripple effects of an economic downturn	[1]
	Change in key project personnel	[1]

**Table 2** Impacts of payment problems in the literature

Category	Description	Sources
Project-related	Increased construction costs as contractors factor in costs of financing, interest charges, and late payment risks in tenders	[4, 13]
	Increased project costs and decreased profitability	[1, 13]
	Delay in work progress and/or completion of project due to interruption or suspension of works	[1, 13, 27, 28, 29]
	Incurred costs on idle equipment and other overheads	[21]
	Poor quality of work performed	[13]
	Abandonment of projects and termination of contract	[1, 13]
	Delay in obtaining expected benefit of project	[13]
Stakeholder-related	Problems acquiring funds from financial institutions which increases financing costs and reduces profitability	[1, 30]
	Cash flow problems and financial hardship which may result in insolvency or liquidation of firms	[1, 7, 10, 13, 15, 27, 30, 31]
	Shifted financial burden from one party to another resulting in negative cascading effects on involved parties	[1, 10, 13, 15]
	Inability to tender for another work	[13]
	Difficulties in procuring materials and equipment	[1, 32]
	Adverse effect on stakeholders' relationships and reputation	[1, 13]
	Payment disputes adding to reduced or delayed payments	[7, 13]

### 2.3 Solutions of Construction Payment Problems

The criticality of payment problems motivated researchers to explore approaches for addressing their root causes. These approaches can be organized in contractual and regulatory measures, bonds and security measures, as well as administrative and management measures (Table 3). Aspects of the contractual and regulatory measures are incorporated in prompt payment legislation. Bonds and security measures are typically considered by stakeholders and included in construction contracts. Administrative and management measures differ based on the stakeholders involved in a project and would require cultural and organizational changes. In all cases, the applicability, feasibility, and effectiveness of such solutions require further analysis and study to identify adequate solutions for projects and jurisdictions.

**Table 3** Proposed solutions to payment problems in the literature

Category	Description	Sources
Contractual and regulatory measures	Implementation of statutory enactment for payment issues	[8]
	Proper implementation of payment provisions in construction acts and standard forms of contract	[8, 21]
	Clearly defining roles and responsibilities under contracts	[1]
	Establishing the right to regular periodic payment, and the removal of contingent payment clauses	[1, 8]
	Establishing the right to lien and expeditious dispute resolution mechanisms for payment disputes	[1]
	Enforcement of contractual payment clauses as charges to overdue payments, work suspension, and reduced work rate	[1, 8]
Bonds and security measures	Payment guarantees by upper tiers and prequalification of their financial status	[21]
	Use of advance, payment, and retention bonds to secure payment	[1, 21]
Administrative and management measures	Training for cashflow and financial management	[1, 8, 14]
	Performing regular audits and mutual discussions about payment problems to identify any financial shortcomings	[1, 8, 18]
	Increasing understanding about payment terms and awareness about the effects of payment issues	[1, 8]
	Contracting with reputable payers	[1]
	Submission of progress work invoicing with adequate documents	[8]

## ***2.4 Prompt Payment Initiatives and Legislation***

The implications of payment problems and disputes motivated the introduction of prompt payment legislation in jurisdictions worldwide to address the timeliness of payments and increase cashflow [3]. The legislation originated in the USA with the federal Prompt Payment Act in 1982 [33]. This was followed by similar movements in other jurisdictions as the UK, New Zealand, Singapore, Ireland, Malaysia, and Canada.

Jurisdictions varied their prompt payment legislation to shape the payment system and stakeholders' interactions. Nevertheless, recurring features include: (1) the right of a contractor and/or subcontractor to make claims for progress payments; (2) the obligation on an owner or a general contractor to evaluate a claim for payment within a reasonable time; (3) the right to give written notice of a disputed claim for payment along with the reasons for it; and (4) the imposition of penalties on late payments. This paper reviews international prompt payment legislation to identify their features and challenges. An overview of payment notices and periods in legislation in seven jurisdictions follows.

The United States Federal Prompt Payment Act was enacted in 1982 and amended in 1988 [33]. Under the Act, the government agency shall return any unsuitable invoice within 7 days of receipt. A prime contractor may withhold payment to its subcontractors after providing a notice prior to the payment due date. Payment to a contractor must be made within 14 days after invoice receipt for progress payments and 30 days after final acceptance for final payment. Payment to subcontractors must be made within 7 days after receipt of payment by contractor.

The United Kingdom Housing Grants, Construction and Regeneration Act 1996 was amended by the Local Democracy, Economic Development and Construction Act 2009 which came into force in 2011 [34, 35]. Under the Act, a payer shall give notice specifying the amount proposed within 5 days after the payment due date and give a notice of intention to withhold or reduce payment (if any) within 7 days before the final date for payment. Progress payments shall become due on the later of 7 days after the end of relevant period or the date of a claim made. Final payment shall become due on the later of 30 days after completion of work, or the date of a claim made. In both cases, the final date for payment shall be 17 days after the payment due date.

New Zealand's Construction Contracts Regulations came in force in 2003 and were amended in 2015 [36]. Under the Regulations, a payer must send a payment schedule to withhold amount by date in contract. If not in contract, then within 20 working days after serving a payment claim. Payment under the default scheme is due and payable 20 working days after serving a payment claim.

In Singapore, the Building and Construction Industry Security of Payment Act came in force in 2005 and was amended in 2019. Under the Act, a respondent to a payment claim shall provide a payment response specifying the amount by the earlier of date in contract or 21 days after serving a payment claim. If not in contract, then within 14 days after serving a payment claim. Payment becomes due and payable on

the earlier of date in contract, 35 days after submitting a tax invoice to respondent, or 35 days after payment response is due. If not in contract, it shall be due by the earlier of 14 days after submitting a tax invoice to respondent, or when payment response is due.

The Malaysian Construction Industry Payment and Adjudication Act was enacted in 2012 and came into force in 2014 [37]. Under the Act, a party who disputes the amount claimed in a payment claim shall serve a payment response within ten working days of payment claim receipt. Payments are to be made monthly within 30 calendar days from the receipt of an invoice.

Ireland's Construction Contracts Act was enacted in 2013 and came into force in 2016 [38]. Under the Act, an executing party shall deliver a payment claim notice to the other party within 5 days after payment claim date. If the other party contests this notice, a payment claim response must be delivered within 21 days after payment claim date. Payment claim dates shall be 30 days after commencement, every 30 days thereafter up to substantial completion, and 30 days after final completion. Payments shall be performed within 30 days after payment claim date.

The Canadian Federal Prompt Payment for Construction Work Act received royal assent in June 2019 but is not yet in force [39]. Under the Act, notice of nonpayment must be submitted after invoice receipt within 21 days for contractors and 28 days for subcontractors. Payment to contractors must be made within 28 days after invoice receipt. Contractors must pay subcontractors for the work that was paid for in the invoice within 35 days after invoice receipt.

Other payment-related aspects follow from the review of prompt payment legislation.

- **Payment notices and period:** The deadlines for notices and payments show that payers need to make their payment decisions on relatively short timelines. They may not withhold payment without providing notice within a stipulated period with the withheld amount and reasons for withholding. Moreover, conditional provisions are expressly prohibited. These procedures may be challenging for payers but enable payees to assess their financial situation and plan accordingly.
- **Level of contract:** The legislation implemented in the jurisdictions applies to all levels of the construction pyramid. However, they may vary by payment regimes from owners to contractors and from contractors to subcontractors. This is more obvious in the US legislation.
- **Payment triggers:** The payment trigger needs to be clearly specified to estimate a payment due date and the date upon which late payment interest penalty shall begin to accrue. Triggering events can include receipt of invoice, contract terms, approval or certification of payment claim, contract terms and/or approval, receipt and/or approval of invoice, contract terms, and monthly payments.
- **Payment dispute resolution:** The expedited dispute resolution is crucial to avoid dispute escalation and limit their impacts. Thus, prompt payment legislation may also establish a right to refer disputes to adjudication to expedite dispute resolution and increase the efficiency for enforcing payments [3]. Among the jurisdictions,

the USA is the only one with no expedited mechanism in its federal and state-based prompt payment legislation. Instead, disputes are resolved through litigation, which can be costly and time consuming. In other jurisdictions, legislation establishes the right for any party in a contract to refer disputes to adjudication.

- Nonpayment remedies: In the event of nonpayment, it is important to provide compensations or remedies to cover any losses and financial stress the parties endured until payment was fully made. Accordingly, prompt payment legislation establishes a right to remedies in the form of interest penalties, work suspension, recovery of costs in a civil action, or other.

### **3 Challenges and Recommendations**

The review and analysis of prompt payment legislation reflects the positive impacts that they may offer to the construction payment system and stakeholders. However, achieving these benefits entails identifying and addressing the challenges that may hinder their effective implementation.

#### ***3.1 Bottlenecks in the Traditional Payment Process***

Bottlenecks in the traditional payment process may need to be restructured. Traditional payment processes require many tasks and procedures for documentation, progress verification, and certification. These tasks may impose bottlenecks in the payment process that increase with the internal bureaucracy of some owners, specifically public owners. However, since the payment legislation requires that payment would not be preconditioned by the certification of payments, this can add pressure on the payment certifier to assess the work performed in a payment application before the deadlines for payment. To manage bottlenecks, stakeholders may need to increase their efficiency or consider the incorporation of new techniques such as automated generation of payment applications and automated progress tracking and verification. Automated systems would be used to support and expedite the payment process and should not be considered as an additional requirement for stakeholders to prove their entitlement to payments.

#### ***3.2 Resistance to Mindset and Culture Change***

Resistance to mindset and culture change may emerge from the perception of stakeholders of the payment legislation and the rights they establish for fair and reasonable payment practices. In many cases, stakeholders may still accept delayed payments

or not use their rights to interest penalties to maintain good relationships with upper-tier stakeholders. However, stakeholders would not recognize the effectiveness of such legislation without changing their mindset towards timely payment and understanding the immense impacts that poor payment practices may cause. A paradigm shift has to happen to move the industry to the point that timely and fair payments become status quo. Stakeholders need to consider the overall performance of projects and not be consumed by their individual gains by deliberately withholding payments. They need to understand the short and long-term impacts of their decisions and that a failure of a project and/or other stakeholders also reflect their own failure. Moreover, lower-tier stakeholders need to use their rights without negatively affecting their reputations or relationships with other stakeholders.

### ***3.3 Inconsistent and/or Inefficient Information Exchange***

Inconsistent and/or inefficient information exchange may be exacerbated as they are data-driven and most legislation requires the submission of proper invoices for triggering the payment period. The adjudication and dispute resolution process typically requires the submission of evidentiary and supporting documents. Therefore, any inconsistencies or delays in information and document exchange would hinder the smooth progress of the payment process and the performance of projects. This challenge can be addressed by offering a transparent process of data gathering and exchange of the actual work performed through the integration of building information models and reality capture techniques. Organizations can also develop automated payment systems based on these data. This can be helpful in shortening investigations and reporting processes, providing tools for closely tying payments to actual site progress, and improving information sharing.

### ***3.4 Unclear Impacts of Prompt Payment Legislation***

Unclear impacts of prompt payment legislation affect the construction industry worldwide. However, jurisdictions address them differently. Moreover, the impacts and cost savings offered by the introduction of prompt payment legislation remain unclear. More research is needed to study prompt payment legislation and identify the lessons learned from their implementation. This includes investigating the nature of the payment problems in light of such legislation and approaches to address them. Such information would be valuable at the industry-level to enable jurisdictions to assess the feasibility of legislation, develop and introduce similar ones, or modify existing rules. It would also be important to stakeholders to understand the potential benefits of abiding by the statutory timeframes and the consequences of payment decisions.

## 4 Conclusions

Payment problems adversely impact the performance of projects worldwide. This research presented a critical review on the problems associated with payments in the construction industry with emphasis on prompt payment legislation in a few global jurisdictions. The review discussed the root causes of payment problems, their impacts, and potential solutions proposed in the literature. This illustrated the criticality of payment problems and the great need to address them. The authors also investigated international prompt payment legislation and their role in addressing payment problems. Based on the review, an analysis was performed to identify the challenges that may hinder the effective implementation of prompt payment legislation. Finally, recommendations were proposed to facilitate the effective implementation of payment legislation including: (1) increasing the efficiency of the traditional payment process through automation; (2) shifting the mindset of stakeholders towards timely payments; (3) supporting a transparent environment for consistent information exchange; and (4) performing extensive research on the impacts and drawbacks of prompt payment legislation. Successfully solving payment problems in the construction industry would not be realized without the informed collaboration of its entities.

## References

1. Peters E, Subar K, Martin H (2019) Late payment and nonpayment within the construction industry: causes, effects, and solutions. *J Legal Affairs Dispute Resolut Eng Constr* 11(3):04519013
2. Ipsos Reid (2015) Trade contractor survey report. Prompt Payment Ontario (PPO)
3. Reynolds B, Vogel S (2016) Striking the balance: expert review of Ontario's construction Lien Act. Ministry of the Attorney General. Government of Ontario
4. Prism Economics and Analysis (PRISM) (2013) The case for payment reform in the construction industry. National Trade Contractors Council of Canada (NTCCC)
5. Ayodele EO, Alabi OM (2011) Abandonment of construction projects in Nigeria: causes and effects. *J Emerg Trends Econ Manag Sci* 2(2):142–145
6. Danuri MSM, Hanid M, Munaaim MEC, Rahman HA (2006) Late and nonpayment issues in the Malaysian construction industry: contractor's perspective. In: International conference on construction, culture, innovations and management (CCIM), pp 613–623
7. Hasmori MF, Ismail I, Said I (2012) Issues of late and nonpayment among contractors in Malaysia. In: Proceedings of the 3rd international conference on business and economics research
8. Mohamad N, Suman AS, Harun H, Hashim H (2018) Mitigating delay and nonpayment in the Malaysian construction industry. *IOP Confer Ser Earth Environ Sci* 117(1):012037
9. Murdoch J, Hughes W (2000) Construction contracts law and management, 3rd edn. E & FN Spon, New York
10. Ramachandra T, Rotimi JOB (2011) The nature of payment problems in the New Zealand construction industry. *Aust J Constr Econ Build* 11(2):22–33
11. Sambasivan M, Soon YW (2007) Causes and effects of delays in Malaysian construction industry. *Int J Project Manag* 25(5):517–526

12. Latham M (1994) Constructing the team: joint review of procurement and contractual arrangements in the United Kingdom construction industry. HM Stationery Office, London
13. Asuquo CF, Effiong EF (2017) Impact of payment problems on the performance of micro, small, and medium size construction contractors. *J Contemp Res Built Environ* 1(1):35–46
14. Azman MNA, Dzulkalnine N, Hamid ZA, Bing KW (2014) Payment issue in Malaysian construction industry: contractors' perspective. *Jurnal Teknologi* 70(1):57–63
15. Ye KM, Rahman HA (2010) Risk of late payment in the Malaysian construction industry. *World Acad Sci Eng Technol* 65:538–546
16. Ansah SK (2011) Causes and effects of delayed payments by clients on construction projects in Ghana. *J Constr Project Manag Innov* 1(1):27–45
17. Johnston S (1999) Debts and interest in the construction industry: a guide to the late payment of commercial debts (interest) Act 1998. Thomas Telford Limited, London
18. Uff J, Thornhill D (2010) Report of the commission of enquiry into the construction sector of Trinidad and Tobago. Ministry of Finance, Trinidad
19. Abdul-rahman H, Kho M, Wang C (2014) Late payment and nonpayment encountered by contracting firms in a fast-developing economy. *J Profess Issues Eng Pract* 140(2):04013013
20. Frimpong Y, Oluwoye J, Crawford L (2003) Causes of delay and cost overrun in construction of ground water projects in developing countries: Ghana's case studies. *Int J Project Manag* 21(5):321–326
21. Ramachandra T, Rotimi JO (2015) Causes of payment problems in the New Zealand construction industry. *Constr Econ Build* 15(1):43–55
22. Wong JT, Hui EC (2006) Construction project risks: further considerations for constructors' pricing in Hong Kong. *Constr Manag Econ* 24(4):425–438
23. Wang D, Hadavi A, Krizek RJ (2006) Chinese construction firms in reform. *J Constr Manag Econ* 24(5):509–519
24. Yunianto I, Rarasati AD (2020) Development of contract management strategy to control late payment in building projects. *J Int Confer Proc* 3(4):10–23
25. Mohammed KA, Isah AD (2012) Causes of delay in Nigeria construction industry. *Interdiscipl J Contemp Res Bus* 4(2):785
26. Wu J, Kumaraswamy MM, Soo G (2011) Regulative measures addressing payment problems in the construction industry: a calculative understanding of their potential outcomes based on gametric models. *J Constr Eng Manag* 137(8):566–573
27. Judi SS, Rashid RA (2010) Contractor's right of action for late or nonpayment by the employer. *J Surv Constr Property* 1(1):65–95
28. Odeh AM, Battaineh HT (2002) Causes of construction delay: traditional contracts. *Int J Project Manag* 20(1):67–73
29. Ramachandra T, Rotimi JOB (2012) Construction payment delays and losses: perceptions of New Zealand. In: *Proceeding PMI New Zealand chapter 18th annual conference*
30. Hou W, Liu Z, Chen D (2011) Payment problems, cash flow and profitability of construction project: a system dynamics model. *World Acad Sci Eng Technol* 58:693–699
31. Wu J, Kumaraswamy M, Soo G (2008) Payment problems and regulatory responses in the construction industry: mainland China perspective. *J Profess Issues Eng Educ Pract* 134(4):399–407
32. Kadir MRA, Lee WP, Jaafar MS, Sapuan SM, Ali AAA (2005) Factors affecting construction labour productivity for Malaysian residential projects. *Struct Surv* 23(1):42–54
33. United States Code (1988) U.S. federal prompt payment act amendments, chapter 39
34. U.K. Legislation (1996) Housing grants, construction and regeneration act 1996
35. U.K. Legislation (2009) Local democracy, economic development and construction act
36. New Zealand Legislation (2003) Construction contracts regulations 2003. Parliamentary Counsel Office



37. Laws of Malaysia (2012) Construction industry payment and adjudication act 2012. Percetakan Nasional Malaysia Berhad
38. Irish Statute Book (2013) Construction Contracts Act 2013. Office of the Attorney General
39. Government of Canada (2019) Federal prompt payment for construction work act. Justice Laws Website, New Delhi

# Gap Analysis and Areas of Improvement for the CCDC 30: 2018 Integrated Project Delivery Contract



Audrey Provost, Érik A. Poirier, and Daniel Forgues

**Abstract** Over the past decades, research has documented the profound changes needed in the way projects are delivered in the architecture, engineering and construction (AEC) industry to improve the outcomes of construction projects. Recent studies show that an increase in collaboration between parties makes it more likely to lead to successful results in terms of project goals, schedule, cost and quality. Within that context, owners have moved toward collaborative approaches to project delivery by using relational contracting methods such as Integrated Project Delivery (IPD). IPD can be distinguished by the use of integrated agreements, being minimally agreed between owner, professional and contractors. It can be defined as a project delivery approach that integrates people and practices by aligning the incentives and goals of the project team through shared risk, shared reward principles, in addition to early involvement of all parties. In response to growing demand, the Canadian Construction Documents Committee (“CCDC”) created the first-ever Canadian tri-party agreement in 2018, the CCDC 30. Although this contract form has already been used a few times in Canada, it remains fairly new, and some owners indicated that it has some shortcomings when compared to other IPD contracts. Through literature review, existing project documentation and semistructured interviews, this research seeks to identify the key differences between the recent CCDC 30 and other existing, more mature IPD contract forms and how these might inform the improvement to CCDC 30. The study reveals that an improvement in the language used within the contractual agreement, coupled with enhanced definitions would make a significant improvement to the standardized form. It also showed that there is a substantial need for additional guidance to support the means and methods needed to successfully perform an IPD project.

**Keywords** Gap analysis • CCDC 30: 2018 integrated project delivery contract

---

A. Provost (✉) • É. A. Poirier • D. Forgues  
École de Technologie Supérieure, Quebec, Canada  
e-mail: [audrey.provost.1@ens.etsmtl.ca](mailto:audrey.provost.1@ens.etsmtl.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_23](https://doi.org/10.1007/978-3-031-34593-7_23)

347

# 1 Introduction

Contractual arrangements traditionally used in the architecture, engineering and construction (AEC) industry clearly define the limits of responsibility of each party involved, while emphasizing the consequences of failure. This instills self-protective and distrustful behaviors toward other stakeholders in a project [1]. It also encourages each party to look out for their own interests rather than those of the project as a whole [2]. Over the years, this traditional model has resulted in numerous problems for project delivery, such as significant inefficiency, low productivity, adversarial relationships, and contractual disputes between stakeholders typically leading to cost overruns, schedule delays, and poor final product quality [3]. In response to these trends, in the last two decades a new contractual model for construction project delivery has been used on a number of infrastructure projects in Australia and has emerged in the USA as Integrated Project Delivery (IPD) [4]. IPD can be defined as a project delivery model that aligns the goals of the team with those of the project as a whole and provides mechanisms for active collaboration among the various parties involved to increase value and effectively achieve the overall project goals.

IPD is based on a relational type of contract, in which the “commercial relationship assumes equal or greater importance compared to the legal agreement” [5], p. 33. The commercial advantage of this contractual approach comes from the sharing of knowledge that requires greater trust and sense of partnership between the parties involved [5]. In recent years, a few Canadian owners have attempted to use a more collaborative approach to project delivery [6] through the use of existing relational contract forms. In response to this growing demand, the Canadian Construction Documents Committee (CCDC), which publishes typical contract documents that serve as recognized standards throughout the construction industry, in 2018 made available to the market the first-ever Canadian tri-party agreement, the CCDC 30: Integrated Project Delivery Contract (hereafter “CCDC 30”). This contract has already been used a few times in Canada but remains fairly new and is still considered unknown and legally untested [6]. As noted in the study by Bhonde et al. [6], some project owners who have had direct experience with the CCDC 30 have mentioned that the Canadian form has shortcomings compared to other existing international contract agreements.

This paper is a summary of a broader research that focused on identifying gaps and improvements to the recent Canadian CCDC 30 form when compared to existing, more mature, American IPD contract forms. The next sections will introduce the key concepts to integrated agreement, followed by the research methodology, results and conclusion.

## **2 Integrated Agreements**

### **2.1 Key Concepts**

Two types of contractual agreements support IPD project: the Tri-Party and the Multi-Party agreements, in which the role, obligations and responsibilities of the primary parties are specified [7]. In IPD, a single contractual agreement is executed by all parties. Since these are bespoke contracts, they vary to represent the project's core values and goals [8]. As [8] summarized, it's through contract workshops, which are attended by the principal management personnel from each IPD team member, that the basic structure of the IPD agreement, governance structure, cash flow and financials, insurance, extent of liability waivers and limitations, are discussed and agreed. These negotiations require trust and commitment from each party, as compensation is directly linked to the project overall success and individual success depends on the contributions of all team members [7].

Although the IPD contract forms vary depending on the project, the AIA Guide [7], p. 32 points out that they all share similar characteristics, which can be summarized as follows:

- All parties, including client that actively takes part of the project, are tied together by a sole agreement;
- A temporary organization is created through the agreement, where project management mechanisms are customized to the project's specific needs;
- Decisions-making processes are focused on "best for project" effect and need to be consensually made;
- Part of compensation is bound to the project's success, rather than individual success; and
- Roles and responsibilities are distributed among team members based on their ability to perform.

### **2.2 General Form of Contractual Agreements**

The AIA has identified three main forms of contractual agreements supporting IPD, namely Projects Alliance, single purpose entities and relational contracts. These three forms are defined by the AIA [7], p. 33 as follows:

- Project Alliances create a virtual organization out of individual entities. Profit, overhead and bonuses of non-owner parties are tied to the project outcome, while their direct costs are covered by the owner. This compensation scheme introduces the principle of winning or losing together, which provides an incentive for collaboration between the parties
- Single Purpose Entities (SPE) create a new independent legal entity whose sole purpose is to carry out a specific project. Key participants hold shares in this entity

based on their individual skills, experience, services, or financial contribution. The SPE generally compensates its shareholders for their services; however, part of the compensation is also tied to the project success.

- Relational contracts are similar to Project Alliances in terms of creating a virtual organization from individual entities. However, relational contracts differ in its approach to compensation, risk sharing and decision-making mechanisms.

### ***2.3 Comparison of Existing Integrated Agreements***

In January 2011, Hanson Bridgett LLP published a table, called “IPD Contract Comparison Spreadsheet,” in which IPD agreement forms known at that time in the US market have been compared, namely AIA C191, AIA C195, ConsensusDocs 300, Sutter Health’s Integrated Agreement for Lean Project Delivery, and Hanson Bridgett LLP [9]. The comparison focused on the following key concepts (1) Decision Making, (2) Target Cost, (3) Compensation, (4) Changes and contingency, (5) Risk Allocation, (6) Documents and Record Access and (7) Dispute Resolution. Table 1 gives a brief overview of each of these existing forms.

In 2020, the Construction Owners Association of Alberta (“COAA”) added four more integrated contracts to the table developed by Hanson Bridgett LLP, including the newest CCDC 30. Table 2 gives a brief overview of CCDC 30.

For the purpose of this paper, a summary of three key concepts, namely (1) Decision Making, (2) Compensation and (3) Target Cost, that supports the results discussed in Sect. 4 is presented. The information presented in the following sections is drawn from the COAA “Contract Structure Comparison Spreadsheet” [10] available online.

#### **2.3.1 Decision Making**

Decision making is ensured through a clear, tiered structure, which is determined and formalized within the contract. The contract types will include what structures are in place and who is part of the teams put in place. For instance, AIA C191 establishes the Project Management Team (“PMT”) and indicates that it is responsible for day-to-day management of the project. The Hanson Bridgett and the CCDC 30 have similar structures. In these structures, the PMT doesn’t have authority regarding decisions that affect cost or schedule. As such an executive or senior management group is formed: the Project Executive Team (“PET”) in the AIA C191 and the Senior Management Team (“SMT”) in the CCDC30. These executive groups are the final decision-makers in IPD projects. Other contract types have similar structures, such as ConsensusDocs which has a Project Management Group (“PMG”) that includes, at a minimum, a representative from the Owner, the Designer, and the Constructor. This PMG is equivalent to the “Core Group” in the Sutter Health IFOA. Decisions

**Table 1** Brief overview of existing IPD agreement forms (Adapted from the contract structure comparison spreadsheet, COAA [10])

IPD agreements	Year	Type of agreement
AIA C191 family standard form multi-party agreement for integrated project delivery (“AIA C191”)	2009	The owner, architect and contractor all sign the main agreement, making the C191 a <b>Tri-Party Agreement</b> . However, there’s a section to identify additional Parties, if any, which gives flexibility to adjust it to a multi-party agreement
AIA C195 family standard form single purpose entity agreement for integrated project delivery (“AIA C195”)	2008	The C195 allows for the creation of a <b>Limited Liability Company (LLC)</b> , in which the Owner, Architect and Construction manager are members. To design and build the project, the Company must contract under separate agreements with the Owner, using AIA C196 and with other non-owner parties, using AIA C197
ConsensusDocs 300 standard multi-party integrated project delivery agreement (“ConsensusDocs”)	2007, rev. 2016	The Owner, Design Professional and Constructor all sign the same agreement, making ConsensusDocs 300 a <b>Tri-Party Agreement</b> . However, all parties, classified into two categories: Risk Pool and Joining Members, sign a joining agreement (ConsensusDocs 396)
Sutter health’s integrated form of agreement for integrated lean project delivery (“IFOA”)	2009	The Owner, Architect and Construction Manager/General Contractor all sign the contract, making it a <b>Tri-Party agreement</b>
Hanson bridgett LLP integrated project delivery agreement (“Hanson Bridgett”)	2009	Owner, Architect and Contractor all sign the Hanson Bridget agreement, making it a <b>Tri-Party Agreement</b>

**Table 2** Brief overview of CCDC 30: 2018 IPD contract form

CCDC 30 integrated project delivery (“CCDC 30”)	2018	Owner, Professional and Contractor all sign the CCDC 30, making it a <b>Tri-Party Agreement</b> . However, additional parties can be identified in the Annex “Other Parties”, and provides flexibility to make it a <b>Multi-Party Agreement</b> if needed
---	------	--

themselves are made consensually and typically have to be made unanimously. In some cases, a majority vote is required, rather than unanimous approval.

### 2.3.2 Compensation

Compensation involves how profits are to be distributed among the project team members. All contract types include mechanisms that define how compensation is to be determined and distributed.

All IPD contract forms stipulate that non-owner IPD members are compensated for their reimbursable costs that are established at contract negotiations with the owner. They additionally all have an incentive compensation that is tied to achieving project goals, be it based on agreed quality or performance criteria. One performance criterion is, however, part of all IPD contract forms, where if the actual cost at project completion is less than the target cost, savings are shared between the owner and non-owner parties. Other incentive criteria are defined by the IPD teams during the contract negotiations and added to the contract form once agreed by all.

Sutter's IFOA, Hanson Bridgett and CCDC 30 all have an at-risk profit pool, where a portion or all the profit of the primary IPD members is retained. It is through this pool that the share pain, share gain reward principle is enforced, as it can be increased or decreased depending on losses or savings throughout the project delivery.

### **2.3.3 Target Cost**

The target cost is developed by the IPD team based on the owner's requirements and once agreed by the owner, is upended to the contractual agreement. It represents the total asset cost that the team is confident enough to meet.

As mentioned in Sect. 2.3.2, the target cost is used as a factor in determining the amount of incentive compensation paid to non-owner IPD parties. All contract forms compared, except ConsensusDocs, set the target cost quite early in the project's development, as it is added as an overall goal to the contract documentation. ConsensusDocs takes more of a traditional approach and develop it based on the construction documents.

## **3 Research Methodology**

To gain a better understanding of IPD as a project delivery model and to understand the basis of certain contractual clauses, a literature review was performed. This research is exploratory as the CCDC 30 is fairly new and not well documented in literature.

Then, to reveal the main differences between the CCDC 30 and other integrated agreements used in Canada, and to gain a better understanding of the impact on project delivery of some variations in contractual clauses, a qualitative approach using semistructured interviews was conducted. A chain-referral-sampling strategy was purposely chosen to identify interviewees with extensive IPD experience in Canada or CCDC 30 knowledge among key stakeholders, the first group of which has been identified through a Canadian collaborator of the GRIDD industrial chair. Additional interviewees were then identified as participants with the capacity to inform the study throughout discussions. A total of ten (10) stakeholders were interviewed, all for approximately one hour, except for the general contractor's representative who was

**Table 3** Summary of respondents' profiles

Respondent category	Count	Specialty
Owner	4	Business development, project management (2), design
Professional	3	Electrical/mechanical engineer, structural engineer and architect
General contractor	1*	Construction
Trade partners	1	Electricity
Legal counsel	1	Lawyer

\*The same representative was met twice, for one hour each

interviewed in two replicates, each lasting one hour. Table 3 summarizes respondents' profiles.

The data collected from these interviews was compiled and coded using NVivo 12 software and an inductive approach, which is defined by Thomas [11] as “a systemic procedure for analyzing qualitative data in which the analysis is likely to be guided by specific evaluation objectives”. The coding was guided by the research objective of identifying areas of improvement for the CCDC 30. It was then possible to identify broad themes to be explored, such as limitations of CCDC 30 and lessons learned from participants, in relation with the key concepts that are grounded in the contractual agreement. As a reminder, those key concepts categories are: (1) Decision Making, (2) Target Cost, (3) Compensation, (4) Changes and Contingency, (5) Risk Allocation, (6) Documents and Record Access and (7) Dispute Resolution. In addition to the interviews, the researchers have been able to get copies of existing documentation from an IPD project in Canada using the CCDC 30. Both sources have been triangulated to ensure greater rigor in their collection and analysis.

## 4 Gaps and Improvements to the CCDC 30: 2018

The analysis revealed five main limitations to the CCDC 30 that are related to the three key concepts reviewed in Sect. 2.3, namely Decision Making, Compensation and Target Cost. The gaps identified are detailed in the following sections and can be summarized in Table 4 as follows.

### 4.1 CCDC 30 Creation Process

Before getting into the specifics of identifying the CCDC 30 gaps, it is important to understand the process by which the Canadian template was created and some of the assumptions made in its development. This process will provide a better understanding of the basis for some of the limitations in the following sections.



**Table 4** Summary of identified gaps

Key concepts	Identified gaps
<i>General</i>	
	Nature of the contract form and lack of guidance
<i>Decision making</i>	
	Composition of project management teams and decision-making mechanisms
<i>Compensation</i>	
	Payment mechanisms during the validation phase
	Definition of overhead
	Ambiguous terminologies: project contingency, risk pool and design/procurement phase
<i>Target cost</i>	
	Missing terminology: total project cost

As summarized on the CCDC website [12], their contract forms and guides are developed by its national joint committee that includes representatives from public and private sectors, as well as representatives from four national organizations: Association of Consulting Engineering Companies-Canada (ACEC), Canadian Construction Association (CCA), Construction Specifications Canada (CSC) and Royal Architecture Institute of Canada (Architecture Canada).

One of the respondents involved in the development of CCDC 30 confirmed that the exercise began in approximately 2016 and was carried out by a group of around 12–14 people, all from the construction industry, with decades of experience in traditional project delivery. That shift in perspective is one of the reasons it took them almost 2 years to develop the CCDC 30. It was also confirmed during that discussion that the creation committee had no real experience with IPD, as it was novel at the time. The CCDC contract forms are well recognized in the Canadian industry, and it was important for the creation committee to maintain that similar format, which has been described by the same respondent as relatively lean in spare verbiage.

## 4.2 Gaps

During the interviews, the CCDC 30 was generally compared by the participants to the Hanson Bridgett American form. More precisely, an Owner confirmed that based on what he'd seen, "it's starting to become more common in Canada to see the CCDC 30 versus the Hanson Bridgett". The use of the latter appears to be more widespread in Canada, which may explain why it was the main reference that participants used for comparison. A version of the bespoke form is also available free of

charge online, which may explain some of its popularity.<sup>1</sup> For these reasons, in the next few sections, the main comparisons of the limitations raised in CCDC 30 are made with the Hanson Bridgett form.

#### **4.2.1 Nature of the Contract Form and Lack of Guidance**

While it is aligned with the CCDC's intent to get a concise IPD contract form, most of the participants suggested that the Canadian template is not sufficiently detailed in terms of tools and processes to be used for a successful IPD project. IPD is based on key principles such as partnership, transparency, collaboration and trust, which are the opposite of traditional principles that focus on risk transfer and accountability for results. In maintaining their usual approach to the creation of the CCDC 30, the CCDC failed to adequately capture and represent the key principles of IPD, that must be implemented through the contractual agreement, which may explain the respondents' perceptions. A public owner summarized the issue in stating that the CCDC 30 was more of a contract that contains the legal clauses, while the Hanson Bridgett was more of an agreement, as it acts as a guide to the process.

The legal counsel representative added "CCDC 30 only gets you so far, it doesn't tell you how to build the project, how to collaborate, how to do pull planning or target value design, it mentions it, but apart from setting out the parties' basic legal relationship, waivers of claims, and other ancillary clauses, that's as far as the contract goes." The respondent that took part in the CCDC 30 creation committee mentioned that they were probably not sufficiently attuned to the fact that the contract is only the hardware, while the implementation of the team, the introduction to lean principles and others, is the whole aspect to doing an IPD project and that both do go hand in hand.

Compared to Hanson Bridgett, there appears to be a significant lack of guidance for conducting an IPD project with the CCDC 30, as it seems to only provide the legal structure.

#### **4.2.2 Composition of Project Management Teams and Decision-Making Mechanisms**

The CCDC 30 provides for two levels of management, the Project Management Team ("PMT") and the Senior Management Teams ("SMT"). The CCDC 30 defines that the PMT "consists of one representative appointed by each party to the Contract to provide management-level guidance for collaborative planning, design and construction of the Project to achieve the Project Objectives" (Definitions, 2018). It also

---

<sup>1</sup> Hanson Bridgett, Integrated Project Delivery Agreement (2009). Available at [https://www.ipda.ca/site/assets/files/1076/ipd\\_standard\\_agreement\\_profit\\_deferred-1.pdf](https://www.ipda.ca/site/assets/files/1076/ipd_standard_agreement_profit_deferred-1.pdf).

specifies that the “PMT shall make decisions by unanimous agreement” (§GC3.2.3). The SMT in turn “consists of one senior executive representative appointed by each party to the Contract” (Definitions).

The general contractor respondent mentioned that having a large PMT can make unanimous decision making very difficult, especially knowing that this leadership group is responsible for all day-to-day management aspect. An owner also added to the issue saying that he was not in favor of a big PMT, because he thought he could get less done. Another owner summarized this first issue by stating, “10 project managers, analysis, paralysis, there are just too many perspectives to try and manage.” The general contractor respondent mentioned that when using CCDC 30 on its IPD projects, the PMT composition clause was often modified in supplementary general condition (“SGC”) to get the wording more aligned with the Hanson Bridgett form that stipulates that the owner, architect and contractor must be on the PMT as a minimum, and then, any other members as required. All interviewees agreed that all signing parties must have an SMT member, but not a PMT member.

Furthermore, the PMT composition has a direct impact on the hierarchical decision-making process. If the PMT doesn’t reach unanimous agreement, the dispute is referred to the SMT (§GC8.1.4), which in turn shall also make its decision by unanimous vote (§GC3.1.4). If the SMT cannot resolve the dispute, it is then submitted to mediation in accordance with the Rules for Mediation and Arbitration of Construction Contract Disputes in CCDC 40 (§GC8.1.6). Finally, if a dispute is not resolved by mediation, the parties can refer it to arbitration (§GC8.1.7).

The general contractor’s respondent confirmed modifying the SMT decision-making process when using the CCDC 30 form, because the way CCDC 30 is set up, by having unanimous decision making at multiple project levels, “it can really drive the team toward dispute resolution process earlier than they might like”. He also added that the team “is setting themselves up for having to escalate issues, which is what they usually want to avoid in the first place.” Again, it appears that the Hanson Bridgett form, which uses unanimous decision making at the PMT level, but changes it to a vote at the SMT level, for which the allocation can be personalized by project teams directly in the form, offer flexibility in the decision-making mechanisms that seems more appreciated by stakeholders. The legal counsel respondent has mentioned seeing IPD teams moving into something more sophisticated or nuanced, such as 80% ratio, or the owner gets more votes than everyone else and such.

#### **4.2.3 Payment Mechanisms During the Validation Phase**

There appears to be some uncertainty as to when the contractual agreement governing IPD projects should be signed, which in turn causes uncertainty regarding the payment mechanisms for the non-Owner members during the validation phase. Some owners mentioned that they prefer having the contract form signed by all parties before the end of the validation phase, while others prefer using letters of intent to cover the interval between the IPD team selection and the end of the validation phase.

The first project phase in IPD project is the Validation, where the owner's objectives are confirmed by the IPD team as being achievable within a set budget and schedule. The results of the validation phase are concealed into a validation report and once accepted by the Owner, incorporated to the Contract Documents, from which the concern seems to stem. A professional respondent stated that by experience, "the contracts are not set up for execution until the end of the validation, because of the documentation that goes in from the validation period, which creates a challenge for doing a lot of work without a definition of a full contract." Another professional mentioned that letters of intent and memorandums of understanding is what the standard has been for some time in Canada, to set the contractual mechanisms during that validation phase.

To the contrary, another professional respondent mentioned that contracts are intended to be signed right at the start of validation and that they are meant to be upended once all project objectives are agreed between the parties. He added that there is a lot of misunderstanding about how the contract actually works from that perspective, and just like any other contract out there, it would be best practice to sign it right at the beginning.

CCDC 30 General Conditions Part 6 mentioned the following: "subject to such innovative financial arrangements as may be set out in the Validation Report, applications for payment [...] including Reimbursable Costs, without profit, incurred during the Validation Phase, may be made monthly by members of the Design/Construction Team." Thus, unless having a different agreement, it appears that non-Owner parties to the agreement are paid at cost during validation.

A public owner representative mentioned that for their projects, Hanson Bridgett or CCDC 30 would not be signed until the validation report is completed and so they created an actual agreement for that phase. He added that "it's pretty simple, nothing revolutionary, but we created that to kind of cover that period off. We'll have those 10 partners come in, sign the validation agreement, we'll go through, if we validate, then we'll go and sign the Hanson Bridgett, if we fail to validate, there are a lot of reasons for that, then we just collapse the agreement and go our separate ways, payout the profit if we decided that it's the city's fault that it didn't proceed or not pay profit if the team failed."

#### **4.2.4 Definition of Overhead**

Another issue that came up a few times among respondents with IPD experience was the definition of overhead. It appears that CCDC 30 does not sufficiently define the types of costs that may be included in the calculation of overhead costs submitted by members in the contractual agreement. One owner stated that the definition of costs, profits and overhead are extremely vague in the CCDC 30 and that any party may come up with very different build-up that makes it difficult to reconcile with the other parties.

A professional also noted that the definition of what actually goes into some categories of overhead aren't as robust as they should be and should be clarified within the CCDC 30. He added that it would be good to have a consistent approach as to what is defined within overhead, because it's already hard to align 10 to 15 companies with the same definitions.

The general contractor mentioned that Hanson Bridgett is pretty explicit on overhead and list what could be considered and what should not, while the CCDC 30 is silent on that aspect and leaves it open to different calculations, depending on the project and what the owners ask.

#### **4.2.5 Ambiguous Terminologies**

The following topic includes a few places where the terms used in CCDC 30 may be confusing and therefore less intuitive when compared to other IPD contract forms or even just when compared to their own definition. For clarification purposes, these issues were revealed primarily by the general contractor's respondent in relation to the English version of CCDC 30.

##### **Project Contingency**

CCDC 30 defines the Project Contingency, as follows: "included in the Base Target Cost, is established during the Validation Phase by the PMT based on an analysis of potential risks and innovations and covers Reimbursable Costs that may arise as a result of that analysis, but does not cover Reimbursable Costs resulting from conditions or circumstances that vary substantially from what was anticipated during the Validation Phase."

Essentially, this contingency is determined following the project risk analysis exercise, for which an amount is established and agreed upon by all parties during the validation phase. The data that stem from the exercise is then conciliated in a document commonly referred to as the "Risk Register." If one of these risks materializes during the project, the predefined amount can then be used to remediate the situation. By the end of the Warranty phase, unspent contingency amounts are distributed between the parties using the established Risk Pool distribution (§GC4.4.4).

During the first interview with the general contractor respondent, he first noted a discomfort with the use of the term contingency, which "seems like an arbitrary number slapped up 'just in case'." During the second interview, he then clarified that the uses of the word contingency can be a bit of a bother because in traditional industry language, contingency could be seen as like "yeah, let's throw a 5% contingency on whatever, just in case, whereas in IPD projects, it's really thought out by all team members, you're putting dollar values to the probability of an event happening and what's the cost of it happening? And then you make a kind of a more scientific and objective approach, [that's why] we tend to use the word "Risk Register" instead of Project Contingency."

It is recognized that in traditional project delivery models, contingency is more often represented by a percentage mark-up related to the margin of error of an estimate as a function of the project's progress. Typically, project teams consider that the closer the project is to the construction phase, the more fixed its scope is and therefore the more accurate the cost estimate is. A smaller perceived margin of error thus leads to lower contingency. Thus, for IPD team there is an important effort to put in the risk analysis exercise to build-up the contingency amount and new parties more accustomed to traditional delivery models may be confused by the term.

It should be noted that the CCDC 30 French translation of Project Contingency can also lead to confusion as it has been translated to "Réserve pour imprévus du projet" which could be interpreted as being related to unforeseen conditions and events, discussed in Part 7 of the form.

### Risk Pool

CCDC 30 defines the Risk Pool as "an amount that is established by the PMT during the Validation Phase, revised upon the addition of Added Parties, if any, and distributed among Design/Construction Team in accordance with the Risk Pool Distribution (set forth in the Contract's Schedule A), as amended in accordance with this Contract" (§A-4). The Risk Pool is also defined as "the sum of the profits of each member of the Design/Construction Team."

The Risk Pool is an incentive compensation layer where, when savings are realized through innovation or sustainable choices, the IPD team shares them at a pre-agreed ratio. This amount represents potential profits for the parties but could also be reduced if problems arise during the implementation of the project. The General Contractor representative mentioned that from a wording standpoint, the Risk Pool could be confused with Project Contingency, evermore if the latter is referred to as "Risk Register." The Hanson Bridgett calls it the Incentive Compensation Layer (ICL), which is not intuitive either. The General Contractor respondent added: "Why not just call it "At-Risk Pool", or "At-Risk Profit Pool", or "Profit Pool"? That's what it is."

It should be noted that the CCDC 30 French translation of Risk Pool can also lead to confusion as it has been translated to "Réserve pour risques" which could be interpreted as being related to the risk analysis exercise instead of the at-risk profit pool.

### Total Project Price

Several cost terms are defined within CCDC 30; there are Base Target Cost, Estimated Final Target Cost, Final Target Cost, Estimated Final Cost, Final Actual Cost and Reimbursable Costs. However, the general contractor respondent pointed out that

none of these terms contain profits and thus, none of them represent the Total Project Price within CCDC 30. He added: “the Base Target Cost covers materials, labor, all direct costs, and then you’ve got your profit on top of that, and CCDC 30 doesn’t have a word for that, it’s kind of weird.”

### Design/Procurement Phase

The second project phase in CCDC 30 is the Design and Procurement Phase, which begins immediately following the owner’s acceptance of the validation report.

It is important to note that incentive compensation allocation may vary within specific phases of completion underway and depending on the agreement reached between the parties that customize the article A-4 within the CCDC 30 form (§A-4.2.2.1 and §A-4.2.2.2). These sections stipulate that “if upon completion of Design/Procurement Phase, the PMT determines that the Final Target Cost, [...] is less than the Base Target Cost, then the Risk Pool shall be increased in accordance.” The proportion is then to be defined, but the difference is established as  $\pm 15\%$ , so if the difference is less than or equal to 15%, the Risk Pool is increased by a certain percentage to be defined.

The general contractor respondent explained that it is more common to have different value model throughout the project phases. For example, it is usual to define a higher percentage of savings going back to the owner during the design phase, as the risk is less for the IPD team, whereas during construction, it is more usual to have a 50–50% split. The Design/Procurement title with CCDC 30 can lead to uncertainties when allocating savings between parties, because putting procurement in the title admits that the team is actually doing a lot of it during the design phase, which is not accurate. The IPD team may do early procurement activity, but most of it is going to happen at the beginning of the construction phase.

The general contractor respondent added that it would be less of an issue if the split was always 50–50, but it’s not necessarily the case. Though he nuanced this statement, mentioning that it is not an issue on most jobs, because they never got 15% below the owner’s budget in the design phase, but it could “get a bit fuzzy for project managers and owners to figure out and, most importantly, you don’t want teams’ holding back on innovation because “wait a minute, we only get 20% right now, so I better hold off on that idea”.”

## 4.3 Identified Improvements

Improvements were identified based on the interview respondents’ recommendations and substantiated with the existing contractual project documentation shared by a collaborator of the GRIDD. Areas of improvements are detailed in the following sections and can be summarized in Table 5 as follows.

**Table 5** Summary of identified Improvements

Gaps	Identified improvements
<i>General</i>	
Nature of the contract form and lack of guidance	Develop a guide to integrated project delivery with the use of CCDC 30
<i>Decision making</i>	
Composition of project management teams and decision-making mechanisms	Modify the PMT composition definition to have at a minimum a representative of the Owner, the Professional and the contractor, while allowing flexibility to add members at the discretion of the SMT Modify the general conditions related to SMT unanimous decision making, to favor majority vote while adding a quorum level to ensure a good decision making
<i>Compensation</i>	
Payment mechanisms during the validation phase	Clarify the moment when the contract is supposed to be signed and/or develop a formal agreement to cover the validation phase, which could be attached to the Contract form afterward. The clarification could be added to the proposed guide in the first improvement
Definition of overhead	Clarify eligible inclusions in overhead costs, by providing a list of eligible items to be submitted for professionals and contractors. This item could also be included in the proposed guide in the first bullet point
Ambiguous terminologies	Change the term “Project Contingency” to “Risk Register Pool” through supplementary general conditions Change the term “Risk Pool” to “At-Risk Profit Pool” or “Profit Pool” through supplementary general conditions Change the title of the design/procurement phase to “design and construction documents phase” to avoid ambiguity and to better represent what tasks take place within this phase
<i>Target cost</i>	
Missing terminology	Add a term to represent the total project price, i.e., the actual amount the client will have spent to obtain the desired good

### 4.3.1 Develop a Guide to Integrated Project Delivery with the Use of CCDC 30

In recent years, various guides have been developed in connection with IPD to facilitate the understanding of this new project delivery model. Specifically, the American Institute of Architects (“AIA”) published a comprehensive guide, entitled “Integrated Project Delivery: A Guide” [7]. More recently, the Action Guide for Leaders, sponsored by the Integrated Project Delivery Alliance (“IPDA”) among others, has been developed and contains information derived from conversations during the IPD Advisory Council workshop between 22 Canadian and U.S. industry professionals with IPD experience, both in public and private sectors (2018). These existing guides,



while thorough and insightful, were developed prior to the CCDC 30 publication, while the Hanson Bridgett form adheres closely to the AIA's principles elaborated in their IPD guide [13].

CCDC has chosen an alternative approach to IPD, and there is a significant need for additional guidance to support a project under CCDC 30. This guide should include explanations of the tools and processes recommended to effectively carry out tasks related to coordination, planning, cost estimation, risk analysis and target value design among others.

#### **4.3.2 Modify the Composition of Project Management Teams and SMT Decision-Making Mechanisms**

Solutions were raised by some respondents, including the general contractor's and one of the owner's representatives who respectively took a similar approach to getting good representation of all trade partners in their PMTs. The general contractor's respondent explained his strategy as including the minimum of one representative from the owner, architect, and general contractor, and then adding one trade partner and one consultant partner to get a nice mix of voices, diversity of experiences and perspectives. The suggestions mentioned during the interviews were well aligned with the Hanson Bridgett form that stipulates that the PMT will include a representative from the Owner's, the Architect's and the contractor's, with the possibility to add other partners. Referring to the existing contractual documentation shared by the GRIDD collaborator, the PMT definition was amended in supplementary general conditions as follows: "The PMT is comprised of one representative from each Owner, Prime consultant and Contractor. Additional members may be added to a maximum of 2 (for a total of 5 PMT Members) as appointed by the SMT."

As mentioned earlier in this paper, it is usual for the PMT to take unanimous decisions, which is what is currently included in the CCDC 30 and the Hanson Bridgett forms. However, it is less usual to have the same mechanism at the SMT level. CCDC stipulates that "The SMT shall make decisions by unanimous vote (§GC3.1.4)". Interview respondents repeatedly mentioned changing this clause through SGC to be more aligned with Hanson Bridgett that stipulates "if the PMT is unable to reach agreement, the PMT will refer the issue to the Senior Management Representative level under Section 4.9, who will first attempt to reach a consensus and only if a consensus is not reached, will decide the issue by majority vote" (§4.7). Referring to the existing contractual documentation shared, the SMT decision making was modified as follows: "Any matters requiring SMT decisions or action will be decided by plurality of votes of the SMT, with Owner having 4 votes regardless of the number of its representatives present at the meeting and the other SMT member firms having only 1 vote each. Quorum shall be achieved by attendance by the Owner and 75% of the other member representatives." Ideally, CCDC 30 would change the unanimous decision making at the SMT level to provide more flexibility for teams to decide

how they wish to govern themselves. Be it majority or plurality rules, in either case it would be interesting to offer the ability to customize the vote split and a percentage representing a good decision, whether by reaching a quorum or an overall percentage level.

Consequently, modifying the PMT composition definition and the SMT decision-making mechanisms through supplementary general conditions to be more aligned with the above suggestions would provide a significant improvement in the management of a CCDC 30 IPD project.

#### **4.3.3 Clarify Timing of Contract Signature and/or Develop an Alternate Agreement for the Validation Phase**

The proposed guide for IPD with CCDC 30 mentioned in Sect. 4.3.1 could have a clear and unequivocal explanation that the contract is meant to be signed prior to the validation phase and that any information arising from the validation report will be upended to the contract once approved by all parties.

To be noted that one of the professionals interviewed mentioned that he had the intent of publishing a validation agreement template through the IPDA in the future. This would also be helpful to try and unify the approach for all IPD projects in Canada.

#### **4.3.4 Clarify the Definition of Overhead**

As Ashcraft et al. [8], p. 150 discussed in the Action Guide for Leaders, overhead costs can be calculated in many ways, but a deliberate and transparent approach must be established. This exercise is particularly important because the IPD team will need to establish billable rates for their resources that can be used to define the project costs.

To this end, it could be worthwhile to generate a list of eligible overhead costs for professionals and general contractors. This proposed list could be added to the guide for IPD with CCDC 30 mentioned in Sect. 4.3.1 or could be included in the next CCDC 30 version, more specifically in Schedule B.

While most respondents mentioned that Hanson Bridgett was much more explicit on overhead costs, the free-of-charge online version does not present specific guidance on this subject, but it should be noted that no schedules are included with this file and that IPD forms are bespoke contracts that are amended for each project. It is therefore possible that owners using Hanson Bridgett have clarified this aspect directly in the forms that respondents used. The Action Guide for Leaders (2018) presents two appendices (27 and 28) as examples of how indirect and overhead costs are calculated for contractors or professionals, which could be useful in defining a list of allowable items.

### 4.3.5 Change Unintuitive Terminologies

Some terminologies used in the CCDC 30, both in the English and French version, are unintuitive and could lead to confusion.

The first term is “Project Contingency,” which represents the amount defined to cover the consequences of a specific risk materializing, which were previously identified through a risk analysis exercise. In both CCDC 30 English and French versions, the term should be modified to align with the risk analysis exercise and the risk register in which the costs, probability and consequence of each risk is conciliated. A proposed change could be “Risk Register Pool” in English and the “Réserve pour le registre des risques” in French.

However, the modification of the first term has a direct impact on the term defining the teams’ profits put at risk. In the English version, that term is “Risk Pool”, while in French it is “Réserve pour risques.” As mentioned by the general contractor respondent in Sect. 4.2.5.2, it would have been more appropriate to use the term “At-Risk Profit Pool” or “Profit Pool” and “Réserve des profits à risque” for the French version.

Then, CCDC 30 seems to be missing a term to define all the cost incurred to complete the project, including the IPD teams’ profits. The Total Project Cost could then represent the actual expenditures the owner had to pay in order to obtain its desired asset.

The final term that respondents raised as potentially confusing is the title of the Design/Procurement Phase. As explained in Sect. 4.2.5.4, the use of the word “procurement” in the title of this phase suggests that much of the procurement occurs within this phase. However, the IPD contract forms allow for modulation of the savings’ sharing between the owner and non-Owner parties. A higher return to the owner is generally expected during the design phase where risks are lower for the IPD team, while a 50–50 split is recommended during construction where risks increase due to the different variables added to the project, including contracting with other subcontractors. Thus, the word procurement confuses the understanding and definition of this sharing of savings and could lead some participants to delay an innovation in order to get a greater return. Instead, the Hanson Bridgett form uses the following titles: Detailed Design Phase (§7.4) and the Implementation Documents Phase (§7.5). It may be more appropriate to adopt a title such as “Design and Construction Documents Phase”, which really represents the activities that will take place, procurement being only a small portion of it.

## 5 Conclusion

From this research, five shortcomings themes emerged in relation to the IPD Canadian contract form CCDC 30, among which the lack of guidance in terms of tools and processes to be used for a successful IPD project, was most significant. Some changes to definitions and terminologies would also make a considerable impact on the comprehensiveness of the Canadian form. Areas of improvement were identified

for each limitation. Of these, three could be resolved through the production of a guide for IPD with CCDC 30, similar to the Hanson Bridgett contract form that follows the AIA IPD Guide (2007). CCDC documents, which also include guides for some of their contract forms, are well recognized in the Canadian construction industry. While one of our respondents mentioned that the CCDC does not have such a user's guide on its agenda, it would certainly help advance the adoption of this new project delivery model in Canada.

Although the research identified high-level gaps currently faced by IPD participants with the use of CCDC 30, it has some limitations, particularly due to the limited number of data sources and its exploratory approach, which deviates from a rigorous scientific framework. The semistructured interviews focused primarily on identifying perceived limitations on the CCDC 30, whereas an equivalent investigation including other contractual forms would have allowed for a more comprehensive comparative approach, which in turn would have allowed for the identification of finer limitations.

**Acknowledgements** This research would not have been possible without the generosity of all industry members who willingly participated in these interviews and shared insightful knowledge on IPD, the authors would like to extend their warmest thanks to them.

## References

1. O'Connor PJ (2009) Integrated project delivery: collaboration through new contracting forms. Faegre and Benson, Minneapolis. <https://www.faegredrinker.com/webfiles/AGC-IPD%20Paper.pdf>
2. Ghassemi R, Becerik-Gerber B (2011) Transitioning to integrated project delivery: potential barriers and lessons learned. *Lean Constr J* 12:32–52
3. Marco AD, Karzouna A (2018) Assessing the benefits of the integrated project delivery method: a survey of expert opinions. *Proc Comput Sci* 138:823–828. <https://doi.org/10.1016/j.procs.2018.10.107>
4. Kent DC, Becerik-Gerber B (2010) Understanding construction industry experience and attitudes toward integrated project delivery. *J Constr Eng Manag* 136(8):815–825. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000188](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000188)
5. Colledge B (2005) Relational contracting: creating value beyond the project 2, 16
6. Bhonde D, Zadeh P, Staub-French S (2020) Owner's perceived barriers to adoption of IPD in Canada. University of British Columbia, Columbia. <https://www.ipda.ca/research-performance/industry-research/owners-perceived-barriers-to-adoption-of-ipd-in-canada/>
7. AIA National, and AIA California Council (2007) Integrated project delivery: a guide. AIA National and AIA California Council. [https://info.aia.org/siteobjects/files/ipd\\_guide\\_2007.pdf](https://info.aia.org/siteobjects/files/ipd_guide_2007.pdf)
8. Ashcraft HW, Allison M, Cheng R, Pease J, Osburn L (2018) Integrated project delivery: an action guide for leaders. Center for Innovation in the Design and Construction Industry. <https://www.ipda.ca/knowledge-competency/tools/integrated-project-delivery-an-action-guide-for-leaders/>
9. Dal Gallo L, Ashcraft Jr H (2011) Comparison of integrated project delivery agreements (summary). Hanson Bridgett LLP. <https://www.hansonbridgett.com/Publications/pdf/2011-1-25-comparison-of-integrated-project-delivery-agreements-summary>
10. COAA (2020) Contract structure comparison spreadsheet/construction owners association of Alberta. Construction Owners Association of Alberta (COAA). <https://www.coaa.ab.ca/library/contract-structure-comparison-spreadsheet/>

11. Thomas DR (2006) A general inductive approach for analyzing qualitative evaluation data. *Am J Evaluat* 27(2):237–246. <https://doi.org/10.1177/1098214005283748>
12. CCDC (2021) À propos du CCDC. <https://www.ccdc.org/fr/a-propos/>
13. Dal Gallo L, O’Leary ST, Louridas LJ (2011) Comparison of integrated project delivery agreements. Hanson Bridgett LLP

# Considerations for Site Layout Planning Decision-Making



A. Marcano Pina and F. Sadeghpour

**Abstract** Site layout planning (SLP) refers to the efficient allocation of temporary facilities on construction sites. SLP especially becomes important when space is scarce in heavily congested sites. Previous research has mainly focused on the optimization of the locations of resources in site layouts. However, despite the achievements made, there is no evidence of the application of these tools in practice. This issue arises from little or no interest in the SLP during the early phases of the projects; a high level of sophistication across optimization tools; difficulty in standardizing optimization models such that they can be used in different projects, and the extensive times required to instruct practitioners in the use of such tools. These causes suggest a potential misalignment between currently available tools and actual requirements from industry practitioners. The long-term goal of this study is to define the requirements and specifications for an SLP tool that is deemed practical to industry practitioners. As an initial step, the objective of this paper is to investigate the decision-making process applied by practitioners during the SLP task. Data will be collected using structured interviews and analyzed through qualitative coding techniques. The results of this study will set the foundations of the SLP decision-making process as done in practice. Ultimately, this will lead to developing the attributes required in an SLP tool that can address the needs of practitioners in the construction industry and potential research opportunities for further developing the SLP tool.

**Keywords** Site layout planning · Decision-making

## 1 Introduction

Site Layout Planning (SLP) refers to the efficient allocation of resources in the site space. Previous academic efforts have focused on providing tools to assist in this task, mostly from optimization perspectives [10, 31, 39]. Despite the academic

---

A. Marcano Pina (✉) · F. Sadeghpour  
University of Calgary, Calgary, Canada  
e-mail: [agnei.marcanopina@ucalgary.ca](mailto:agnei.marcanopina@ucalgary.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_24](https://doi.org/10.1007/978-3-031-34593-7_24)

367

contributions made in the field, there is no evidence of the application of these tools in real-life settings. Extensive initial interviews with industry experts exhibited that practitioners mainly use experience and judgment to allocate resources to the site space and do not consider existing models. In practice, the site planning task involves multiple variables with different levels of difficulty. This suggests that optimization-based tools have not addressed all aspects and constraints of the problem. Therefore, there is still a need to understand the practical approach taken by practitioners and how a tool that would provide benefit to this process.

The development of functional tools for decision-making is based on a proper investigation of the requirements, practices, and user perspectives [5]. To conduct such an investigation, the process undertaken by users must be understood and replicated. The long-term goal of this study is to determine the requirements for a site planning tool that addresses the needs of industry practitioners. This paper presents the first step in this investigation by focusing on the decision-making process of practitioners while conducting the planning task. Ultimately, comprehending the thinking process behind the way they approach and solve the problem will contribute to developing functional tools for site planning decision-making.

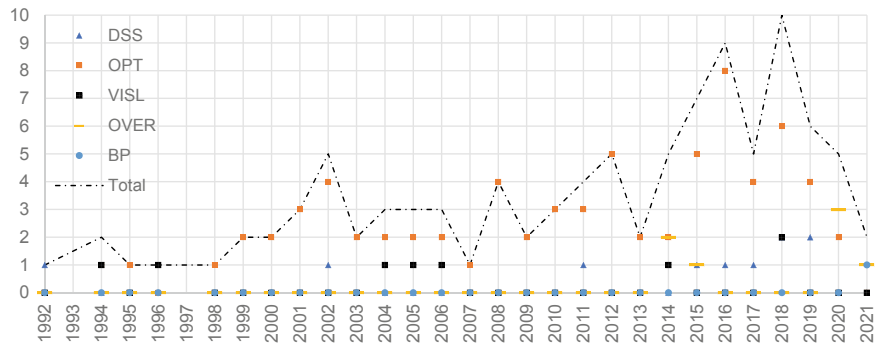
Structured interviews were conducted to identify the thought process of practitioners during the site planning task. The interview questions were designed iteratively, based on the constraints that impact the problem in the literature and extensive initial interviews with industry experts. The interview participants comprised superintendents and senior project management professionals who have self-identified to have the main responsibility for the site planning task. All the participants currently perform planning tasks for the highest-ranked contractor companies in Canada in terms of revenue [16].

## 2 Literature Review

The SLP process has been approached mostly from optimization (OPT) perspectives [23, 39, 41]. Other perspectives include decision support systems (DSS) to assist the planning task [24, 36] and layout visualization tools (VISL) [7, 13, 33]. In addition, there have been review papers (OVER) on different aspects of the SLP literature [12, 28] and best practices (BP) [37]. Figure 1 presents a time series of the journal publications categorized by area of knowledge.

### Optimization Models

Optimization models use mathematical, heuristics, and meta-heuristics techniques to find optimum locations for site resources. This group of studies aimed to reduce the values of objective functions, decrease computation time, and improve the modeling of multiple constraints. The majority of studies optimized the layout by minimizing weighted distances between site resources [9, 15, 39]. To minimize the weighted



**Fig. 1** Time series of SLP journal publications ( $n = 99$ , 1990–2021)

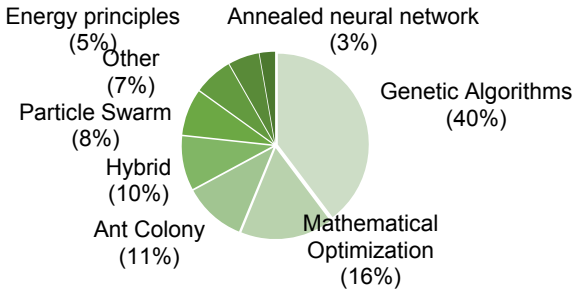
distance, studies applied different optimization techniques, such as genetic algorithms [6, 21, 27], particle swarm optimization [20, 40], and physics energy principles [2, 15]. Figure 2 shows the different optimization techniques that have been used in the SLP literature.

While most studies have used genetic algorithms to optimize the locations of site resources, other approaches explored exact and approximate methods. Examples of optimization approaches used for site layout planning include genetic algorithms (40%), mathematical optimization (16%), and hybrid methods (14%), which combine multiple approaches for enhanced performance of the algorithm (e.g., whale optimization and colliding bodies, max–min ant systems, and genetic algorithms). Other approaches, such as probabilistic methods, have been used less frequently (7%).

**Decision Support Systems**

A number of studies have developed systems to assist users in the SLP [24, 35, 36]. These tools are designed to assist the user during the input and evaluation stages built within the system, resulting in selecting the best layout based on pre-established costs and safety metrics [24, 26]. While the main objective of these studies is not the optimization process, a few of them have included optimization procedures within the system [19].

**Fig. 2** Techniques used in site layout planning optimization ( $n = 73$ )





## Site Layout Visualization Tools

Research in this area focuses on using available technology for enhanced site layout visualization. Examples of software used to visualize the layout include MS Excel, computer-aided drawing [25, 29, 30], building information modeling [11, 18, 34], geospatial information systems [1, 7], and virtual reality [22, 32]. These studies typically do not include optimization. However, a few examples have integrated visualization and optimization features [13, 25]. Some of the technologies used in site layout visualization tools include:

- Geographic information system (GIS)
- Computer-aided drawing (CAD)
- Hybrid models
  - Augmented reality and building information modeling integration
  - 3D models and schedule integration.
- Closed-circuit television (CCTV)
- Spreadsheets.

## Overview Papers

Table 1 presents the themes identified across SLP review articles categorized based on the SLP elements reviewed by each study. While all the review articles included optimization models in their sources, two studies presented categorizations of the approaches used in these models [8, 38]. In these articles, authors categorized SLP optimization models using different terms, which for the scope of this paper have been labeled as exact, heuristics, and metaheuristics optimization. The multiple benefits of the optimization approach are discussed, and their conclusions predict the higher use of multi-objective metaheuristic algorithms in future models.

A recent study presented a review of the spatio-temporal planning of construction sites, where SLP models were categorized from optimization approaches and based on the other methods, such as decision-support and visualization or simulation models [4]. In a more specific perspective, time dimension approaches are discussed in a comparative study while demonstrating their differences in space and time utilization [3]. This article highlights the enhanced performance of dynamic time modeling when achieving optimum minimal solutions with less space use and demonstrates the limitations of phased and static approaches. In contrast with the previously mentioned review articles, three studies in the literature aimed to describe the elements required to develop an SLP tool, which is partially considered in the design of an interview questionnaire in the present study [12, 17, 28]. These articles outline site planning elements, such as optimization algorithms, temporary facilities, site space, planning goals, and spatial relationships. While these contributions provide a general guide for future model development, these do not propose specific actions to increase industry applications of SLP tools.

**Table 1** Thematic categorization of SLP review articles ( $n = 7$ )

Overview theme	Study	Categorization criteria used in the paper	Thematic categorization
Approaches among optimization algorithms	El-din et al. [8]	Metaheuristics	Review articles focused on explaining the advantages of each optimization approach when applied to the SLP problem
		Hybrid: optimization + visualization tools	
	Xu et al. [38]	Metaheuristics	
		Heuristics	
Characteristics of SLP models	Ardila and Francis [4]	Exact	Categorization of SLP models based on the methods and techniques used to solve the problem
		Optimization: metaheuristics and exact models	
		Decision support systems	
Space and time approaches in SLP models	Andayesh and Sadeghpour [3]	Visualization and simulation tools	Quantitative comparison of time dimension approaches and their impact on the solution of SLP models
		Static	
		Phased	
		Dynamic	
SLP elements for model development	Kim et al. [17]	Examples and time-space requirements for each approach	The study built upon a review of the literature to categorize two SLP elements (spatial constraints and temporary facilities) to develop a typology model
		Spatial relationships	
	Hawarneh et al. [12]	Temporary facilities	Review article focused on categorizing static and dynamic SLP models based on their design variables
		Time dimension	
		Static	
		Dynamic	
		Optimization approach	
		Model formulation (# of objectives)	
		Decision-making parameters	
		Pathway representation	
		Representation of facility shape	

(continued)

**Table 1** (continued)

Overview theme	Study	Categorization criteria used in the paper	Thematic categorization
	Sadeghpour and Andayesh [28]	Site space	Review article focused on outlining the elements required for SLP model development based on their design variables
		Site layout objects	
		Time dimension	
		Planning goals and objectives	
		Spatial relationships	
		Search approach	
		Optimization techniques	

There is a large body of literature on SLP tool development; however, initial interviews with industry experts indicate that the available tools are not used in practice, showing a gap between what is currently available and the needs of the industry. A previous study addressed this issue [37]. The objective of this paper was to determine the best practices of SLP in the industry and propose a set by step procedure for layout planning in practice. The authors surveyed the site planning practice in the construction industry with questions about SLP elements, which included facilities considered, the software utilized, and design aspects. For each question, a fixed number of multiple choice answers was provided. The participants included project and construction managers (40%), executives and directors (31%), superintendents (5%), and other roles (24%). While the survey technique efficiently collects large amounts of data relatively quickly, the options and categories provided to participants are often narrowed to fit the limited survey space [14]. In practice, the SLP problem involves multiple variables and levels of difficulty that are difficult to narrow down, creating challenges to providing useful tools for practitioners.

The present study aims to build upon previous research on site planning best practices [37] to understand the decision-making process of practitioners when planning a construction site. To achieve this objective, this study will use structured interviews to capture knowledge that would not be collected otherwise.

### 3 Methodology

#### 3.1 Method

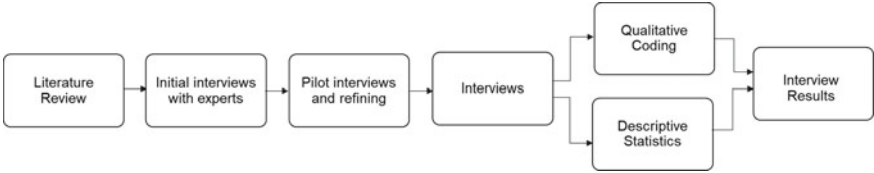
This study uses structured interviews to understand the decision-making process of site layout planners. A questionnaire was developed based on themes from review articles (Table 1) and extensive initial interviews with industry experts. Elements from SLP themes include planning goals and objectives, temporary facilities considered, spatial relationships between facilities, time dimension, and relocation. Questions developed from the initial interviews with experts include site planning responsibilities, definition, SLP development for construction projects, levels of effort in the planning, SLP sequence, resources (documents and tools) used in the planning, efficiency measures, and requirements for an SLP tool. Elements related to optimization approaches were intentionally not included based on the initial interviews that indicate these tools are unknown to practitioners. These two steps facilitated refining the interview questions and conducting pilot interviews to verify that questions were understandable and perceived with the same meaning by all participants as intended by researchers. While there is a high level of structure, questions seek open-ended answers in the same order for each participant. A fixed set of follow-up questions and probes was set in advance to expand or clarify the responses.

#### 3.2 Data Collection

The study sample was selected based on current roles, focusing on participants who work as superintendents or project management professionals on construction sites, with a geographic limitation to Calgary, Alberta. The sample used for data analysis includes twenty (20) superintendents and senior project management practitioners who have self-identified as responsible for the planning of construction sites. A search on LinkedIn displayed that across the highest revenue contractor companies in Canada [16], the local and provincial superintendent practitioners population is approximately 224 and 2600, respectively. The current sample represents 8.92% of the superintendent population at the local level.

#### 3.3 Data Analysis

The collected data was analyzed using qualitative coding techniques and quantification for descriptive statistics. The data was categorized using criterion and labels to generate decision-making considerations. In the context of this study, *criterion* refers



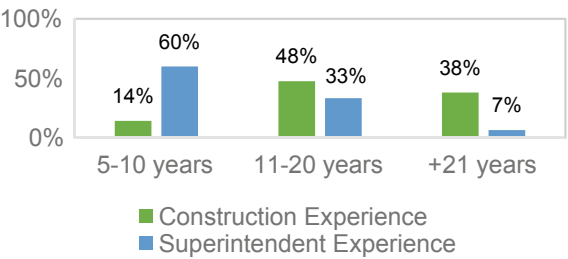
**Fig. 3** Steps in the research methodology

to the standards by which the responses were analyzed, while *labels* allowed to distinguish similar information by providing traceable keywords. Figure 3 summarizes the research methodology and sequence used in this investigation.

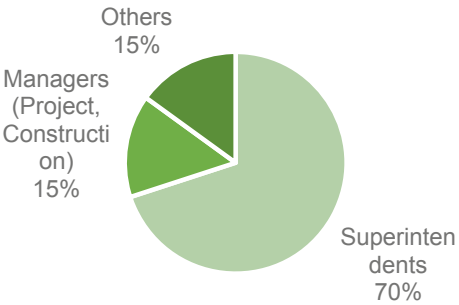
**3.4 Demographics**

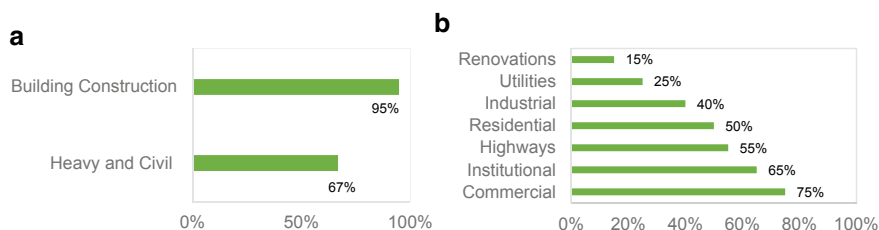
The first section of the interview seeks to provide a better understanding of the experience and knowledge of participants regarding the site planning task. Participants were asked about their professional background, in terms of their level of experience in the construction industry and superintendent responsibilities, previous and current roles, experience in construction subsectors and groups, range of contract amounts and project size. The results of this section are summarized in Figs. 4, 5 and 6.

**Fig. 4** Level of experience among participants



**Fig. 5** Current positions



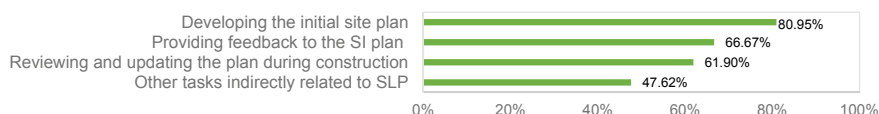


**Fig. 6** Level of experience in construction projects **a** construction subsectors, **b** industry groups

Most of the participants have more than 11 years of experience in the construction industry (86%, Fig. 4), followed by a level of experience in superintendent roles between 5 and 20 years (93%, Fig. 4). Additionally, results show that 70% of participants work as a superintendent with complete responsibility for the site plan. In comparison, 30% of the participants hold senior management, project coordination, and field engineering positions with different levels of involvement in the site planning (Fig. 5). Following the current north American industry classification system, participants were asked about the specific types of construction they have worked on in their experience. 95% of the participants indicated having experience in buildings, and 67% have worked in heavy civil construction (Fig. 6a). Among these two subsectors, participants have worked in various industry groups: residential, commercial, institutional, industrial, highways, and bridge construction, along with renovations of institutional or commercial buildings (Fig. 6b). Participants were asked about their current responsibilities regarding site planning. While different themes appeared among the answers, three main responsibilities were identified and summarized in Fig. 7, along with sub responsibilities indirectly related to the planning task.

The majority of responsibilities held by participants in their current roles are related to the development of site plans, team members providing feedback to the plan developed by the superintendent, and a collaborative review and update process of the plan during construction. Other tasks were teaching new superintendents into the planning, peer-reviewing the plans developed for other projects, and communicating the plan to stakeholders.

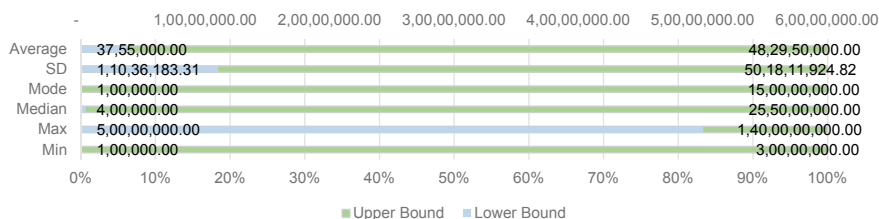
Finally, the range of contract amounts in the experience of participants was assessed, as shown in Table 2 and Fig. 8. Contract amounts start from CAD \$100,000 up to CAD \$1,400,000,000 and statistical values are estimated for lower and upper bounds. This exhibits a wide range of values in the sample, which translates into a diverse group of construction sites in the experience of participants.



**Fig. 7** Distribution of site planning responsibilities performed by participants

**Table 2** Range of contract amounts (\$M)

Measure	Lower bound	Upper bound
Min	100,000.00	30,000,000.00
Max	50,000,000.00	1,400,000,000.00
Median	400,000.00	255,000,000.00
Mode	100,000.00	150,000,000.00
SD	11,036,183.31	501,811,924.82
Average	3,755,000.00	482,950,000.00

**Fig. 8** Range of contract amounts (\$M)

The summary shows a wide range of expertise in construction and site planning responsibilities among participants. The level of experience in construction subsectors, industry groups, and their professional background exhibits a group of highly-skilled individuals in the site planning practice. The different levels of complexity, site sizes, contract amounts, and context ensure collecting broad, open-ended answers for general questions (Sect. 4.1) and distinct answers for project-specific questions (Sect. 4.2).

## 4 Results

### 4.1 Definitions

The purpose of the next set of questions is to define the planning of construction sites from the perspective of practitioners. The instrument was designed using open-ended questions so that responses by participants were not narrowed to a finite number of given choices.

#### 4.1.1 Site Planning Characterization

Participants were asked to define what site planning is in their opinion. The answers referred to different terms and areas, and the categorization criterion method was

**Table 3** Categorization criterion approach

Term	Definition
Definition	The SLP definition according to participants
Goals	The results to achieve (improve, minimize, maximize)
Method	How the process is done
Requirement	Elements that should/must be included in the process
Impact	The effect of the process
Factors	Potential situations/elements that influence the planning
Keyword	Single concepts mentioned to define or describe the process

used as described (Sect. 3.3) to enable summarizing the different styles of answers. The resulting labels were categorized as definition, goals, methods, requirements, impact, and factors. Additionally, related keywords were also mentioned as part of the responses. Table 3 presents the labels used to categorize the responses, and Table 4 shows a sample of the final answers after labeling them.

**Table 4** Sample of labels to characterize the site planning definition

Definition	Goals	Method
Meeting the requirements of the client	Minimizing movement of materials	Planning for routes, power, people access, locations of trailers
Planning for the success of the job	Maximizing efficiency	Planning the materials to be close to the working areas
Understanding the movement within the site	Saving money Ensuring safety	Involving the whole team Identifying all items that are going to be part of the project and the site
Pre-execution activity	Ensuring a safe and productive flow of material and equipment	Setting the perimeter
Accommodating for inevitable changes		Putting infrastructure together
Communication tool	Protecting the public	Planning for access and storage
Communication of the requirements of the project	Having a good flow of traffic (not too many reversing vehicles)	Determining locations of site resources, access, haul routes, and cranes
Starting point for any project		Breaking down the entire project from start to finish



**Table 5** Site planning responsibilities matrix

Tasks in site planning	Project management roles						
	Superintendent	Project manager	Project coordinator	Safety coordinator	Trades	Client	Consultant
Providing requirements and guidelines						X	X
Developing the initial site plan	X	X	X	X	X		
Reviewing and updating the plan during construction	X	X	X				
Providing input or feedback to the SI plan		X	X	X	X	X	X

SI superintendent

#### 4.1.2 Roles and Tasks in the Site Planning

Participants were asked about the roles and tasks involved in planning a construction site, specifically, what tasks are done and who is in charge of completing them. Seven different roles among project stakeholders kept repeating, such as superintendent, project manager, and project coordinator, along with four (4) main tasks as summarized in Table 5.

Other tasks mentioned by the participants include drawing the formal site layout, communicating the site layout to stakeholders, and peer-reviewing site plans developed by other superintendents. Additionally, participants mentioned that experienced superintendents participate in internal trainings within the companies to teach new team members about the site planning task.

#### 4.1.3 Site Planning Benefits

Participants were asked about the benefits of developing the site plan for a construction project. They mentioned various terms and keywords labeled using the categorization criterion approach. The results are compiled as goals and benefits in Table 6.

When labeling the responses, researchers came upon patterns of answers that correspond to goals from the site planning literature [28]. However, other terms were related to specific project elements that have not been categorized as a goal in the

**Table 6** Categorization of site planning benefits

Productivity	Safety	Security	Functionality
Increases efficiency	Improves site morale	Keeps materials and equipment secure from thefts	Improves site organization
Avoids rework	Helps emergency response	Helps implement measures for violence or disruption on the site	Site coordination
Reduces movement	Improves safety indicators		Promotes efficient use of space
Reduces distance	Facilitates risk mitigation		Provides smooth flows
Facilitates loss mitigation			
Improves costs savings			

past, such as schedule compliance, stakeholders, and project visualization. While some of these benefits are tangible and measurable, the main benefit extracted from these answers establishes that having a site plan is critical for the success of the other aspects of the construction project.

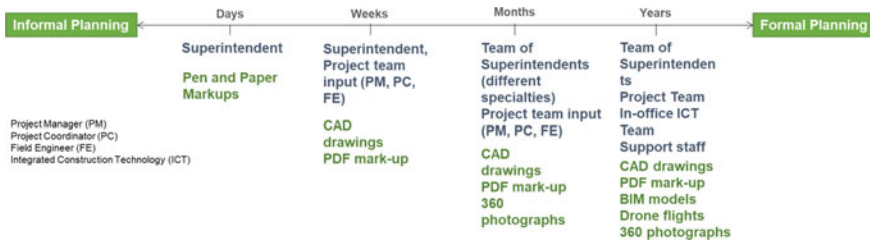
**4.1.4 Development and Levels of Effort of the Site Plan in Construction Projects**

Participants were asked if the development of a site plan is used for any construction project in their practice and the levels of effort and factors that influence that planning. All the participants confirmed developing a site plan for every construction project. They also indicated that the level of effort varies for different projects. From their explanations, themes appeared in the data, where different stakeholders, decision-makers, and points in time were identified. For categorization purposes, responses were grouped in informal and formal planning, based on the decision-makers involved, resources used, and time ahead planned. The summary of the responses is shown in Fig. 9.

With regard to the specific factors that influence the planning of a construction site, some examples include:

- Level of experience of the project team
- Site location context
- Project type (industry group)
- Space availability
- Availability of utilities and services.

All these, along with unforeseen situations, can trigger potential changes in the level of effort required to plan the site.



**Fig. 9** Levels of effort to plan a construction site layout (time, decision-makers, and resources involved)

## 4.2 Project Specific Questions

### 4.2.1 Sequence to Plan a Construction Site Layout

Participants were asked about the planning sequence followed in planning a construction site. For this purpose, they were shown an example project with a number of location and context constraints (downtown area, congested traffic, large site) to answer while considering an example institutional or commercial project from their experience. The steps mentioned as part of the sequence are summarized as follows:

#### I. Preliminary investigation

Participants concur on starting the site planning as early as possible (tender phase, post-award, before mobilization). The initial planning would involve a preliminary investigation of:

- Site areas and surroundings, soil conditions, utilities available.
- Site resources required at different times in the schedule (based on labor curves).
- Logistics around access and egress of people and equipment, deliveries, garbage disposal.
- Revision of contract documents, regulations, and available information on the project.
- Emission of required permits as established in bylaws or regulations.
- Conversations with stakeholders to understand needs and limits and establish agreements if needed.
- Measures related to safety (public protection, employee protection), environment (species protection, nearby rivers or wetlands, protected areas), and security (fencing, lighting).

#### II. Developing the plan

At the early stage of the planning (before contract award or mobilization), the superintendent would make decisions based on experience from past projects and the amount of information available. Typically, input from the project team (manager, coordinator, estimator, planner, trades, field engineers, and safety representatives)

would also be considered. The information is focused on understanding if locations are:

- (a) logical
- (b) feasible (dimensions fit, no collisions)
- (c) not on the way of future items
- (d) if a resource is in the way, the need for a relocation plan is considered.

The point of view of others in the team (coordinators, managers) focuses on providing ideas, insights, or potentially missed situations regarding the locations of resources. In this sample, all the participants concurred on the importance of having an experienced superintendent in charge of the site planning and highlighted that this role makes the final decisions concerning the site layout.

### **III. Reviewing and updating the plan**

100% of the responses mentioned that the whole schedule of the project is planned during the early phases. However, revisions and updates to the plan are made if:

- (a) New information is available.
- (b) Changes must be implemented to maintain smooth operations.
- (c) Change orders or unforeseen situations changed the initial plan.

Updates to the plan can be performed regularly (weekly, monthly). However, these are impacted by the nature of the project (simple or complex), the experience of the team, and the value placed by stakeholders in the site layout plan. Once the site layout is marked-up, an updated layout is developed. These processes occur iteratively during the execution of the project.

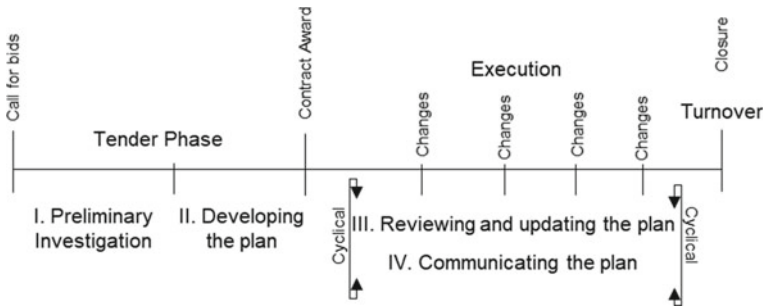
### **IV. Communicating the plan**

Both the written plan and the resulting site layout are used as a communication tool for the project and serve multiple purposes:

- (a) Onboarding document
- (b) Guidelines for personnel
- (c) Showing emergency services how to access the site
- (d) Leading suppliers on how to access the site
- (e) Showing the trades the areas to work in the day or week.

The graphical result of these four steps is summarized in Fig. 10.

This sequence (Fig. 10) represents the different stages of the site planning and highlights how this task starts even before the contract is awarded to the contractor. Reasons provided by participants explain that developing the plan at an early stage is a strategy to ensure different scenarios are considered and discussed with the client before the start of execution.



**Fig. 10** Sequence to plan a construction site

#### 4.2.2 Documents Used in the Site Planning Task

Participants were asked about the documents used during the site planning task. The examples provided cover twelve physical or digital documents considered during this task. More specifically, the purpose of using these documents during the site planning is:

- **Schedule:** Helps understand the sequence of activities and the order in which resources should arrive on the site. From the schedule, practitioners can obtain labor curves or histograms that indicate the number of people and resources required by the project at any point in time. This information is important to estimate the number of resources, capacity (allocating people, tools, materials), and arrival and departure times (queue). Based on this information, superintendents can choose long-term or short-term locations for site resources. The schedule also provides foresight regarding potential changes and relocations that facilitate planning them with anticipation. 94% of participants mentioned using the schedule as part of the planning.
- **Drawings:** Participants mentioned using general, structural, or utilities drawings to better understand the underground areas of the site and the ratio between available site space and the building footprint. 86% of participants mentioned drawings as part of the planning. Additionally, 39% of participants stated the overlay of these drawings on top of Google Earth images for a higher understanding of site conditions.
- **Regulations and bylaws:** These documents provide the requirements and restrictions that the site should comply with to execute the project. Key regulations include provincial safety codes, traffic codes, and environmental regulations. 71% of participants mentioned using regulations as part of the planning.
- **Specifications:** The bid package and the specifications provide the initial information that allows superintendents to plan the site. 39% of participants mentioned the specifications as part of the planning. While specifications typically refer

to execution details and regulations, participants mentioned that site restrictions could also be shown in these documents. Some examples include:

- Requirements regarding the layout.
- Space provided by the client (also if the client is dictating how to set the site).
- Constraints (environmental, pedestrians, neighbors).
- Access or traffic in the area.

Other documents include digital images (site photographs, google earth views, drone footage), environmental reports, internal manuals or checklists, safety plans, city maps, survey data, labor reports, and flying path information.

### **4.2.3 Site Resources**

Participants were asked about the site resources considered during the planning of the example projects. The examples provided by practitioners confirm many of the resources seen in the literature while adding a number of resources that are typically not considered during model development (e.g., heliports, turnstiles, environmentally protected areas, barriers, public protection facilities, and utilities). On average, participants thought of 12 resources during the interviews and provided multiple parameters for their allocation, which are summarized in Table 7.

It is important to note that these parameters can be translated into if–then rules for tool development, which is one of the sub-objectives of the long-term investigation. Additionally, some of these parameters can be measured using traditional methods (distance measurement) or through binary conditions as established in academic site layout models, facilitating its integration into a tool.

### **4.2.4 Relocation of Resources**

Participants were asked about perspectives of relocation (negative, positive), examples of relocation, and parameters considered during this process. 63% of participants consider that relocation could provide potential benefits or give space to new space opportunities; nevertheless, relocation is perceived negatively. Table 8 presents example parameters and relocations provided by participants.

In all the examples provided, there are costs or schedule impacts. Participants plan to avoid or reduce these impacts by selecting fixed-long term locations, among other parameters. Additionally, it was mentioned that relocation happens when other alternatives are not feasible; hence the movement is planned with as much anticipation as possible to decrease the potential impact. Other times, it occurs in a reactionary way (at the moment, no planning, increased impact).

**Table 7** Sample of parameters for allocation of site resources

Resource	Parameters
Space	How much site space is available after considering the building footprint?
	Is there enough space for laydown areas?
	Is additional laydown space needed?
Utilities	Are utilities available on the site? (permanent or temporal power, water, internet)
Materials	Which kind of materials is going to be used?
	What kind of storage is required? (on-site, off-site)
	Are materials storage and laydown areas accessible to trucks?
Trailers and washrooms	How many trailers are required for the number of people on the site?
	How many washrooms are required?
	Can the trailers be stacked?
	Can the trailers be combined for additional office or storage space?
Workflow	How much flow of equipment traffic?
	How much flow of people?
	How much flow of materials?
Tower crane	Can the tower crane reach materials?
	Is there any internal or external interference to the tower crane operations?
Protection	Are security protection measures required?
	Are environmental protection measures required?
	Are public protection measures required?
	Is there any sort of interference in the area? (equipment, pedestrians)
Parking	Is there a parking lot for employees available on the site?
	Is there an off-site space available for parking?
Relocation	Will the resource need to be moved in the future?
	How far in advance can the relocation be planned?
	Is the resource moveable through wheels or accessible to movement equipment?

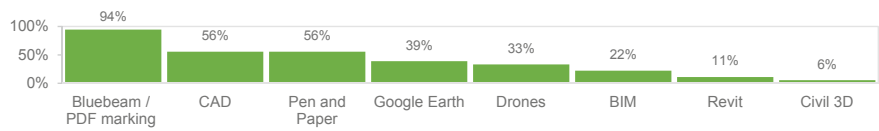
**4.2.5 Tools Used in the Site Planning Task**

Participants were asked about the specific tools used during the planning process. While participants mentioned multiple resources and software, these are used to assist in the process (check drawings, check construction activities, evaluate site conditions or topography) and not to develop or measure the site plan itself. Figure 11 summarizes the tools mentioned by participants during the interviews.

CAD is used to draw the layout and its changes, while BIM checks flow and collisions in the site operations. 94% of practitioners use Bluebeam, a PDF marking software that facilitates drawing changes in the layout and communicating them to the team. 56% of practitioners keep using pen and paper to draw changes in the layout.

**Table 8** Sample of parameters for relocation of resources

Parameters	Examples
Budget available for different alternatives, e.g., off-site storage	Receiving more materials than needed, which were then in the way of other activities and required resources to move them across the site
Cost-based decisions Actual improvements for the site layout	Relocating office space into built parts of the building
Work sequence, the progress of execution Availability of space use Planned in advance	Dealing with haul routes relocation with significant costs impact and additional operational issues
Ensuring resources are as moveable as possible: on wheels, easily dissembled, or accessible to equipment (cranes)	Moving excavation ramps to execute different areas of the project. The relocation was part of the plan
Planning for long-term locations as much as possible	Planning for the location of the office complex to last three years, planned relocation one month before finishing



**Fig. 11** Tools for site planning

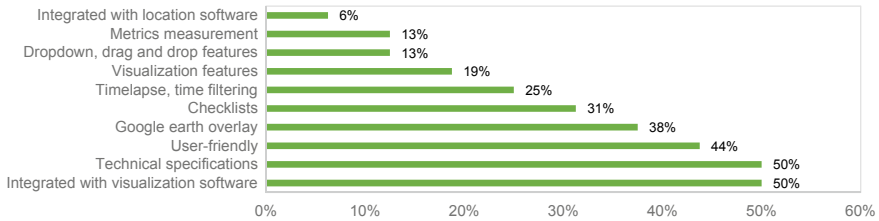
The same percentage mentioned the use of CAD to draw the changes more formally. Still, participants highlighted that these drawings are typically made by others in the team (project coordinator, project manager, engineering students, or in-office staff). Google Earth was also mentioned as a helpful tool by 39% of participants. Revit, BIM, Civil 3D, or drone technology would be used in-office by a team skilled in these tools. The outcome of this process is usually recommendations for the project in terms of collisions, how well is the flow of elements in the site, and potential issues identified through the software.

**4.2.6 Requirements for a Site Planning Tool**

For the final question of the interview, participants were asked to state what a potential tool for site planning would look like to be helpful in this task. While various responses were obtained on this question, participants referred mainly to two areas: software integration features and decision-support tools. The results of this question are summarized in Fig. 12.

50% of participants mentioned the potential benefits of having CAD or BIM drawings as part of the tool but adding the capacity of modifying the locations of





**Fig. 12** Requirements for a tool for site planning

resources in real time, and 50% of participants would prefer to have technical specifications, resource dimensions, and equipment available in the interactive layout to assist the decision-making; 44% of participants highlighted the importance of having simple, user-friendly tools that consider the IT skills of superintendents. As mentioned in previous sections, different decision-makers have different roles in the planning process. While the superintendent takes operational and strategic decisions, a support team like project coordinators or field engineers draws the formal layout. Hence, different tool features should consider the actual users of each function. Google Earth overlay, checklists, and time filtering also have a significant consideration (38%, 31%, and 25%, respectively), underlining potential features for tool development.

## 5 Summary and Concluding Remarks

General definitions provided by practitioners offer a high-level understanding of the site planning process from a practical perspective. Contrary to the initial hypotheses, site planning is considered a critical task for the success of the project, which is both influenced by multiple factors and has an impact on different project areas, such as schedule compliance, logistics, and productivity metrics. While the value invested in the site layout plan can vary based on the unique characteristics of the project, 100% of the participants considered that a site layout plan must be in place for any project, with different levels of effort as required.

The interview results show multiple and diverse responses to project-specific questions, where the answers were provided by practitioners using a specific project from their experience as an example. Despite the broad range of industry subsectors and construction groups, 94% of participants declared planning the entirety of the project schedule during the early phases and, in some cases, even before the contract was awarded. Some of the reasons state using this time to capture anything not disclosed in the bid documents. 6% of the participants mentioned that the project had to be planned in a phased approach, addressing different sections at different times due to specific constraints. In contrast with the academic problem definition, where static, phased, and dynamic approaches have been explored, the practical

perspective evaluates the problem in a “single-take” manner with slight changes as required by the project.

Participants concur with having a superintendent in charge of this process regarding the decision-makers. While others in the project team can contribute by providing feedback or information, the superintendent is the primary decision-maker. Decisions are made based on experience and judgment, and no standard guideline or tool is utilized to select the locations of site resources. Commercially available software (e.g., CAD, BIM, Revit, Bluebeam) is used by the project team to draw the layout or marking-up changes. However, the IT skills of superintendents vary across the sample, and different roles in the project team will use the software for various purposes. This situation creates a constraint for tool development, which should address the needs of the users while considering their limitations.

When asking the participants what would be their ideal tool, the majority of responses were focused on three areas: integrating the tool with commercially available software (drones, GPS, VR), having technical specifications of the site and its resources available in the tool, and having a simple, user-friendly tool that can be used without a significant time commitment or complexity. Some of these requirements have been presented in past studies. However, the level of sophistication of the tools offers a challenge for the users. Decision support features also had a significant portion among the results, with checklists, dropdowns, drag and drop, and metrics information available to the superintendent to consider when planning the site. Although developing a single tool that works for any construction project is challenging, the practical considerations presented in this paper provide a general outline of site planning for any project. These steps and information can be modeled artificially to develop efficient tools for site layout planning.

This paper presented the first step into a long-term investigation focused on reducing the gap between currently available site planning tools and actual requirements from industry practitioners. As an initial step, this paper presented the findings of structured interviews with a highly-experienced group of industry practitioners in Calgary, Alberta. The results provide practical considerations regarding the site planning decision-making process and potential requirements for tool development, which will be addressed in the subsequent phases of the investigation.

**Acknowledgements** The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) to fulfill this research. The authors also express their deepest gratitude to all the practitioners who participated in this study.

## References

1. Abune'meh M, El Meouche R, Hijaze I, Mebarki A, Shahrou I (2016) Optimal construction site layout based on risk spatial variability. *Autom Constr* 70:167–177. <https://doi.org/10.1016/j.autcon.2016.06.014>
2. Andayesh M, Sadeghpour F (2012) Dynamic site layout planning through minimization of total potential energy. *Autom Constr* 31:92–102. <https://doi.org/10.1016/j.autcon.2012.11.039>

3. Andayesh M, Sadeghpour F (2014) The time dimension in site layout planning. *Autom Constr* 44:129–139. <https://doi.org/10.1016/j.autcon.2014.03.021>
4. Ardila F, Francis A (2020) Spatiotemporal planning of construction projects: a literature review and assessment of the state of the art. *Front Built Environ* 6:1–13. <https://doi.org/10.3389/fbuil.2020.00128>
5. Charoenngam C, Maqsood T (2001) A qualitative approach in problem solving process tracing of construction site engineers. In: *Proceedings of the seventh annual association of researchers in construction management conference*, 1 Apr 2015, pp 5–7. [http://www.arcom.ac.uk/-docs/proceedings/ar2001-475-483\\_Charoenngam\\_and\\_Maqsood.pdf](http://www.arcom.ac.uk/-docs/proceedings/ar2001-475-483_Charoenngam_and_Maqsood.pdf)
6. Chau KW (2004) A two-stage dynamic model on allocation of construction facilities with genetic algorithm. *Autom Constr* 13(4):481–490. <https://doi.org/10.1016/j.autcon.2004.02.001>
7. Cheng MY, O'Connor JT (1996) ArcSite: enhanced GIS for construction site layout. *J Constr Eng Manag* 122(4):329–336. [https://doi.org/10.1061/\(ASCE\)0733-9364\(1996\)122:4\(329\)](https://doi.org/10.1061/(ASCE)0733-9364(1996)122:4(329))
8. El-din MN, Hesham B, Khaled S (2015) Existing site layout planning models and approaches. *Int J Sci Eng Res* 6(8):997–1003
9. Elbeltagi E, Hegazy T, Hosny AH, Eldosouky A (2001) Schedule-dependent evolution of site layout planning. *Constr Manag Econ* 19(7):689–697. <https://doi.org/10.1080/01446190110066713>
10. Hammad AWA (2019) A multi-objective construction site layout planning problem solved through integration of location and traffic assignment models. *Constr Manag Econ* 38(8):756–772. <https://doi.org/10.1080/01446193.2019.1659510>
11. Hammad AWA, Akbarnezhad A, David R, Travis S (2015) A computational method for estimating travel frequencies in site layout planning. *J Constr Eng Manag* 142(5):1–13. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001086](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001086)
12. Hawarneh AA, Salaheddine B, Firas G (2021) Construction site layout planning problem: past, present and future. *Expert Syst Appl* 168:114247. <https://doi.org/10.1016/j.eswa.2020.114247>
13. Hegazy T, Elbeltagi E (1999) EvoSite: evolution-based model for site layout planning. *J Comput Civ Eng* 13(3):198–206. [https://doi.org/10.1061/\(ASCE\)0887-3801\(1999\)13:3\(198\)](https://doi.org/10.1061/(ASCE)0887-3801(1999)13:3(198))
14. Holloway I, Galvin K (2016) *Qualitative research in nursing and healthcare*, 4th edn. Wiley-Blackwell, Chichester
15. Kaveh A, Mostafa K, Mohammad R, Mohammad R (2018) Charged system search and magnetic charged system search algorithms for construction site layout planning optimization. *Period Polytech Civ Eng* 62(4):841–850. <https://doi.org/10.3311/PPci.11963>
16. Kennedy D (2021) Canada's top contractors. *On-Site Magazine*, June 2021. [https://www.on-sitemag.com/wp-content/uploads/2021/06/ONSITE\\_JUNE21\\_LAZ-TOPCONTRACTOR-DE-1.pdf](https://www.on-sitemag.com/wp-content/uploads/2021/06/ONSITE_JUNE21_LAZ-TOPCONTRACTOR-DE-1.pdf)
17. Kim M, Ryu H-G, Kim TW (2021) A typology model of temporary facility constraints for automated construction site layout planning. *Appl Sci*
18. Kumar SS, Cheng J (2015) A BIM-based automated site layout planning framework for congested construction sites. *Autom Constr* 59(59):24–37. <https://doi.org/10.1016/j.autcon.2015.07.008>
19. Lam KC, Tang CM, Lee WC (2005) Application of the entropy technique and genetic algorithms to construction site layout planning of medium-size projects. *Constr Manag Econ* 23(2):127–145. <https://doi.org/10.1080/0144619042000202834>
20. Lien LC, Cheng MY (2012) A hybrid swarm intelligence based particle-bee algorithm for construction site layout optimization. *Expert Syst Appl* 39(10):9642–9650. <https://doi.org/10.1016/j.eswa.2012.02.134>
21. Mawdesley MJ, Al-Jibouri J, Hongbo Y (2002) Genetic algorithms for construction site layout in project planning. *J Constr Eng Manag* 128(5):418–426. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2002\)128:5\(418\)](https://doi.org/10.1061/(ASCE)0733-9364(2002)128:5(418))
22. Muhammad A, Ibrahim Y, Sepehr A, Tolga C (2020) Adoption of virtual reality (VR) for site layout optimization of construction projects. *Tek Dergi* 9833–9850. <https://doi.org/10.18400/tekderg.423448>

23. Ning X, Lam KC (2013) Cost-safety trade-off in unequal-area construction site layout planning. *Autom Constr* 32:96–103. <https://doi.org/10.1016/j.autcon.2013.01.011>
24. Ning X, Lam KC, Lam M (2011) A decision-making system for construction site layout planning. *Autom Constr* 20(4):459–473. <https://doi.org/10.1016/j.autcon.2010.11.014>
25. Osman HM, Maged E, Moheeb E (2003) A hybrid CAD-based construction site layout planning system using genetic algorithms. *Autom Constr* 12(6):749–764. [https://doi.org/10.1016/S0926-5805\(03\)00058-X](https://doi.org/10.1016/S0926-5805(03)00058-X)
26. RazaviAlavi S, AbouRizk S (2017) Genetic algorithm-simulation framework for decision making in construction site layout planning. *J Constr Eng Manag* 143(1):1–13. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001213](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001213)
27. RazaviAlavi S, AbouRizk S (2017) Site layout and construction plan optimization using an integrated genetic algorithm simulation framework. *J Comput Civ Eng* 31(4):1–10. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000653](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000653)
28. Sadeghpour F, Andayesh M (2015) The constructs of site layout modeling: an overview. *Can J Civ Eng* 42(3):199–212. <https://doi.org/10.1139/cjce-2014-0303>
29. Sadeghpour F, Moselhi O, Alkass ST (2004) A CAD-based model for site planning. *Autom Constr* 13(6):701–715. <https://doi.org/10.1016/j.autcon.2004.02.004>
30. Sadeghpour F, Osama M, Sabah TA (2006) Computer-aided site layout planning. *J Constr Eng Manag* 132:871–881. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2006\)132](https://doi.org/10.1061/(ASCE)0733-9364(2006)132)
31. Said H, El-Rayes K (2013) Optimal utilization of interior building spaces for material procurement and storage in congested construction sites. *Autom Constr* 31:292–306. <https://doi.org/10.1016/j.autcon.2012.12.010>
32. Singh AR, Venkata K (2018) User behaviour in AR-BIM-based site layout planning. *Int J Prod Lifecycle Manag* 11(3):221–244. <https://doi.org/10.1504/IJPLM.2018.094715>
33. Singh AR, Vinayak S, Venkata K (2017) Augmented reality (AR) based approach to achieve an optimal site layout in construction projects. In: *ISARC 2017—proceedings of the 34th international symposium on automation and robotics in construction*, pp 165–172. <https://doi.org/10.22260/isarc2017/0022>
34. Singh AR, Patil Y, Delhi VSK (2019) Optimizing site layout planning utilizing building information modelling. In: *Proceedings of the 36th international symposium on automation and robotics in construction, ISARC 2019, July 2019*, pp 376–383. <https://doi.org/10.22260/isarc2019/0051>
35. Song X, Zhe Z, Xu J, Zeng Z, Shen C, Peña-Mora F (2018) Bi-stakeholder conflict resolution-based layout of construction temporary facilities in large-scale construction projects. *Int J Civ Eng* 16(8):941–964. <https://doi.org/10.1007/s40999-017-0233-4>
36. Tommelein ID, Levitt RE, Hayes-Roth B (1992) SightPlan model for site layout. *J Constr Eng Manag* 118(4):749–766
37. Whitman J, Deshpande A, Zech W, Perez M (2021) Construction site utilization planning: a process based upon industry best practices. *CivilEng* 2(2):309–324. <https://doi.org/10.3390/civileng2020017>
38. Xu M, Zhongya M, Siyu L, Yi T (2020) Optimization algorithms for construction site layout planning: a systematic literature review. *Eng Constr Archit Manag*. <https://doi.org/10.1108/ECAM-08-2019-0457>
39. Yi W, Chi H, Wang S (2018) Mathematical programming models for construction site layout problems. *Autom Constr* 85:241–248. <https://doi.org/10.1016/j.autcon.2017.10.031>
40. Zhang H, Wang JY (2008) Particle swarm optimization for construction site unequal-area layout. *J Constr Eng Manag* 134(9):739–748. [https://doi.org/10.1061/\(ASCE\)0733-9364\(2008\)134:9\(739\)](https://doi.org/10.1061/(ASCE)0733-9364(2008)134:9(739))
41. Zhang JP, Liu LH, Coble RJ (2002) Hybrid intelligence utilization for construction site layout. *Autom Constr* 11(5):511–519. [https://doi.org/10.1016/S0926-5805\(01\)00071-1](https://doi.org/10.1016/S0926-5805(01)00071-1)

# Irregular Dynamic Site Layout Optimization Model



Heba Kh. Gad, Mostafa H. Ali, Aya Eldesouky, Alshimaa Abdullatif,  
Mos. Serry, Hosny Ossama, and Yasmeeen A. S. Essawy

**Abstract** When site space is limited, construction site layout planning plays a critical role in the safety and productivity of all operations. As the construction progresses, dynamic planning for the site layout becomes a need to match the project progress. Proper layout planning should take into account the productivity of operations, minimizing the in-situ travel time, the dynamic nature of site operations, and safety regulations implementation on site. This paper is a continuation of the previous work tackling irregular site layout planning, where it presents a new model approach for optimizing the dynamic planning of site layout with safety consideration. This model's algorithm shows high performance in tackling complex projects with limited run time compared to other work in literature. The model showcases an automated mapping tool in conjunction with a dynamic scheduling observing safety to produce the optimized site layout using genetic algorithms. To demonstrate the benefits of the proposed model approach, the case study presented in the “Dynamic Layout of Construction Temporary Facilities Considering Safety” paper is taken as the performance reference to verify and validate the model's output. Examination of the results and the comparative analysis is performed to demonstrate the variance between the outputs of the existing reference model, and this model performed herein.

**Keywords** Irregular site layout optimization · Genetic algorithms · Automation · Construction management

---

H. Kh. Gad (✉) · M. H. Ali · A. Eldesouky · A. Abdullatif · Mos. Serry · H. Ossama ·  
Y. A. S. Essawy

Department of Construction Engineering, The American University in Cairo (AUC), Cairo, Egypt  
e-mail: [heba.kh@aucegypt.edu](mailto:heba.kh@aucegypt.edu)

H. Ossama  
e-mail: [ohosny@aucegypt.edu](mailto:ohosny@aucegypt.edu)

Y. A. S. Essawy  
e-mail: [yasmeeen.sherif@eng.asu.edu.eg](mailto:yasmeeen.sherif@eng.asu.edu.eg); [y\\_essawy@aucegypt.edu](mailto:y_essawy@aucegypt.edu)

Y. A. S. Essawy  
Department of Structural Engineering, Ain Shams University (ASU), Cairo, Egypt

## 1 Introduction

The process of properly allocating site facilities during the course of a construction project life cycle with the least cost is known as site layout planning. Several site layout models have been developed with different aspects in mind, such as safety, productivity, and security in an attempt to maximize the overall operation and improve efficiency. It was proven that proper site layout planning can impact the project's objectives; cost, time and quality; therefore, it has gained much attention and the need to be studied at more depth and use diverse methodologies to overcome these challenges.

When planning a site layout there are two types; static and dynamic models. Static model entails that the site's Temporary Facilities (TF) are allocated a specific site space from the beginning until the end of construction. In contrast, the dynamic model allows for the movement of TF throughout the project lifecycle, which could act as an edge to speed up the construction, as the facilities are moved to serve the construction activities more efficiently. The size, shape, and function of TF varies from one construction project to another, and those TF could include but are not limited to warehouses, job offices, workshops, batch plants and equipment such as tower cranes. Site layout planning is not only limited to placement of TFs as close as possible to their desired locations, but it also considers the safety measures. Safety measures are taken into consideration through adopting several strategies; like keeping a minimum safe distance between the fixed facilities and the temporary facilities, minimizing the intersections of the paths to minimize possible construction accidents that could take place.

This paper tackles the challenge of developing a model that would optimize the dynamic construction site layout design for irregular shaped facilities. This paper attempts at filling the gap identified in literature, by presenting a dynamic and time-efficient irregular site layout optimization tool that considers multiple orientations of facilities incorporating safety considerations.

## 2 Literature Review

Site layout planning has attracted researchers' attention long ago; it has too many variables and considerations that one should keep in mind to plan a site layout that would actually enhance the efficiency. A paper by Elbeltagi et al. [2] is one of the early papers to tackle the site layout planning issue, the authors were one of the first researchers to incorporate safety considerations to the model. They provided safety zones around each facility following the OSHA regulations. Nevertheless they also were the first to incorporate scheduling plan changes throughout the project lifecycle. Elbeltagi et al. used discrete model formulation and modeled regular and irregular shapes, adopting genetic algorithms for the optimization process. This paper has been cited more than 100 times to date.

Abotaleb et al. [1] used genetic algorithm to optimize the site layout problem using mathematical formulations for dynamic shapes like shapes with curves and freeform irregular shapes. Their algorithm shapes facilities in accordance with the available space, which entails that a facility could take several shapes depending on the nearest available free space. The model had some limitations and mainly due to that it only considers cost when finding the near optimal solution. It's also very time consuming due to the high number of constraints, therefore computation is a time intensive task.

Since 2010 safety concern has gained momentum, most recent papers are inclined to include safety considerations in the site layout planning. As Farmakis [3] showed us by incorporating in their objective of minimizing cost while maximizing safety. Also as presented by Xu et al. [4], where they introduced a hybrid multiobjective simulated annealing model that works with two objective functions, one to minimize cost and the other to maximize safety.

Furthermore, this paper addresses the gap referred to by Elbeltagi et al. [2], suggesting the enhancement of a mapping tool for the site layout. Not to fall short of the ethical aspect of taking safety into consideration, we have incorporated the necessary safety measures. Our paper also presents a new approach to defining irregular shapes that accepts modifications.

### 3 Methodology

#### 3.1 Model General Logic

This model tackles some of the gaps identified in literature as it presents a dynamic and time-efficient irregular site layout optimization tool that considers multiple orientations of facilities and incorporates safety considerations. Genetic algorithms approach is used to solve this optimization problem while maintaining all constraints. The proposed model utilizes excel macros embodied in Evolver to optimize the site layout problem. The model is created in a user-friendly format to ease the user's experience. The proposed model is dynamic in the sense that it can accommodate different schedules for the same project and plans the layout based on the existing fixed facilities at the time and the needed temporary facilities. The model shows good time efficiency in the case study compared to the reference model. This is mainly due to the simplicity of the algorithm in mapping the site layout and defining the constraints. Several models in literature neglected the possible orientations of the Temporary Facilities (TFs) unlike this model, which considers four different orientations of the TFs with 90 degrees variance: (0°, 90°, 180°, 270°).

### 3.1.1 Safety Considerations

Incorporating safety aspects in site layout planning is a concept first introduced by Elbeltagi et al. [2]. Safety considerations in the model are derived from Elbeltagi's work and are incorporated through two aspects. The first aspect is in defining negative relationships between facilities that constitute a safety hazard if placed close to each other. The user can choose due to safety consideration to place two facilities as far from each other as possible. The second aspect is in defining safety zones. Based on the project requirements, the user can define certain areas to be restricted safety zones (a barricade for example), where no facilities would be placed.

## 3.2 Model Logic Definition

### 3.2.1 Site Perimeter Definition

In this proposed model, the site area is modeled as a mesh of square units where every facility is plotted by occupying the square units that best fit the shape of the facility. Sizing the mesh units is important as it determines the accuracy of shape representation (the smaller the mesh unit the more accurate curves and irregularities are presented). However, the smaller the mesh unit size, the more computational capacity is needed when the optimization process is performed. The model utilizes the method of Greatest Common Divisor (GCD) mentioned in literature to determine the smallest applicable unit size. The GCD is the largest integer that divides without remainder all facilities areas.

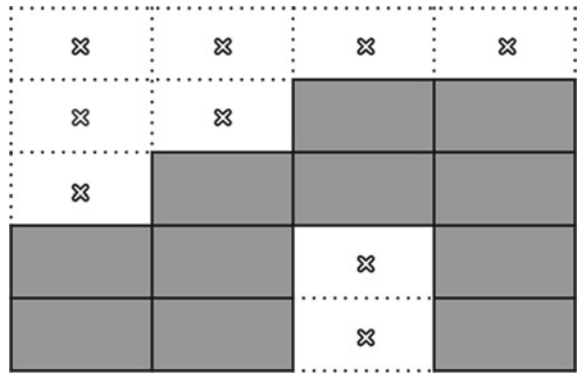
### 3.2.2 Irregular Shapes Definition

In formulating the logic of the model, one of the main challenges that surfaced was to accurately model the complexities and irregularities of site facilities shapes. This model developed a unique and a simplified approach to define irregular shapes. Each shape is defined as rows of mesh units and the shape is referenced to the upper left corner of the shape location. Each row is defined through several parameters: The shift of the row starts from the above row's start, number of gaps in the row, first, second ... etc. part width (depending on the number of gaps) and width of each gap (depending on the number of gaps). To demonstrate the approach, take a look at the shape defined in Fig. 1. This shape is defined using the proposed model in Table 1.

With this simple shape definition technique, any shape can be defined easily with the four orientations. A simple macro on AutoCAD or any similar software can automatically transform a facility or building drawing into such a definition by counting row by row the mesh units (blocks) that the facility's area is occupying and exporting such information on excel sheet to be the data entry of the model. The scope of developing such a macro is, however, not included in this paper. With the



**Fig. 1** Irregular shape example



**Table 1** Shape definition example of shape in Fig. 1

Shape row number <sup>a</sup>	Shift from above row (units)	No. gaps (units)	1st part width (units)	1st gap width (units)	2nd part width (units)	2nd gap width (units)
1	0	0	2	0	0	0
2	− 1	0	3	0	0	0
3	− 1	1	2	1	1	0
4	0	1	2	1	1	0

<sup>a</sup> Table presenting an irregular shape definition example

shapes being defined, irregular shapes locations can be plotted with determining the index (row No. and column No.) of the upper left corner of each shape.

3.2.3 Distance Calculations

Distance between facilities is calculated as the straight line distance between the Center of Gravity CoG of the irregular shapes (1). Hence the shapes take varying geometries, a method aligning with the shape definition technique is chosen to allocate the centroid of each shape.

Distance Equation

$$d_{ij} = \sqrt{(y_i - y_j)^2 + (X_i - X_j)^2} \tag{1}$$

in which

$d_{ij}$  = travel distance between facilities  $i$  and  $j$   
 $y_i$  and  $y_j$  = the  $y$  coordinate with respect to the total site mesh for facilities  $i$  and  $j$  respectively

$X_i$  and  $X_j$  = the  $y$  coordinate with respect to the total site mesh for facilities  $i$  and  $j$  respectively.

In order to accurately calculate the center of gravity for each of the facilities, the calculations were done in two phases, the first phase was to calculate the center of gravity of each scanned row that forms the facility individually with respect to the total site layout mesh upper left corner Eq. (2) then the same process was done in the vertical direction on all columns forming the shape Eq. (3). The second phase was to calculate the CoG of the entire shape using the sum product of all the subareas multiplied by their respective CoGs, and the sum product is divided by the summation of the areas Eq. (4).

C.o.G (X) Direction Equation for part  $n$  of shape  $i$

$$\text{C.o.G}(x) = \frac{\text{Part width}}{2} + \text{column index} \quad (2)$$

C.o.G (Y) Direction Equation for part  $n$  of shape  $i$

$$\text{C.o.G}(y) = \frac{1}{2} \text{Unit} + \text{Row index} \quad (3)$$

Shape C.o.G Equation

$$\text{C.o.G}(\underline{X_i}) = \frac{\sum_{n=1}^i \underline{X_i} \cdot A_i}{\sum_{n=1}^i A_i} \quad (4)$$

in which

$\text{C.o.G}(\underline{X_i})$  = the center of gravity of facility  $i$

$A_{i,n}$  = area of part  $n$  in facility  $i$ .

### 3.2.4 Closeness Matrix

A proximity matrix is introduced such that all the relations between the facilities and each other are assigned weights depending on how far or near the facilities need to be which was adopted from the literature [2]. This was done based on the weights scale shown in Table 2.

## 3.3 Model Formation

### 3.3.1 Problem Definition and Variables

This model optimizes the site layout with a dynamic nature according to different project milestones. In each milestone the fixed facilities are defined as Available or

**Table 2** Proposed closeness relationship weights (adopted from Elbeltagi et al. [2])

Desired closeness relationship	Weight
Necessary to be close	1000
Better to be close	100
May be close	10
Indifferent	0
May be apart	− 10
Better to be apart	− 100
Necessary to be apart	− 1000

Not available indicating whether the facility is built yet or not, and hence should be included in the site plan optimization or not. Temporary facilities also differ according to the milestone and are also considered in the optimization if identified to be needed in the site layout for the particular milestone planned. The model operates by generating possible solutions for the variables, where valid solutions are recorded and enhanced upon. There are two sets of variables in site layout planning. The first is related to the location index of the temporary facility (row No. and column No.) which can take any integer value between 0 and the total number of rows and columns forming the mesh. The second set is related to the orientation of the facility which can take a value of {0, 1, 2, 3} for the four possible orientations.

3.3.2 Objective Function

Site layout plan targets utilizing the site area properly to minimize in-situ travel time and cost while maintaining safety regulations. In the majority of literature, this target is achieved by minimizing the proximity score defined as the multiplication of the facilities distance matrix by the proximity relations matrix (1).

Proximity Score Equation

$$\text{Min} \sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij} R_{ij}$$

(5)

in which

- $n$  = total number of temporary facilities (TFs)
- $d_{ij}$  = travel distance between facilities  $i$  and  $j$
- $R_{ij}$  = a relative proximity weight reflecting the required closeness between facilities  $i$  and  $j$ .

3.3.3 Layout Mapping and Constraints

In site layout planning, there are two main constraints that control whether a solution is valid or not; all facilities lay within the site boundaries and that no overlapping exists between facilities (no facility is placed above another). To overcome the complexity of defining such constraints, macros were developed to ease the representation of facilities on the mesh.

Macros are created to graphically represent the site layout. First, one of the developed macros creates the mesh and assigns a value of 1 to all the cells forming the site perimeter. After which another macro is used to plot the fixed facilities, roads, and safety zones available at the chosen time of site layout planning. This is done by incrementing the values of the location cells by 1. A third macro is used with each Evolver trial to test the fitness of each solution and the validity of the constraints. The Temporary facility mapping macro plots the temporary facilities by incrementing the cells of the location of each facility by the value of 10. This results in having the site mesh with the possible values for listed in Table 3. Coding the cells in such a way eases the definition of constraints of site boundaries and overlapping (6) and (7). Figure 2 showcases invalid values representation in site layout that do not meet the required constraints.

Facilities within Site Boundaries Constraint

$$\sum \text{cells with count of } 10 = 0 \tag{6}$$

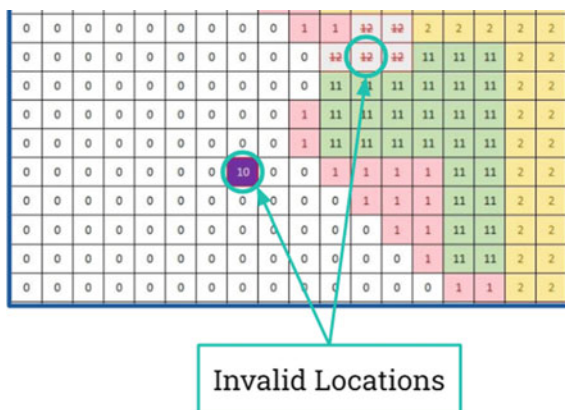
No overlapping facilities Constraint Equation

$$\sum \text{cells with value } \geq 12 = 0 \tag{7}$$

**Table 3** Site layout cell possible values

Cell value	Indication
0	Cells outside the layout of the site
1	Cells within the layout of the site but unoccupied by any facility
2	Cell within the layout of the site occupied by a permanent facility
10	Cell outside the layout of the site occupied by a temporary facility
11	Cell within layout of the site and occupied by a temporary facility that is not clashing ( <b>valid location</b> )
+ 12	Cell within the layout of the site occupied by a temporary facility that overlaps with one or more other facility/ies (temporary or permanent)

**Fig. 2** Invalid locations for temporary facilities



## 4 Case Study

### 4.1 Case Study Definition

The case study illustrated here is the same one applied in the paper “Dynamic Layout of Construction Temporary facilities Considering Safety” for the purpose of having a reference and since the model is continuation of the literature work done in that paper [2]. The project is “Tanta University Educational Hospital” with a footprint area of 28,500 m<sup>2</sup> and a scope of three multistory buildings. There are 8 permanent facilities and 18 temporary facilities within different milestones of the project’s execution.

### 4.2 Closeness Matrix

For the case study to be valid, the same closeness (proximity) relationships matrix should be applied on both cases. Looking at the reference of the model in Elbeltagi’s work, some of the relationships between facilities were clearly identified. However, some relationships were not mentioned. Hence the missing relationships were selected by industry professionals and incorporated with the known relationships in forming the proximity matrix for the case study (Fig. 3). Elbeltagi’s final optimized locations of facilities were tested against the developed proximity matrix to test comparability. The optimized layout score of Elbeltagi’s model was in the magnitude of 540,639, while the score applying our proximity matrix was in the magnitude of 764,568. This was considered sufficient similarity of magnitude to adapt the proximity matrix in the case study comparison between the models.

Facility Name	Proximity Matrix																										
	B1	B2	B3	R1	R2	R3	GH1	GH2	BP	LD1	LD2	CW	LRA	OFF	SSL	CS	WH	P	RFY	WC	WSH	SOFF	FA	MR	T	SL	
Building 1	B1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Building 2	B2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Building 3	B3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Road 1	R1	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	1000	0	0	0	0	
Road 2	R2	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	
Road 3	R3	0	0	0	0	0	0	0	1000	1000	0	1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Guard house 1	GH1	0	0	0	0	0	0	0	-1000	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	
Guard house 2	GH2	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	1000	0	0	0	0	
Batch plant	BP	0	0	0	0	0	1000	-1000	0	0	0	1000	0	-1000	0	0	0	0	0	0	0	-1000	100	0	0	100	
Laydown area 1	LD1	0	0	0	0	1000	1000	0	0	0	0	0	0	-100	0	0	0	0	1000	0	0	-100	100	0	0	0	
Laydown area 2	LD2	0	0	0	0	0	0	0	0	0	0	0	0	-100	0	0	1000	0	0	0	0	-100	100	0	0	0	
Cement warehouse	CW	0	0	0	0	0	1000	0	0	1000	0	0	0	-1000	0	-100	0	0	1000	0	0	-1000	0	0	0	0	
Labors rest area	LRA	0	0	0	0	0	0	0	0	0	0	0	0	-100	-100	0	-100	0	1000	0	-100	0	-100	0	0	0	
Offices	OFF	0	0	0	0	1000	0	0	1000	-1000	-100	-100	-1000	-100	0	0	-100	1000	-1000	0	-1000	1000	100	0	1000	0	
Scaffold storage yard	SSL	0	0	0	0	0	0	0	0	0	0	0	0	-100	0	0	100	0	0	0	0	0	0	0	0	0	
Carpentry shop	CS	0	0	0	0	0	0	0	0	0	0	0	-100	0	0	0	1000	0	0	0	0	0	1000	0	0	0	
Warehouses	WH	0	0	0	0	1000	0	1000	0	0	0	1000	0	-100	100	1000	0	-100	1000	0	-100	0	0	0	0	0	
Parking	P	0	0	0	0	0	0	0	0	0	0	0	-100	1000	0	0	-100	0	0	0	0	1000	0	0	0	0	
Rebar fabrication yard	RFY	0	0	0	0	0	0	0	0	1000	0	1000	0	-1000	0	0	1000	0	0	0	1000	-1000	1000	0	0	0	
Toilet on site	WC	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	-100	0	0	
Welding shops	WSH	0	0	0	0	0	0	0	0	0	0	0	0	-1000	0	0	0	0	1000	0	-1000	1000	0	0	0	0	
SubContractor Site office	SOFF	0	0	0	0	1000	0	0	1000	-1000	-100	-100	-1000	-100	1000	0	-100	1000	-1000	0	-1000	0	0	0	1000	0	
First aid	FA	0	0	0	0	0	0	0	100	100	100	0	0	100	0	1000	0	0	1000	-100	1000	0	0	0	0	0	
Machine room	MR	0	0	0	0	0	0	0	0	0	0	0	0	-100	0	0	0	0	0	0	0	0	0	0	100	0	
Tank	T	0	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	1000	0	100	0	0	
Sampling lab	SL	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig. 3 Developed proximity matrix

4.3 Results Analysis

A number of trials were carried out with the aim of minimizing the objective function of the model, with the runtime used as a stoppage criteria for the algorithm. Additionally, the solution that was obtained from the reference paper was also fed to the model in order to measure the fitness of the proposed solution for comparison purposes. Table 4 shows a summary of the optimization results obtained from the final run as well as the reference paper solution compared to the model’s solution in Fig. 4. There is a substantial improvement in the model’s results compared to that of literature. The model’s score reached almost 10% of the literature value indicating a major change. The model runtime is also improved; the explanation to that is the simplicity of shape definition in the developed model which reduced the computational time significantly.

Table 4 Comparison between reference model and model developed

Comparison aspect	Developed model	Literature result (Elbeltagi)
Constraints achieved	Yes	Yes
Fitness score	70,765	764,568

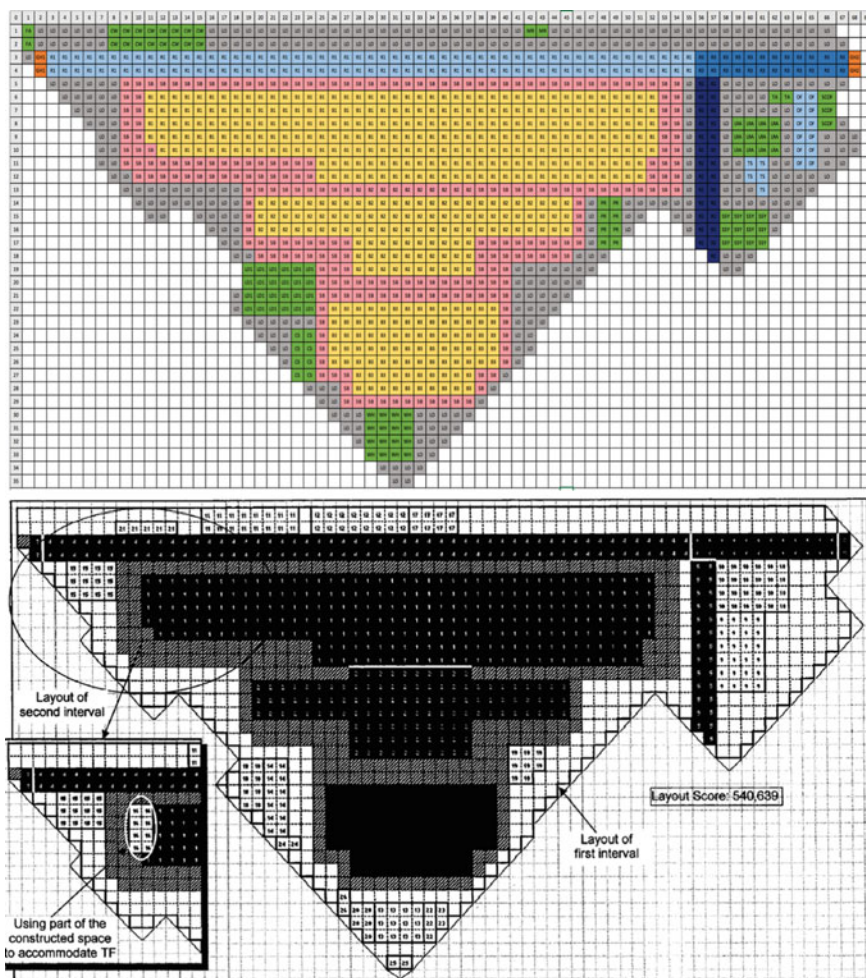


Fig. 4 Developed model's optimized solution (top), Elbeltagi's optimized solution (bottom)

## 5 Conclusion

From all of what was discussed, it can be concluded with confidence that the developed model is successful in optimizing site layout and shows improvement over that of the reference model. The shape definition technique developed in the paper is simple and effective in capturing the irregularities of facilities and can be modified in its accuracy based on the mesh unit size. Safety elements were considered in the planning of site layout through constructing safety zones and in the consideration of the proximity matrix. However, there is still room for improvement in the proposed model, the model is far from reaching its full potential. There is room to improve the

model's flexibility and ease of application. Linking the model to AutoCAD or other drawing software and developing a code to analyze drawings to detect the shape definitions parameters is considered the most useful expansion for the model.

## References

1. Abotaleb I, Nassar K, Hosny O (2016) Layout optimization of construction site facilities with dynamic freeform geometric representations. *Autom Constr* 66:15–28
2. Elbeltagi E et al (2004) Dynamic layout of construction temporary facilities considering safety. *J Constr Eng Manag* 130(4):534–541. [https://doi.org/10.1061/\(asce\)0733-9364\(2004\)130:4\(534\)](https://doi.org/10.1061/(asce)0733-9364(2004)130:4(534))
3. Farmakis PM (2018) Genetic algorithm optimization for dynamic construction site layout planning. *Organ Technol Manag Constr Int J* 10(1):1655–1664
4. Xu J, Liu Q, Lei X (2016) A fuzzy multi-objective model and application for the discrete dynamic temporary facilities location planning problem. *J Civ Eng Manag* 22(3):357–372



# Gordie Howe International Bridge Managing a Pile of Paper on a Paperless Project



Randy Pickle

**Abstract** The Gordie Howe International Bridge (GHIB) construction actively commenced for all components in Summer, 2019. On all projects and especially large one's like GHIB, there needs to be a place for all documents and that place needs to be managed. This is typically accomplished by an Electronic Document Management System (EDMS). In 2021, Physical Filing Cabinets and Storage Rooms have been replaced with Digital Folders on Cloud Based Servers. This 'data in the sky' is by far more secure than any physical filing of yesteryear. On GHIB, there are the typical documents, Design Drawings, Specifications, RFI's, Working Drawings, Record Drawings, etc. Bridging North America (BNA) has utilized the Autodesk BIM360 platform to manage notifications and documents. BNA and AECOM have collaborated and developed workflows to process and manage each type of document. Interactive forms were developed by AECOM to facilitate documentation flows. As with any project there is the need to find the right solution. The digital world is constantly changing, and adaptation is ongoing. During the construction phase of any project, there are expectations of the need to process contractor's questions, review of third-party drawings for the undertaking of the works, procedures, and other plans. The magnitude and complexity of the scope of the GHIB project generates a significant volume of construction related submissions to be processed by the respective Engineers of Record. There are two genres of submittals: Field Clarification Requests and Working Drawings. The Working Drawing submittals are further subdivided into multiple sub-categories. The AECOM Construction Phase Services Leads are the first point of contact for submissions and provide triage for determining the needed resources from the applicable discipline Engineer of Record. AECOM utilizes Bentley's ProjectWise to provide the repository for submissions to be processed and then archived. The volume of submissions increased in direct proportion to the extent of works being undertaken on the various components. Confirmation that the correct process has been undertaken is certified by the Construction Phase Services Design Quality Manager.

---

R. Pickle (✉)  
AECOM, Ontario, Canada  
e-mail: [randy.pickle@aecom.com](mailto:randy.pickle@aecom.com)

**Keywords** Gordie Howe International Bridge · Paperless project

## **1 Introduction**

### ***1.1 The Project***

The Gordie Howe International Bridge Project consists of four components—Canadian Port of Entry, U.S. Port of Entry, a new interchange on I-75 in Detroit and the Bridge itself. Each component is a major heavy construction undertaking. The Owner of the completed works will be the Windsor–Detroit Bridge Authority (WDBS), a Canadian Crown Corporation. The project is being delivered through a Public–Private Partnership Agreement (P3), by the Design-Build Joint Venture (DBJV), Bridging North America. The design phase of the works commenced in July, 2018, and site preparation works by the DBJV began on both sides of the Detroit River in January, 2019.

### ***1.2 The Components***

#### **1.2.1 Canadian Port of Entry**

The Canadian Port of Entry (CA POE) is a 53 ha (130 acres) campus that will include the Canadian inbound border inspection facilities for both passenger and commercial vehicles, Canadian outbound inspection facilities, tolling operation for both the US-bound and Canada-bound traffic, and maintenance facilities. Once constructed, this port will be the largest Canadian port along the Canada–US border and one of the largest anywhere in North America. It includes eight buildings, four canopies, US turn-around bridge, extensive site works and circulation, utilities, and security. The CA POE connects directly to the western terminus of the MacDonald–Cartier Freeway, Hwy 401, in Windsor.

#### **1.2.2 U.S. Port of Entry**

The US Port of Entry (US POE) is a 60 ha (145 acres) campus that will include the US inbound border inspection facilities for both passenger and commercial vehicles, US outbound inspection facilities, and commercial exit control booths. It includes six buildings, eight canopies, an enclosed pedestrian bridge, parking, circulation roadways and other site works.

### **1.2.3 I-75 Interchange**

The I-75 Interchange will provide the primary connecting ramps to and from the US Port of Entry to I-75 through Detroit and associated local road improvements required to fit the new ramps into the interstate system. It includes four flyover ramps, four new crossing road bridges, five new pedestrian bridges, two bridges crossing the railway and connecting I-75 to the US Port of Entry, and relocation/reconstruction of service roads and other local road improvements.

### **1.2.4 Bridge**

The Bridge—the project includes a six-lane cable stayed (CS) bridge, providing three Canada-bound lanes and three US-bound lanes over the Detroit River. The bridge will have a clear span of at least 853 m (2798 ft.) across the Detroit River with no piers in the water. Two approach bridges (one on each side) will connect the main bridge to the Canadian Port of Entry and the US Port of Entry. The crossing (bridge and approaches) will be approximately 2.5 km (1.5 miles) in length. This CS bridge will be the 6th longest in the world, the longest in North America, and it is the longest composite CS in the world, with concrete towers of 218 m (715') height, out of the river and drilled shaft foundations. The approach bridges will be all concrete design (11 spans on US side, 10 spans on CA side) with precast concrete girders and driven pile foundations.

## **2 Design Services During Construction**

### ***2.1 Design Services Agreement***

As per agreement with the DBJV, the Engineers of Record are providing services during the construction phase responding to Field Clarification Requests, review of working drawings and the review of materials to be incorporated into the works, either temporarily or permanently.

### ***2.2 Field Clarification Requests***

Design documents, including drawings and specifications, sometimes have unintended discrepancies, conflicting information, or missing information. In other cases, the Construction teams may desire a change for convenience, necessity, or process improvement. In such cases, it is necessary that missing information be obtained, discrepancies/conflicts be resolved, or the request for change be approved by the

Engineer of Record before construction proceeds on the affected activities. The FCR is the standard method for dealing with the above situations. The FCR provides the interface between BNA Construction Management and the respective Design Team, to address questions and minor requests arising during construction. A FCR is not intended to be used in lieu of a design change during construction unless it is determined that a Design Change Notice (DCN) is required.

Field Clarification Requests (FCR) are processed prior to, or after, the Released for Construction (RFC) Documents have been distributed within the DBJV organization for the construction of the Gordie Howe International Bridge Project (GHIB) works.

## **2.3 Working Drawings**

Working Drawings on the GHIB Project have different classifications and are expected to be classified by the DBJV correctly when they are transmitted to the Engineers of Record to insure a proper review is undertaken and the response is provided in a timely fashion. Work applicable to a working drawing cannot commence until the review is completed to the satisfaction of the EOR and ultimately the Owner. Working Drawings (WD) include Shop Drawings (SD) that require detailed review and response while other WDs may be for information only and may not require a response.

The defined classifications of Working Drawings for GHIB are as shown in Table 1.

It has been agreed that any shop drawings, as per Item No. 1 in Table 1, based entirely on the relevant signed and sealed design drawings do not require the seal of the shop drawing preparer. The exception to this is when engineered information is added to the drawing by the shop drawing preparer.

Working drawing review by the Engineer of Record is for the sole purpose of ascertaining conformance with the Project Agreement (PA) requirements, plans, and specifications and as a precaution against oversight or error. This review shall not mean that the Engineer of Record approves the detail design inherent in the working drawings, responsibility for which shall remain with the DBJV submitting same, and such review shall not relieve the DBJV of its responsibility for errors or omissions in the working drawings or of its responsibility for meeting all requirements of the Contract Documents.

The DBJV is responsible for dimensions to be confirmed and correlated at the job site, for information that pertains solely to fabrication processes or to techniques of construction and installation, and for coordination of the work of all subcontractors. The DBJV shall log, check for completeness, stamp, and sign and make notations it considers necessary on working drawings before each submission to the Engineer of Record for review.

**Table 1** Working drawing classification

Item	Working drawings	Review required	Information only
1	Shop drawings, including specially prepared technical data for this project, calculations, reports, and similar information not in standard printed form for general application to a range of similar projects	Y	N
2	Product data including manufacturer's standard catalogues, pamphlets and other printed materials that show and describe materials and items, including product specifications; installation instructions; colour charts; catalogue cuts; rough-in diagrams and templates; wiring diagrams; performance curves; operational range diagrams; and mill reports	Y	N
3	Samples including both fabricated and non-fabricated physical examples of materials, products and units of work; both as complete units and as smaller portions of units of work; normally used for limited visual inspection	Y	N
4	Certificates of conformance or compliance, including documents attesting that a product complies with a specified standard	Y	N
5	Certified test (or inspection) reports: documents attesting that a product meets a specified level of performance or quality when a prototype specimen is tested or inspected in accordance with a specified procedure, consisting of a certified statement by the product supplier or contractor accompanied by a complete report of the inspection or test	Y	N
6	Submittals related directly to the work (non-administrative) including warranties, maintenance agreements, workmanship bonds, project photographs, survey data and reports, copies of industry standards, field measurement data, manufacturer's installation instructions, operating and maintenance materials, spare parts/data, product data samples, certificates of conformance or compliance, or certified test reports, and similar information, devices and materials applicable to the project work and not processed as shop drawings	TBD	N
7	Erection plans	N	Y
8	Fabrication quality control and quality assurance plans	N	Y
9	Fabrication plans	N	Y
10	Stress sheets	N	Y
11	Lift plans	N	Y
12	Bending diagrams for reinforcing steel	N	Y
13	Falsework plans	N	Y
14	Similar data required for the successful completion of the DB work	N	Y

The DBJV shall package and submit the working drawings to the Engineer of Record in a well-organized manner that facilitates efficient review by the Engineer of Record. Where possible all submissions must be electronic with the exception of physical samples. The working drawings shall show, but not necessarily be limited to the following:

- Clear and obvious notes of any proposed changes from drawings and specifications.
- Fabrication and erection dimensions.
- Provisions for allowable construction tolerances and deflections provided for live loading.
- Details to indicate construction arrangements of the parts and their connections, and interconnections with other work.
- Location and type of anchors, and exposed fastenings.
- Materials and finishes.
- Descriptive names of equipment.
- Mechanical and electrical characteristics when applicable.
- Information to verify that superimposed loads will not affect function, appearance, and safety of the work detailed as well as of interconnected work.
- Assumed design loadings, and dimensions and material specifications for load bearing members.
- Dimensions and dimensioned locations of proposed chases, sleeves, cuts, and holes in structural members.
- Certify working drawing submittal review as per Quality Management Plan.

The Reviewers responsibilities include the following:

1. The Reviewers should review submitted working drawings to the necessary degree to assure themselves of consistency with the intent of the Project Agreement (PA) requirements, plans and specifications.
2. Working drawings reviews are for general conformity only and do not include detailed checking of dimensions or extensive checking of calculation.
3. Designing through the working drawings review process is to be avoided. The working drawings review should not be used as a conduit for the DBJV/ Subcontractor to suggest changes to design. Design Changes should be requested.
4. Agreed to Issued DCN/IFC documents updated before receiving working drawings for review that incorporates design changes.
5. Working drawings with major deficiencies should be rejected with the major deficiencies noted. Reviewers should not continue with a full review of the working drawings. Notwithstanding the working drawings with major deficiencies, multiple resubmittals are to be avoided with working drawings being returned stamped 'As Noted' instead of 'Revise and Resubmit' whenever possible.
6. Working drawings provided to the Engineer of Record without completed DBJV sign off shall be rejected without further comments and counted as one review iteration.

7. Working drawings that require extensive correction or are in substantial disagreement with the intent of the contract documents will be rejected without further comments.
8. Working drawings for items required to be sealed by a Professional Engineer (or as otherwise indicated as engineered) that are not sealed by a Professional Engineer before submission to the Engineer of Record for its review will be rejected without further comments.

## ***2.4 Material Acceptance Requests***

With reference to Table 1, Item's 2–5, inclusive, are **Material Acceptance Requests**. Product data and applicable product samples are reviewed by the Engineer of Record for the sole purpose of ascertaining that the material meets the technical and performance requirements in accordance with the project requirements and the signed and sealed plans and specifications.

Review and acceptance of the proposed material, by the EOR, is not approval of the material for use in the project but a recommendation through the DBJV to Owner approval of the materials being incorporated into the Project. Review by the EOR and approval by the Owner also includes materials from suppliers on the Ministry of Transportation Ontario Designated Sources of Materials list and the Michigan Department of Transportation Qualified Products List, as relevant.

DBJV shall submit to the EOR all relevant information to allow the EOR to be able to ensure that the material will meet the specified technical and performance requirements. DBJV shall submit all necessary information, such as, but not limited to:

- Site application
- Product specifications
- Installation instructions
- Colour charts
- Catalogue cuts
- Performance curves
- Mill reports
- Certificates of conformance or compliance
- Certified test or inspection reports
- Samples
- Fabrication quality control and quality assurance plans.

Material data and information submitted that does not adequately satisfy the Reviewer, whether it meets the technical and performance requirements of the Project Agreement (PA) requirements, plans and/or specifications, may cause the material to be rejected.

## **3 Document Management Systems**

### ***3.1 Contract Requirements***

Contractually, the DBJV is required to prepare and implement a Document Management Plan for the management of the flow of documents during all phases of the project, including creation, authentication and versioning of electronic and hard copy documentation and samples and models.

The need for a Document Management Plan is also paramount in AECOM's control of submissions to the EOR's. The process for document filing is detailed in the AECOM Document Management Plan. Document Control is required to keep a copy of all pertinent files that originate from AECOM staff and subconsultants or are received from an external party. These files can be either a hardcopy (Document Control filing cabinet/binders) or electronic (on SharePoint and ProjectWise), as noted in Table 2. Maintenance of parallel hardcopies of electronic documents is not required. Hardcopies may be printed if required but are supplemental and do not serve as part of the master project quality record.

### ***3.2 ProjectWise Site***

ProjectWise (PW) is being used for storage, collaboration and management of CPS files and submittals. A site dedicated to GHIB exists for this purpose. Construction Support, for the processing of FCR's, working drawings and MAR's is in the 600 Series folders.

### ***3.3 BIM360***

The DBJV utilizes the Autodesk BIM360 platform to manage notifications and documents. All communication with DBJV is through their BIM360 Application.

## **4 Forms**

### ***4.1 FCR's***

To facilitate and standardize the submission of and response to an FCR, a fillable form, in .pdf format, was crafted to include relevant, basic information in order to put the FCR into the proper context within project scope, the ask and the response.



**Table 2** Document storage

Document type	Location		Comments
	CPS subsite SharePoint	CPS folder ProjectWise	
<i>Correspondence</i>			
Internal	✓		<ul style="list-style-type: none"><li>• In respective subfolders by document type, i.e. letter, memo (includes memos to file and photos), attachments, transmittals</li><li>• Access available to Project Team only</li></ul>
External	✓		
<i>Reports</i>			
Daily site	✓		<ul style="list-style-type: none"><li>• In respective subfolders by component</li><li>• Access available to all</li></ul>
Certification		✓	<ul style="list-style-type: none"><li>• In respective subfolders by component and submission type</li><li>• Access available to Project Team only</li></ul>
Commissioning		✓	
Other		✓	
<i>Submissions for EOR review</i>			
Field clarification requests		✓	<ul style="list-style-type: none"><li>• In respective subfolders by component and submission type</li><li>• Access available to Project Team only</li></ul>
Material acceptance requests		✓	
Shop drawings		✓	
Working drawings		✓	
Other		✓	
<i>Drawings</i>			
Released for construction		✓	<ul style="list-style-type: none"><li>• In respective subfolders by component</li><li>• Including revisions</li><li>• Access available to Project Team only</li></ul>
As-recorded/ as-built		✓	<ul style="list-style-type: none"><li>• In respective subfolders by component</li><li>• Access available to Project Team only</li></ul>

The relevant information includes:

- Relevant dates: Originating, Needed by
- FCR number
- Title
- Component Segment: 1—Canadian POE; 2—Canadian Bridge Spans; 3—US Bridge Spans; 4—US POE; 5—I75 Interchange; 6—Corridor Services
- Discipline: Civil, Mechanical, Electrical, Architectural, etc.
- Reference documents: Project Agreement; (Design) Drawing, Spec, Working Drawing, Other

- Origin: DBJV Owner and Owning Component.

This information provides the CPS Team with the ‘where’ and ‘what’ behind the ‘ask’, or issue to be resolved.

In consultation with the DBJV Owner, all FCR’s are categorized as being **INTERNAL**, **MINOR**, or **MAJOR**.

An FCR is **INTERNAL** when a clarification or minor modification can be provided with minimal or no changes to the Release for Construction (RFC) drawings or specifications. The modification is recorded for traceability purposes and later incorporated into the Record Drawings, if applicable. There is no commitment to share these issues and resolutions with the Project owner.

An FCR is **MINOR** when a change or modification to the RFC drawings or specifications requires calculations, complex sketches, or redline drawings to define the modification. These changes or modifications are required to be shown on the Record Drawings. These issues and resolutions are submitted to the Project Owner for their information.

An FCR is **MAJOR** when a major change or modification requiring the issuance of a new revision to the RFC drawings or specifications. The answer to this category of FCR either requires significant calculations or requires changes that cannot be clearly depicted with only sketches and/or redline. These issues and resolutions are submitted to the Project Owner for their information as well as notifying the Project Owner that a revised drawing(s) will be submitted.

The remainder of the form provides for the response from the Engineer of Record, including noting any necessary attachments, identifying who is providing the response and verification of the response by the EOR, including relevant dates. The response is entered using a suitable .pdf editor. Where a response may not be able to be fully captured in the space provided, it is possible to craft the response and include it as an attachment.

Finally, the DBJV is asked to provide concurrence of the final response.

The two-way transmission of the FCR form was undertaken utilizing BIM360. As such, the workflow required the Originator to fill in the form, forward to the DBJV Document Control to upload to BIM360 for forwarding to the CPS Document Control. CPS Document Control would then download the form and save it in ProjectWise. CPS Document Control would then advise the CPS Team that an FCR was available for action. The CPS Team would then retrieve the FCR and provide the required response. The process would then be reversed to return the FCR to the DBJV.

Even though this fillable form served the process well initially, the inefficiencies became quite evident as the volume of FCR submissions increased. Using the form, as it was, resulted in the document having to be handled numerous times which led to increased effort and time in the life cycle of the FCR.

As a means to mitigate time and effort associated with the life cycle of an FCR, the DBJV migrated to a fully electronic form based in BIM360 Document Management. The previous form was fully replicated electronically.

To create an FCR, the DBJV responsible staff provided, online, the same basic information as previously, including the ask.

This online version created an active **Assigned to** field. This field identified who the next responsible person would be. Record of the FCR being assigned is captured on the **ACTIVITY** tab of the FCR. It was no longer necessary to complete all of the fields, save the document, and relay via BIM360 or email.

Further, to facilitate the deliverance of the FCR, roles were created. One particular role, relevant to FCR’s, is **@AECOM FCR Coordinator**. This is the gateway for FCR’s to be delivered to the EOR. Attached to the role are the electronic addresses for Document Control, CPS Manager, and the responsible CPS Lead.

Once the CPS Lead has reviewed the FCR, then utilizing the **Assigned to** field and/or the **ACTIVITY** tab, the CPS Lead initiates responding to the FCR as required.

Besides being more efficient in the handling of the FCR document, another significant benefit is that all actions associated with processing the FCR, as well as noting who participated in the process, are captured by the **ACTIVITY** tab and thereby creating a virtual paper trail.

4.2 Working Drawings and MAR’s

The submission of all working drawings and MAR’s, as identified in Table 1, was by traditional means, except that e-files were utilized as opposed to hard copies.

To start the process, DBJV would prepare a submission transmittal as per usual convention. A condition of submission from DBJV was that the submission had undergone quality control, and, in this regard, the following statement appears on the transmittal:

The submittal has been checked and reviewed for general conformance with the design concept of the project and general compliance with the overall design and requirement of the project agreement.

To facilitate the review of the working drawings by CPS, DBJV provides an indication of the type of working drawing that is being submitted as per the following:

Document category codes			
As-built	AB	Operating and maintenance manuals	OM
Certificate of compliance	CC	Personnel certification	PC
Contract drawings	CD	Photographs	PH
Delivery ticket	DT	Procedure qualification record	PQ
Design calculations	DC	Product data	PD
Design drawings	DD	Reports	RP
Equipment lists	EL	Shop drawing	SD
Erection plans	EP	Spare part list	SP

(continued)

(continued)

Document category codes			
Field measurement and survey data	FS	Temporary works (including major)	TW
General quality submittal	GQ	Test reports or test results	TR
Installation manual	IM	Warranties	WR
Material certification	MC	Welding performance qualifications (PQR)	WP
Material or product samples	MS	Welding procedure specification (WPS)	WS
Mix design	MD	Work plan	WP

Upon deliverance of the working drawing submission package, Document Control prepared a fillable Shop Drawing Review Form.

Information transferred from the DBJV transmittal included:

- Submission No.
- Subcontractor/Supplier
- File Name
- Component
- Applicable division of the specifications
- Primary review discipline.

Notice of the submission being available for review is provided to CPS Document Control by email. If necessary, the package is forwarded to the primary EOR/Discipline Reviewer and then subsequently to any other EOR/Discipline Reviewer(s).

The submitted drawings or other documents are redlined with review comments, using a suitable .pdf editor and the drawings are stamped as per the usual review procedures. Other comments can be entered into the review form.

The same concerns with the processing of FCR's also arose in the processing of working drawings. The process involves several uploads and downloads of the submission, overhead to the effort to review. Also, the review of MAR's was found to be a variant to the review of the other genres of working drawings. To mitigate, a multi-use fillable form, with relevant drop-down menus, was created to further the review process upon receipt of the package from DBJV.

The first action, in using the form, is to choose the form type from a drop-down menu: Working Drawing Review Form, Material Acceptance Request—Review Form, by the CPS Document Control.

Identifying the Document Classification, based on the tables on the DBJV transmittal form is accomplished through a drop-down list.

Upon completion of the identification part, of the form, as with FCR's, the CPS team determines the disposition for review. If the CPS team determines that EOR review is required, then CPSDC advises the EOR, by email, in BIM360, including a link to the location of the review package in ProjectWise.

The relevant Discipline Reviewer would undertake the required review and as appropriate redline the documents using a suitable .pdf editor. Additional comments can be entered into the review form.

Upon completion of all relevant discipline reviews, the CPSDC is advised, by email, in BIM360, that the reviewed documents are ready to be returned to DBJV.

#### 4.2.1 Disclaimer

The role of the EOR is to review all MAR submissions for compliance with the contract specifications. The role of the EOR is not to approve the use of any particular material. That is solely the responsibility of the Owner.

In this regard there is a disclaimer drop-down option. Two disclaimers were available:

This MAR has been reviewed with respect to design intent and is technically compliant, subject to Owner approval,

or

This MAR has been reviewed with respect to design intent and is technically compliant.

As noted previously, the MTO Designated Sources of Materials list and the MDoT Qualified Products List provided a quasi-pre-approval of materials to be incorporated into the works. In this regard, any MAR submitted for a material associated with either of these lists was considered to be approved by the Owner and in that case the second disclaimer was identified.

#### 4.2.2 Review Stamp

To indicate the status of the review of any working drawing noted in Table 1, a stamp is affixed to the documents by the reviewer as shown.

As shown a reviewed document may be returned:

- Reviewed—no comment
- Reviewed—revise and resubmit
- Reviewed—as noted
- Review by consultant not required

Initially, this stamp was affixed to the reviewed documents and completed using a suitable .pdf editor.

To mitigate the effort to affix the review stamp to every drawing and fill it in, an e-stamp was created. This stamp is dynamic and fillable. Once completed on the first sheet or page, to be stamped, it can be cut and pasted to subsequent sheets or pages, as required.

A progressive series of drop-down menus allow the reviewer to enter information in all fields. For the review status, the options are available by toggling through the options until the desired status is found and then an 'X' is placed in the data field. Similarly, the other fields of the stamp are filled. Upon completion and application of the stamp, all text is flattened and cannot be altered. The stamp can be locked to ensure that its size and placement on the document cannot be altered.

For reports, procedures, calculations and other similar submissions, the stamp is applied on the first page indicating the review status of the entire document. For fabrication plans, shop drawings and other similar submission, the stamp is applied to every drawing indicating the review status of the drawing.

5 Workflows

5.1 FCR's

The following is the initial FCR process that enabled the DBJV field team to request information and for the EOR to respond in timely fashion, utilizing the fillable form and email (Fig. 1).

Subsequently, with the implementation of the online FCR form, the process became as in Fig. 2.

An active log is maintained of the date of submission of the FCR and duration of time that the FCR has been with the CPS Team to deal with. Interim processing durations are calculated, and a final duration noted when the FCR is returned to the client.

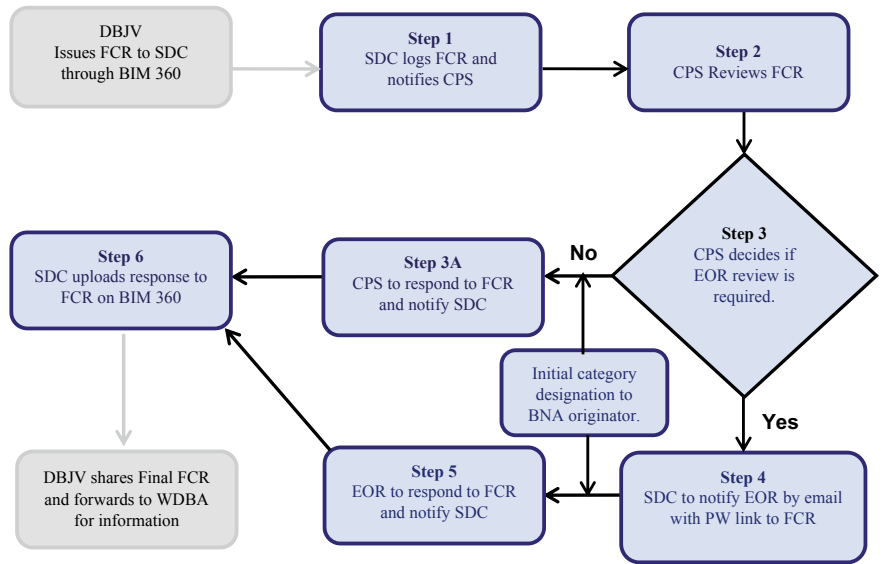


Fig. 1 FCR workflow

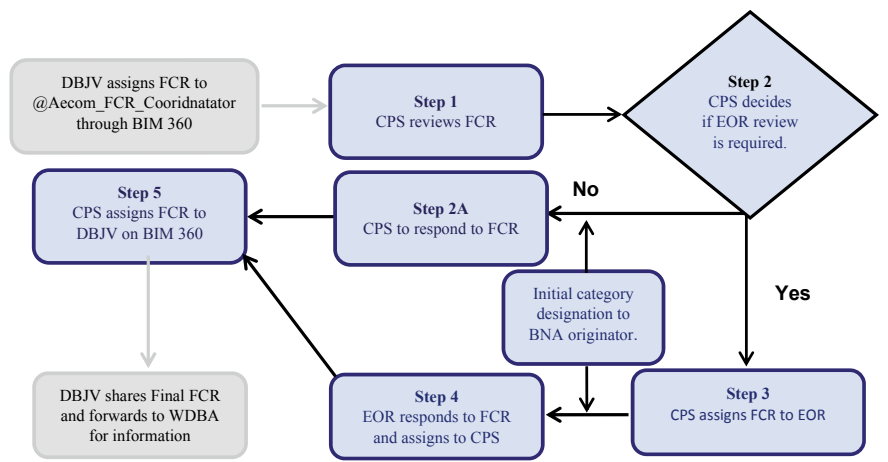


Fig. 2 FCR workflow—online form

5.2 Working Drawings

The working drawing process enables the DBJV field team to submit and for the EOR to respond in timely fashion (Fig. 3).

An active log is maintained of the date of submission of the working drawing(s) and duration of time that the working drawing(s) has been with the CPS Team to deal with. Interim processing durations are calculated and a final duration noted when the working drawing(s) is returned to the client.

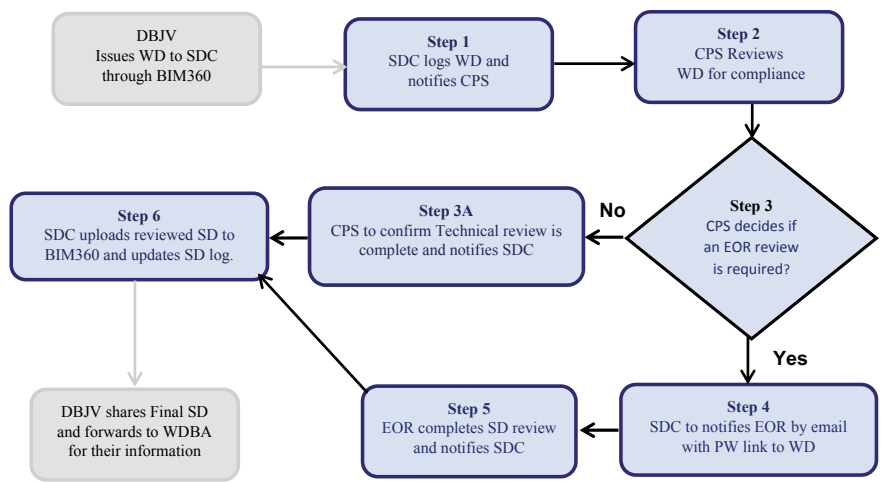


Fig. 3 Working drawing workflow

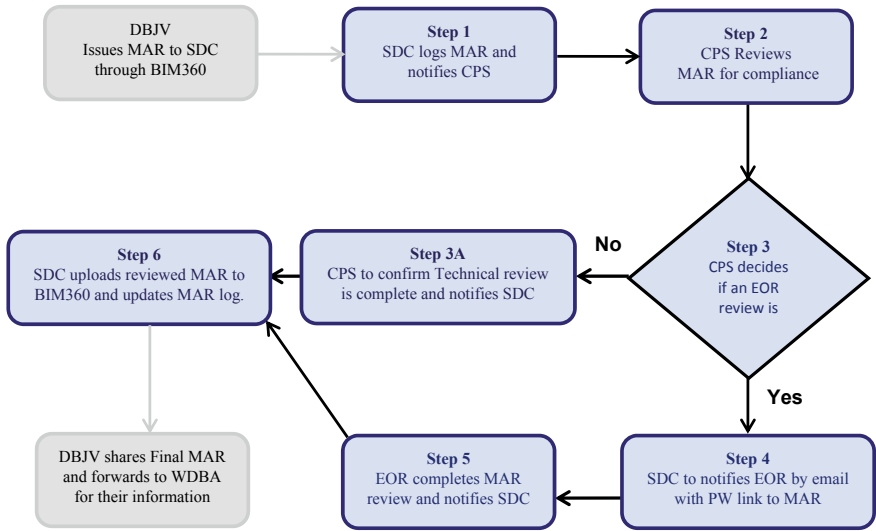


Fig. 4 MAR workflow

5.3 MARs

The working drawing process enables the DBJV field team to submit and for the EOR to respond in timely fashion (Fig. 4).

An active log is maintained of the date of submission of MAR’s and duration of time that a MAR has been with the CPS Team to deal with. Interim processing durations are calculated, and a final duration noted when the MAR is returned to the client.

6 System Evolutions

In the course of managing the expected submissions from a contractor on a project, the use of fillable forms and the multiple handing of documents, as experienced in the onset of GHIB, would be effective, efficient and economical. The volume of FCR, working drawing and MAR submission traffic increased exponentially as the subcomponents of the project ramped up. For the volumes of submissions being experienced, time management and expenditure, of the CPS team, including the EOR’s, became issues. Significant volumes of submissions, daily, resulted in unacceptable backlogs. It became necessary to develop mitigating measures to handle the submissions targeting the non-technical review effort being expended in handling submission be it ensuring compliance to established submissions processes for the project, saving files upon receipt from DBJV and subsequently from CPS and undertaking reviews.



As discussed, means to facilitate and mitigate the ‘paperwork’ associated with submissions but incidental to the technical reviews were sought and developed to meet the needs of the five major components of the GHIB project.

# Low Regret-Based Design and Corrosion Management for Steel Roadway Bridges



M. Barkhori, S. Walbridge, and M. Pandey

**Abstract** This paper describes a low regret-based adaptive decision-making methodology for evaluating alternative design and corrosion management strategies for steel roadway bridges despite limited site-specific knowledge of the actual corrosiveness of the environment. To illustrate the method, an example is provided, in which design options include whether using metallizing or not as a corrosion protection measure and its time of application. While in a mild environment, unpainted weathering steel might show little degradation, in case of realization of a severe environment, for example, due to extensive use of de-icing salt during winter, using this material without protection can be problematic as doing so might result in higher management costs or even catastrophic failures. On the contrary, the realization of a mild or moderate environment after utilization of the expensive galvanized/metallized steel or early adaptation of preventive measures can also be viewed as suboptimal. Another strategy is resorting to adaptive solutions, beginning with a less expensive option until the actual environment is more evident. Willing to minimize the “maximum sense of loss”, in the methodology applied here, regret is combined with the real option methodology and implemented in the decision-making framework. In this way, the methodology quantifies the desire of decision-makers to minimize the sense of loss associated with having made the wrong decision.

**Keywords** Low-regret based design · Corrosion management · Steel roadway bridges

---

M. Barkhori (✉) · S. Walbridge · M. Pandey  
Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada  
e-mail: [mbarkhor@uwaterloo.ca](mailto:mbarkhor@uwaterloo.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_27](https://doi.org/10.1007/978-3-031-34593-7_27)

421

## 1 Introduction

Composite concrete and steel slab-on-girder systems are commonly used in bridges throughout North America. In recent years, some bridge owners have become aware of an unexpectedly large amount of corrosion degradation of some of their weathering steel highway structures. Weathering steel contains small amounts of nickel, chromium, and copper; it is available as Type A or Type AT, as designated in CAN/CSA G40.21 [1]. Under repeated cycles of wetting and drying, weathering steel forms a thin, adherent oxide patina in 18–36 months [1], which afterward protects it from further penetration of oxygen and moisture that leads to corrosion. However, it appears that under certain conditions, this patina does not form as previously supposed. In these cases, the patina tends to flake off, sometimes in large pieces, exposing a new steel surface to corrosion [2, 3].

This extensive corrosion is believed to be due to microclimate conditions resulting extensive exposure of the steel element to a combination of moisture from melting snow, road salt, and sulfur dioxide. Typically, the corrosion degradation in weathering steel structures is concentrated at the piers and abutments, where leaky joints permit water contaminated with salt to run onto the weathering steel girders, or at the mid-span, where the draft of traffic passing underneath the overpass splashes the girders with salt-contaminated water. Both cases can be troubling especially for simply supported girder bridges. For these types of bridges, the highest shear and moments forces occur at the supports and mid-span, respectively. While falling debris from extensive corrosion of this type may endanger vehicles passing underneath the structures [2, 3], in this study, only the danger of collapse of the structure is considered.

In a new construction project, the impacts of factors that result in the formation of adverse microclimates such as clearance between the overpass and the road below and air circulation are unknown. However, over time, one can observe the corrosion performance of the bridge and speculate on the actual microclimate. In this situation, resorting to fixed strategies with no consideration of information update may not be the best choice. In these “predict and then design” approaches, while the costs of a rust protection measure are clear and immediate, the benefits are uncertain and less obvious. Considering available corrosion projections in practice will therefore require engineering judgment.

In engineering, to identify investment options with good performance under a range of potential futures, the applications of new decision-making tools such as robust decision-making (RDM) and real options (RO) analysis rather than optimizing against a single projection have been gaining wider acceptance [4, 5]. RO handles uncertainties in infrastructure investments by providing managerial flexibility, where adaptive design options are available to cope with the evolving demand. In this way, RO provides means for deciding the extent and timing of applying managerial options to ensure an expected level of performance over an assessment horizon.

The focus of the study is on providing a decision framework for determination of the optimum strategy for preservation of weathering steel highway structures when the possibility of formation of adverse microclimates is undetermined. Here a strategy is comprised of choosing among a set of engineering options (e.g., whether metalizing or not) and determining an action time for exercising that choice. The developed framework is a combination of low regret decision-making, RDM, and RO. RDM allows for providing a solution that works on a wide range of possible futures instead of preparing for an expectation point, while low regret decision-making makes the integration of RO and RDM meaningful.

In what follows, first, an overview of the earlier-mentioned terms is provided. Later in another section, the framework developed is explained. Afterward, to illustrate the framework, an analysis is performed for finding the optimal corrosion management strategy for a typical weathering steel highway structure and the results are interpreted. Finally, a conclusion is provided, based on the presented work.

## 2 Management Under Uncertainty

The uncertainty about microclimates poses challenges to bridge managers in investing on corrosion protection measures for safeguarding an asset of vulnerable bridges. Deterministic approaches are appropriate for defining optimum management plans for a clear and relatively certain future, which is not the case in this problem with very different and uncertain projections of corrosion processes possible due to the range of microclimate possible. In the face of deep uncertainties, to identify investment options with good performance, the applications of new decision-making tools such as RDM and RO rather than optimizing against a single projection have been gaining wider acceptance in engineering [4, 5].

While optimization for the maximum utility may not necessarily yield a preferable outcome in the worst probable scenarios, a robust decision works satisfactorily under a broad range of scenarios. Robust decisions can be optimized against a range of future scenarios without the need for a full description of the probability of occurrence of these future scenarios. In this context, Minimax and Maximax are among the most widely used non-probabilistic, robust optimization formulations [6]. Minimax formulation plans for the worst possible outcome based on a pessimistic outlook of future and takes the following form:

$$\begin{aligned} \text{Objective: } & \max_{\Phi, T} \left( \min_{\Psi} \omega_{km\tau} \right) \\ \text{Find: } & m^*, \tau^*, \omega^* \\ \text{Given: } & \Phi, T, \Psi, \mathbb{C} \end{aligned} \tag{1}$$

where  $m^* \in \Phi$  is the optimal action among the set of potential actions  $\Phi$ ;  $\tau^*$  is the optimal action time from the time interval  $T$  for employing  $m^*$ ;  $\omega_{mk}$  represents the

payoff of decision  $m$  under scenario  $K \in \psi$ ;  $\Psi$  is the set of future scenarios; and  $\mathbb{C}$  is the set of costs for the potential adaptation actions. The Maximax formulation, however, represents an optimistic view of the future and plans for the best outcome according to the following equation:

$$\begin{aligned} \text{Objective: } m^* &= \max_{\Phi, T} \left( \max_{\Psi} \omega_{km\tau} \right) \\ \text{Find: } m^*, \tau^*, \omega^* \\ \text{Given: } \Phi, T, \Psi, \mathbb{C} \end{aligned} \quad (2)$$

Attributes of RDM render this approach a suitable candidate for finding “low regret” or “no regret” options. Regret metrics are comparative and originate from Savage [7]. Defining “regret” as the difference between the expected performance function of an option under the uncertain set of scenarios and the best possible performance under perfect information, no regret options are solutions that are optimized toward the worst scenario [8]. Denoting “Regret” with  $R$ , it can be determined through Eq. (3). In this respect, to estimate the regret associated with each potential strategy and determine its robustness, a payoff table for each strategy in each potential future scenario is required [9]. Currently, RDM has been deemed a suitable tool in water management projects in the face of climate change. Furthermore, in transportation infrastructure adaptation, robust prioritization has also been integrated into the decision-making framework. Despite its benefits, however, robust designs may still be found uneconomical as the future unfolds.

$$R_{mk\tau} = \max_{\Phi, T} (\omega_{km\tau}) - \omega_{km\tau} \quad (3)$$

RO analysis is suitable where adaptive management strategies are available to cope with the evolving demand. RO analysis originated from options analysis developed in finance [10] and received a Nobel Prize in Economic Science in 1997. Since the 1990s, it found its way into design and management of infrastructure systems in the presence of uncertainty due to analogies of infrastructure management decisions and financial market interventions [11–14]. This new perspective of RO in (re)design of infrastructure systems is known as “real ‘in’ options (RIO) analysis” [15], which is slightly different from its traditional application.

The traditional prospective of RO in finance treats the system configuration as a black box. For instance, it can help decide on investing in a stock now or in the future, whereas the new perspective of RO in (re)design of infrastructure systems manipulates technical characteristics of system in response to reduction of uncertainty through future learning. Therefore, for RIO the availability of technical characteristics of system including technical details of options and their interdependencies is necessary. RIO provides the rational means of quantifying the value of flexibility built into infrastructure systems, helping with identifying worthwhile flexibilities, and expanding the horizon for considering flexible designs [16]. In the case of the protection of a bridge against corrosion, for example, a robust decision can involve

employing an expensive protection measure against the worst possible scenario. On the contrary, an adaptive solution may include employing a less expensive protection measure for a less severe scenario and considering enhancing the protection level, or even replacement of the whole bridge if found economical in the future in light of new information. In the context of RO, option value,  $F$ , is the value of delaying an uncertain investment. On the contrary to decision-making based on NPV suggesting investing when the total profit,  $V$ , is greater than investment cost,  $I$ , in the context of RO, an investment is recommended if  $V$  is greater than  $I$  plus a hidden cost of  $F$  for losing the option to invest later.

### 3 Methodology

The methodology employed herein considers deep uncertainty in choosing the optimum management strategy against corrosion. Using established theories in decision-making, this research provides a framework for investigating low regret and flexible decision-making in the design and management of a system. The framework involves a five-stage procedure as follows:

1. Identify the potential strategies, or flexibility, in the considered system and quantify their cost and possible benefits. In-service flexibility is achieved by adaptation through system configurations that can be changed after establishment of an initial configuration.
2. Evaluate economic indicators of each strategy based on perfect future foresight.
3. Compares each strategy across all possible scenarios in a matrix that is the foundation for choosing an adaptation strategy.
4. Formulate the optimization problem in terms of its objectives, constraints, and decision variables. Here, the objective function is to minimize the maximum regret for the managed/adaptive strategy.
5. Establish and run the optimization model.

The result of this framework is the presentation of a set of flexible adaptation strategies. More detail of each stage is given in what follows.

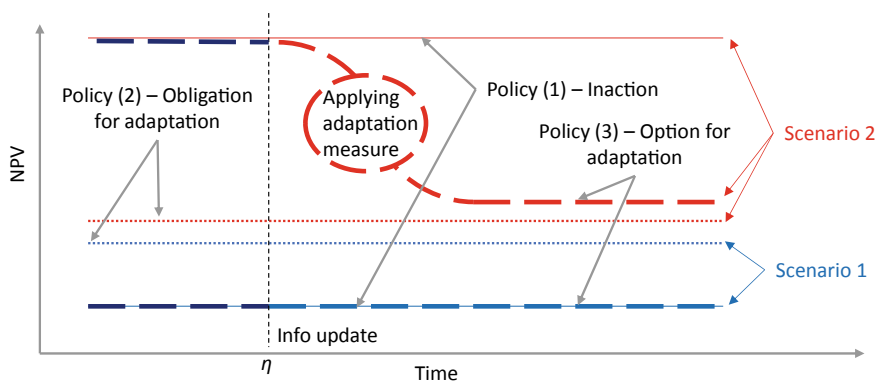
#### *Costs and Benefits of Adaptation*

The first stage identifies adaptation strategies for mitigating vulnerabilities from possible future scenarios. In the considered problem for example, metallizing is a possible adaptation strategy. Subsequently, the cost and risk of these strategies are used for evaluation of their potential benefit. In doing so, three steps are completed, including (1) identifying possible scenarios, (2) evaluating the cost and benefit of the strategies for each scenario, and (3) assessing risk and compiling and summarizing the total costs.

### *Economic Analysis of Adaptation Strategies*

In the second stage, economic indicators are evaluated for each of the scenarios. The evaluation results in a different value under each scenario, as each scenario produces different levels of risk and vulnerability. Vulnerability cost is defined as the amount of risk considering the adaptation strategy. Having the vulnerability and cost of adaptation determined in the previous stage, the total value of each strategy will be identified in terms of net present value (NPV). NPV is the summation of discounted adaption and vulnerability costs accruing over a planning horizon. The strategy with the lowest NPV is the most desirable, resulting in the lowest life cycle cost. For each strategy, NPV may significantly vary across various future scenarios. Furthermore, assumptions about the possibility of new information affect adaptive strategies and their NPVs. To consider the possibility of information update, two assumptions are considered. The first assumption is no update of knowledge of possibilities of various scenarios over the planning horizon. The outcome of this assumption can be regarded as a low regret robust decision, whereas the second assumption employs RO by assuming full scenario knowledge after an assumed time,  $\eta$ . The concept has been illustrated in Fig. 1 considering only two possible scenarios: (1) non-corrosive environment and (2) intensely corrosive environment.

Here, three policies have been considered: (1) inaction, (2) blind early action for early adaptation assuming no information update, and (3) assuming information update and applying the adaptation measure after an information update indicating more possibility of Scenario 2 (intense environment). Policy 1 will result a large rift among NPV for Scenarios 1 and 2. In other words, NPV has a low value under Scenario 1 and a very high value under Scenario 2. Policy 2 reduces the rift by decreasing NPV under Scenario 2 in the cost of increasing NPV under Scenario 1. In this case, should possibility of Scenario 2 be eliminated after the investment, the decision-maker would be regretful. Policy 3 waits for information update before implementing an adaptation measure in the cost of some extra risk. In this situation, after time  $\eta$ , the decision-makers have the option to whether invest for adaptation



**Fig. 1** Demonstration of the effect of these assumptions on a flexible adaptation strategy

in case of Scenario 2, or not to do so in case of Scenario 1. According to this explanation, Policy 2 is an obligation to a certain action at a specific time among all possible strategies. On the contrary, Policy 3 forgoes the possible strategies before  $\eta$ , in favor of having the option to choose the best possible strategy after  $\eta$ .

#### *Comparison Across All Future Scenarios and Policies*

Once the NPVs of various strategies under all possible scenarios are determined, a comparison should be performed to identify the most appropriate strategy across all future scenarios. In so doing, maximum regret across all possible scenarios is chosen as the performance measure of the various strategies. Accordingly, the strategy resulting the minimum of maximum possible regret has the best performance. This analysis has different forms for Policies 2 and 3, which are hereafter called as obligation- and option-based policies, respectively. The obligation-based policy takes the form of (4), whereas the option-based policy is according to (5). In both cases, max and min conditions are interchanged in comparison with (1), as regret is inversely related to utility. Another important observation in (5) is that maximization across scenarios has moved to the outer layer. This is so as the information update at  $\eta$  allows the decision-maker to opt among the possible engineering options,  $\Phi_\eta$ , over a time interval  $T_\eta$  beginning from  $\eta$ . In this way, the optimal strategy becomes a function of  $\eta$ .

$$\begin{aligned} \text{Objective: } & \min_T \left( \min_\Phi \left( \max_\Psi R_{km\tau} \right) \right) \\ \text{Find: } & m^*, \tau^*, R^* \\ \text{Given: } & \Phi, T, \Psi, \mathbb{C} \end{aligned} \quad (4)$$

$$\begin{aligned} \text{Objective: } & \max_\Psi \left( \min_{T_\eta} \left( \min_{\Phi_\eta} R_{km\tau} \right) \right) \\ \text{Find: } & m_\eta^*, \tau_\eta^*, R_\eta^* \\ \text{Given: } & \Phi_\eta, T_\eta, \Psi, \mathbb{C}, \eta \end{aligned} \quad (5)$$

An important outcome of comparing result of these two is a monetary value of information at  $\eta$  as

$$V_\eta = R^* - R_\eta^*. \quad (6)$$

The information value indicates that in order for the option policy to be proficient over obligation policy, obtaining the information should be possible, and its cost at  $\eta$  should be less than  $V_\eta$ .



## 4 Problem Description

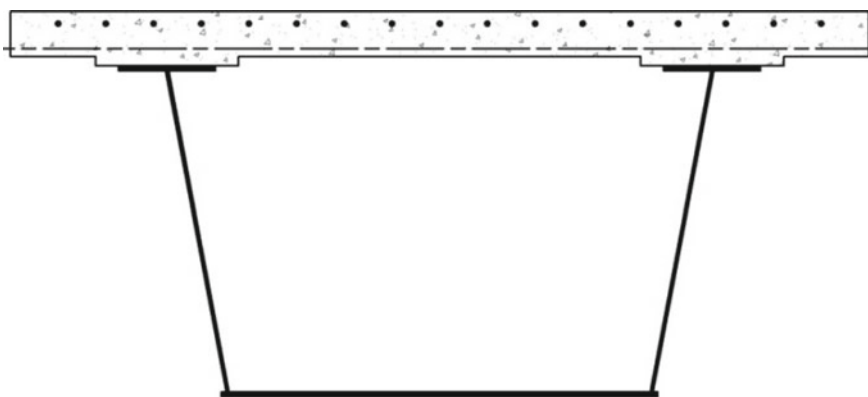
To illustrate the framework, an analysis is performed for finding the optimal corrosion management strategy for a typical weathering steel highway structure in the province of Ontario, Canada. The properties of the bridge are briefly described here. For a more detailed description of the loads, resistance, and reliability calculations, the reader is referred to [2, 3].

### 4.1 Bridge Model

The considered bridge is a typical single-span simply supported box girder overpass. A typical section of bridge girder is shown in Fig. 2. Using a macro-based Excel program, the bridge was designed in a previous research study on the reliability of corroding roadway bridges [2, 3].

In the program, at predetermined time intervals, corrosion penetration is calculated, and the geometrical and structural properties of the bridge are modified accordingly. Here, it is assumed that corrosion occurs on the web and flange together, and the general corrosion model is intended to replicate the thickness loss across the whole surface of a structural plate. In so doing, over time, when the resistance falls below the load effect, the program stops and records the time of failure. A total of four failure modes are considered here; each of these failure modes or “limit states” is identified and quantified by [17]:

- (1) Shear ( $V_f/V_r \leq 1$ ; Clause 10.10.5.2 (a) in [17])
- (2) Moment ( $M_f/M_r \leq 1$ ; Clause 10.10.5.2 (b) in [17])
- (3) Shear + moment ( $727M_f/M_r + 0.455V_f/V_r \leq 1$ ; Clause 10.10.5.2 (c) in [17])



**Fig. 2** Cross section of a typical box girder

(4) Bearing ( $B_f/B_r \leq 1$ ; Clause 10.10.8.1 in [17]).

These limit states are checked at eleven equidistant points along the length of the bridge (except for bearing, which is only checked at the supports). Typically, dead and live (i.e., traffic induced) loads dominate the design of short- and medium-span bridges. Therefore, only these load types are considered in the current study. Two kinds of dead load are distinguished: dead load and superimposed dead. The former refers to the slab and steel girder self-weight, while the latter refers to the sidewalk and wearing surface. The live load is the greater of a five-axle truck load and a truck-plus-lane load.

## 4.2 Corrosion Models

The corrosion rate of weathering steel is mainly affected by three factors: the presence of chloride pollution, the presence of sulfur dioxide pollution, and the time of wetness of the steel. Based on measurements on field specimens, Albrecht and Naeemi [18] categorized corrosion rates into three main regimes: rural, urban/industrial (henceforth referred to as urban), and marine. In this study, these three categories are considered as the potential corrosion rates due to microclimates undetermined during the design phase, and it is assumed the possibility of each scenario is unassigned.

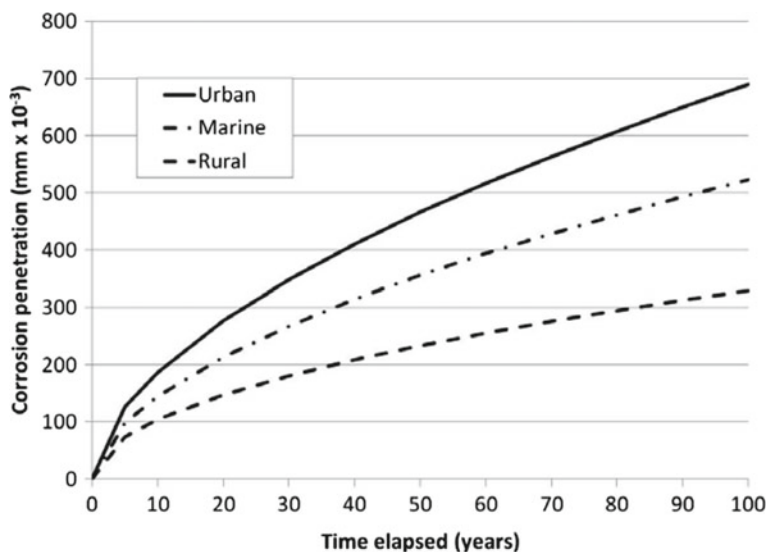
As described in [18], for modeling corrosion penetration, thickness loss over time,  $t$ , (in years) is assumed to follow the power function:

$$C = A \cdot t^B, \quad (7)$$

where  $C$  is the thickness loss in  $\text{mm} \times 10^{-3}$  and  $A$  and  $B$  are constants. The values used for these constants in this study are given in Table 1 and are based on corrosion penetration tests performed in England, Germany, and the USA [18, 19]. Using these values, a projection of thickness loss according to the three scenarios is presented in Fig. 3.

**Table 1** Corrosion rates [19]

Environment	Parameter	Distribution type	Mean	Coefficient of variation
Rural	A	Lognormal	33.3	0.34
	B	Lognormal	0.498	0.09
Urban	A	Lognormal	50.7	0.30
	B	Lognormal	0.567	0.37
Marine	A	Lognormal	40.2	0.22
	B	Lognormal	0.557	0.10



**Fig. 3** Mean corrosion penetration over time for different environments

### 4.3 Cost Evaluation

Considering a set of possible corroding environments, the probability of failure of the bridge and the life cost of failure are determined using the developed program described in [2, 3]. Then, the life costs were used in the developed frameworks to determine the benefits of postponing a corrosion protection measure. The results were used to compare these frameworks and draw conclusions regarding appropriate management strategies. Based on the prospective of the decision-maker, different failure costs could be considered. In order to illustrate the methodology, monetary values are presented as portions of the failure cost and presented in Table 2. It is assumed that the corrosion protection measure with a relative cost of 0.01 involves a metallizing process, which hinders further corrosion of the steel components for up to 40 years. The accessibility cost involves field application costs such as permits, scaffolding, sand blasting, and transportation of the equipment, which is mostly avoided for a construction phase application.

**Table 2** Assumed values for cost and life of a corrosion protection measure

Parameters	Failure cost	Corrosion measure cost	Accessibility cost	Discount rate	Protection life years
Values	1	0.01	0.01	0.02	40

## 5 Results and Discussion

Based on the probability of failure of the bridge and values of Table 2, the NPVs of various strategies under various scenarios are determined and presented in Fig. 4, where the horizontal axis represents the implementation time of the protection measure. The first jumps represent the accessibility cost for field application of the protection measure at time more than zero. Over time, NPV is declining for rural and marine scenarios due to the discounting effect, while under the urban scenario, it is increasing due to the increased risk. To illustrate the three policies explained in Sect. 3, Fig. 5 is presented for an information update occurring in the 20th year. Here, obligation policy lines represent NPVs of obligation to metalize at  $\tau = 40$ . Later it will be shown that this is the best strategy under obligation policy. For option policy, an information update at  $\eta = 20$  is assumed. Under this policy, the optimal decision is metalizing at  $\tau = 20$  in case of the urban scenario and not metalizing in case of the other scenarios.

The procedure of finding the optimum strategies under obligation and option policies can be as follows through Figs. 6 and 7, respectively. In both figures, first,  $R_{kmt}$  values are computed according to the described methodology and presented. Regret of not metalizing for the various scenarios is presented with the black dash and dash-dot lines, while the regret of metalizing under urban and marine and rural scenarios is shown with red and blue curves, respectively. In Fig. 6, the thick dash blue curve shows the minimum possible regret under the worst scenario at a given  $\tau$ . The position of this curve in respect to the other curves at a given  $\tau$  indicates the optimum action at that action time. The minimum of this curve is at  $\tau = 40$  and determines the objective of obligation policy as defined in (4). Therefore, under obligation policy, the optimum strategy is metalizing at  $\tau = 40$ . As depicted in Fig. 7, on the contrary to obligation policy, optimum strategy under option policy depends on the time of information update,  $\eta$ . The optimum action and its timing under option

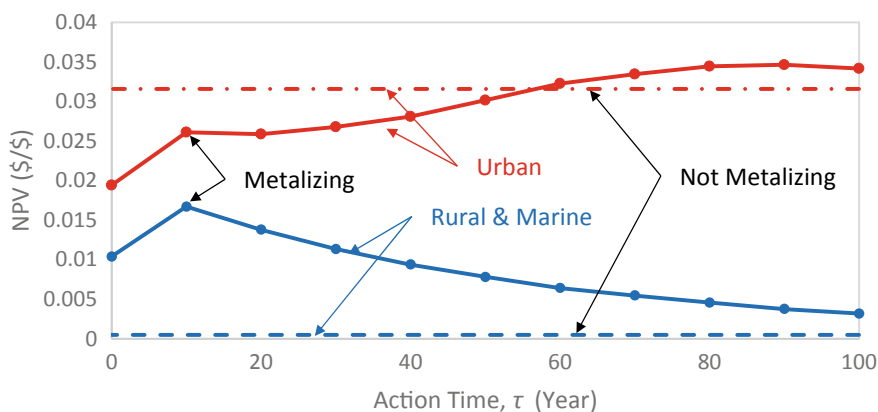
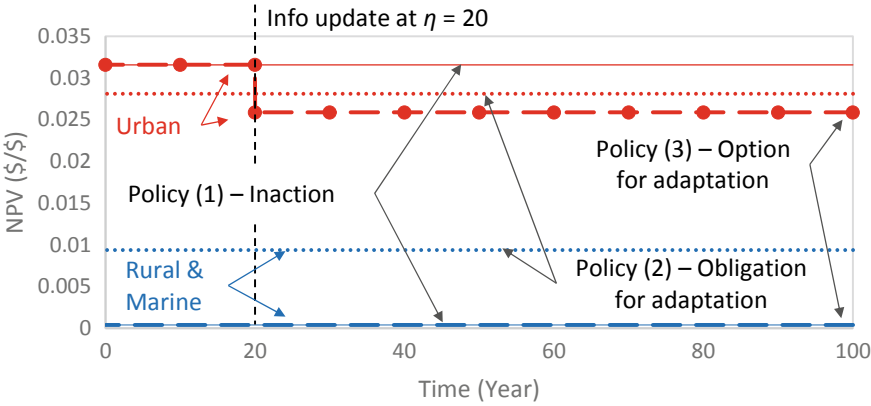


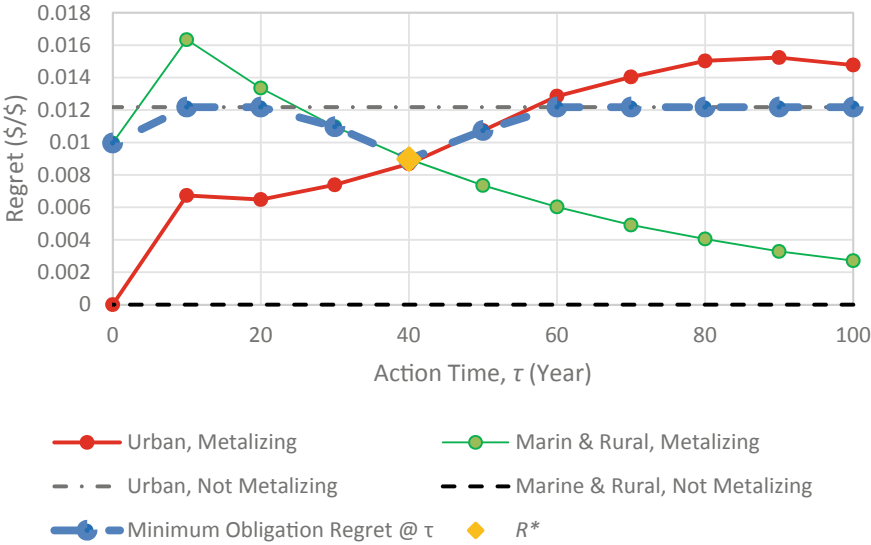
Fig. 4 NPV of various strategies under various scenarios



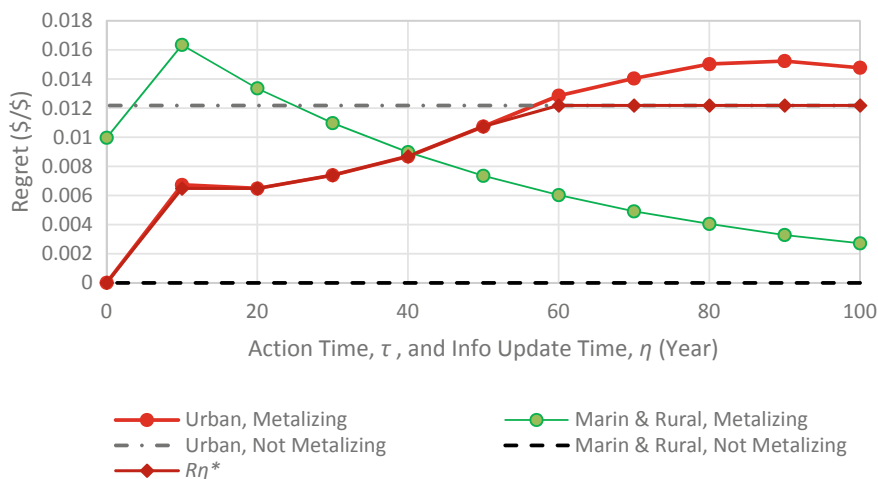
**Fig. 5** Perception of NPV of the different policies under urban and rural conditions

policy are shown in Figs. 8 and 9, respectively. For instance, from these figures it can be concluded that for  $\eta = 10$ , the optimum strategy involves delaying metalizing until  $\tau = 20$ , which lowers the costs through discounting.

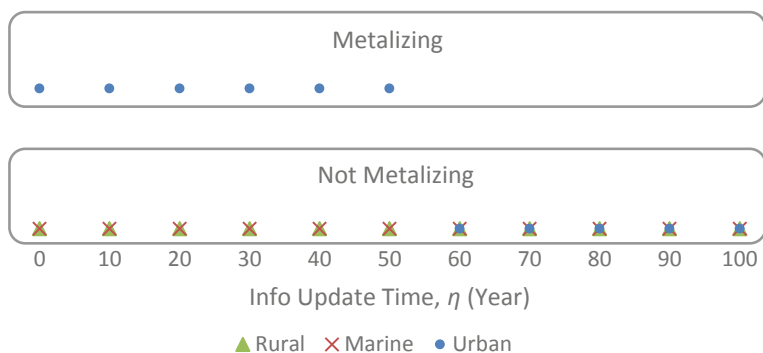
For decision-making on the policy, a comparison of the outcomes of employing obligation and option policies is shown in Fig. 10. In summary, employing obligation policy results in an obligation to metalize at the  $\tau = 40$  with  $R^* = 0.0090$ , while employing option policy provides an option but no obligation to metalize sometime after  $\eta$  with a maximum possible regret of  $R^*_\eta$ . While the presence of such an option



**Fig. 6** Obligation policy—optimum obligation regret determination based on  $R_{km\tau}$  values

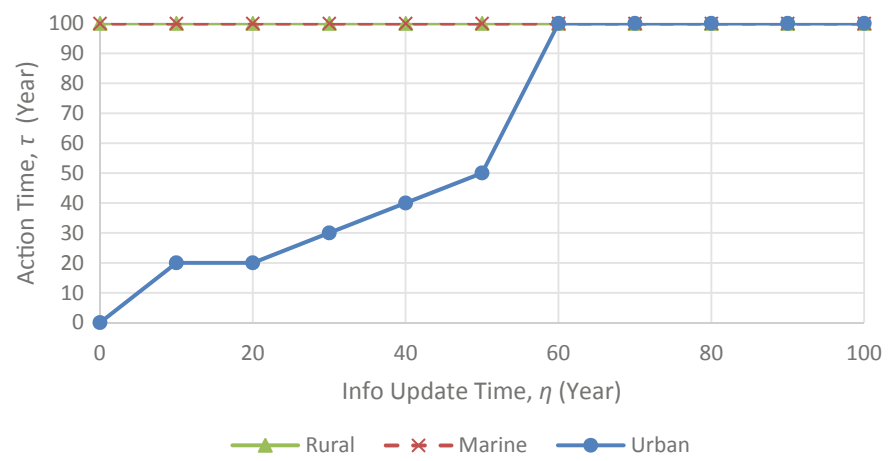


**Fig. 7** Obligation policy—optimum option regret determination for various times of information update,  $\eta$ , based on regret of various strategies under different scenarios

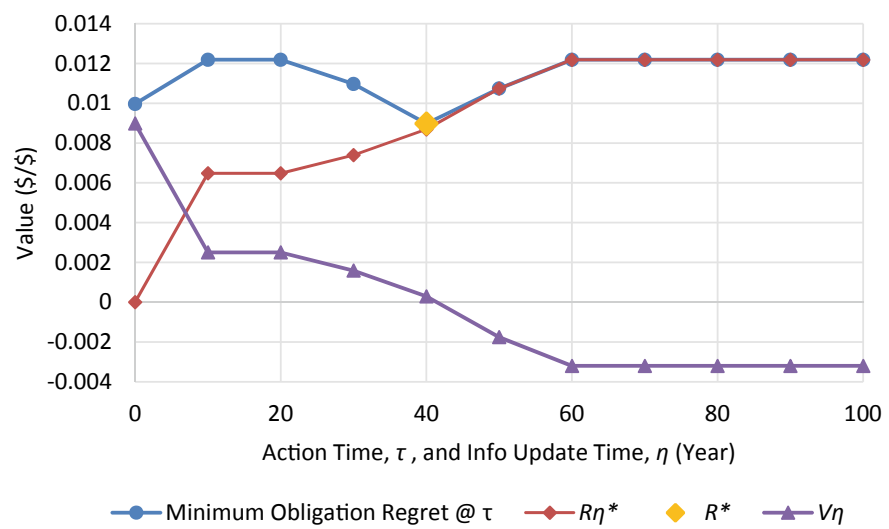


**Fig. 8** Optimum action for each scenario and given time of information update under Policy 3—option for adaptation

can reduce regret, its availability depends on the prospective of the decision-maker or availability of a test that provides the information on the condition of future. The monetary value of such information is determined according to Eq. (6) and presented in Fig. 10. For the considered problem, the information value is positive for around 40 years. In order for a site test to be beneficial, it should be able to provide the information on the state of the microclimates sooner than 40 years with a cost lower than  $V_\eta$ .  $V_\eta$  is a decreasing curve with a first down jump due to accessibility costs of metalizing right after bridge installation. In this figure,  $V_\eta$  is constant from  $\eta = 10$  to 20 years as delaying metalizing to  $\tau = 20$  even in the case of urban scenario is beneficial.

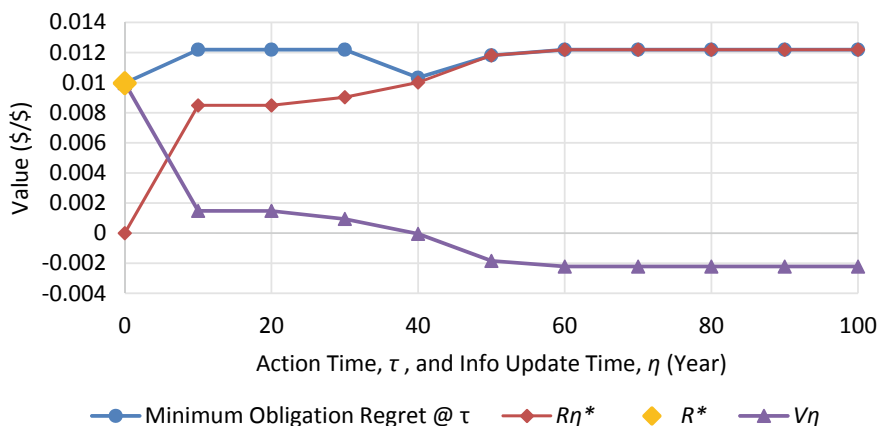


**Fig. 9** Action time for each scenario and given time of information update under Policy 3—option for adaptation



**Fig. 10** Regret of various strategies and value of information on condition of microclimates

Given the condition of the example problem, decision-making on metalizing during the construction phase is similar for active and passive bridge owners. A passive bridge owner will metalize the bridge at  $\tau = 40$  based on obligation policy. Similarly, an active bridge owner considering option policy would avoid metalizing and assign resources for exploring the state of the bridge sooner than 40 years considering the cost of site test and information value. It should be noted that this similarity during the construction phase is not always the case. For instance, Fig. 11 depicts the



**Fig. 11** Regret of various strategies and value of information on condition of microclimates after increasing the accessibility cost to 0.013

regret of various strategies and value of information after increasing the accessibility cost to 0.013. In this situation, a passive bridge owner will metalize the bridge in the workshop, while an active bridge owner may avoid doing so.

## 6 Conclusion

The paper provides a framework for evaluation of various strategies for protection of steel bridges against corrosion. Composite concrete and steel slab-on-girder bridges are commonly used throughout North America. In recent years, some bridge owners have become aware of an unexpectedly large amount of corrosion degradation occurring to a number of their weathering steel highway structures. While in a mild environment, weathering steel might show little degradation, in case of realization of a severe environment, for example, due to extensive use of de-icing salt during winter, using this material without protection can be found problematic as doing so might result in higher management costs or even catastrophic failures. On the contrary, the realization of a mild or moderate environment after utilization of the expensive preventive measures would also be considered to be suboptimal.

On this basis, the proposed method is founded on an assumption that the information on the state of the microclimate somehow becomes available over time, which makes delaying application of a protection measure valuable. The framework resorts to adaptive solutions, beginning with a less expensive option until the actual condition of the environment is more evident. Willing to minimize the maximum sense of loss, regret is combined with the real option methodology and implemented in the decision-making framework. The method is illustrated using a problem of a single-span simply supported box girder bridge. A comparison of the outcomes for two



cases of having and not having an expectation of a future information of microclimates is provided. The comparison yields a monetary information value for making a robust decision for protection against corrosion. Furthermore, although the analysis was performed on a single data point, the results show that a blind application of a protection measure on an existing bridge just based the potential of extensive corrosion can be regretful in comparison with decision-making based on option value. In the future, a sensitivity study could be done to determine the boundary cost for effectiveness of various strategies with consideration of different bridge types and protection measures. In this way, the provided framework can help bridge owners more effectively target their resources on safety investments. Although the methodology was applied to corrosion management of steel bridges, it is equally applicable to other aspects of infrastructure management such as considering deep uncertainties in climate change effects on environmental loads.

## References

1. Commentary on CAN/CSA-S6-00 (2000) Canadian highway bridge design code. Canadian Standards Association
2. Damgaard NR (2009) Prediction and prolongation of the service life of weathering steel highway structures. University of Waterloo
3. Damgaard N, Walbridge S (2009) Service life prediction for weathering steel highway structures. In: Proceedings of the 2009 structures congress—don't mess with structural engineers: expanding our role, pp 2189–2198. [https://doi.org/10.1061/41031\(341\)240](https://doi.org/10.1061/41031(341)240)
4. Dittrich R, Wreford A, Moran D (2016) A survey of decision-making approaches for climate change adaptation: are robust methods the way forward? *Ecol Econ* 122:79–89. <https://doi.org/10.1016/j.ecolecon.2015.12.006>
5. Vallejo L, Mullan M (2017) Climate-resilient infrastructure—getting policy right, no 121
6. Marchau V, Walker W, Bloemen P, Popper S (2019) Decision making under deep uncertainty. From theory to practice
7. Savage LJ (1951) The theory of statistical decision. *J Am Stat Assoc* 46(253):55–67. <https://doi.org/10.1080/01621459.1951.10500768>
8. Mondoro A, Frangopol DM, Liu L (2018) Bridge adaptation and management under climate change uncertainties: a review. *Nat Hazards Rev* 19(1):04017023. [https://doi.org/10.1061/\(asce\)nh.1527-6996.0000270](https://doi.org/10.1061/(asce)nh.1527-6996.0000270)
9. Espinet X, Schweikert A, Chinowsky P (2017) Robust prioritization framework for transport infrastructure adaptation investments under uncertainty of climate change. *ASCE-ASME J Risk Uncertainty Eng Syst Part A Civ Eng* 3(1). <https://doi.org/10.1061/ajrua6.0000852>
10. Myers SC (1984) Finance theory and financial strategy. *Interfaces (Providence)* 14(1):126–137. <https://doi.org/10.1287/INTE.14.1.126>
11. Dixit RK, Pindyck RS (1994) Investment under uncertainty. <https://doi.org/10.1515/9781400830176>
12. Voeks R (1997) Real options: managerial flexibility and strategy in resource allocation. *J Bank Finance* 21(2):285–288. [https://doi.org/10.1016/s0378-4266\(97\)85585-9](https://doi.org/10.1016/s0378-4266(97)85585-9)
13. Amram M, Kulatilaka N (1999) Real options: managing strategic investment in an uncertain world. *Choice Rev Online* 36(10):36-5767. <https://doi.org/10.5860/choice.36-5767>
14. Copeland TE, Antikarov V (2001) Real options: a practitioner's guide. Texere. [Online]. Available: <https://books.google.ca/books?id=fnhPAAAMA AJ>
15. Neufville R (2003) Real options: dealing with uncertainty in systems planning and design. *Integr Assess* 4(1):26–34. <https://doi.org/10.1076/IAIJ.4.1.26.16461>

16. Gersonius B, Ashley R, Pathirana A, Zevenbergen C (2013) Climate change uncertainty: building flexibility into water and flood risk infrastructure. *Clim Change* 116(2):411–423. <https://doi.org/10.1007/s10584-012-0494-5>
17. CAN/CSA-S6-06 (2006) Canadian highway bridge design code. Canadian Standards Association
18. Albrecht P, Naeemi A (1984) National cooperative highway research program report 272: performance of weathering steel in bridges. Transportation Research Board, National Research Council, Washington, DC
19. Kayser J (1988) The effects of corrosion on the reliability of steel girder bridges

# Agent-Based Modeling and Simulation of Congested Sites



Raghda M. Moharram, Yasmeen A. S. Essawy, Abdelhamid Abdullah, and Khaled Nassar

**Abstract** Site planning is always performed on construction sites to arrange the locations of facilities to minimize travel distances of the different resources. However, this type of planning and optimization does not consider the clashes and the congestions that may happen due to the presence of multiple types of resources. This results in safety hazards and injuries and loss of productivity in congested areas, leading to delays in the original schedule. The objective of this paper is to track the actual paths where resources are followed, starting from entering the site, moving to the desired location, performing work, moving to another location, and exiting the site. A simulation model is created using AnyLogic software to track the path of every single resource and produce a density map combining all of them, highlighting the different congestion areas. The model focuses on three types of resources, namely (1) labor, (2) equipment, and (3) material. After analyzing the congestions and conflicts between the resources, the model produces a safety analysis. In addition, the model helps identify the actual productivity rates of each resource, taking into consideration the travel time and delays on the site, hence producing an updated schedule. Finally, during the execution of the work, online tracking of resources may be used to compare it against the model to identify loss of productivity between the resources on-site. The main aim of this paper is to design the framework and the model that will further be used for the analysis.

**Keywords** Agent-based modeling · Congested sites

---

R. M. Moharram (✉) · Y. A. S. Essawy · A. Abdullah · K. Nassar  
Department of Construction Engineering, The American University in Cairo, Cairo, Egypt  
e-mail: [raghdammoharram@aucegypt.edu](mailto:raghdammoharram@aucegypt.edu)

Y. A. S. Essawy  
Department of Structural Engineering, Ain Shams University, Cairo, Egypt

A. Abdullah  
Department of Architectural Engineering, Faculty of Engineering at Mataria, Helwan University, Helwan, Egypt

## 1 Introduction

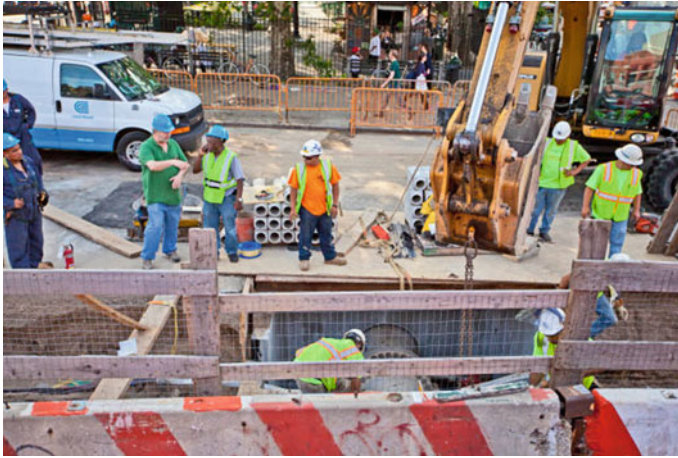
Construction industry generally has a competitive nature where planning and optimization in every aspect have to be undertaken to improve productivity, safety matters, quality, etc. Construction sites usually have congestion that results in delays and loss of productivity. To be able to generate a realistic schedule, the crowd of the different types of resources on-site should be tracked to find out any places that should be free for safety reasons, for example, or to define that a particular activity will not be able to take place because there is specific equipment moving in this space; there is no space for labors, etc. This paper will present a framework to track the resources moving on-site from when they enter the site, perform the work, move between different activities, and leave the site. AnyLogic software was used to produce an agent-based simulation using the pedestrian library. The results will be a density map for the resources on-site that will be further analyzed to find out different congested areas and safety hazards. The construction schedule is, then, generated, highlighting the construction sequence along with the elements' duration (considering the actual speed of workers as well as the travel time). Finally, in further research, the model will be used to compare it against online tracking of resources to determine any loss in productivity on-site.

## 2 Literature Review

### 2.1 *Construction Site Crowd*

The crowd in construction sites usually results from limited working areas, overdesign of crews needed, fast-track projects with very high productivity rates, etc. Crowded sites are commonly happening in most construction sites, as shown in Fig. 1, which shows the availability of different types of resources in the same place, which may be more crowded than needed and may result in safety hazards.

According to Spillane et al. [6], a study was conducted on case studies to identify the common factors arising health and safety issues due to the crowd on-site. The main factors were “(1) lack of space, (2) problem of co-ordination and management of site personnel, and (3) overcrowding of the workplace.” Another study was conducted by Spillane et al. [7]; it concluded after analyzing case studies that the main issues of the congested construction sites were as follows: “(1) accidents due to an untidy site, (2) one contractor holding up another because of the lack of space, (3) a risk to personnel because of vehicular traffic on-site, (4) difficult to facilitate several contractors at one work location, and (5) numerous personnel working within the one space.”



**Fig. 1** Crowded construction site

## ***2.2 Real-Time Monitoring Systems (RTMSs) in Construction***

Real-time monitoring systems are emerging in the construction industry with efforts to monitor project performance and control it. RTMS is now used in safety fields. According to Soltanmohammadlou et al. [5], a study was conducted to monitor labor and equipment to enhance safety on-site, which yielded accident prevention.

## ***2.3 Agent-Based Modeling***

Agent-based modeling (ABM) is a computational modeling research tool studying organizational behavior in general. And since individual members' interactions and behavior are very complex systems, they need different tools to be analyzed [3, 4]. ABM uses virtual agents that can imitate different interactions and behaviors of the various individuals within the system. The system can also be used against what-if scenarios and other alternatives. Previously, ABM has been used in the construction industry in production problems, supply chains, dispute resolutions, and some safety issues [8]. Ahn et al. [1] studied "the effects of workers' social learning on absenteeism and concluded that cultivating a good culture was more effective than regulating individuals for worker attendance control," Kiomjian et al. [2] simulated the evolution of collaboration in temporary project teams and indicated that the system was less efficient when more efforts were needed to form relationships.

### **3 Methodology**

#### **3.1 Framework**

The model will start by identifying different types of resources and the elements' data. Elements data includes their name, position (x- and y-coordinates), quantity, and required number of resources. Then, the model will first start to check if we have available resources on-site; if yes, it will assign the required number of resources to the element waiting. Afterward, the resource will move toward the element's coordinates, and the model will track all the trips. When the element is done, the resource will return to the beginning to be assigned to a new element to start working on. If the resources available are less than the needed resources, the model will be on hold for this element until a resource is back so that we can start the work on the element again. The procedure is repeated for the different types of resources assigned to the different model element activities. The framework is shown in Fig. 2.

#### **3.2 Model**

The model consists of two main items: agents and resource pools. This was done to create a relationship between the items, resources, and their activity. A sample of the model logic is as in Fig. 3.

##### **3.2.1 Agents**

Two libraries for the agents were used in the model. The first one is the pedestrian library, where pedestrian agents were generated in the model with the different resources. This type of agent will help us generate a density map at the end of the model. Each resource will have a separate pedestrian source, whether labor, equipment, or material. The second one is the process modeling library, where the source generates agents that are the activities that will take place, for example, a column constructed on-site.

##### **3.2.2 Linking the Model**

To link the model that consists of two different types of loops (which will further be discussed in the next section), the resources will take two other forms; the first one is the pedestrian agent, as mentioned before, and the second one is the resource pool. The resource pool's primary purpose is to produce a model based on the availability of resources to monitor the productivity of each resource afterward. And the pedestrian agent is used mainly to assign locations and track the paths these agents will follow.

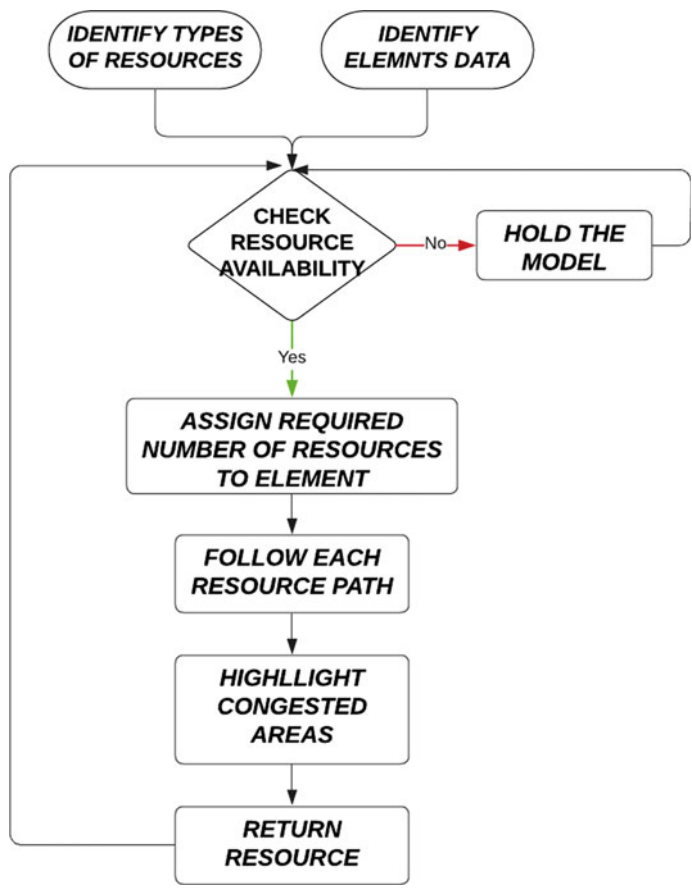


Fig. 2 Proposed framework

3.2.3 Loops

The first chart is a loop where pedestrians are generated to follow a particular loop. The loop describes the flow they are moving in. First, the pedestrian enters the queue to wait until needed; then, the hold will be unblocked to pass each element’s required number of resources. Afterward, the resources will be asked to go to the location of the elements where they will work on. Finally, the resources will go back to wait in the queue once again to be assigned to a new element to work on. The second chart defines the sequence of activities. First, the source will get the elements data from the Excel file that has the user’s inputs. The elements then will enter the activities. Each activity is represented by one line. The activities will enter the first queue to wait, then enter the second queue, which has a capacity of 1; to get the data, we need to decide to start the element or not and assign their data to the variables. Then, on seize block, the model will un-hold the block on the resource to allow the required

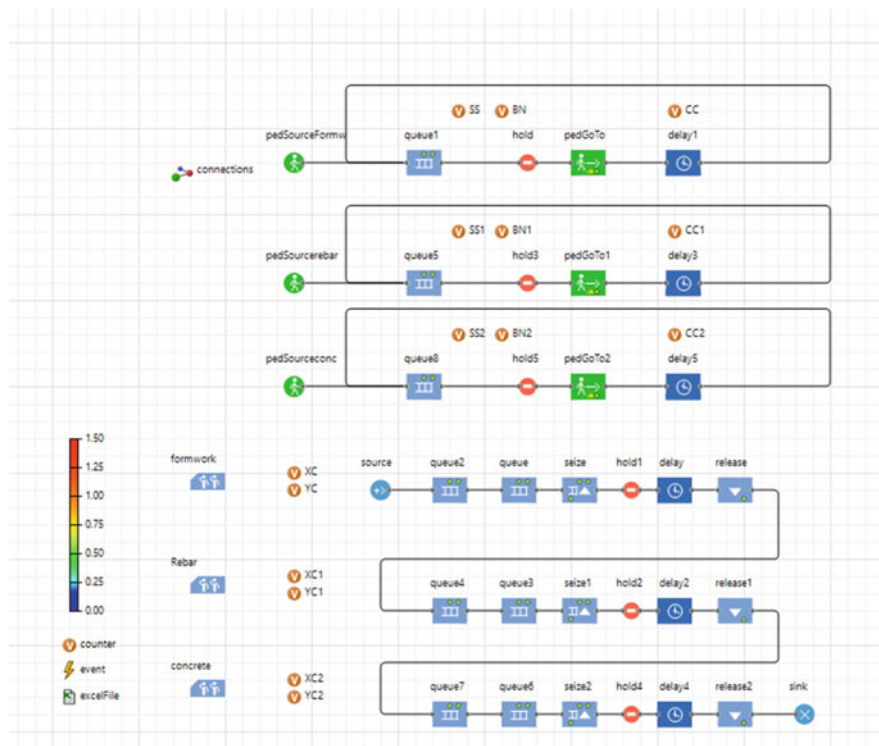


Fig. 3 Model logic

number of resources to pass. Then the element will remain on seize until the next hold is unblocked. This hold is unblocked after the arrival of the required number of resources to the position we require. The loop variables define the required number of resources needed for this specific activity (variables SS and BN). Another variable was added to link the two flow models together (variable CC). This variable will update once one resource arrives on time. After the required number of resources arrives, it will un-hold the hold in the activities model to pass the element to the delay to take the time necessary. After the element enters the delay, the variable will be set to zero again to start new counting. The loop will be repeated as many times as the number of different resources we have on-site. And the number of activities will determine the number of lines or delays we have in the model.

First Loop

See Fig. 4.



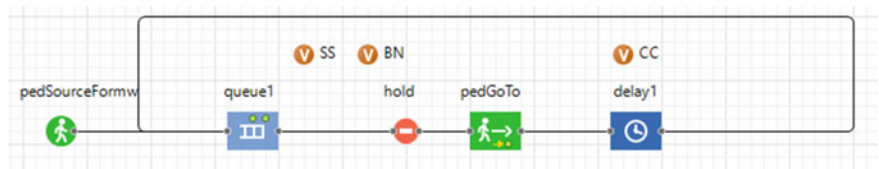


Fig. 4 Loop1

▼ Action

```
pedSourceFormw.inject(15);  
pedSourceconc.inject(15);  
pedSourcecobar.inject(15);
```

Fig. 5 Actions on event

Mode:

Block automatically after N agents (use unblock()) ▼

Fig. 6 Hold mode-Loop1

Source

The source generates the pedestrians, which are the resources, with an inject function from an event with the number of resources available on-site for each type. The source can adjust the speed of the resources based on their type and the applicable speed on-site. A new pedestrian source is needed for each type of resource with all the loop. The details of the actions on the block are shown in Fig. 5.

Queue

The primary purpose of this queue is the place where all agents will wait until moving to their assigned task.

Hold

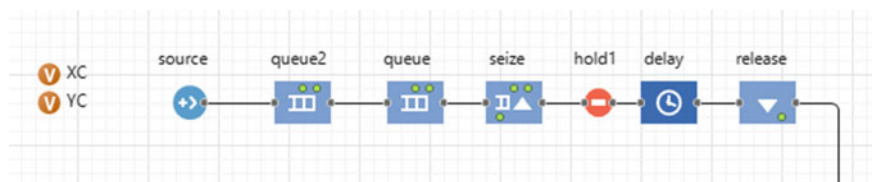
The hold here is to block any movement of any resources (pedestrians) that are not assigned to a specific task. It is unblocked based on the seize in the second loop. It unblocks only to pass a particular number of resources needed in a particular location. The mode of the hold is shown in Fig. 6.

Ped-Go-To

This assigns the location where the pedestrian will move to it based on the variables XC and XY that have values from the second loop actions.

Delay

The delay time is the same time needed by the delay in Loop1 that is calculated based on the productivity equation (to be discussed in the coming sections).



**Fig. 7** Second loop

**On exit:**

```

agent.nr = excelFile.getCellNumericValue(1, counter , 6);
agent.nr1 = excelFile.getCellNumericValue(1, counter , 7);
agent.nr2 = excelFile.getCellNumericValue(1, counter , 8);
agent.x = excelFile.getCellNumericValue(1, counter , 3);
agent.y = excelFile.getCellNumericValue(1, counter , 4);
agent.q = excelFile.getCellNumericValue(i, counter , 5);
counter++;

```

**Fig. 8** Action on source-Loop2

## Second Loop

See Fig. 7.

### Source

The source, as mentioned, is generating activities to be executed on-site. The generated source is with a predefined new agent that has a set of new variables: *nr*, *nr1*, *nr2*, *x*, *y*, and *q*, which indicates the number of resources needed from source 1, number of resources required from source 2, number of resources required from source 3, *x*-coordinate, *y*-coordinate, and quantity, respectively. The action here is to assign all the variables of the agent to its user-entered variables in the Excel file. The details of the actions on the block are shown in Fig. 8.

### Queue 1

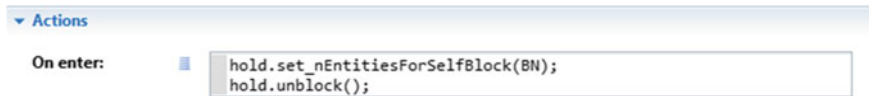
The first queue's primary purpose is to have a space where all produced agents are arranged before starting the activity loop. This may be further used afterward to arrange them according to specific deadlines for the different components or elements to be executed on-site.

### Queue 2

The second queue's primary purpose (with a capacity of 10) is to assign the variables *XC* and *YC*, the *x*- and *y*-coordinates of the component that will be executed to be sent to the first loop.

### Seize

The purpose of the seize is to assign (if available) the required number of resources to these activities. Moreover, it allows the hold in the first loop to pass the required number of resources needed. The details of the actions on the block are shown in Fig. 9.



**Fig. 9** Action on seize-Loop2

### Hold

The primary purpose of this hold is to pass one agent at a time to ensure variables values are not overwritten.

### Delay

The delay's time is the time needed for this component to be done with this number of resources. It is calculated using the productivity equation:

$$\text{Duration} = \frac{\text{Quantity}}{\text{Productivity Rate} * \text{Number of Resources}}$$

### Release

This block is releasing all the seized resources for this agent so that it can go back to the resource pool to be assigned to a new component.

## 4 Model Output

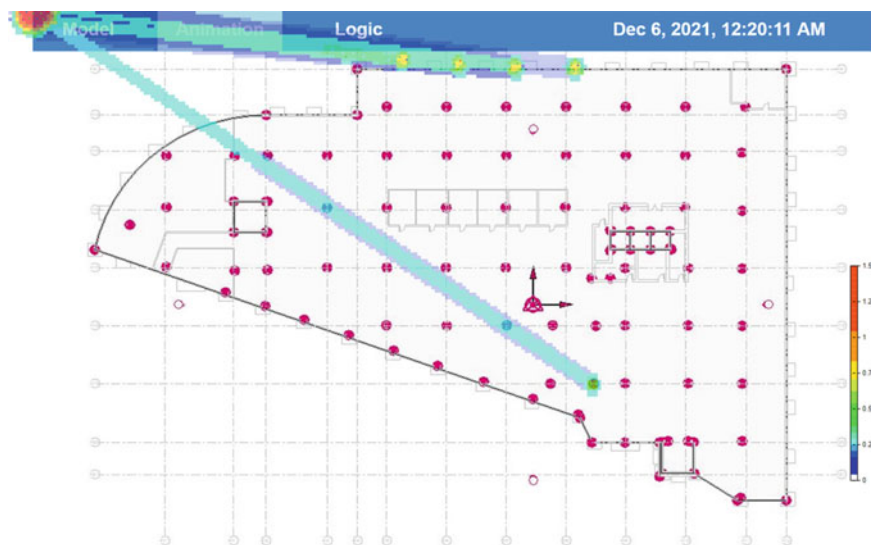
The model was tested on real case project: commercial building. This project is used as an illustrative case to trace the resources paths. Below is a snapshot of the heat map produced by the model shown in Fig. 10. It shows that the resources are initially at the origin (the top left corner with the highest density), and then they started moving to their desired locations.

It is shown on the heat map (Fig. 10) two things:

1. The path each resource is taking which highlights the movement of the resources.
2. The location of the crowded place: In this case, most of the resources are at the location of the columns, and the yellow color highlights the crowd.

This heat map can be used as a basis for analyzing the crowdedness of the construction site due to the presence of the different types of crews. It can also include material flow and can be further extended to include the movement of cranes and other equipment. This will help

1. Monitor productivity of crew on-site and
2. Identify and monitor technical hazards on-site.



**Fig. 10** Heat map

## 5 Conclusion and Recommendation

The proposed model produced a heat map where each resource, with all their types, is tracked with their actual speed on-site. The model was able to execute the activities on-site with the required resources and calculate the time according to the number of resources available. Moreover, the construction sequence was done based on the availability of resources. The primary recommended analysis from this model is to highlight the congested areas to analyze whether the crowd will be applicable or not; moreover, tracking equipment with high hazards is recommended. Furthermore, the model could be used to produce a more accurate schedule with better timing. Finally, RTMs could be used and compared to the generated simulation model to study the progress and the actual productivities on-site.

## References

1. Ahn S, Lee SH (2015) Methodology for creating empirically supported agent-based simulation with survey data for studying group behavior of construction workers. *J Constr Eng Manag* 141(1):04014065. [https://doi.org/10.1061/\(asce\)co.1943-7862.0000918](https://doi.org/10.1061/(asce)co.1943-7862.0000918)
2. Kiomjian D, Srour I, Srour FJ (2020) Knowledge sharing and productivity improvement: an agent-based modeling approach. *J Constr Eng Manag* 146(7):04020076. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001866](https://doi.org/10.1061/(asce)co.1943-7862.0001866)

3. Mahjoubpour B, Nasirzadeh F, Mohammad Hosein Zadeh Golabchi M, Ramezani Khajehghiasi M, Mir M (2018) Modeling of workers' learning behavior in construction projects using agent-based approach. *Eng Constr Archit Manag* 25(4):559–573. <https://doi.org/10.1108/ecam-07-2016-0166>
4. Navon R, Goldschmidt E (2003) Can labor inputs be measured and controlled automatically? *J Constr Eng Manag* 129(4):437–445. [https://doi.org/10.1061/\(asce\)0733-9364\(2003\)129:4\(437\)](https://doi.org/10.1061/(asce)0733-9364(2003)129:4(437))
5. Soltanmohammadlou N, Sadeghi S, Hon CKH, Mokhtarpour-Khanghah F (2019) Real-time locating systems and safety in construction sites: a literature review. *Saf Sci* 117:229–242. <https://doi.org/10.1016/j.ssci.2019.04.025>
6. Spillane JP, Oyedele LO, Meding JV, Konanahalli A, Jaiyeoba BE, Tijani IK (2011) Confined site construction: a qualitative investigation of critical issues affecting management of health and safety. *J Civ Eng Constr Technol* 2(7):138–146
7. Spillane JP, Oyedele LO, von Meding J, Konanahalli A, Jaiyeoba BE, Tijani IK (2013) Critical factors affecting effective management of site personnel and operatives in confined site construction. *Int J Inf Technol Proj Manag* 4(2):92–108. <https://doi.org/10.4018/jitpm.2013040106>
8. Zhang P, Li N, Jiang Z, Fang D, Anumba CJ (2019) An agent-based modeling approach for understanding the effect of worker-management interactions on construction workers' safety-related behaviors. *Autom Constr* 97:29–43. <https://doi.org/10.1016/j.autcon.2018.10.015>

# Selecting Most Appropriate Delay Analysis Technique Using Quantitative Approach



Mostafa Farouk and Ossama Hosny

**Abstract** The increasing complexity and magnitude of the projects impose greater impact of delays on stakeholders. Construction delays are a major source of disputes in construction projects. Since a construction project depends on interactions and shared responsibilities among parties, research works were directed towards identifying delay causes, quantifying their impacts, and proposing ways to deal with them. Different delay analysis techniques (DATs) applied to the same project's delays provide different results, and thus, the selection of technique to use in evaluating delays becomes vital. Reviewing the literature, it has been realized that often there are disagreements, which lead to escalating a claim into a dispute, between engineers and contractors on the technique to be used to evaluate the delay. A dispute results in additional costs, time, and in some cases negatively impacts the relation between the parties as it goes up the ladder of dispute resolution. Some research was conducted to gather experts' opinions on the best technique to be used; however, little research was done to quantify factors behind the selection and transform them into a numerical model. This research is an attempt to support different parties in selecting the most appropriate DAT to be used for a delay by building a model based on quantifying experts' opinions to score different factors influencing the selection of DATs. Moreover, a survey based on the Egyptian market was conducted and used to build the model. Results of the survey were compared to surveys, from different countries, that tackle those different factors. This research helps in integrating the efforts that were exerted to tackle this challenge while providing analysis of how different factors are perceived through different law systems. Delay analysts, contract administrators, and other parties can use the model to validate their chosen DAT for a claim.

**Keywords** Delay analysis technique · Quantitative approach

---

M. Farouk (✉) · O. Hosny  
The American University in Cairo, New Cairo, Egypt  
e-mail: [mostafakanza@aucegypt.edu](mailto:mostafakanza@aucegypt.edu)

O. Hosny  
e-mail: [ohosny@aucegypt.edu](mailto:ohosny@aucegypt.edu)

## 1 Introduction

Various statistics indicate growth in the Egyptian construction industry. This growth can be attributed to increasing population, a growing market need for buildings and infrastructure, and technological advances which allows building higher, bigger, and more complex structures. In its 3Q2019 Global Construction Outlook report, GlobalData, data and analytics company, forecasts around 11.3% annual increase in the Egyptian construction industry between 2019 and 2023. With more complex projects, occurrence of delay is critical and may have severe financial implications on either, if not both, parties to a contract. The inability to meet a projects deadline, construction delay, often needs to be analysed through examining the alternating impact of that delay on the critical path of the project. The critical path is the longest path in terms of duration to arrive to the end date of the project. This analysis is vital as there is usually a dispute regarding the accountability of the delay by the different parties [1].

Construction delay is defined to be the late completion of the construction work than expected or than planned and not on the proper time. Delay in construction projects is a common occurrence that may lead to significant losses for the project parties. Many studies have focused on analysing aspects that cause delays and proposing ways to minimize delays in construction projects (see, e.g. [2, 3]). Other studies tackled more functional issues to highlight weakness of existing DATs by developing modified DATs [4].

Delays' impacts remain significant and, the literature, although rich in some areas, still lacks contribution in other areas regarding delays [5]. Delay analysis investigates events using CPM scheduling methods to establish the cause and extent of delays and to resolve construction delay claims through negotiations or legal proceedings [1]. Construction delay claims are a common occurrence in projects. When they arise, they need to be evaluated quickly and managed efficiently. Delays may be classified into four major groups, critical and non-critical, excusable, and non-excusable, compensable, and non-compensable, and concurrent delays [6]. Delay analysis is conducted to evaluate the expected effect of a delay and to determine the amount to be claimed for whether as extension of time or as liquidated damages associated with that additional time.

According to AACE's "Recommended Practice on Forensic Schedule Analysis", RP 29R-03, there are several techniques to perform the delay analysis; however, it breaks the methods into four major families: The As-Planned versus As-Built (As-Planned vs. As-Built/MIP 3.2), the Contemporaneous Period Analysis (CPA/MIP 3.5, sometimes commonly called the "Windows" method), the Retrospective Time Impact Analysis (TIA/MIP 3.7), and the Collapsed As-Built (CAB/MIP 3.9).

Different delay analysis techniques load the delays on a project schedule, a baseline programme, or a schedule reflecting the actual chronology of activities that were performed on site, an as-built schedule. A sound baseline programme should follow the contract requirements, contain all the scope of works, and be validated through an agreed upon set of standards such as AACE Source Validation Protocol

or DCMA 14-point assessment. It is essential for a forensic schedule analyst to have professional judgement that is derived from both experience and knowledge.

## 2 Objectives

This research aims at evaluating different factors that influence the selection of the DAT from an Egyptian construction industry perspective. This was further elaborated by comparing the data of similar research in different countries to provide further insight to the global construction industry regarding the factors influencing the selection of delay analysis techniques. Moreover, this research is an attempt to transform from a siloed approach in selecting delay analysis technique based on experts' opinions to a more quantifiable approach by building a model that uses mathematical equations to score five different DATs. Although the selection of most suitable delay analysis technique (DAT) is addressed in the literature, only one paper was identified that utilizes experts' opinions in developing a decision-making mechanism by simple additive weighting method [7]. Having such a model can guide parties with no experience on selection criteria to guide them into selecting an appropriate DAT for their claim. In using coherent criteria for the selection, the model eliminates subjectivity to a great extent which results in less room for disputes. This model can serve as base for developing a more advanced and user-friendly model that uses artificial intelligence in determining which method should be selected.

## 3 Literature Review

Generally, the techniques can be grouped under non-critical path method (CPM)-based techniques and CPM-based techniques. This research focuses on five primary CPM-based DATs which have been identified in the literature (e.g. [5, 8–13]). Those are impacted as-planned, collapsed as-built, as-planned versus as-built, and time impact analysis. Window analysis technique is an arguable derivative of the primary techniques, but since it is widespread it has become recognized as a primary method on its own. Each of the primary techniques has acquired different names over the course of its application [9]. Secondary derivatives are highlighted through the work of Keane (2008). It shows that impacted as-planned to have two derivatives, while TIA can be applied in four different ways, collapsed as-built in three different ways and as-planned vs as-built in five different ways. It is important to highlight that different delay analysis techniques (DATs) applied to the same project's delays provide different results [9, 14].

In addition to the published articles, there are four recent industry references and standards on delay analysis, namely the Recommended Practice No. 29R-03 (RP-FSA) of the American Association of Cost Engineers (AACE) [15], Delay and Disruption Protocol of the Society of Construction Law (SCL) 2nd edition [16],



ASCE/ANSI Standard “Schedule Delay Analysis” [17], and Construction Schedule Delays by Dale and D’Onofrio [18].

From an Egyptian perspective, Marzouk et al. [3] identified and ranked 22 sub-causes of delays. Those 22 sub-causes were grouped under three main categories, design development, workshop drawings preparation and/or approval; and project parties’ changes. A thorough ranking of delay causes after the Egyptian revolution was conducted by Aziz [19] who identified ninety-nine (99) causes. He categorized them into nine major categories, namely consultant, contractor, design, equipment, external, labour, material, owner, and project-related categories.

In their work, Deep et al. (2017) highlight critical delaying variables from research works conducted in various countries between 1971 and 1998. They conducted a survey from the Indian perspective in which they categorized fifty-one delay factors into seven of the categories highlighted by Aziz [19], and they did not include the external and project-related categories. Within the UAE construction industry, Faridi [2] conducted a survey that included 93 construction professionals. It was noted the most significant causes of delay were (ranked in order of significance) (i) preparation and approval of drawings, (ii) slowness of the owner’s decision-making process; (iii) financing by contractor during construction; (iv) shortage of manpower; (v) inadequate early planning of the project; (vi) skill of manpower; (vii) non-availability of materials on time; productivity of manpower; (ix) poor supervision and poor site management; (x) obtaining permit/approval from the municipality/different government authorities; (xi) unsuitable leadership style of construction/project manager; and (xii) delays in contractor’s progress payment (of completed work) by owner.

Delays may be categorized as excusable, non-excusable, compensable, and non-compensable. When demonstrating that a delay is both excusable and compensable, the delay must be shown to be critical, by reference to a reliable critical path analysis. The tests which must be satisfied for a delay to be considered excusable and compensable are described and discussed in the work of Keane (2008).

Different factors influencing the selection of the most appropriate DAT have been highlighted in several research. Many of the scholars have based their research survey and/or interviews on a country’s construction industry practitioners. Through the literature review, a list of most dominant factors was compiled to be included in a survey. A secondary aim of this research is to put several research works in comparison with examine the distinction, if any, between common and civil law countries regarding the factors impacting the selection of DAT.

In the second edition of SCL delay and disruption protocol, nine factors that affect the selection of DAT are identified. Arditi and Pattanakitchamroon [20] identified four factors towards the selection of DAT. AACE [15] has identified 11 factors that influence the selection. Braimah and Ndekugri [8] identified 18 factors that influence the selection of DAT. Those same factors were used in the works of Enshassi and Jubeh [21], Perera and Sudeha [22] and Parry [23] with the exception that Perera and Sudeha included an additional seven factors. Each of the four works conducted separate surveys and interviews and arrived at different rankings of the factors from the different parties’ perspectives.

On the one hand, those sources along with SCL delay and disruption protocol, AACE and Arditi and Pattanakitchamroon [20] can be considered to represent the views of common law countries. On the other hand, civil law countries' perspectives can be perceived through the works of Perera et al. [7] in which 13 factors which were expanded to 23 sub-factors that affect the selection were identified, Abdelhadi [24] who identified ten factors in selecting the DAT, Abouorban [5] who identified 14 factors, and Magdy et al. [25] who identified eight factors. Summary of the factors gathered from each source is represented in Table 1.

It can be noted how the availability of records factor is mentioned in all the 11 sources. This follows that some DATs cannot be utilized in the absence of records. DATs can provide unreliable results when the records are not updated or more prone to interpretation. The project complexity factor shows up in nine of the sources while having the AACE relates the dispute complexity to the selection of the DAT. In five of the sources, project size is mentioned along with project complexity which contributes to the idea that complexity can be perceived as a factor with sub-factors such as project size, project type, interdependency, and interaction between project parts [26]. Perera [7] related complexity to mechanical and electrical works difficulty. While this can be true for one project type or one specific contractor, it cannot be generalized. In seven of the sources, time availability for conducting the analysis is established as a factor that contributes to the selection. Some DATs like impacted as-planned take less time in comparison with more sophisticated ones like window analysis [5, 11]. In eight of the sources, budget availability to conduct the analysis is established as a factor. Of those sources, Arditi [20] and Abouorban [5] set the time and cost availability as one factor assuming that they carry equal weights. Some of the sources referred to the time and cost factors as those required, while others referred to them in terms of availability. Both are different ways to arrive at the same conclusion. Nine sources clearly established that the skill of the analyst is a factor that contributes to the selection. Having a more capable analyst makes more DATs options for selections.

Contractual requirements factor is set as a major factor that affects the selection. Although cases in the literature that showcase the overturning of a DAT selection clause were not found, in some cases the selected method can be prone to legal challenges in courts [27]. El Nemr elaborates on four legal challenges that can be raised against the use of TIA which are the accuracy of the as-planned schedule and updates, the analysis can be considered hypothetical, use of prospective TIA after-the-fact, and susceptibility of a TIA to manipulation. Livengood (2017) highlights that retrospective TIAs have suffered certain challenges in US courts, which begs the question whether the selection of DAT should be influenced by law systems. Moreover, Keane (2008) states that many of today's larger international engineering and construction projects and the contracts specify which method of analysis will be used. Although Parry [23] suggested that one DAT, namely, window analysis can be fit for all cases, several researches like the ones mentioned above show that one DAT is not fit for all cases or claims.

**Table 1** DAT's selection factors in common versus civil law perspectives

	Civil law					
	Common law	UK	USA	UAE	Egypt	Egypt
Arditi [20]	UK, Gaza, Sri Lanka	America	UK	UAE	Egypt	Egypt
	Brainah [18] Enshassi [21] Parry [23] (18 factors) Perera [22] (25 factors) <sup>(b)</sup>	AACE [15] (11 factors)	SCL 2nd Edition, Delay and Disruption Protocol [16] (10 factors)	Abdelhadi et al. [24] (10 factors)	Perera et al. [7] (23 factors)	Magdy et al. [25] (8 factors)
Availability of information	Updated programme availability; records availability	Source data availability and reliability	Nature, extent, and quality of records available	Available records	Availability of other records (e.g. daily records)	Availability of programme updates; availability of project records
	Complexity of the project; size of project	Complexity of the dispute	Nature of the project	Project complexity	Complexity of the project; size of the project; value of the project; obscurity and sophistication of issues in prolongation claims	Project type (size, location, complexity)
Time-cost effort	Time availability for delay analysis *	Time allowed for forensic schedule analysis	Time available		Concern for time to be spent for analysis	Time and cost accompanied required to carry out the analysis
	Skills of the analyst	Expertise of the forensic schedule analyst and resources available	Implied in the text	Skills of the analyst	Experts skills	Availability of skilled resources for carrying out delay analysis
	Type of contract	Contractual requirements	Relevant conditions of contract	Contractual requirements	Analysis method defined in the contract	Contract particular conditions

(continued)



**Table 1** (continued)

Common law				Civil law			
	UK, Gaza, Sri Lanka	America	UK	UAE	UAE	Egypt	Egypt
	Time of the delay					Time of delay occurrence	Time of occurrence
Time of analysis				Status of project	Status (prevailing stage of the project)		
	Nature of baseline programme		Nature, extent, and quality of the programme information available		Baseline program type		
	Baseline programme availability				Baseline program availability		
	The other party to the claim			Attitude of the opponent party			
	Applicable legislation	Legal or procedural requirements; custom and usage of methods on the project or the case					
	Nature of proof required <sup>a</sup> ; public project or private project <sup>a</sup> ; type of project <sup>a</sup> ; simplicity <sup>a</sup> ; fast track project or not <sup>a</sup> ; level of exposure of the responsible party <sup>a</sup>			Ownership of the float	Concurrency and float ownership defined in the contract; high quality of transparency (clearly established causation); as-built periodical updates of program; as-built periodical updates of program (mutually agreed); need to illustrate isolated delay effects; need of sequential (chronological) analysis	Delay type; party carrying out the analysis; materialization of delay impact; reliability of project schedules; reasons behind delay	Adequacy and skills of project management office (PMO) personnel; availability of a robust construction program

<sup>a</sup> Wordings are used as is from their corresponding reference and interpreted to the best ability of the author

Six of the seven sources that are based on common law countries set the dispute resolution forum in which the delay analysis arguments are going to be presented by both parties as a factor in selecting the appropriate DAT. When considered in the light of the work of El Nemr [27], this can relate to how important it is to remember how presenting a delay analysis in front of a judge in litigation can be different from presenting it to the engineer/consultant for a claim. Also, how important it can be to survey the previous rulings of a certain country regarding issues as concurrency, float ownership, etc. None of the civil law sources include dispute resolution form as a factor. When considered along with time of delay analysis factor, time of delay occurrence factor can tell whether the analysis is occurring retrospectively or prospectively.

Capabilities of the method were mentioned in two sources, while purpose of delay analysis was mentioned in another two. Although it is not clearly stated, it was inferred that purpose of delay analysis had a direct correlation with the capabilities of each method. In total, six sources considered duration of projects as a factor in their works. Perera [7] referred to amount of time claimed to be a factor in the selection. Perera's reasoning is that the use of TIA or CAB requires high expertise which would not be economical and time efficient if the claimed value/time is not significant enough to justify the cost of delay analysis. The same reasoning applies to the amount in dispute which was set as a factor in three common law sources, while it was mentioned in [7] a civil law source as a factor.

Six of the sources include time of delay occurrence (presumably referring to impact) as a factor. Three other sources include time of delay analysis or current status (phase) of project (assuming current phase means that delay analysis is conducted at this instance). Two sources remain silent on this factor, and no sources mention both factors together to draw the relation of having retrospective or prospective analysis; however, Abouorban [5] includes materialization of delay impact as a factor which deals with this aspect.

Concurrency of delay events and its relation to the DAT selection was referred to as a separate factor in the four civil law sources, while it was not given the same level of clarity or focus when referred to as an example illustrating the nature of delay events factor in the common law sources. Although several research tackles concurrent delay events in common law cases, the need to properly show concurrency of delay events is not considered in the common law sources as a factor in the selection of DAT. Of the four civil law sources, two sources highlighted that float ownership affects the selection as well. According to Arditi and Pattanakitchamroon (2005), the varied positions concerning who owns float can influence the result of delay analysis. In their article, they showcase how different float ownership scenarios can lead to different results. Whether or not a specific scenario of float ownership affects the selection of DAT is not addressed in their work. Perera [7]<sup>1</sup> highlights that float ownership does not affect the outcome of the analysis when APvAB is used. Both

---

<sup>1</sup> It is worth noting that in Perera [7], a possible revision might be necessary as the overall ranking of the factors does not reflect both parties' rankings. It is the same as those of the contractor ranking and weights.

nature of delaying events and number of delaying events are mentioned as factors in eight sources. Four sources include the attitude of the other parties to the claims in their lists.

## 4 Methodology

After conducting the literature review to highlight as many research that tackles the factors influencing the DAT selection from both common and civil law perspectives, it is believed that all the researches that ranked Braimah's 18 factors up to date are included in this research.

Next step was to conduct a few structured interviews with some construction experts to filter the factors identified through the literature review. The filtration aimed mostly at excluding factors that are not dominant in the literature or consistently scored relatively low. Also, factors with impacts that were hard to quantify were not included in the survey. Also, the filtration was done so that conducting a survey to rank the factors is time efficient as surveying 40 factors can be time consuming.

The aim of both, the two surveys and the interviews, was to come up with factors' weights and sub-weights for the possible values in each factor so that a decision support tool can be developed.

### 4.1 *Semi-structured Interviews*

Through the literature review, forty factors were identified. Rankings of factors used in Braimah [8], Enshassi [21], Perera [22], Parry [23], Perera [7], and Abouorban [5] were set for each factor as given in Table 2. **Average** of rankings for each factor was calculated, and combined overall rankings were given. Through semi-structured interviews, experts gave their opinions on the factors, eliminating 19 of the factors out of the selection.

The following factors were included in the survey as the experts to impact the selection of DATs by the experts. Also, they were mentioned in several sources and received varying rankings:

- |                                    |  |
|------------------------------------|--|
| 1. Records availability            | 12. Nature of delaying events                |
| 2. Baseline programme availability | 13. Number of delaying events                |
| 3. Nature of baseline programme    | 14. Amount in dispute                        |
| 4. Updated programme availability  | 15. Cost of using the technique              |
| 5. Complexity of the project       | 16. Time availability for delay analysis     |
| 6. Size of project                 | 17. Time of the delay occurrence             |
| 7. Skills of the analyst           | 18. Status (prevailing stage of the project) |

- |  |                                |
|--|--------------------------------|
| 8. Type of contract (contract particular conditions)   | 19. Other parties to the claim |
| 9. Concurrency and float ownership defined in contract | 20. Dispute resolution forum   |
| 10. Need of showing concurrent delay/mitigation        | 21. Applicable legislation     |
| 11. Reliability of project schedules                   |                                |

The following factors were not included in the survey as the experts did not identify them as impactful factors in the DAT selection. Also, each of these factors was mentioned in only one source and had a relatively low ranking.

- |   |  |
|---|--|
| 1. Amount of cost (of prolongation) claimed | 7. Level of exposure of the responsible party  |
| 2. Amount of time claimed                   | 8. Reasons behind delay                        |
| 3. Value of the project                     | 9. Need of sequential (chronological) analysis |
| 4. Public or private project                | 10. Obscurity of issues in prolongation claims |
| 5. Simplicity                               | 11. Reason for delay analysis                  |
| 6. Fast track project or not                | 12. Need to illustrate isolated delay effects  |

Seven other factors were removed due to other reasons. Type of project factor is mentioned clearly in one source only in which it ranked significantly low. Abouorban [5] included it with other aspects as one factor. When asked about it in the interview round, three experts deemed it unfit as a factor, so it was decided to not be included in the survey. Duration of the project and applicable legislation factors were mentioned in several sources but given very low rankings in all sources. Duration of the project factor was not included in the survey; however, according to Parry [23], the selection of the appropriate delay analysis methodology is influenced by common law. English common law supports the windows analysis methodology. When considered in the light of the point above-mentioned by El Nemr [27], it was determined that such a factor is worth including in the survey to gain further insight from the Egyptian perspective, a civil law country. Party carrying out the delay analysis factor was mentioned in one source only. Although the experts identified this issue as a factor that influences the selection of DAT. This factor accounts for the bias which the party carrying out the analysis has towards the different DATs, so it was decided that it should not be included in the tool or future research as the primary purpose of this research is to create a tool that overcomes this bias and personal judgement. High quality of transparency and nature of proof required factors were determined to reflect similar aspects to reliability of project schedules factor, and thus, only the reliability of project schedules factor was included in the survey. Similarly, as-built periodical



**Table 2** Overall rankings of factors gathered from literature

Factor	References							Average	Ranks
	Braimah [8]	Enshassi [21]	Perera [22]	Parry [23]	Perera [7]	Abouorban [5]	Al Ghory (2015) (Braimah's)		
Records availability	1	1	1	5	1	2	1	1.83	1
Baseline programme availability	2	2	6	2	4	–	2	3.20	2
Updated programme availability	5	4	1	4	3	–	3	3.40	3
Nature of baseline programme	4	12	1	1	2	–	4	4.00	4
Party carrying out the analysis	–	–	–	–	–	4	8	4.00	4
High quality of transparency (clearly established causation)	–	–	–	–	5	–	–	5.00	6
As-built periodical updates (mutually agreed)	–	–	–	–	5	–	–	5.00	6
Complexity of the project	7	8	6	7	13	3	11	7.33	8
Type of contract	11	11	13	3	7	1	9	7.67	9
Skills of the analyst	8	2	1	14	11	10	7	7.67	9
Concurrency and float ownership defined in the contract	–	–	–	–	8	–	–	8.00	11
Reliability of project schedules	–	–	–	–	–	8	–	8.00	11
Nature of the delaying events	9	6	11	6	–	8	5	8.00	11
Need of showing concurrent delay/mitigation	–	–	–	–	9	11	–	10.00	14
Need to illustrate isolated delay effects	–	–	–	–	10	–	–	10.00	14
Number of delaying events	6	14	14	8	17	5	12	10.67	16
Amount in dispute	3	10	6	14	21	–	10	10.80	17
Time availability for delay analysis	–	–	–	–	20	7	–	13.50	18

(continued)

**Table 2** (continued)

Factor	References						Average	Ranks
	Brainmah [8]	Enshassi [21]	Perera [22]	Parry [23]	Perera [7]	Abouorban [5]	Al Ghory (2015) (Brainmah's)	
Materialization of delay impact	–	–	–	–	–	11		11.00
Reasons behind delay	–	–	–	–	–	11		11.00
Time of the delay	14	12	17	8	–	6		11.40
Cost of using the technique	12	15	6	10	19	7	19	11.50
The other parties to the claim	17	16	1	14	–	–	13	12.00
Need of sequential (chronological) analysis	–	–	–	–	12	–		12.00
Size of project	15	5	12	14	23	3	15	12.00
Type of project	–	–	22	–	–	3		12.50
Reason for the delay analysis	10	7	20	14	–	–	14	12.75
Duration of the project	16	8	14	14	18	11	16	13.50
Obscurity and sophistication of issues in prolongation claims	–	–	–	–	14	–		14.00
Dispute resolution forum	13	17	14	14	–	–	18	14.50
Status (prevailing stage of the project)	–	–	–	–	15	–	6	15.00
Value of the project	–	–	–	–	15	–		15.00
Applicable legislation	18	18	18	14	–	–	17	17.00
Nature of proof required	–	–	19	–	–	–		19.00
Public project or private project	–	–	21	–	–	–		21.00
Amount of cost (of prolongation) claimed	–	–	–	–	21	–		21.00
Amount of time claimed	–	–	–	–	22	–		22.00
Simplicity	–	–	23	–	–	–		23.00
Fast track project or not	–	–	24	–	–	–		24.00
Level of exposure of the responsible party	–	–	25	–	–	–		25.00

updates (mutually agreed) factor reflects the same aspect that updated programme availability reflects. It was decided that this factor should not be included in the survey with the assumption that updated programme availability entails agreement between parties. Materialization of delay impact factor is mentioned in one source only. It was decided this factor should not be included in the survey with the understanding that it was reflected through other factors related to time of delay occurrence and time of delay analysis (status of the project is assumed as the phase at which the delay analysis takes place). Experts were asked about two other factors that were not included in literature. First factor was contract type as in lump sum, unit price, and cost-plus contracts. Second factor was country's arbitration law if the analysis is presented in front of an arbitral tribunal. Both factors were deemed to have no influence over the selection of the DAT.

During the interviews, experts were asked whether contracts specify which method of analysis will be used. Only one expert stated that he has encountered one project implemented in Egypt where the DAT was predefined in the contract. Two of the experts indicated that it is uncommon for a DAT to be specified in the contract in Egypt, but they believed that in the USA several projects are having TIA as the specified DAT.

Experts in the interview agreed that another party to the claim factor is one that is not often taken into account; however, when current and future business between the parties are of great significance, acceptance of a DAT by one party may be faced with more leniency from the other parties.

During interviews, three experts agreed that there may be a relation between number of delaying events and the time it takes to conduct the analysis and thus relates to the time availability to conduct analysis factor. It was determined that quantifying time availability and complexity factors would suffice rather than having number of delay events as a factor in the model.

## 4.2 *Survey One*

The first survey aimed at asking experts in the field to score each of the factor contribution to the selection of a delay analysis technique. In total, 28 experts, whose major experiences are divided between contractors and consultants, responded to the survey.

The survey questionnaire provided a measuring scale based on a five-point Likert scale, to assess the relative importance of the attributes in a ranking order. For each factor, the values given by all experts were added, divided by the maximum value a factor could have obtained which equals 140, transformed to percentages and ranked as given in Table 3.

Compared to the combined ranking of the factors based on several research works showcased in Table 2, the top three factors remain almost the same with the exception that the most important factor is determined to be whether a contract condition specifies one or not. Although several works highlight how this factor is a major

**Table 3** Filtered factors after ranking according to survey

Factor	Weight %	Ranking
Type of contract (contract particular conditions)	97.14	1
Records availability	96.43	2
Baseline programme availability	95.71	3
Updated programme availability	95.00	4
Nature of baseline programme	95.00	5
Complexity of the project	93.57	6
Status (stage at which delay analysis takes place)	92.86	7
Reliability of project schedules	92.14	8
Dispute resolution forum	89.29	9
Nature of the delaying events	87.86	10
Skills of the analyst	87.14	11
Number of delaying events	86.43	12
Time availability for delay analysis	85.71	13
The other party to the claim	80.00	14
Time of the delay occurrence	76.43	15
Amount in dispute	73.57	16
Concurrency and float ownership defined in the contract	72.14	17
Need of showing concurrent delay/mitigation	71.43	18
Applicable legislation	71.21	19
Size of project	68.57	20
Cost of using the technique	65.00	21

one in the selection, its rank was never the highest. Complexity of the project factor achieved a higher rank than it was in the overall ranking of the literature. This factor importance was highlighted by Enshassi [21] who established that the selection of APvAB as the most common DAT in Gaza Strip was due to the low complexity nature of projects being conducted there. Status (stage at which delay analysis takes place) factor achieved a high importance which aligns with Abouorban [5] study that at different project stages, different parties would choose different DATs. This was later emphasized through results of survey two.

### 4.3 Survey Two

The second survey asked experts to score the contribution of each possible criterion to each of the factors regarding the selection of each of the five DATs. Some criteria were considered based on the work of Perera [7] and Abouorban [5]. Other criteria were established based on the semi-structured interviews. Two factors were not used in the

development of the tool. Concurrency and float ownership defined in contract and applicable legislation factors were removed as several experts could not distinguish how would the factor influence the selection of DAT (Table 4).

**Table 4** Criteria set for each factor

Factor	Criteria
Type of contract (contract particular conditions)	Contract does not specify a DAT Contract specifies IAP Contract specifies APvAB Contract specifies CAB Contract specifies TIA Contract specifies WA Contract specifies a retrospective tech Contract specifies a prospective tech
Records availability	Frequently Occasionally Rarely
Baseline programme availability	Available Not available
Updated programme availability	No updated programme available Updated programme but not most recent Most recent programme available
Nature of baseline programme	CPM Non-CPM
Complexity of the project	Simple (typical project undertaken before) Complex (specialized project not familiar)
Status (stage at which delay analysis takes place)	Preconstruction phase (i.e. design) Amid of construction phase Close-out phase (i.e. testing)
Reliability of project schedules	Reliable Not reliable
Dispute resolution forum	Engineer/project management office DAB/arbitration/court
Nature of the delaying events	Non-excusable Excusable/compensable Excusable/non-compensable
Skills of the analyst	Novice Intermediate Advanced
Number of delaying events	Few ( $\leq 5$ ) Moderate ( $> 5$ to 25) Many ( $> 25$ )
Time availability for delay analysis	Time is not a constraint Time is limited

(continued)

**Table 4** (continued)

Factor	Criteria
The other parties to the claim	Lenient/seeks fair judgement Not lenient/seeks own interest
Time of the delay occurrence	Preconstruction phase (i.e. design) Amid of construction phase Close-out phase (i.e. testing)
Amount in dispute	Small (< 5 million EGP) Moderate (> 5 to 50 million) Large (> 50 million)
Need of showing concurrent delay/mitigation	Showing concurrency is required Showing concurrent is not required
Size of project	Small (BUA: $\leq 5000 \text{ m}^2$ ) Medium (BUA: > 5000 to 50,000 $\text{m}^2$ ) Large (BUA: > 50,000 $\text{m}^2$ )
Cost of using the technique	Cost is not a constraint Cost is limited

## 5 Model Design

Factor	Selected criterion	Scores				
		IAP	CAB	APvAB	TIA	WA
Type of contract (contract particular conditions)	Contract does not specify a DAT	0.000	0.000	0.000	0.000	0.000
Records availability	Frequently	0.011	0.012	0.012	0.012	0.012
Baseline programme availability	Available	0.015	0.008	0.015	0.015	0.015
Updated programme availability	Most recent programme available	0.008	0.011	0.008	0.011	0.012
Nature of baseline programme	CPM	0.010	0.010	0.013	0.011	0.011
Complexity of the project	Complex (specialized project not familiar)	0.013	0.012	0.009	0.012	0.011
Status (stage at which delay analysis takes place)	Close-out phase (i.e. testing)	0.008	0.010	0.011	0.013	0.012
Reliability of project schedules	Reliable	0.008	0.011	0.010	0.012	0.012
Dispute resolution forum	Engineer/project management office	0.008	0.006	0.008	0.011	0.010
Nature of the delaying events	Excusable/non-compensable	0.010	0.011	0.011	0.011	0.010
Skills of the analyst	Advanced	0.009	0.013	0.009	0.011	0.011

(continued)

(continued)

Factor	Selected criterion	Scores				
		IAP	CAB	APvAB	TIA	WA
Number of delaying events	Many (> 25)	0.008	0.009	0.009	0.009	0.009
Time availability for delay analysis	Time is not a constraint	0.007	0.009	0.008	0.008	0.008
The other parties to the claim	Lenient/seeks fair judgement	0.009	0.009	0.011	0.011	0.010
Time of the delay occurrence	Amid of construction phase	0.009	0.008	0.009	0.010	0.009
Amount in dispute	Moderate (> 5 to 50 million)	0.009	0.008	0.009	0.009	0.010
Need of showing concurrent delay/mitigation	Showing concurrency is required	0.006	0.007	0.009	0.009	0.010
Size of project	Medium (BUA: > 5000 to 50,000 m <sup>2</sup> )	0.008	0.009	0.009	0.009	0.009
Cost of using the technique	Cost is not a constraint	0.007	0.009	0.008	0.008	0.008
<b>Total score</b>		<b>0.162</b>	<b>0.172</b>	<b>0.175</b>	<b>0.190</b>	<b>0.188</b>
<b>DAT selected</b>		<b>TIA</b>				

## 6 Conclusion

The mixed-method approach in collecting data through semi-structured interviews as well as two thorough surveys allowed the research to benefit from the expert's judgements in selection of DATs.

The inputs from the case study applied by Perera [7] were used with the weights given by the experts in this research work. Similarly, time impact analysis technique was the one with the highest total score and thus the one that should be utilized for that case study. This outcome aligns with Abouorban [5] conclusion that TIA is the DAT suitable throughout and towards the end of the project based on the survey conducted from an Egyptian construction industry perspective. Still, a close second to the TIA was window analysis technique which according to many experts is considered a very worthy technique which is applied by many practitioners in the Egyptian construction industry. As suggested by Parry [23], the choice of methodology is overwhelmingly in support of the windows analysis technique, which he proposes to be the best technique that should be adopted in any delay analysis. Different from Perera [7], APvAB had the third highest score instead of CAB. According to Enshassi [21], there was a consent that the as-planned versus as-built is the most used method in Gaza Strip. The reason for this is the simplicity of this method since it relies on common sense and the environment of construction projects in Gaza Strip where approximately no

complex projects were executed that could force practitioners to resort to the other DAMs. Still the project applied in this case study was a complex one and thus the reasoning behind choosing APvAB may not apply, however, it ranked third rather than first or second. IAP ranks fifth which aligns with the characteristics of IAP and the characteristics of this case study. It is worth noting that IAP ranking fifth aligns with Arditi and Pattanakitchamroon [20] conclusion that impacted as-planned method is the least favoured method as it has theoretical flaws, however, during interviews, experts confirmed Abouorban [5] conclusion that IAP is a common DAT used in the construction industry despite its major drawbacks. Conclusions of this research can be summarized in the below points:

- The developed tool can help delay analysts, arbitrators, contract administrators, and other parties support and validate their DAT selection.
- The tool was applied on the same case study in the work of Perera [7]. The highest scored DAT was time impact analysis in both works.
- Namely, records availability, baseline programme availability, and updated programme availability factors are the most influential three factors in the selection of the DAT. The exception according to this research and the one conducted by Abouorban [5] is that the most important factor is whether a contract includes a provision that specifically requires a specific DAT. Both works are based on Egyptian industry construction practitioners. A question remains what happens if the contract specifies a DAT that is not suitable for the analysis.
- Works from common law countries highlight that the dispute resolution forum is a factor in selecting the appropriate DAT. It is worth noting that presenting a delay analysis may require a different strategy in front of a judge in litigation than to the engineer/consultant for a claim. Also, how important it can be to survey the previous rulings of a certain country regarding issues as concurrency, float ownership, etc. Further analysis on differences between the factors highlighted in common and civil law works can provide more insight on this aspect.
- To the knowledge of the researcher, available guidelines on delay analysis selection are mostly developed by entities based in common law systems.
- No one DAT is fit for all cases or claims as indicated through the discussion with the experts as well as the several research mentioned above.
- It is worth noting that the author believes that project complexity factor has sub-factors that contribute to it like project size. This may be a point worth examining on its own.

As highlighted by Perera [7], tools such as the one in this research are mainly based on experts' psychological constructs. Thus, further research is recommended to ensure continuous improvement of the tool in its database of experts. Also, the findings of this research should be validated through a series of case studies to confirm the findings of this research and to provide a more reliable ranking to guide different parties on how to select the most appropriate DAT.



## 7 Future Work

As highlighted by Magdy et al. [25], artificial intelligence techniques such as artificial neural networks (ANNs) can be employed to develop a more advanced decision support system to the selection of DAT. From both surveys conducted in this research, a scoring system will be set to give a score to each DAT when answers to the factors are set. With the proposed factors and criteria established from this research, there can be millions of scenarios that would result in selecting one of the five DATs. From this huge set, a considerable training set can be extracted to train the model on the selection. Another set can be extracted to test the model, however, as a matter of validating the model, several cases from literature can be extracted to test the model against them.

This research utilized the rankings of previous research to provide an overall ranking of the factors found in the literature. By comparing weights assigned to factors based on number of responses of contractors to number of responses of consultants, a more thorough analysis of the works found in literature along with the survey conducted in this research may provide further insight on the DAT selection from parties' perspectives.

**Acknowledgements** This research is part of an MSc thesis for which my family, university, and advisor are to be thanked. The authors of the sources that were used in the comparison part of this research are immensely appreciated for their work as it paved way for this humble contribution to the literature on the delay analysis techniques selection factors topic.

## References

1. Chambers RL (2017) Methods of forensic schedule delay analysis—pros and cons. Smith Currie, 13 Mar 2017, [www.smithcurrie.com/publications/common-sense-contract-law/methods-of-forensicschedule-delay-analysis-pros-and-cons/](http://www.smithcurrie.com/publications/common-sense-contract-law/methods-of-forensicschedule-delay-analysis-pros-and-cons/). Web 10/02/2022
2. Faridi A, El-Sayegh S (2006) Significant factors causing delay in the UAE construction industry. *Constr Manag Econ* 24:1167–1176. <https://doi.org/10.1080/01446190600827033>
3. Marzouk M, El-Dokhmasey A, El-Said M (2008) Assessing construction engineering-related delays: Egyptian perspective. *J Prof Issues Eng Educ Pract* 134(3):315–326
4. Braimah N, Ndekugri I (2009) Consultants' perceptions on construction delay analysis methodologies. *J Constr Eng Manage.* [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000096.1279-1288](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000096.1279-1288)
5. Abouurban H, Hosny O, Nassar K, Eltahan R (2018) Delay analysis techniques in construction project. In: *Proceedings of the Building Tomorrow's Society*, Fredericton, Canada
6. Trauner T, Manginelli W, Lowe J, Nagata M, Furniss B (2009) Types of construction delays. <http://doi.org/10.1016/B978-1-85617-677-4.00002-7>
7. Perera N, Sutrisna M, Yiu T (2016) Decision-making model for selecting the optimum method of delay analysis in construction projects. *J Manage Eng* 32(5):04016009
8. Braimah N, Ndekugri I (2007) Factors influencing the selection of delay analysis methodologies. *Int J Project Manage* 26(8):789–799
9. Braimah N (2013) Construction delay analysis techniques—a review of application issues and improvement needs

10. Farrow T (2001) Delay analysis—methodology and mythology. Society of Construction Law
11. Keane PJ, Caletka AF (2008) Analysis of construction delays. Blackwell Publishing, Hong Kong
12. Palaneeswaran P, Kumaraswamy M (2008) An integrated decision support system for dealing with time extension entitlements. *Autom Constr* 17(4):425–438
13. Zack J Jr (2006) Delay and delay analysis—isn't it simple? In: 1st CEC & IPMA Global Congress on project management, Ljubljana, Slovenia, 26 Apr 2006
14. Al-Gahtani KS, Mohan SB (2011) Delay analysis techniques comparison. *J Civ Eng Archit* 5(8):740–747
15. AACE International (2011) Recommended practice No. 29R-03, <http://www.aacei.org> (18 June 2012)
16. Society of Construction Law (SCL) (2017) Protocol for determining extensions of time and compensations for delay and disruption. SCL, Burbage, UK
17. “Schedule delay analysis.” Standards, American Society of Civil Engineers, 8 Aug 2017. <https://doi.org/10.1061/9780784414361>
18. Dale WS, D’Onofrio RM (2018) Construction schedule delays. Thomson Reuters
19. Aziz R (2013) Ranking of delay factors in construction projects after Egyptian revolution. *Alex Eng J* 52:387–406. <https://doi.org/10.1016/j.aej.2013.03.002>
20. Arditi D, Pattanakitchamroon T (2006) Selecting a delay analysis method in resolving construction claims. *Int J Project Manage* 24(2):145–155
21. Enshassi A, Jubeih AI (2008) Delay analysis methods and factors affecting their selection in the construction industry in Gaza strip
22. Perera K, Sudeha H (2013) A framework to select the most suitable delay analysis technique for building construction through a consideration of utility factors. *Bhumi Plann Res J* 3:11. <https://doi.org/10.4038/bhumi.v3i2.16>
23. Parry A (2015) The improvement of delay analysis in the UK construction industry
24. Abdelhadi Y et al (2018) Factors influencing the selection of delay analysis methods in construction projects in UAE. *Int J Constr Manag* 19(4):329–340
25. Magdy M, Georgy M, Osman H, Elsaid M (2019) Delay analysis methodologies used by engineering and construction firms in Egypt. *J Legal Affairs Dispute Resolut Eng Constr* 11. [http://doi.org/10.1061/\(ASCE\)LA.1943-4170.0000293](http://doi.org/10.1061/(ASCE)LA.1943-4170.0000293)
26. Wood H, Gidado K (2008) Project complexity in construction. In: COBRA 2008—construction and building research conference of the Royal Institution of Chartered Surveyors
27. El Nemr W, Mohamed HE (2019) Legal and practical challenges to the implementation of the time impact analysis method

# Axiomatic Design-Based Optimization Framework for Factory Logistics Design in Precast Concrete Construction



Shuai Liu, Asif Mansoor, Ghulam Muhammad Ali, Ahmed Bouferguene, and Mohamed Al-Hussein

**Abstract** Precast concrete construction has shown great potential as an alternative to traditional in situ construction methods, as it supports a high level of industrialization, low dependency on labour, and significant benefits in terms of sustainability and environmental performance. In spite of these widely recognized advantages, though, it is still in its infancy compared to other manufacturing sectors. Numerous studies have been conducted to improve management in precast concrete construction through scheduling models, supply chain optimization, information technology applications, and integration of lean. However, the design of the production factories in which most of the work associated with precast concrete construction is carried out has been overlooked in the existing literature in this area. Due to the bulky size of prefabricated products and related parts, their flowing inside the factory requires equipment and manpower. Well-designed in-plant logistics can boost production efficiency and reduce material handling costs (which have been reported to account for a significant proportion of production cost in precast concrete construction). In the research presented in this paper, axiomatic design (AD) is applied to optimize the factory logistics in precast concrete production to achieve the concept of “right part mix to the right place at the right time”. The aim of this paper, then, is to propose a framework that seeks complete optimization via axiomatic design by converting customer attributes into functional requirements and deriving the design parameters based on the functional requirements in a “zigzagging” process. In such a way, factory logistics activities such as parts feeding, material handling, and even the addition of functional departments can be designed and integrated based on the actual production requirements. To the best of the authors’ knowledge, the holistic design of precast concrete construction factory logistics has not been investigated in previous studies; moreover, the evaluation results presented in this paper demonstrate the effectiveness of the proposed optimization framework.

---

S. Liu (✉) · A. Mansoor · G. M. Ali · M. Al-Hussein  
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada  
e-mail: [sl5@ualberta.ca](mailto:sl5@ualberta.ca)

A. Bouferguene  
Campus Saint-Jean, University of Alberta, Edmonton, Canada

**Keywords** Precast concrete construction • Factory logistics design

## 1 Introduction

Prefabricated construction has demonstrated superb production efficiency, construction quality, environmental benefits as well as sustainable performance compared to the traditional cast-in situ construction methods [1–3]. It transferred most of the workload from the uncontrollable site to the factory as a form of industrialization. Today, because of its preponderance over the traditional construction, it has received extensive attention among mature economies around the world including the USA, the UK, Sweden, Australia, and Singapore [4]. In the meantime, for some developing countries, such as China, Malaysia, and Turkey, prefabricated construction has also been strongly promoted as a means of addressing the increased demand brought about as a result of rapid urbanization. Take China as an example, the State Council in China announced in 2016 that the proportion of prefabricated buildings in all new constructions should reach 30% in ten years [5]. Furthermore, precast concrete (PC) structures, as one of the three major prefabricated construction structures (the other two structures: prefabricated steel structure and prefabricated wood structure), has been adopted worldwide because of its superior durability, excellent insulation and fire resistance, remarkable design flexibility as well as outstanding sustainability [6]. Its global market size is estimated to grow from USD 130.6 billion in 2020 to USD 174.1 billion by 2025, with a compound annual growth rate of 5.9% [7]. Thus, it is conspicuous that PC construction has a great development potential and market value in the whole world for the following decades.

PC construction refers to splitting a building into numerous elements (e.g. beams, columns, floor slabs, walls) that are prepared, cast, and cured off-site, normally in a controlled factory environment, using reusable moulds. Then, they are transported to the construction site for installation. As an alternative to the traditional cast-in situ construction methods, it brings most of the workload into a climate-controlled structure with the elimination of rain, dust, cold, or heat, utilizing specialized equipment and machines to produce PC components on the ground rather than at height, implementing repetitive operations of the same products. In this way, PC constructions are expected to be a representation of construction industrialization and to boost the production efficiency and increase the project quality. Nevertheless, the well-documented advantages of PC construction have often been hampered if the production of the PC components is not properly managed [8]. As a more meticulous approach, it requires an advanced level of management, integration, and scheduling. The PC component production should be dynamically balanced with the storage, transportation, and erection to achieve higher efficiency and lower waste. Deviations in any parts of the production process could result in late delivery of the PC components which is unacceptable for a construction project because it may delay the scheduled installation, prolong the project duration as well as increase the labour cost [9]. On the other hand, early delivery due to improper production management

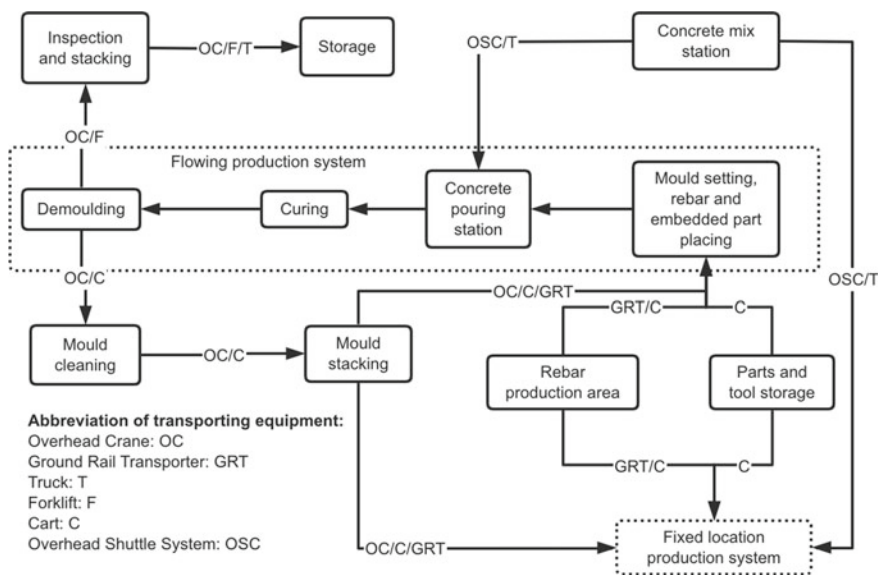
or schedule is neither a good performance. To be more specific, PC components that are delivered early to the construction site require additional storage space which is costly and even intractable for most of the urban projects [10]. Moreover, this will also lead to double handling of the bulky PC components and consequently waste the manpower and equipment. Due to this, especially for some small and medium construction enterprises, they are still not willing to adopt PC construction because of the difficulty to reduce cost, the unfamiliar operation and management mode and the attenuated advantages [11]. Most companies adopting PC, however, have relatively low production efficiency, a disorganized production process, and production imbalance, all of which results in a low production capacity [12].

To summarize, the production and management of PC construction are still in their infancy compared to that in the manufacturing industry. Therefore, comprehensive optimization for production inside the plant becomes a must-do to maintain the superiority of PC construction. Due to the bulky size of the prefabricated units, their work-in-progress (WIP), and related parts (shown in Fig. 1), a well-designed in-plant logistics can boost the production efficiency and help to avoid the production interruption or stoppage which will lead to the aforementioned delivery problems. Furthermore, it can also reduce the material handling cost which is reported as 20–70% of the total operating expenses in a manufacturing system [13, 14]. Figure 2 displays a typical material and WIPs flow by different means of transporting equipment in the PC plant. Before the beginning of the PC production, the required rebar should be manufactured and bound in the rebar production area and transported to the storage. The moulds which have been removed from last production cycle need to be cleaned in the cleaning area and stacked in mould storage. The concrete which is the most important material for PC component production needs to be transported from concrete mix station to the concrete pouring station by truck or automated shuttle system. Right before the production begins, the cleaned mould, finished rebar, embedded parts, and mixed concrete are transported to the corresponding workstations waiting for their treatment. Before, during, and after the production, with the exception of the WIP flows along the automated production line, all procedures require the movement of material, products, and equipment among different areas. Thus, improper design of the in-plant logistics will ultimately lead to inefficient material handling system, improper positioning of the material, long processing time for a specific workstation, etc. However, studies regarding this area are scarce.

Therefore, in this study, an AD-based optimization framework is proposed to provide a complete precast factory design roadmap in terms of parts feeding policy, transport of parts/material, and functional areas design which are the three main components in a factory logistics system. By utilizing AD, a systematic design can be achieved with reduced complexity based on logical and rational thought processes and tools [15]. The requirements from stakeholder or customer can be well conveyed and reflected into the design process. Moreover, by deploying the hierarchical design process in AD, the whole design problem is decomposed into independent or partial independent sub-problems which are easy to handle and require less information. The rest of the paper is organized as follows. Section 2 summarizes the related work.



**Fig. 1** Bulky PC products and WIPs



**Fig. 2** Overview of the product, material, and WIP flow in relation to various transporting equipment in a PC factory

Section 3 introduces basic principles of axiomatic design (AD) and explains the AD-based design framework. Section 4 presents a practical case and the application of the proposed optimization framework to demonstrate its effectiveness, and, finally, Sect. 5 concludes the study.

## 2 Related Work

### 2.1 Construction Factory Logistics Design

Factory logistics design gains great concern because of the increasing variability of product types and demand, and the decreasing profit margin also forces enterprises to further minimize the product-related cost. It is reported that 20–70% of the total operating expenses in a manufacturing system are attributed to material handling cost [13, 14], which is closely related to the in-plant logistics. Therefore, efficient design of the in-plant logistics is vitally important especially for the PC factory case which has complicated material handling system and WIP flow. Some important manufacturing performance indicators, such as lead time and productivity, are also affected by the logistics design to a large extent [16]. In terms of lean production which attracts many researchers and practitioners' attention by its state-of-the-art waste elimination concept, in-plant logistics can be regarded as waste since they do not add any value to the final product [17]. However, these are necessary activities which should not be totally removed. Thus, in order to gain benefits of lean philosophy, comprehensive management of material handling is needed to compensate the "waste". Moreover, the success of lean systems heavily depends on "right part mix to the right place at the right time" which is also the representation of a well-designed logistics system. Abundant studies were dedicated with this respect in the construction industry. However, most of them were focused on the optimization of the construction site. Li et al. [18] developed a dynamic non-centralized model for the allocation of construction components in the storage area. Al Hawarneh et al. [19] proposed a grid layout model to plan the facilities dynamically in the construction site in response to the site changes and with the goal of overall safety and cost. In terms of bi-level planning, Song et al. [20] developed an optimization model to integrate the construction site layout planning with construction material logistics planning to solve the interdependent and conflicts of the layout planner and logistics planner. With regard to the construction plant, Chen et al. [21] proposed a framework to plan the department layout of the automated guided vehicle (AGV)-based prefabricated manufacturing system and design the path of AGVs. However, given the current automation level of most PC factories, this planning framework is not able to achieve large-scale application. Hong et al. [22] developed an algorithm to plan the layout of composite PC components for the in situ production where a temporary production plant is built to eliminate the transportation cost, and this only represents a tiny portion of PC construction since most PC projects require the in-plant production. Recently, Hyun et al. [23] proposed a multi-objective optimization model to design the configuration and layout of continuous modular unit production line with the goal of minimizing the production duration and labour cost. Similarly, the global design of the plant is not considered in this study. Hence, it is essential for researchers and practitioners to think about the holistic design of the in-plant logistics and layout



so that the common issues in a PC factory can be solved, such as production stoppage due to material shortage or excessive material acquisition near the production line which would affect the labour operations.

## 2.2 *Axiomatic Design for Complicated Systems*

AD is a rational framework that leads and guides the designers along the design phases, and the ultimate goal of it is to develop a scientific basis for design and to improve the whole design process by providing the designer a theoretical foundation based on logical and rational thought procedures instead of empiricism, intuition, or trial and error [15]. AD principles help define “what” should be achieved based on the input from customers and “how” it should be achieved in a systematic manner. It is a valuable tool that aids to design a complex system which comprises both independent and dependent activities. By its scientific hierarchical design structure, the main problems are separated into smaller independent and/or partially independent sub-problems in order for the complexity reduction. Besides, AD is capable of providing a complete view point to take the holistic design problems into consideration and offering partial design revision when the requirements of a specific layer in the functional domain change [24]. Although there are other design methods like Taguchi, design for manufacturing and assembly (DFMA), theory of inventive problem solving (TRIZ), etc., in approaches of designing complex systems with the use of algorithms, AD is a general design framework that is easy and effective to apply for both practitioners and researchers [25]. Therefore, AD is widely applied to for manufacturing system design. Ertay and Satoğlu [26] utilized information axiom in AD to select design parameters for new products in hybrid manufacturing systems. Matt [27] tested the validity of AD by incorporating a case study of a mixed-model assembly system and proposed an AD-based approach to monitor and control the time-dependent complexity of the manufacturing system for guidance of management of system re-initialization. Vinodh and Aravindraj [28] proposed an AD-based model for lean manufacturing system design with respect to management responsibility leanness, manufacturing management leanness, workforce leanness, technology leanness, and manufacturing strategy leanness. Han et al. [29] used AD procedures to build a complete virtual cellular manufacturing system by selecting the most suitable techniques and resources to minimize system cost and maximize system efficiency. Chakraborty et al. [30] developed a hierarchy model to evaluate the design alternatives of product re-manufacturability based on AD and fuzzy analytical hierarchy process (AHP). Hager et al. [31] developed a combined production line design that includes different configuration parameters using structured analysis and design techniques (SADT) and AD. Rauch et al. [32] defined the guidelines for the design of flexible and agile manufacturing and assembly system using AD with a special focus on small and medium-sized enterprises (SMEs). To the best of the authors’ knowledge, AD has been widely used and validated by researchers within the system design domain and thus can be applied in PC construction to provide



effective guidance for factory logistics design, whether at the early stages of factory design or for the purpose of optimization during the operational stage.

### 3 Research Methodology

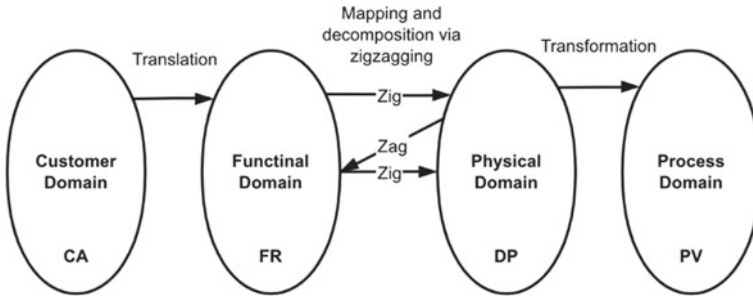
#### 3.1 Principles of Axiomatic Design

There are four domains, or hierarchical levels, in AD, namely the customer domain, the functional domain, the physical domain, and the process domain. The customer domain is the top level of the hierarchy, and it has to do with customer requirements. In the context of a production system, the customer may be an internal party such as the engineering department, management, or workers on the production line. On the other hand, external parties can also be defined as customers [15]. The satisfaction of customer requirements, in turn, is governed by the functional domain in which functional requirements (FRs) are defined. Similarly, to achieve the FRs, the design parameters (DPs) in the next level of the hierarchy (the physical domain) must be defined. The last domain is the process domain, where the corresponding DPs are pursued by defining the process variables (PVs). These PVs represent the bottom layer of the design and may include assigned machines, labour personnel, etc. Figure 3 illustrates these four domains, where a “zigzagging” process links the functional and physical domains in such a manner that the DPs are defined based on the FRs of the same level (i.e. “zig”), whereas the lower-level FRs are derived with the higher-level DPs (i.e. “zag”). This iterative process continues until the design reaches the final stage where further decomposition is not required. Since the aim of this study is to design the overall PC factory logistics, PVs are not necessary, and hence, they are excluded from the process. Furthermore, in AD, two axioms—an “independent axiom” and an “information axiom”—must be satisfied, and satisfying these axioms is the most important procedure in the AD process. The former axiom guarantees the independence of FRs such that the design for each FR will not affect the realization of other FRs, while the latter axiom guarantees that the design of DPs will be representative of the least information content that ensures the clear implementation of the requirements. In terms of manufacturing system design, the information axiom prescribes that the development of DPs follow the “KISS” concept, which is to “keep it simple and smart” [32].

During the mapping and decomposition of FR-DP, the independence axiom should be checked in each iteration. Equation (1) defines the mathematical relationship between FR and DP below:

$$\{\text{FR}\} = [A]\{\text{DP}\}, \quad (1)$$

where  $\{\text{FR}\}$  is the FR vector,  $\{\text{DP}\}$  is the DP vector, and  $[A]$  is the design matrix (DM) expressing the relationships between the various pairings of FRs and DPs. The



**Fig. 3** Four domains of AD and the FR-DP mapping

rows in  $[A]$ , it should be noted, represent FRs, while the columns represent the DPs. “X”, meanwhile, represents the relationship between corresponding FR and DP, and it can be numeric or binary (“0” and “X”). In order for the independence axiom to be satisfied, the DM must be diagonal or triangular. When each FR is satisfied by only one DP, the design equation has an exact solution such that the DPs can be defined in any sequence. The corresponding DM is diagonal, and this is representative of “uncoupled design” [as shown in Eq. (2)]. On the other hand, a triangular DM [shown in Eq. (3)] is reflective of a “decoupled design”, where the DPs are defined according to a specific order that satisfies the independence axiom. However, if the DM is neither diagonal nor triangular [shown in Eq. (4)], the design is considered “coupled” and the independence axiom cannot be satisfied. It should be noted that if a coupled design such as that expressed in Eq. (4) is encountered, the decomposition sequence should be adjusted to determine whether it can be decoupled. If so, the iteration can proceed; if, on the other hand, it cannot be decoupled by adjusting the decomposition sequence, then the functional DPs should be redesigned in such a manner as to satisfy the independence axiom.

$$[A] = \begin{bmatrix} X & 0 & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & 0 & X \end{bmatrix} \quad (2)$$

$$[A] = \begin{bmatrix} X & 0 & 0 & 0 \\ X & X & 0 & 0 \\ X & 0 & X & 0 \\ X & 0 & X & X \end{bmatrix} \quad (3)$$

$$[A] = \begin{bmatrix} X & X & 0 & 0 \\ 0 & X & X & X \\ X & 0 & X & X \\ X & X & 0 & X \end{bmatrix} \quad (4)$$

3.2 AD-Based Optimization Framework for PC Factory Logistics Design

As represented in Fig. 4, the aim of this research is to develop a design framework for PC factory logistics via AD. The process begins with identifying the customer attributes (CAs) using a questionnaire and interviews. Then, a literature review is carried out in order to capture a more complete view of “what does the customer really want to and need to achieve”. This step ensures the right input from stakeholders and defines the orientation of the design. Next, the CAs defined in the customer domain need to be mapped into the functional domain in which the corresponding FRs are defined. The realization of the FRs within the functional domain, meanwhile, is governed by the definition of the DPs within the physical domain. In the meantime, the related DM must be constructed in such a manner as to ensure that the design is compliant with the independence axiom (either an uncoupled design or a decoupled design). If not, the mapping from FRs to DPs must be restructured until the desired DM is met. The derived DPs are decomposed into FRs of the lower level to further refine the design purpose, and the defined FRs are mapped again into DPs for physical settlement. The zigzagging process continues until the DPs of the lowest level form a clear and executable design framework. In addition, the dynamic nature of PC factory logistics is considered in the framework since the logistics design is subject to change. Because some of the PC products are tailor-made and require different parts, configurations of production lines are adjustable and new production lines can even be established during operation in response to changing market demand. In this regard, the defined CAs are checked regularly, and if a change within the customer domain results in a revision in corresponding FRs, then the AD procedures are repeated again to arrive at a revised design that captures the changes in the customer domain.

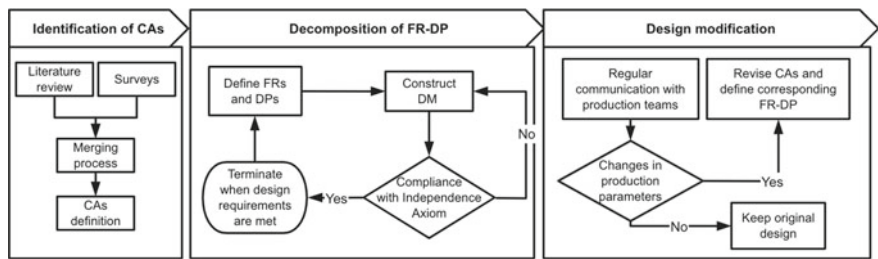


Fig. 4 AD-based optimization framework for PC factory logistics design

## 4 Case Study

To validate the proposed design framework, a site visit was made to Henan D.R. Fair-faced Concrete Technology Co., Ltd., which was established in 2017 and is located in Yanjin, China. As a wholly owned subsidiary of Henan D.R. Construction Group Co., Ltd., it is invested and established to build a whole industry chain for PC buildings. By communicating with the top management team in the company, we found that the “Phase 1” production plant of the company, which produces various PC components, including half-slabs, sandwich shear walls, and precast stairs, suffers from frequent production interruptions due to inadequate production planning and resource allocation. According to the production manager, this is a significant issue in the precast industry, particularly in China, where PC companies often receive significant injections of capital and benefit from favourable policies and incentives, since the adoption of PC construction is a priority of the national government. Companies typically use this capital investment to establish one or more production lines, often with inadequate planning, and with basic factory planning procedures such as the design and location selection of supporting facilities and means of transporting equipment having been implemented hastily or even overlooked entirely. Assistance from the production line supplier, meanwhile, is also of little help, since the supplier tends to have limited knowledge about customer requirements. This will normally cause productivity or performance variations when the production is executed. Furthermore, a large amount of PC companies restructured or renovated their plants after several years’ operation which is costly and time consuming. To this end, we applied the design framework to the “Phase 1” production plant so that the factory logistics can be redesigned based on the current need of the company and the company can choose to adopt some of the design parameters and update the current logistics system accordingly. The following subsections describe the detailed design procedures.

### 4.1 Identification of CAs

Surveys and semi-structured interviews are conducted with the management team of the company regarding their thoughts and design need. Since the overall design goal is relatively clear and the participants of survey and interview are all real practitioners or professionals, literature review is not implemented. That is to say, the management team which acts as the internal stakeholder is defined as customers so that their need forms the CAs. Their input is summarized in Table 1. Subsequently, the merging process is conducted to combine the attributes with the similar design goal. Finally, three CAs are defined, namely “optimize material preparation process” (CA-1, CA-6, and CA-7), “optimize material handling efficiency” (CA-2 and CA-4), and “optimize the factory layout” (CA-3 and CA-5). They are converted to the first-level FRs in the following process.

**Table 1** Summary of CAs from the top management team

Customer attributes (CAs)	Content	Merging results
CA-1	Reduce production line stoppage frequency	CA-1, CA-6, and CA-7
CA-2	Increase the material handling efficiency	
CA-3	Optimize the factory layout	CA-2 and CA-4
CA-4	Reduce the material handling time	
CA-5	Organize the whole system	CA-3 and CA-5
CA-6	Make production site cleaner	
CA-7	Avoid material depletion	

4.2 *Decomposition of FR-DP*

The highest level FR is defined as FR0, “Improve the logistics system performance of phase 1 production plant” which represents the research aim. The corresponding DP is defined as DP0, “Redesign the logistics system of phase 1 production plant”. CAs defined in the previous section are conveyed into the FR1, FR2, and FR3, respectively, in the next hierarchy, namely “improve material feeding mechanism”, “reduce material handling cost and time”, and “increase the efficiency of the whole system”. Based on the defined FRs, the corresponding DPs consist of material feeding policy selection, transporting equipment selection, and additional functional area design. The material feeding policy deals with the arrangement of the material and methods of delivery materials to the point of use [33]; the aim of transporting equipment selection, meanwhile, is to achieve “right vehicle, in the right place, at the right time, with the right method and for the right part” which is substantial for the efficiency and effectiveness of the whole logistics system [34]; additional functional area design incorporates the addition of different functional departments that can support the production and improve the production efficiency. It should be noted that none of these DPs are independent of each other. The application of certain material feeding policy is based on the utilization of specific material handling equipment; the selection of material handling equipment can also affect the functional area design. Therefore, DM should be constructed to determine the sequence of the FR-DP mapping such that the design is compliance with the independence axiom. The DM shown in Equation (5) indicates the design is decoupled, and DP1 (material feeding policy selection), DP2 (transporting equipment selection), and DP3 (additional functional area design) should be defined in sequential order.

$$\begin{bmatrix} \text{FR1} \\ \text{FR2} \\ \text{FR3} \end{bmatrix} = \begin{bmatrix} X & 0 & 0 \\ X & X & 0 \\ X & X & X \end{bmatrix} \times \begin{bmatrix} \text{DP1} \\ \text{DP2} \\ \text{DP3} \end{bmatrix}$$

(5)

The iterative process continues until the lowest levels of DPs have a clear and simple meaning of implementation that does not require any further FRs. The point of termination is usually determined by the customers (the top management team in this case) once their need is totally fulfilled. The detailed decomposition of DP1, DP2, and DP3 is illustrated in the following sections. Due to the scope of the paper, the iterative processes for DP2 and DP3 are not interpreted in detail.

#### 4.2.1 Material Feeding Policy Selection

Based on the current production systems of the phase 1 production plant, continuous supply-based hybrid policy and kitting-based hybrid policy are selected for the flowing production system and stationary production system, respectively. Continuous supply policy is also known as “line side stocking” which is one of the oldest material feeding policies [35]. In this policy, each type of material is distributed from the warehouse and stored in its own container near the production area. The major advantages of this policy are to provide continuous supply for the production [33]. However, if there are various parts and material for the product, the amount of stacking will be too high and occupy a wide area which would result in the productivity loss when the labours spend time on searching the right material [36]. On the other hand, kitting policy can be defined as the activity of feeding required material to the assembly lines in predetermined quantities that are put together in specific containers, even the internal layout of the container can be designed by the kitting workers [37]. A variety of superiorities can be achieved using kitting policy such as decrease of level of stock near the production area, less time for searching the required parts, easier replenishments schedule, and ergonomic improvements for workers on the production line [38]. Drawbacks of kitting include extra work for kitting preparation and handling. Moreover, errors in the kit can production interruptions and quality problems [35]. Therefore, the continuous supply is adopted as the main policy for the flowing production system where standard products are produced that requires less types of material. And kitting policy is selected as the main policy for the stationary production system where high-variety tailor-made products are produced that require various parts. The next levels of FRs are defined based on these two policies, while reaching the lowest DP, automated concrete distribution system, and flow rack parts stacking is suggested, some other supply policies are adopted for different production lines such as sequential supply policy [39] and Kanban-based supply policy [35]. Figure 5 shows the detailed zigzagging procedures. It is worth noting that the proposed methodology provides a scientific and systematic design roadmap for material supply policy selection based on customer requirements and the existing body of knowledge presented in the literature. By comparison, the case company’s practice has been to schedule the material feeding according to the manager’s experience, an approach that is inherently unreliable and inaccurate.

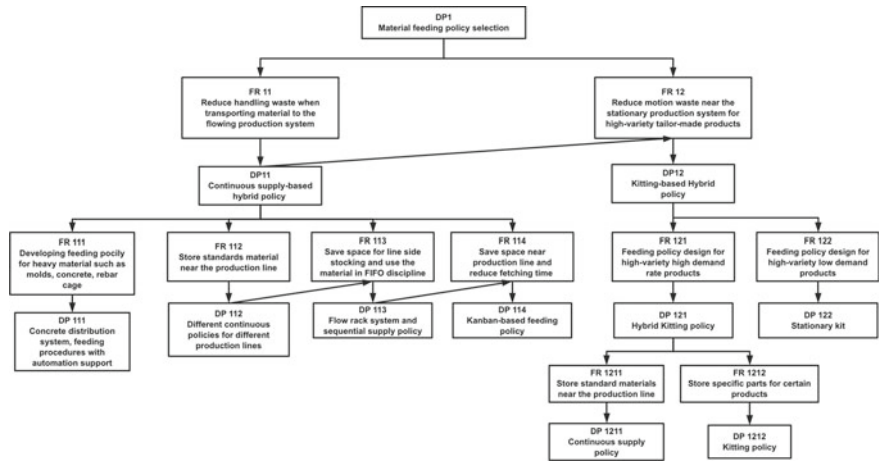


Fig. 5 Decomposition of DP1

4.2.2 Transporting Equipment Selection

Based on the designed material feeding policies, overhead shuttle system and ground shuttle are suggested for fresh concrete transportation. Furthermore, overhead crane, mobile crane, forklift, ground rail transporter, etc., are selected for the handling of different products, WIP components, and production materials. The model type and number for each transporting equipment are considered as PVs in the process domain which are out of the scope of this study. They are determined by the specialists in the company afterwards. Figure 6 shows the detailed roadmap of DP2 decomposition.

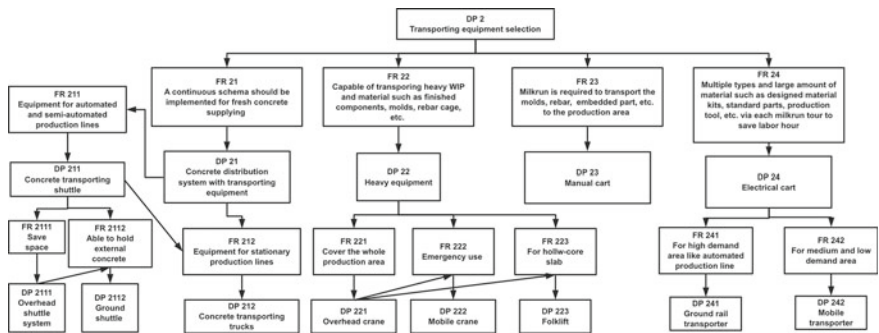


Fig. 6 Decomposition of DP2

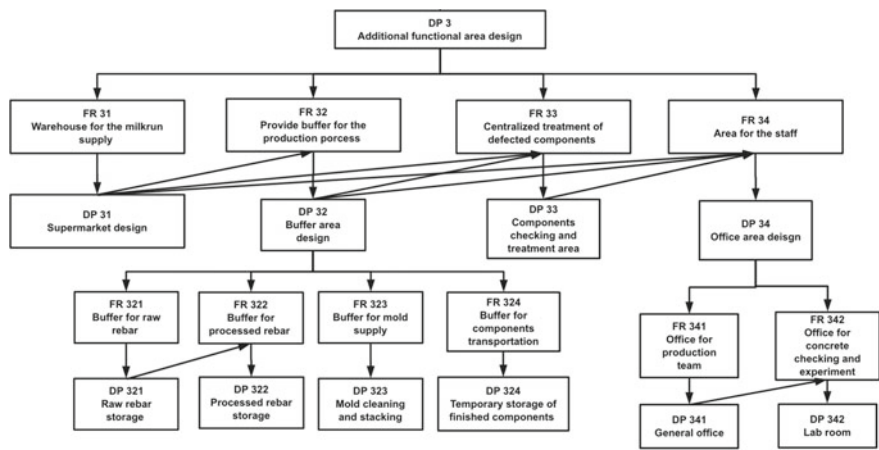


Fig. 7 Decomposition of DP3

4.2.3 Additional Functional Area Design

Central supermarket, buffer area, components checking, treatment area, and offices are added into the factory layout based on the determined part feeding policy and transporting equipment. The size and location of them are considered as PVs and are not covered in this study (Fig. 7).

4.3 Validation Based on Empirical Data

Since phase 1 production plant is running for several years and some of the existing logistics structures are difficult and costly to change, the design is partially adopted after the internal consultation of the top management team. Table 2 describes the key changes adopted by the company according to the design results.

To validate the effectiveness of the design, empirical study is conducted regarding the production stoppage frequency and the perception of workers. The production stoppage frequency is the key performance indicator of the factory logistics performance, and it should be calculated based on total production quantity. It should be mentioned that for flowing production lines, any ongoing activities that are stopped for more than 10 min are considered as production stoppage. And the stop duration set for stationary production lines is 20 min since the requirements for production rhythm in the stationary production system are not as high as in the flowing system. Table 3 shows the production stoppage frequency over the total production quantity and average stoppage duration (30 days before and 30 days after the logistics design changes); the related data of production systems involved in the design change are



**Table 2** Summary of DPs adopted by the company for phase 1 production plant

DPs adopted	Description
Kitting policy	Applied for sandwich shear wall and balcony production
Continuous policy	Applied for half-slab production
Flow rack system	Rolling racks are purchased and used to store the supporting elements such as insulation connector, grout sleeve, PVC conduit, etc. The rack can be moved along the production line and between workstations
Manual cart	Purchased to transport light parts from the storage to the production line through milk run
Central supermarket	Designed for kitting preparation and supporting elements storage
Processed rebar storage	Designed for finished rebar storage with an organized structure

recorded in this table. As can be seen from Table 3, improvements are made regarding both stoppage frequency and duration.

The design proposed by the management team should be validated by the workers as well since they are the real practitioners for the production process. In order to summarize the perceptions of workers for the above key changes, a questionnaire survey is designed which contains six questions concerning the effectiveness level of each adopted DP using a five-point Likert-type scale from 1 (very little effect) to 5 (very high effect). Eventually, 37 workers in Henan D.R. Fair-faced Concrete Technology Co., Ltd. participate in the survey. The importance of each defined problems can be determined using the index of relative importance IRI. In addition, since the PC-related knowledge is strongly correlated to the years of experience [40], the level of working experience of respondents was taken into consideration when finalizing the overall IRI for each critical problem. The index was first calculated for every set of respondent with different experience level using Eq. (6) [41]:

$$IRI_k(\%) = \frac{5(n5) + 4(n4) + 3(n3) + 2(n2) + n1}{5(n5 + n4 + n3 + n2 + n1)} \times 100, \quad (6)$$

**Table 3** Production stoppage comparison

Production system	Stoppage frequency/total production quantity		Average stoppage duration (min)	
	Before	After	Before	After
Half-slab	0.91	0.87	17.6	14.4
Sandwich shear wall	1.34	0.99	12.9	10.0
Balcony	2.23	2.35	31.2	25.8
Composite beam	3.27	2.62	26.5	26.1

**Table 4** Overall IRI for each DP adopted by the company

DPs adopted	Overall IRI (%)
Kitting policy	96.1
Continuous policy	89.7
Flow rack system	91.6
Manual cart	84.5
Central supermarket	87.3
Processed rebar storage	93.6

where  $IRI_k(\%)$  is the working experience percentage of the IRI of each factor, which is calculated separately for each corresponding working experience set ( $k$ ) of respondents;  $k$  is a number that represents the working experience of respondents (when working experience is less than 5 years,  $k = 1$ ; when working experience is between 5 and 10,  $k = 2$ ; when working experience is between 10 and 15,  $k = 3$ ; and when working experience is longer than 15 years,  $k = 4$ ); and  $n1, n2, n3, n4$ , and  $n5$  represent the number of survey respondents who selected “1” (very little effect), “2” (little effect), “3” (average effect), “4” (high effect), and “5” (very high effect), respectively. Consequently, Eq. (7) is used to calculate the overall IRI for each adopted DP among all respondents, which is calculated as a weighted average of  $IRI_k(\%)$  obtained from Eq. (6) [42]:

$$\text{Overall IRI}(\%) = \frac{\sum_{k=1}^4 (k \times IRI_k)}{\sum_{k=1}^4 k} \times 100 \tag{7}$$

where Overall IRI(%) is the total weighted average percentage of the IRI of each factor, which is calculated based on all of the sets of different levels of working experience of the respondents. Table 4 shows the overall IRI score for each DP adopted. The high IRI scores are indicative of a high degree of recognition among workers, meaning that they are likely to meaningfully experience the effects of the logistical changes during production operations.

5 Conclusion

In this study, a novel AD-based framework for PC factory logistics optimization is proposed. It provides an entire design and optimization roadmap in a systematic and logical manner. It can be used by practitioners either in early stages for holistic PC factory logistics design or in later stages for logistics system update or optimization. Through surveys/questionnaires, stakeholder input can be captured and the design oriented accordingly to accommodate it. The FR-DP mapping combines existing knowledge in the literature with stakeholder input in order to define further design details in each hierarchy of the decomposition until the lowest level DPs are clear,

implementable, and aligned with the stakeholder requirements. Furthermore, the independence axiom ensures that the design activities are well organized and that the fulfilment of a given FR does not affect others. This also allows room for subsequent design revisions.

The case study in this paper validates the effectiveness of the design framework by which for PC practitioners to design or optimize factory logistics. Material feeding policy, material handling equipment selection, and functional area design are considered in this case study based on the designs need of the case company. The scope can be broadened (by adding other components such as vehicle routing and storage design) or narrowed depending on stakeholder requirements. In this regard, the proposed framework offers great design flexibility to meet the expectations of customers.

This study contributes to the body of knowledge on PC construction by filling the gap with respect to optimizing PC factory logistics, which is vitally important for efficient PC component production and on-time delivery. It can also be applied to prefabricated steel construction factories and prefabricated wood construction factories. As for future research directions, PVs can be added to the framework to further concretize the design and concept of lean can be integrated to achieve better performance.

**Acknowledgements** This study would not have been possible without the generous support and participation of Henan D.R. Fair-faced Concrete Technology Co., Ltd., particularly the assistance provided by Mr. Zhang Yongju (General Manager), Ms. Lu Qin (Chief Engineer), and Mr. Wang Fuguang (Chief of the Technical Department).

## References

1. Jiang W, Wu L (2021) Flow shop optimization of hybrid make-to-order and make-to-stock in precast concrete component production. *J Clean Prod* 297:126708
2. Kim T, Kim YW, Cho H (2020) Dynamic production scheduling model under due date uncertainty in precast concrete construction. *J Clean Prod* 257:120527
3. Xiong F, Chu M, Li Z, Du Y, Wang L (2021) Just-in-time scheduling for a distributed concrete precast flow shop system. *Comput Oper Res* 129:105204
4. Dan Y, Liu G, Fu Y (2021) Optimized flowshop scheduling for precast production considering process connection and blocking. *Autom Constr* 125:103575
5. General Office of the State Council of China (2016) Guidance of the General Office of the State Council on the rapid development of prefabricated buildings. State Office of Development [2016] 71 [EB/OL]
6. Elliott KS (2017) Precast concrete structures, 2nd edn. CRC Press, Boca Raton
7. Marteksandmarkets (2021) Precast concrete market by element (columns and beams, floors and roofs, girders, walls and barriers, utility vaults), construction type (elemental, permanent modular, relocatable), end-use sector (residential, non-residential)—global forecast to 2025. Available at: <https://www.marketsandmarkets.com/Market-Reports/prefabricated-construction-market-125074015.html>. Accessed 11 Oct 2021
8. Wang Z, Hu H, Zhou W (2017) RFID enabled knowledge-based precast construction supply chain. *Comput Aided Civ Infrastruct Eng* 32:499–514

9. Wang Z, Hu H, Gong J (2018) Framework for modeling operational uncertainty to optimize offsite production scheduling of precast components. *Autom Constr* 86:69–80
10. Wang Z, Hu H (2018) Dynamic response to demand variability for precast production rescheduling with multiple lines. *Int J Prod Res* 56(16):5386–5401
11. Phang TCH, Chen C, Tiong RLK (2020) New model for identifying critical success factors influencing BIM adoption from precast concrete manufacturers' view. *J Constr Eng Manag* 146(4):04020014
12. Wang S, Tang J, Zou Y, Zhou Q (2019) Research on production process optimization of precast concrete component factory based on value stream mapping. *Eng Constr Archit Manag* 27(4):850–871
13. Ahmadi A, Pishvae MS, Akbari Jokar MR (2017) A survey on multi-floor facility layout problems. *Comput Ind Eng* 107:158–170
14. Epstein E (2011) Facilities planning. Industrial composting
15. Suh NP (2008) Axiomatic design: advances and applications. Oxford University Press, Oxford
16. Keller B, Buscher U (2015) Single row layout models. *Eur J Oper Res* 245(3):629–644
17. Womack JP, Jones DT (2010) Lean thinking: banish waste and create wealth in your corporation. Simon & Schuster, New York, NY
18. Li K, Luo H, Skibniewski MJ (2019) A non-centralized adaptive method for dynamic planning of construction components storage areas. *Adv Eng Inf* 39:80–94
19. Al Hawarneh A, Bendak S, Ghanim F (2019) Dynamic facilities planning model for large scale construction projects. *Autom Constr* 98:72–89
20. Song X, Xu J, Shen C, Peña-Mora F (2018) Conflict resolution-motivated strategy towards integrated construction site layout and material logistics planning: a bi-stakeholder perspective. *Autom Constr* 87:138–157
21. Chen C, Tran Huy D, Tiong LK, Chen IM, Cai Y (2019) Optimal facility layout planning for AGV-based modular prefabricated manufacturing system. *Autom Constr* 98:310–321
22. Hong WK, Lee G, Lee S, Kim S (2014) Algorithms for in-situ production layout of composite precast concrete members. *Autom Constr* 41:50–59
23. Hyun H, Yoon I, Lee HS, Park M, Lee J (2021) Multiobjective optimization for modular unit production lines focusing on crew allocation and production performance. *Autom Constr* 125:103581
24. Houshmand M, Jamshidnezhad B (2006) An extended model of design process of lean production systems by means of process variables. *Robot Comput Integr Manuf* 22(1):1–16
25. Botsaris PN, Anagnostopoulos KP, Demesouka O (2008) Using axiomatic design principles for designing a simple and innovative product: a case study. *Int J Des Eng* 1(3):300
26. Ertay T, Satoğlu SI (2012) System parameter selection with information axiom for the new product introduction to the hybrid manufacturing systems under dual-resource constraint. *Int J Prod Res* 50(7):1825–1839
27. Matt DT (2012) Application of axiomatic design principles to control complexity dynamics in a mixed-model assembly system: a case analysis. *Int J Prod Res* 50(7):1850–1861
28. Vinodh S, Aravindraj S (2012) Axiomatic modeling of lean manufacturing system. *J Eng Des Technol* 10(2):199–216
29. Han W, Zhao J, Chen Y (2013) A virtual cellular manufacturing system design model based on axiomatic design theory. *Appl Mech Mater* 271(Part 1):1478–1484
30. Chakraborty K, Mondal S, Mukherjee K (2017) Analysis of product design characteristics for remanufacturing using fuzzy AHP and axiomatic design. *J Eng Des* 28(5):338–368
31. Hager T, Wafik H, Faouzi M (2017) Manufacturing system design based on axiomatic design: case of assembly line. *J Ind Eng Manag* 10(1):111–139
32. Rauch E, Spena PR, Matt DT (2019) Axiomatic design guidelines for the design of flexible and agile manufacturing and assembly systems for SMEs. *Int J Interact Des Manuf* 13(1):1–22
33. Hanson R, Brolin A (2013) A comparison of kitting and continuous supply in in-plant materials supply. *Int J Prod Res* 51(4):979–992
34. Tuzkaya G, Gülsün B, Kahraman C, Özgen D (2010) An integrated fuzzy multi-criteria decision making methodology for material handling equipment selection problem and an application. *Expert Syst Appl* 37(4):2853–2863

35. Faccio M (2014) The impact of production mix variations and models varieties on the parts-feeding policy selection in a JIT assembly system. *Int J Adv Manuf Technol* 72(1–4):543–560
36. Usta SK, Oksuz MK, Durmusoglu MB (2017) Design methodology for a hybrid part feeding system in lean-based assembly lines. *Assem Autom* 37(1):84–102
37. Caputo AC, Pelagagge PM, Salini P (2018) Selection of assembly lines feeding policies based on parts features and scenario conditions. *Int J Prod Res* 56(3):1208–1232
38. Limère V, Landeghem HV, Goetschalckx M, Aghezzaf EH, McGinnis LF (2012) Optimising part feeding in the automotive assembly industry: deciding between kitting and line stocking. *Int J Prod Res* 50(15):4046–4060
39. Johansson E, Johansson MI (2006) Materials supply systems design in product development projects. *Int J Oper Prod Manag* 26(4):371–393
40. Chen JH, Yan S, Tai HW, Chang CY (2017) Optimizing profit and logistics for precast concrete production. *Can J Civ Eng* 44(6):393–406
41. Zhao ZY, Chen YL (2018) Critical factors affecting the development of renewable energy power generation: evidence from China. *J Clean Prod* 184:466–480
42. El-Gohary KM, Aziz RF (2014) Factors influencing construction labor productivity in Egypt. *J Manag Eng* 30(1):1–9

# Agent-Based Modeling and Simulation of Project Schedule Risk Analysis in the Construction Industry



Mohamed ElGindi, Sara Harb, Abdelhamid Abdullah,  
Yasmeen A. S. Essawy, and Khaled Nassar

**Abstract** The risk of project delay is a common phenomenon with an adverse effect on the performance of projects in the construction sector. The effect of its negative impacts in terms of cost overruns, reduced quality, and productivity extends to the owner, consultant, and contractor. The goal of this paper is to introduce an agent-based simulation model for the risk analysis of the project schedule component of construction projects, based on three risk management decisions. In the simulated model, the authors indicate four main phases in the construction process, along with their approval stages: (1) handing over; (2) engineering; (3) procurement; and (4) construction, which are commonly subjected to delays in the completion of the required activities. In addition, the developed simulation model should allow decision-makers to explore the impact of risks on the project schedule, in terms of schedule and cost overrun, based on two risk-response controls: acceptance or mitigation. As such, three simulation models are formulated: (A) no risks; (B) with risks; and (C) with mitigation. The model has been run based on a 70% mitigation value. The results indicate rational values. Since the duration for the risks for the ‘with risk’ scenario resulted in the highest time, followed by the ‘with mitigation’ scenario, the lowest time is recorded for the ‘no risks’. Similarly, the highest cost is recorded for the ‘with mitigation’ scenario, followed by the ‘with risks’, ending with the ‘no risks’ scenario. Further validation tools signified the effectiveness of the mitigation decision on the recorded results. This is demonstrated by the sudden drop as the mitigation value has been decreased based on the user’s input.

---

M. ElGindi (✉) · S. Harb · A. Abdullah · Y. A. S. Essawy

Department of Construction Engineering, The American University in Cairo, New Cairo, Egypt  
e-mail: [m\\_elgindi@aucegypt.edu](mailto:m_elgindi@aucegypt.edu)

A. Abdullah

Department of Architectural Engineering, Faculty of Engineering at Mataria, Helwan University, Helwan, Egypt

Y. A. S. Essawy

Department of Structural Engineering, Ain Shams University (ASU), Cairo, Egypt

K. Nassar

Industrial Partnerships and Extended Education, Department of Construction Engineering, The American University in Cairo (AUC), New Cairo, Egypt

© Canadian Society for Civil Engineering 2023

493

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_31](https://doi.org/10.1007/978-3-031-34593-7_31)

**Keywords** Agent-based modelling • Project schedule risk analysis

## 1 Introduction

Projects within the construction sector are associated with a high degree of risk and uncertainty [17]. Project risk is defined by International Standards as a ‘combination of the probability of an event occurring and its consequences for project objectives’ [19], whereas the Project Management Institute (PMI) discusses the concept of risk in more detail and raises the concept of uncertainty and the perception of risk as a negative event, in its definition of risk as ‘an uncertain event or condition that, if it occurs, has a positive or negative effect on one or more project objectives’ [26]. The risk of project delay is a common phenomenon with an adverse effect on the performance of projects in the construction sector [5, 29]. Construction project risks impact approximately 56% of projects, at varying extents [8]. Its negative impacts in terms of cost overruns, reduced quality, and productivity [35] lead to issues of litigations, disputes, and arbitrations, as its effect extends to the owner, consultant, and contractor [4]. Moreover, the impact of poor management and analysis of construction risks extends to affect the country’s GDP [13].

The rapid increase in the size and level of complexity of projects in the construction industry increases the likelihoods and associated effects of project risks and uncertainties [17]. Thompson and Perry conclude that the lack of an effective risk management system is directly related to the failure of construction projects [33]. The literature pertaining to risk management systems focuses on the significance of identifying risk factors, analyzing their probability of occurrence and resultant impact, to reduce the probability and severity of risks on the attainment of project objectives [1, 28]. Moreover, the effective management of a project involves the efficient allocation of resources in each project phase, which would otherwise incur schedule and cost overruns [27]. Productivity in the construction sector has been at a considerable decline through the past years [6]. As such, risk management techniques are essential to an effective construction management process [30], as risks, resources allocation, and improvement of production productivity performance can be managed more proactively [31]. The purpose of this paper is to introduce an agent-based simulation model for the risk analysis of the project schedule component of construction projects. The developed simulation model should allow decision-makers to explore the impact of risks on the project schedule, in terms of schedule and cost overrun, based on two risk-response controls: acceptance or mitigation.

## 2 Research Background

This section first reviews current studies regarding the importance of the effective risk management in the construction industry and the attributed different phases and variables. Subsequently, it discusses the use of the Monte Carlo simulation in assessing construction project risks. The section then concludes with an introduction to the basic principles of agent-based simulation modeling.

### 2.1 Risk Management in the Construction Industry

The successful delivery of construction projects is defined by the intrinsic relationship between the ability to meet the assigned project schedule within the budgeted amount and at the proposed quality [11]. In Egypt, thirty-seven percent of projects do not meet their cost constraints, as unexpected, incurred costs arise. In addition, ninety-eight percent of contractors fall behind the planned project time schedule [12]. The assessment of risk factors should be performed in the context of risk management. In their study, Ha et al. [17] express the value of risk factors (RV) by the below equation, governed by four variables (RV,  $P$ ,  $I$ , and  $D$ ). The four variables are defined over the range of 1–5, with RV over the range of 1–125. Yet the range values can be adjusted according to each construction project requirements and context [17].

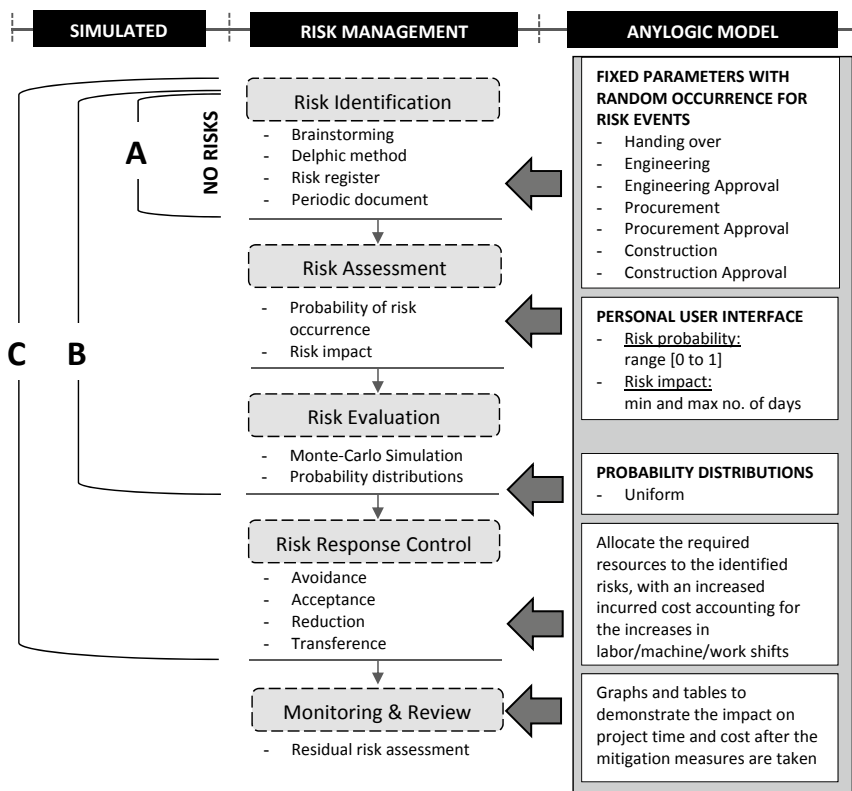
$$RV = P \times I \times D,$$

- $P$  referred to as the probability of the risk occurrence.
- $I$  referred to as the level of risk impact on project schedule and project objectives.
- $D$  indicates the difficulty of the risk detection, control, and management process.

Project risk management is the formal, systematic process of identifying, assessing, and responding to any arising risks. There are different definitions and number of steps for the risk management process across the literature. The PMI [26] categorizes four main steps: (1) identification, (2) assessment, (3) response; and (4) management. Yet more comprehensive models are developed, Baloi and Price [9] propose a seven-step risk management model: (1) planning, (2) identification, (3) assessment; (4) analysis; (5) response; (6) monitoring; and (7) communication. The authors propose a five-step risk management process, illustrated in Fig. 1.

Literature reveals that delays in construction projects vary across different countries [14, 18, 23]. This is due to different variables such as construction environment, location, design requirements, and laborer's levels of expertise [20]. The most contributing factors of delay risks in construction projects in Egypt could be categorized into nine categories: (1) consultant-related, (2) contractor-related, (3) design-related; (4) equipment-related; (5) external-related; (6) labor-related; (7) material-related; (8) client-related; and (9) project-related [7, 24]. In the simulated model,





**Fig. 1** Proposed methodology

the authors indicate four main phases in the construction process, along with their approval stages which are commonly subjected to delays in the completion of the required activities. Table 1 identifies possible causes of delay for each of the four phases.

## 2.2 Monte Carlo Simulation

The Monte Carlo simulation-based approach is a sampling-based method commonly used in the evaluation of construction project risks [34]. During a Monte Carlo simulation, random values are generated from a variety of specific probability distributions [15, 27], as such a behavior is allocated to the specific range of random values with an inherent uncertainty [16]. The model is then iterated thousands of times, each time using different random values based on each input values specific probability distribution, producing distributions of all possible outcome values [21].

**Table 1** Possible delay causes in the construction industry

Delay cause	References
<i>1. Handover</i>	
Project variation	[25]
Inability of client to pay the contractor	[25]
Inaccurate site investigation	[7]
Unsuitable subsurface site conditions (soil, high water table, etc.)	[23]
New government policies	[25]
<i>2. Engineering</i>	
Incomplete design scope	[17]
Design errors	[17]
Inadequate specifications	[17]
Design changes by client	[7]
Inadequate details in design drawings	[7]
<i>3. Procurement</i>	
Delay in manufacturing materials	[7]
Late delivery of materials	[7]
Shortage of construction materials	[7]
Poor procurement of construction materials	[7]
Unreliable suppliers	[7]
<i>4. Construction</i>	
Insolvency during the construction process from the contractor and/or client	[22]
Adverse weather conditions	[22]
Untrained construction staff	[22]
Unsustainable timeframe	[22]
Unavailability of utilities in site	[22]
Equipment failure	[17]

The AnyLogic software incorporates approximately 25 distributions. In the literature, there are several different probability distributions used in the applications of the Monte Carlo simulation in project risk management. Agarwal and Mahajan [3] proposed the use of three probability distribution functions, PERT distribution, normal distribution, and triangular distribution to assess the risks associated with delays in construction activities which affect the overall construction project schedule. Similarly, AbouRizk and Halpin conducted a study for seventy-one samples of construction activity duration patterns and determined that the beta distribution is the most appropriate function for the representation of construction activity durations [2].

## **2.3 Agent-Based Modeling Simulation in the Construction Industry**

Agent-based modeling (ABMS) is based on three principal aspects: the combination of agents; agent relationships; and the agent environment, to generate the required outputs. Agents are active decision-making entities, executing various behaviors varying from simple to complex, depending on the represented situation and system [10, 32]. The identification of agents is the initial step in the in the ABMS, in which each agent has a distinct attribute, responsible for its behavior in the system. Agent relationships signify both the interactions between all agents in a system and between the agents and their surrounding environment, which represent the agent's environment [32].

## **3 Methodology**

In this section, the proposed process for developing a simulation-based risk management model is presented, as identified in Fig. 1. The model has been developed for repetitive projects for constructing medium-sized buildings. For simplicity, the project activities are grouped to have the following sequence.

- Stage 1—Site handing over: Client to hand over the site to the contractor as such, the commencement for the project would be considered.
- Stage 2—Engineering: Developing the engineering deliverables by the contractor.
- Stage 3—Engineering approval: Engineer to approve the engineering deliverables.
- Stage 4—Procurement: Contractor to perform the procurement activities for the materials that would be use during the construction.
- Stage 5—Procurement approval: Engineer to approve the procured material deliverables.
- Stage 6—Construction: Contractor to perform construction activities.
- Stage 7—Construction approval: Engineer to approve the constructed works.
- Stage 8—Project finish: Contractor to hand over the project to the client.

### **3.1 Model Objectives**

The objectives of developed model are to:

1. Establish a tool to estimate the total duration and the needed relative resources during the bidding phase.
2. Provide stakeholders with a flexible user interface based on their expert judgment to define the risk probability and impacts.
3. Provide a visual representation of outputs.
4. Allow project stakeholders to forecast the cost for all simulated project scenarios.

5. Provide the flexible adjustment of the risk management techniques based on the constraints of time and cost.

### **3.2 Scenario Description**

The developed model considers three scenarios:

1. The first scenario has no risks at all.
2. The second scenario has the risks occurred.
3. The third scenario has the risks occurred, and the mitigation actions have been taken.

The risks are presented via random event. Once the event happens, a consequential delay and costs shall be added for each stage. This consequential delay shall follow a uniformly distributed delay time with a minimum and a maximum duration inserted by the model's user for each stage. The mitigation control is introduced, through a decision by the model user, as a percentage which 100% means fully mitigated, and in this case, the duration would be close to the first scenario. While 0% means no mitigation at all, as such, the duration would be like the second scenario.

The cost for each stage is concluded by the following formula:

The cost per stage = the total cost per day inserted by the user \* the stage duration.

The total cost per day could be assumed to be equal to the total contracted value divided by the total planned duration. The cost for the mitigation is assumed to be uplifted over the regular cost per day with the same mitigation percentage introduced by the user.

### **3.3 Simulation Model Formulation**

The three developed models for the three different scenarios indicated above are shown in Fig. 2.

### **3.4 Simulation Libraries and Parameters**

The risk event is developed based on several variables as illustrated in Fig. 3. The ONOFF variable is a randomly generated number between 0 and 1 and is generated for every simulation case. If this variable is greater than the risk probability provided by the user, another variable is introduced, identified in the second column of variables ex. HO\_P\_E, which is equivalent to a uniform value between the minimum

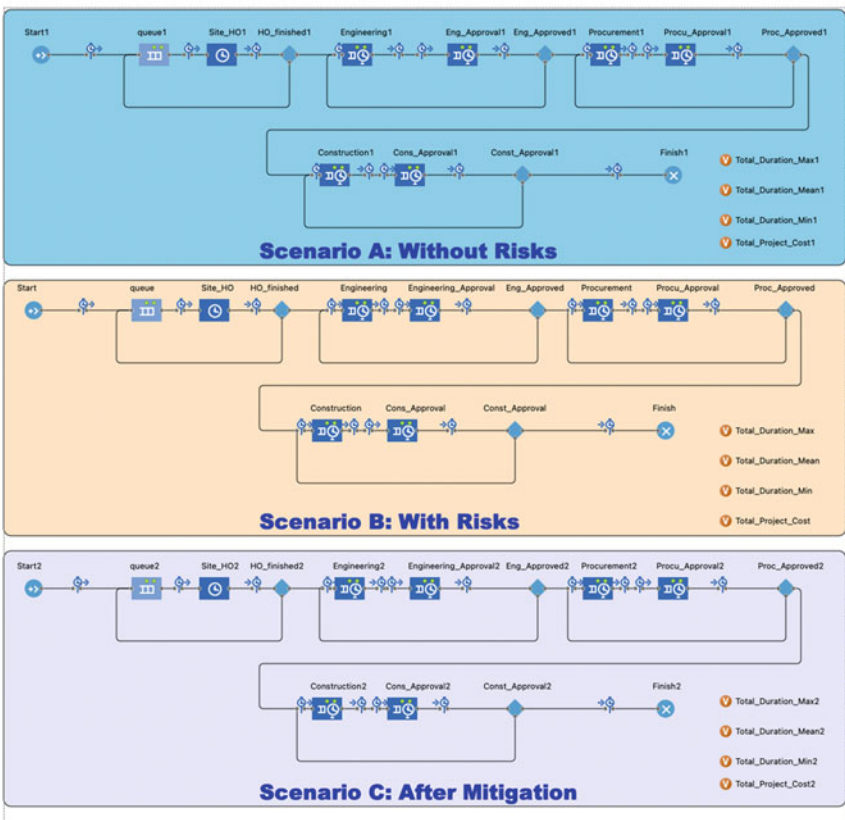


Fig. 2 Scenarios in the developed simulation model

and maximum risk impact, which is also introduced by the user. Otherwise, if the ONOFF variable is less than the risk probability provided by the user, the associated risk impact shall equal zero, and no risk occurs.



Fig. 3 Risk event modeling criteria

As indicated in Table 2, each stage is assumed to use the following duration distribution:

It is assumed that 70% of the handing over, engineering, procurement, and construction activities would grant the engineering approval. In case the engineering has rejected the activity, it should be subject to a rework until it receives approval.

The simulation steps and input values for each of the three scenarios are illustrated in Tables 3, 4, and 5.

**Table 2** Duration distributions

Stage	Duration distribution in days
Stage 1—Site handing over	Uniform(0, 7)
Stage 2—Engineering	Triangular(5, 30, 60)
Stage 3—Engineering approval	Uniform(1, 7)
Stage 4—Procurement	Uniform(2, 90)
Stage 5—Procurement approval	Uniform(1, 7)
Stage 6—Construction	Pert(7, 360, 90)
Stage 7—Construction approval	Uniform(1, 7)
Stage 8—Project finish	1

**Table 3** Simulation steps and input values (Scenario A)

Simulation components	Identification	Input values
Source	Start1	Calls of inject () function <i>After finishing each simulation case</i>
Delay	Site_HO1	Uniform(0, 7)
Output	HO_Finished1	0.7
Service	Engineering1	Triangular(5, 30, 60)
Service	Engineering_Approval1	Uniform(1, 7)
Output	Engineering_Approved1	0.7
Service	Procurement1	Uniform(2, 90)
Service	Procu_Approval1	Uniform(1, 7)
Output	Procu_Approved1	0.8
Service	Construction1	Pert(7, 360, 90)
Service	Const_Approval1	Uniform(1, 7)
Output	Const_Approved1	0.8
Sink	Finish1	None

**Table 4** Simulation steps and input values (Scenario B)

Simulation components	Identification	Input values
Source	Start	Calls of inject () function <i>After finishing each simulation case</i>
Delay	Site_HO	Uniform(0, 7) + HO_P_E
Output	HO_Finished	0.7
Service	Engineering	Triangular(5, 30, 60) + Eng_P_E
Service	Engineering_Approval	Uniform(1, 7) + Eng_Appr_P_E
Output	Engineering_Approved	0.7
Service	Procurement	Uniform(2, 90) + Proc_P_E
Service	Procu_Approval	Uniform(1, 7) + Proc_App_P_E
Output	Procu_Approved	0.8
Service	Construction	Pert(7, 360, 90) + Const_P_E
Service	Const_Approval	Uniform(1, 7) + Const_App_P_E
Output	Const_Approved	0.8
Sink	Finish	None

**Table 5** Simulation steps and input values (Scenario C)

Simulation components	Identification	Input values
Source	Start2	Calls of inject () function <i>After finishing each simulation case</i>
Delay	Site_HO2	Uniform(0, 7) + HO_P_E – (HO_P_E * Mitigation)
Output	HO_Finished2	0.7
Service	Engineering2	(Triangular(5, 30, 60) + Eng_P_E) – (Eng_P_E * Mitigation)
Service	Engineering_Approval2	(Uniform(1, 7) + Eng_Appr_P_E) – (Eng_Appr_P_E * Mitigation)
Output	Engineering_Approved2	0.7
Service	Procurement2	(Uniform(2, 90) + Proc_P_E) – (Proc_P_E * Mitigation)
Service	Procu_Approval2	(Uniform(1, 7) + Proc_App_P_E) – (Proc_App_P_E * Mitigation)
Output	Procu_Approved2	0.8
Service	Construction2	(Pert(7, 360, 90) + Const_P_E) – (Const_P_E * Mitigation)
Service	Const_Approval2	(Uniform(1, 7) + Const_App_P_E) – (Const_App_P_E * Mitigation)
Output	Const_Approved2	0.8
Sink	Finish2	None

A. Risk Probability %

B. Risk Effect Boundaries in days

		Min	Max
1- Handing Over	<input type="text" value="0.3"/>	<input type="text" value="1.0"/>	<input type="text" value="15.0"/>
2- Engineering	<input type="text" value="0.3"/>	<input type="text" value="5.0"/>	<input type="text" value="40.0"/>
3- Engineering Approval	<input type="text" value="0.3"/>	<input type="text" value="3.0"/>	<input type="text" value="15.0"/>
4- Procurement	<input type="text" value="0.3"/>	<input type="text" value="5.0"/>	<input type="text" value="40.0"/>
5- Procurement Approval	<input type="text" value="0.3"/>	<input type="text" value="3.0"/>	<input type="text" value="15.0"/>
6- Construction	<input type="text" value="0.3"/>	<input type="text" value="7.0"/>	<input type="text" value="60.0"/>
7- Construction Approval	<input type="text" value="0.3"/>	<input type="text" value="0.3"/>	<input type="text" value="20.0"/>

Mitigation %

Cost / day

Run

Developed by Sara Harb (M.Sc.) and Mohamed ELGINDI (M.Sc.)  
December 2021

Fig. 4 User interface in simulated model

3.5 Running the AnyLogic Model

As shown in Fig. 4, the user interface allows the user to enter the following values:

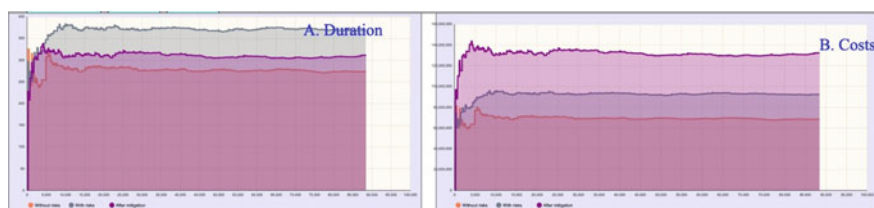
- 1. Risk probability percentage within the range of 0–1, with 0.1 intervals
- 2. Risk effect minimum and maximum values, in days
- 3. Initial targeted mitigation percentage
- 4. Cost per day.

The model runs 10,000 simulation case, where each represents the duration of one project.

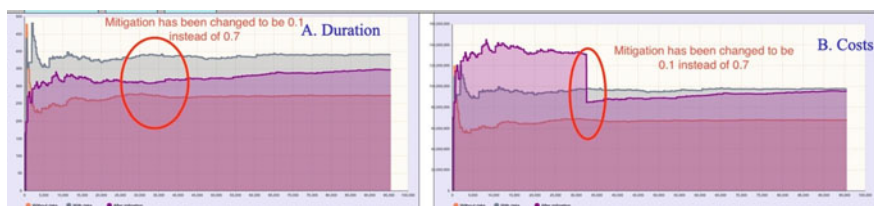
3.6 Model Verification and Validation

The model has been run based on a 70% mitigation value. As shown in Fig. 5, the results indicate rational values. Since the duration for the risks for the ‘with risk’ scenario is the highest time, followed by the ‘after-mitigation’ scenario, the lowest time is recorded for the ‘without risks’. Similarly, the highest cost is recorded for the ‘after-mitigation’ scenario, followed by the ‘with risks’, ending with the ‘without risks’ scenario.





**Fig. 5** Duration and costs at 70% mitigation value



**Fig. 6** Duration and costs at 10% mitigation value

Another validation tool is represented in Fig. 6, as it signifies the effectiveness of the mitigation decision on the recorded results. This is demonstrated by the sudden drop as the mitigation value has been decreased based on the user's input.

## 4 Results

Once the model is run, a dashboard of graphs appears to the user, with three tabs at the top. One shows the processing logic, one for the graphs, and the third tab for some tables.

The mean durations for each of the four activities and their associated approvals are illustrated in Figs. 7, 8, 9, 10 and 11. Figure 12 presents the duration project distribution function (PDF) and cumulative distribution function for the three scenarios. The total duration is illustrated in Fig. 13, while the total costs are presented in Fig. 14. A slider presents the mitigation percentage and is introduced to the model as shown in Fig. 15. At any given point, the user can adjust the percentage, which updates the results instantly. Worth noting that this slider is used to validate the model's credibility. The tables tab introduces another presentation for the model results as shown in Fig. 16. It gives an idea about the duration and the associated costs for each group of activities in the three scenarios and the overall cost figures as well. It also has a slider for mitigation, as described before.

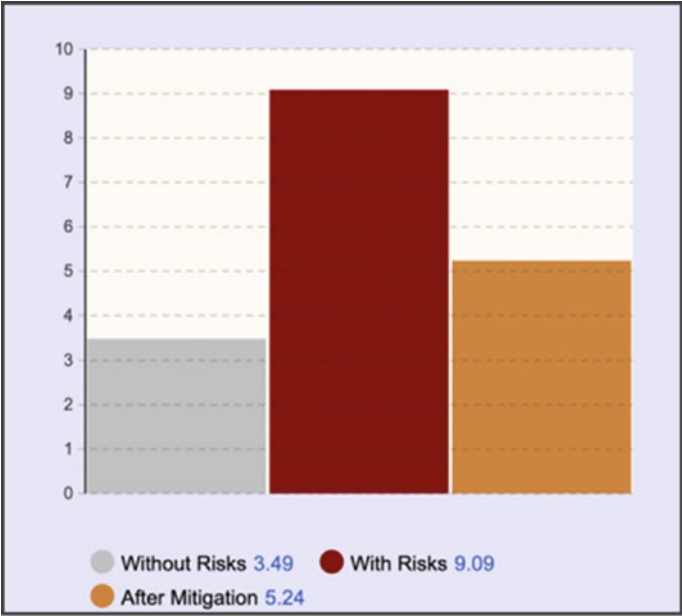


Fig. 7 Handing over activity mean durations/days

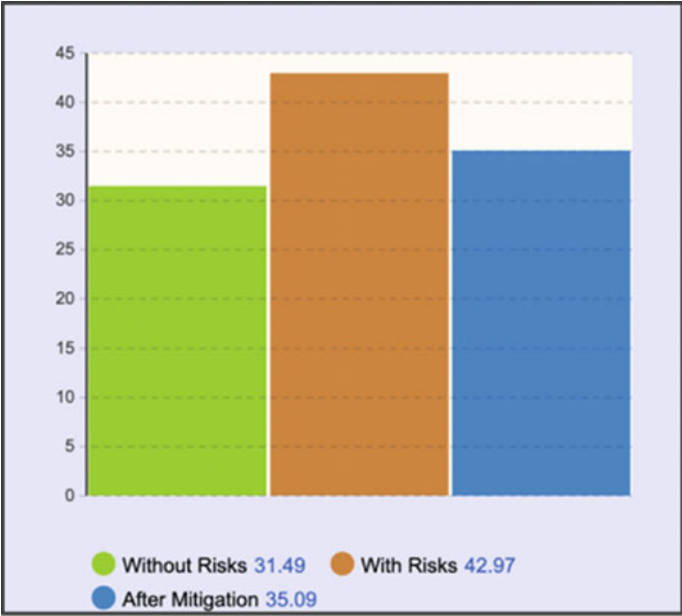


Fig. 8 Engineering activity mean durations/days

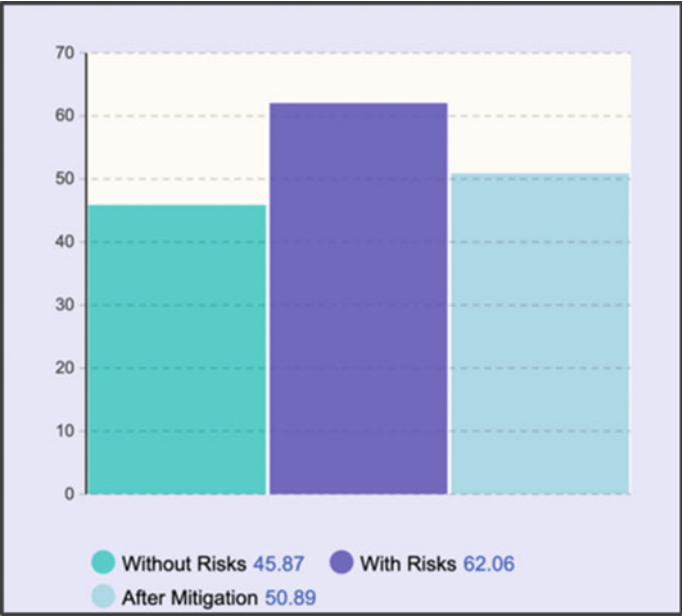


Fig. 9 Procurement activity mean durations/days

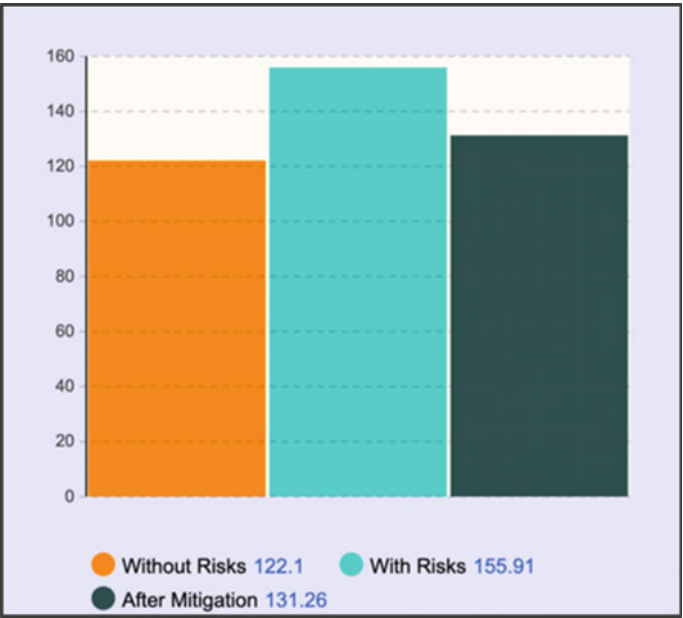


Fig. 10 Construction activity mean durations/days

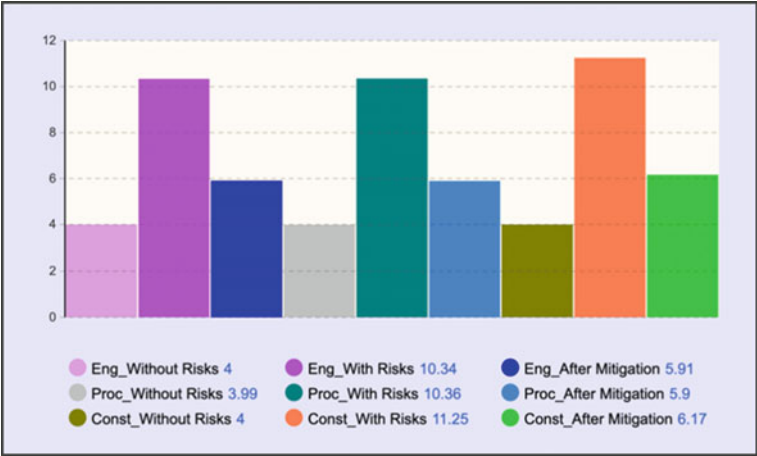


Fig. 11 Approval durations for engineering, procurement, and construction activities/days

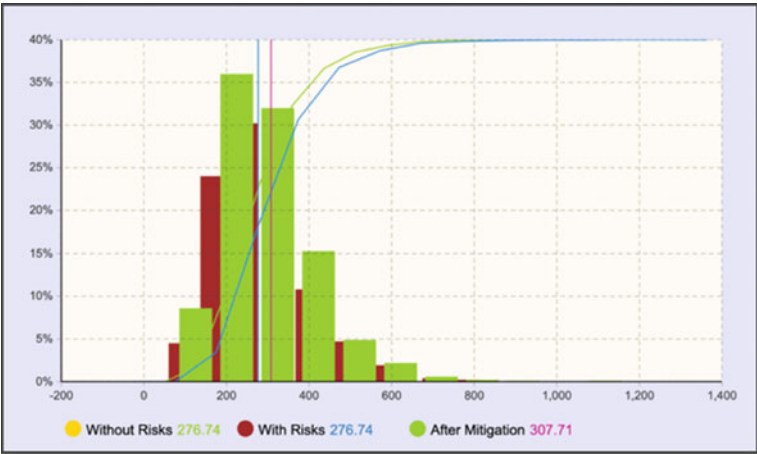


Fig. 12 Total project duration distribution for the three scenarios/days

5 Conclusions

In construction industry, projects are usually vulnerable for prolongation risks more than the planned duration. The prediction of the risks before it occurs is a proactive technique that would assist the project managers to control the impacts resulting from thereof. The authors introduce an agent-based simulation model using AnyLogic software for the risk analysis of the project schedule.

The developed model provides reliable and comprehensive results that give the user a chance to decide the proper control over the occurred risks. The model through

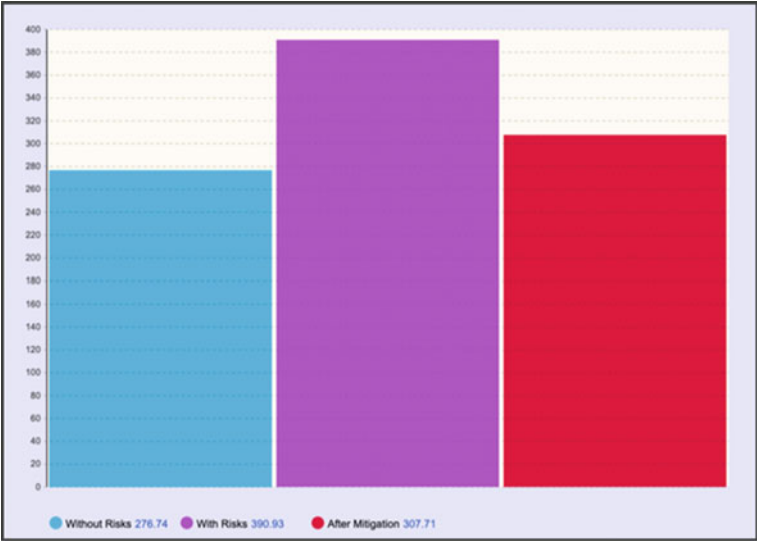
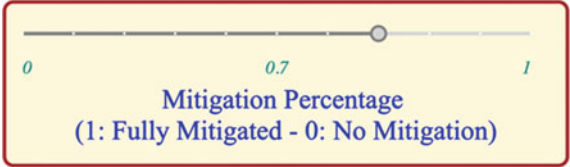


Fig. 13 Total duration for the three scenarios/days

Fig. 14 Total cost for the three scenarios/days



Fig. 15 Slider tool to adjust the risk mitigation percentage



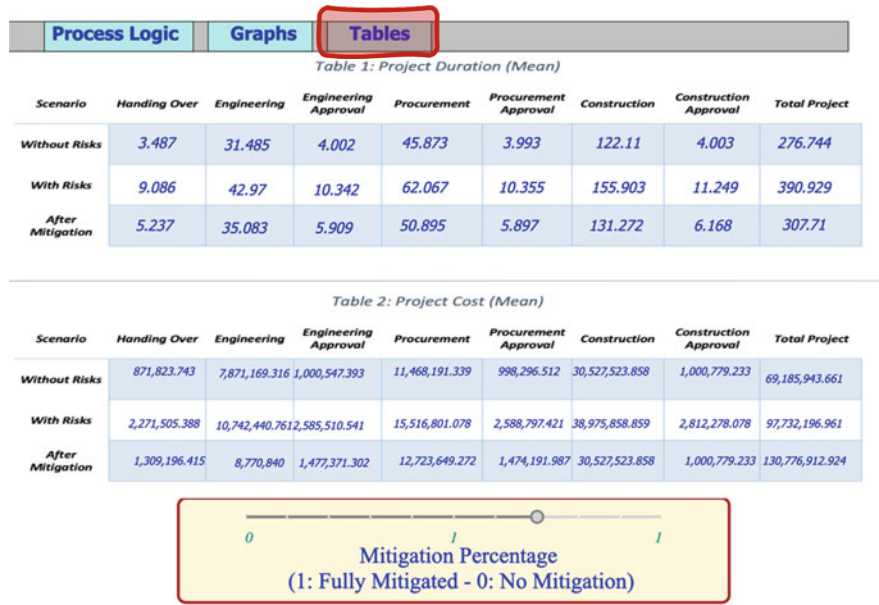


Fig. 16 Snapshot from the tables tab in the model simulation

10,000 runs shall simulate the project duration with the predictable risks. The user shall decide the mitigation protocol he would use according to his priority to overcome the occurred risk impacts; duration over costs, costs over the duration, or a balance between both.

The model has proven to have the following benefits:

- It can be used in the bidding phase to estimate the total duration and the needed relative resources.
- Forecasting the costs for all three scenarios.
- Flexible adjustment of the mitigation techniques. Time versus cost.
- Provide user with a flexible user interface based on their expert judgment to define the risk probability and impacts.
- Visual representation of outputs which is beneficial to project stakeholders.
- Allow decision-makers to explore the impact of risks on the project schedule and cost.

5.1 Suggestions for Further Research

For extending the given work and increasing the applicability of its framework, it would be possible to further break down the construction activities and introduce other risks in the risk management process, rather than just the schedule-related

ones. Linking the model to planning softwares, as such Primavera, or Microsoft Projects could transfer the input into the AnyLogic via JavaScript. This model used a uniform distribution for risk impacts to simplify the developed model, whereas the activities delays have been considered for the use of different distributions, such as uniform, PERT, and triangular as indicated in Tables 3, 4 and 5. Further research could include additional probability distributions for both risks and activities delays. The introduction of resource pools and its associated costs per each resource type would provide more accurate cost calculations.

## References

1. Abdelgawad M, Fayek AR (2010) Risk management in the construction industry using combined fuzzy FMEA and fuzzy AHP. *J Constr Eng Manag* 136(9):1028–1036
2. AbouRizk SM, Halpin DW (1992) Statistical properties of construction duration data. *J Constr Eng Manag* 118(3):525–544
3. Agarwal AL, Mahajan DA (2017) A probability analysis of construction project schedule using risk management tool. *Int J Sci Technol* 3(1):104–109
4. Aibinu AA, Jagboro GO (2002) The effects of construction delays on project delivery in Nigerian construction industry. *Int J Proj Manag* 593–599
5. Assaf SA, Al-Hejji S (2006) Causes of delay in large construction projects. *Int J Project Manag* 349–357
6. Aziz RF, Hafez SM (2013) Applying lean thinking in construction and performance improvement. *Alexandria Eng J* 679–695
7. Aziz RF (2013) Ranking of delay factors in construction projects after Egyptian revolution. *Alexandria Eng J* 387–406
8. Ballard G, Howell G (1998) Shielding production: essential step in production control. *J Constr Eng Manag* 124(1):11–17
9. Baloi D, Price AD (2003) Modelling global risk factors affecting construction cost performance. *Int J Project Manage* 21(4):261–269
10. Bonabeau E (2002) Agent-based modeling: methods and techniques for simulating human systems. *Proc Nat Acad Sci U S A (Nat Acad Sci)* 99(10):7280–7287
11. Caldas C, Gupta A (2017) Critical factors impacting the performance of mega-projects. *Eng Constr Archit Manag* 24(6):920–934
12. Davies JP, Clarke BA, Whiter JT, Cunningham RJ (2001) Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban Water* 3(1):73–89
13. Dixit S, Mandal SN, Thanikal JV, Saurabh K (2019) Evolution of studies in construction productivity: a systematic literature review (2006–2017). *Ain Shams Eng J* 10(3):555–564
14. Doloi H, Sawhney A, Iyer KC, Rentala S (2012) Analysing factors affecting delays in Indian construction projects. *Int J Project Manag* 479–489
15. Feldman RM, Flores CV (1995) *Applied probability and stochastic processes*. Springer, London
16. Fox WP, Burks R (2019) Monte Carlo simulation and agent-based modeling (ABM) in military decision-making. In: Price CC (ed) *International series in operations research and management science*. Springer, Berlin, pp 395–453
17. Ha LH, Hung L, Trung LQ (2018) A risk assessment framework for construction project using artificial neural network. *J Sci Technol Civ Engi* 12(5):51–62
18. Hamzah N, Khoiry M, Arshad I, Tawil N, Ani Che A (2011) Cause of construction delay—theoretical framework. *Procedia Eng* 490–495
19. IEC (2013) IEC 62198: managing risk in projects—application guidelines. International Electrotechnical Commission

20. Jeong W, Chang S, Son J, Yi SJ (2016) BIM-integrated construction operation simulation for just-in-time production management. *Sustainability* 8(11):1–25
21. Kong Z, Zhang J, Li C, Zheng X, Guan Q (2015) Risk assessment of plan schedule by Monte Carlo simulation. In: International conference on information technology and management innovation. Atlantis Press, Shenzhen, pp 509–513
22. Lo TY, Fung IW, Tung KC (2006) Construction delays in Hong Kong civil engineering projects. *J Constr Eng Manag* 636–649
23. Marzouk MM, El-Rasas TI (2014) Analyzing delay causes in Egyptian construction projects. *J Adv Res* 49–55
24. Nassar AH (2016) Causes and effects of delay in the construction industry in Egypt. *Int J Eng Res Technol* 5(4):447–452
25. Oke AE, Olatunji SO (2016) Factors affecting construction project handover and feedback mechanism. In: Joint international conference. Palstar Concepts & JIC Local Organising Committee, Nigeria, pp 841–850
26. PMI (2017) A guide to the project management body of knowledge (PMBOK guide). Project Management Institute Inc., Pennsylvania
27. Qazi A, Simsekler MC (2021) Risk assessment of construction projects using Monte Carlo simulation. *Int J Manag Proj Bus* 14(5):1202–1218
28. Sadeghi N, Fayek AR, Pedrycz W (2010) Fuzzy Monte Carlo simulation and risk assessment in construction. *Comput Aided Civ Infrastruct Eng* 25(4):238–252
29. Sambasivan M, Soon YW (2007) Causes and effects of delays in Malaysian construction industry. *Int J Project Manag* 517–526
30. Schieg M (2006) Risk management in construction project management. *J Bus Econ Manag* 77–83
31. Shehab L, Ezzeddine A, Hamzeh F, Power W (2020) Agent-based modelling and simulation of construction crew performance. In: 28th annual conference of the International Group for Lean Construction (IGLC28), Berkeley, pp 1021–1032
32. Taylor S (2014) Agent-based modeling and simulation. Brunel University. Springer, Berlin
33. Thompson P, Perry J (1992) Engineering construction risks: a guide to project risk analysis and risk management. Thomas Telford, London
34. Tong R, Cheng M, Zhang L, Liu M, Yang X, Li X, Yin W (2018) The construction dust-induced occupational health risk using Monte-Carlo simulation. *J Clean Prod* 598–608
35. Yaseen ZM, Ali ZH, Salih SQ, Al-Ansari N (2020) Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability* 12(4):1514



# Automated Resource Scheduling for Construction Projects Using Genetic Algorithm



Raghdha M. Moharram, Yasmeen A. S. Essawy, and Osama S. Hosny

**Abstract** Site advance methods for planning and scheduling are essential for effective project management. Typically, construction managers tend to produce schedules based on two constraints: (1) Minimize Project Duration and (2) Allocate Minimum Resources. Moreover, deficits in the cash flow result in reduced profits at the end of the project and delays if financing problems arise, which results in damages (usually additional costs). Traditional scheduling tools like the Critical Path Method (CPM) and Time Constrained Project Scheduling Problem (TCPSP) do not show high efficiency in achieving the required objectives, as scheduling gets complicated. Advanced planning and scheduling methods will be used to produce a feasible schedule that meets specific objectives. This paper proposes a multi-objective model, (1) minimum project duration, (2) resource availability, and (3) minimum cash flow deficit. The model is divided into two modules. The first module produces an optimized, automated schedule achieving the objective of minimum project duration. The second module applies an optimized resource constraint scheduling using user input data of the available resources to allocate the resources on each activity while maintaining the maximum number of resources on site. The model is optimized using an evolutionary algorithm, namely: Genetic Algorithm (GA).

**Keywords** Automated resource scheduling · Genetic algorithm · Construction projects

---

R. M. Moharram (✉) · Y. A. S. Essawy · O. S. Hosny  
Department of Construction Engineering, The American University in Cairo, New Cairo, Egypt  
e-mail: [raghdammoharram@aucegypt.edu](mailto:raghdammoharram@aucegypt.edu)

Y. A. S. Essawy  
Department of Structural Engineering, Ain Shams University, Cairo, Egypt

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_32](https://doi.org/10.1007/978-3-031-34593-7_32)

## 1 Introduction

Construction scheduling is one of the most essential and critical parts of the project. It affects all the sequence of a project, starting from the engineering, the manufacturing, the procurement, the construction itself...etc. Moreover, the lack of a well-structured schedule for any project may result in severe impacts, like cost overruns and not meeting the project's milestones. An effective schedule can determine the feasible work sequence within a specific requirement. It also helps reduce surprises that may arise due to lack of good planning. Generally, there are many aspects a project could be scheduled on. However, construction managers tend to schedule and plan projects based on time or cost most of the time. Nowadays, these are the two main factors of a project; thus, managers usually base all their scheduling and resources on making a time cost trade-off; where they are trying to produce the project in the least time in the most cost-effective ways. Unfortunately, this technique opens the space for some other problems, like unmet productivity due to lack of scheduling based on resources, additional unexplained costs because they did not take into account an extra cost in a particle activity that may take place, or additional shifts for resources that need to be paid...etc. Cash flows are an item that is usually in projects left behind until the end of the planning process. Construction managers typically look at the cash flow at the end to determine how much they will need to finance from the bank. However, this always led to a decreased profit that was not taken into consideration.

## 2 Literature Review

### 2.1 Automated Scheduling

According to research done by Chua et al. [6], researchers are always trying to develop construction activities sequences to be automated; however, there is a need for a scheduling model that attempts to integrate the functional requirements to be reasoned automatically into temporal constraints. Moreover, as Amer [3], Amer et al. [4] mentioned, atomization of the construction scheduling research has been focused on in the literature. Even though many researchers have achieved automatic sequencing between scheduling activities and establishing breakdown structure, most construction projects still prefer working using the annual solutions; the problems were typically: "(1) lack of flexibility in how construction knowledge is stored in existing construction method model templates for sequencing algorithms; (2) the dependency of current automated scheduling methods on manually formed and maintained work templates; (3) lack of learning methods to automate learning of construction knowledge from existing records without extensive human input; (4) limited validation on the applicability of existing automated planning systems on real-life construction projects; and (5) the decoupled nature of research on automated planning versus schedule optimization."

## ***2.2 Construction Delays***

As Assaf and Al-Heji [5] defined, construction delays are overrun the time beyond the project's completion date. Research done by Sambasivan and Soon [8] and Agyekum-Mensah and Knight [2] found that more than 40% of construction projects experience delays or time overruns globally. Moreover, according to the research, the minority of the projects of building contracts were completed on time, and the projects that experienced delays had an average delay of 40%.

## ***2.3 Major Causes of Construction Delays***

Construction delays occurrence were due to some common factors: (1) poor management, (2) schedule pressure, (3) lack of owner's financing/payment for completed work, (4) design modification, and (5) shortage of materials. This literature will focus on schedule pressure and financing problems in detail.

### **2.3.1 Schedule Pressure**

Site managers usually produce highly pressured schedules to meet specific deadlines and maintain the project on schedule. Moreover, from the client or employer's side, they pressure the contractor by setting more tight deadlines. The main problem is the lack of understanding of the consequences and effects of these pressures on the schedule [7].

### **2.3.2 Financing Problems**

Four leading causes are contributing to the problem happening in the construction industry. The factors are typically: (1) Late payment by the client, (2) financial market stability, (3) insufficient financial resources, and (4) poor cash flow management.

## ***2.4 Genetic Algorithm***

Genetic algorithm is a search algorithm inspired by Darwin's Theory of Evolution in Nature. The main idea is to simulate the nature of mutation and reproduction of the genes. Genetic algorithm searches for the best genes to generate an algorithm optimized to achieve the optimal solution. The main idea is that the algorithm maintains a population where the better-adapted solution to the environment, the higher chance it will survive.

3 Methodology

3.1 Model Generation

The model will be generic; it will be applicable for any number of elements. For the time being, the model will be suitable for concreting activities, and it may be developed in the future for more activities.

The model starts with user input. The user will enter all the elements they have with all their predecessors. They will need to enter the element, its type, data (area and volume), and its predecessors, as shown in Fig. 1.

Then, the user will have to enter the productivity rates for the different types of activities (Typical for concreting activities), as shown in Fig. 2.

Then, the user will have to input the percentages that will be used for the cash flow, as shown in Fig. 3.

Element Number	Element	Type	Area (m2)	Volume (m3)	Predecessor 1	Predecessor 2	Predecessor 3	Predecessor 4	Predecessor 5
1	C1	Column	100	500	0	0	0	0	0
2	C2	Column	100	500	0	0	0	0	0
3	C3	Column	100	500	0	0	0	0	0
4	C4	Column	100	500	0	0	0	0	0
5	B1	Beam	100	500	C1	C2	0	0	0
6	B2	Beam	100	500	C3	C4	0	0	0
7	B3	Beam	100	500	C2	C3	0	0	0
8	B4	Beam	100	500	C3	C4	0	0	0
9	F1	Floor	100	500	B1	B2	B3	B4	0
10	F2	Floor	100	500	B1	B2	B3	B4	0

Fig. 1 User entered data 1

Fig. 2 User entered data 2

Type	Activity	Productivity Rates
CONCRETE	Formwork	5
	Steel rebar	3
	Concreting	1
	Curing	2
	Removing formwork	2

Fig. 3 User entered data 3

INDIRECT COST	10%
OVERHEAD	12%
MARKUP	10%
ADVANCE PAYMENT	10%
RETENTION	5%
DIFICIT EXPENSES	1%

Element Number	Element	Type	Area (m2)	Volume (m3)	Predecessor 1	Predecessor 2	Predecessor 3	Predecessor 4	Predecessor 5	SS	Count	Ref
1	C1	Column	100	500	0	0	0	0	0	1	1	11
2	C2	Column	100	500	0	0	0	0	0	1	2	12
3	C3	Column	100	500	0	0	0	0	0	1	3	13
4	C4	Column	100	500	0	0	0	0	0	1	4	14
5	B1	Beam	100	500	C1	C2	0	0	0	2	1	21
6	B2	Beam	100	500	C3	C4	0	0	0	2	2	22
7	B3	Beam	100	500	C2	C3	0	0	0	2	3	23
8	B4	Beam	100	500	C3	C4	0	0	0	2	4	24
9	F1	Floor	100	500	B1	B2	B3	B4	0	3	1	31
10	F2	Floor	100	500	B1	B2	B3	B4	0	3	2	32

Fig. 4 Coding 1

After the user has entered all the inputs, it is time to generate the chromosomes. Each gene in the chromosome will represent an element. The chromosome is always generated to be achieving all the predecessors; to achieve better results that always all the trials will be considered.

First, we will have a sample of 10 genes equal to the number of elements we have. The sequence step for the activities is first calculated, and the count cell will show the numbering of the element in this sequence step in the column. Then, this is converted into a code (Ref column) for each element to have a unique code. The formulas are shown in Fig. 4.

The first gene available options are calculated by searching for all the activities with a sequence step of 1; it is arranged to get the first sequence step, then the second sequence step, then the third...etc. This was done by searching for the code using the sequence step number and gene number in the first column (now the code is generated). The formulas are shown in Fig. 5.

All the following genes' available options are done using the same technique. All activities are listed, then the model searches for its predecessors if it was fulfilled before or not. If it is done, it will take a zero value; if it is not yet done, it will take one value. Then, we calculate the predecessor score, which is the sum of all the predecessors' columns; this shows the activities that are ready to start with a zero and the others with numbers more than zero (which is the same idea as the sequence step). Afterward, to generate the code or the reference, the same ranking idea was

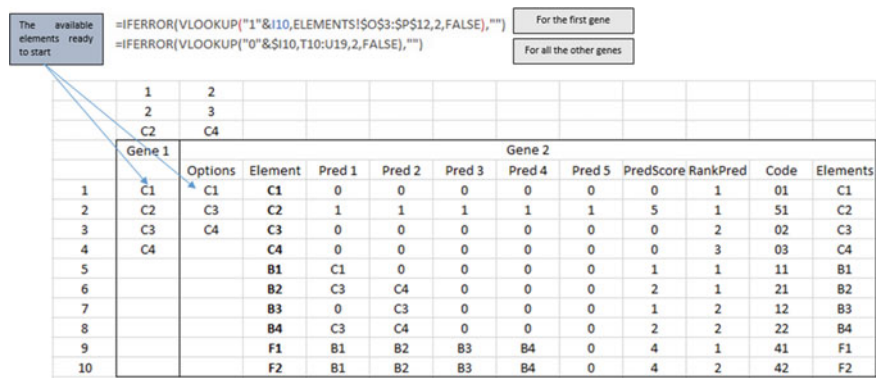


Fig. 5 Formulas 1

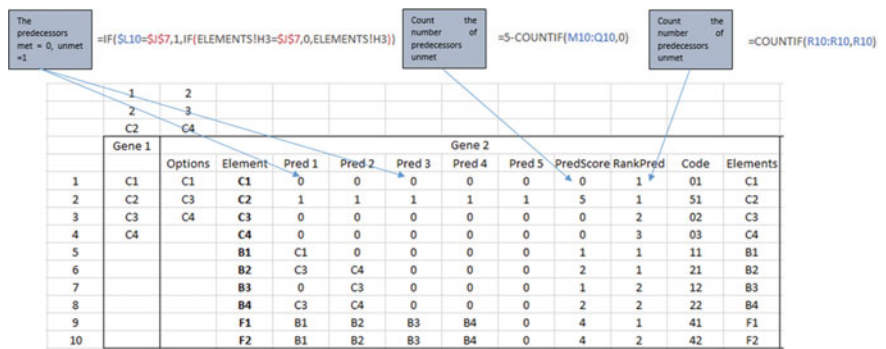


Fig. 6 Formulas 2

generated; the RankPred cell will show the numbering of the element in this sequence step in the column. The formulas are shown in Fig. 6.

In previous steps, the model constantly searches for the available elements that could be selected in this gene; now, it is the time to choose from the available elements. It will be simply a rand between function between the available elements. The formulas are shown in Fig. 7.

The duration is calculated by using the productivity equation:  $\text{Duration} = \text{Quantity}/(\text{No. of Crews} * \text{productivity rates})$ .

The schedule is calculated by defining all the predecessors' finish dates and taking the maximum date. The formulas are shown in Fig. 8.

As for the cash flow, it was calculated by first identifying the cost in each day for each activity by the schedule produced and the cost per day for each activity (it is assumed to be linear).

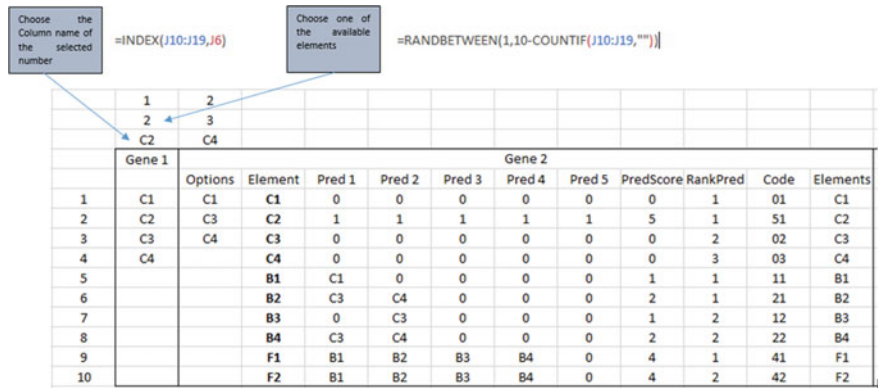


Fig. 7 Formulas 3

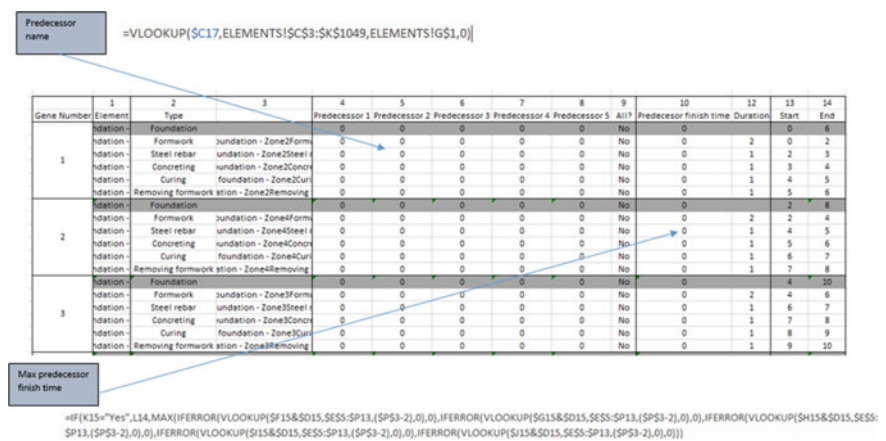


Fig. 8 Formulas 4

Then, the model will do the cash flow analysis monthly; it will first calculate the total direct cost for each month by summing each thirty-day (month) cost. Then, the indirect cost and overheads will be the percentages that the user has inputted first. Finally, the total cost is calculated as well as the cumulative cost. The formulas are shown in Fig. 9.

Afterward, the price will be calculated based on the markup from the user. Then, for the analysis, the cash in should be invoices that will be received two months (assumption) after submitting it (cash out). Also, deductions of the advance payment amount and the retention should be deducted from the payment. Now, cumulative cash

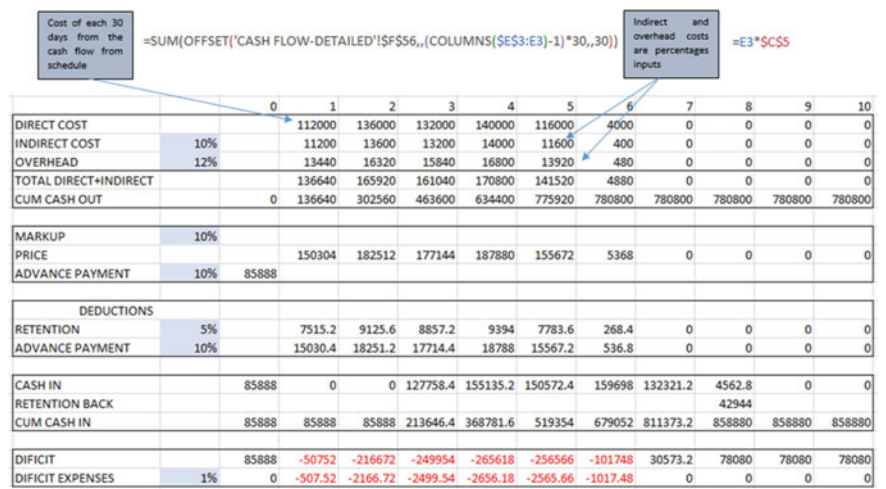


Fig. 9 Formulas 5

Advance Payment paid will be input from the user =C12\*SUM(E11:I11)

Price after adding markup percentage =E7\*(1+SC\$10)

Deductions will take place which is the retention and the advance payment received =E11-E15-E16

Deficit will be the difference between the cumulative cash in and out and the expenses is from the percentage for the negative values =E20-E8  
=IF(E22<0,SC\$23\*E22,0)

		0	1	2	3	4	5	6	7	8	9	10
DIRECT COST			112000	136000	132000	140000	116000	4000	0	0	0	0
INDIRECT COST	10%		11200	13600	13200	14000	11600	400	0	0	0	0
OVERHEAD	12%		13440	16320	15840	16800	13920	480	0	0	0	0
TOTAL DIRECT+INDIRECT			136640	165920	161040	170800	141520	4880	0	0	0	0
CUM CASH OUT		0	136640	302560	463600	634400	775920	780800	780800	780800	780800	780800
MARKUP	10%											
PRICE			150304	182512	177144	187880	155672	5368	0	0	0	0
ADVANCE PAYMENT	10%		85888									
DEDUCTIONS												
RETENTION	5%		7515.2	9125.6	8857.2	9394	7783.6	268.4	0	0	0	0
ADVANCE PAYMENT	10%		15030.4	18251.2	17714.4	18788	15567.2	536.8	0	0	0	0
CASH IN		85888	0	0	127758.4	155135.2	150572.4	159696	132321.2	4562.8	0	0
RETENTION BACK										42944		
CUM CASH IN		85888	85888	85888	213646.4	368781.6	519354	679052	811373.2	858880	858880	858880
DIFCIT		85888	-50752	-216672	-249954	-265618	-256566	-161748	30573.2	78080	78080	78080
DIFCIT EXPENSES	1%	0	-507.52	-2166.72	-2499.54	-2656.18	-2565.66	-1617.48	0	0	0	0

Fig. 10 Formulas 6

in is calculated. Finally, the cash flow deficit is calculated by finding the difference between the cumulative values of the cash in and the cash out; this will indicate the cumulative financing value needed for the project; then, the expense of financing is calculated. The formulas are shown in Fig. 10.

### 3.2 Model Stages

The model has three stages; the first is to produce a schedule using the rand() function to test the model's validity. The second stage is producing an optimized schedule based on the minimum project duration only after assuming only one resource is available on site. The second stage is resource scheduling, where lags are applicable added as a variable, and the number of resources is based on the maximum availability of on-site resources.

### 3.3 Genetic Algorithm

All the rand() functions will be removed. These become the variables that will be used in the genetic algorithm process with the aid of software (evolver) to optimize the model and reach the optimum solution.



### **3.4 Outputs**

The model will produce an optimized schedule achieving three objectives: (1) minimum project duration, (2) achieving resource constraints, and (3) minimum deficit. Moreover, the model will produce graphs for further analysis, namely resource histogram and cash flow.

### **3.5 Limitation**

1. The model accepts only finish to start relationships
  - We could overcome this limitation by changing the relationship to work around with finish to start.
  - For example, SS activities will be FS with the previous activity.
2. The model assumes the cost of an activity is linear with the time
  - We could overcome this limitation by trying to make the activity details as detailed as possible to make sure the cost of an activity is affected only by one factor.
  - For example, concreting is divided into formwork assembly, steel repair, pouring, curing, formwork disassembly.

## **4 Conclusion and Recommendation**

The proposed framework can serve as a generic model that can be applied on any type of projects regardless of the number of elements and accordingly the number of activities. The framework was verified, in the first stage, using random numbers in order to make sure that the model performs correctly. And, then, the model was used in order to optimize the construction duration (the second stage), while taking into account the resources availability (the third stage).

The proposed model utilizes Genetic Algorithm which proved efficient for such large-scale combinatorial optimization problems. Where the objective function could either be: minimum project duration, minimum cash flow deficit, and/or resource availability on-site.

The model was tested using an illustrative case. It is recommended to validate the model with a real case study and compare the resulting construction schedule.

## References

1. Abdul-Rahman H, Takim R, Min WS (2009) Financial-related causes contributing to project delays. *J Retail Leis Prop* 8(3):225–238. <https://doi.org/10.1057/rlp.2009.11>
2. Agyekum-Mensah G, Knight AD (2017) The professionals' perspective on the causes of project delay in the construction industry. *Eng Constr Archit Manag* 24(5):828–841. <https://doi.org/10.1108/ecam-03-2016-0085>
3. Amer WH (1994) Analysis and evaluation of delays in construction projects in Egypt. Master thesis, Zagazig Univ., Zagazig, Egypt
4. Amer F, Koh HY, Golparvar-Fard M (2021) Automated methods and systems for construction planning and scheduling: critical review of three decades of research. *J Constr Eng Manag* 147(7). [https://doi.org/10.1061/\(asce\)co.1943-7862.0002093](https://doi.org/10.1061/(asce)co.1943-7862.0002093)
5. Assaf SA., Al-Hejji S (2006) Causes of delay in large construction projects. *Int J Proj Manag* 24:349–357. <https://doi.org/10.1016/j.ijproman.2005.11.010>
6. Chua DKH, Nguyen TQ, Yeoh KW (2013) Automated construction sequencing and scheduling from functional requirements. *Autom Constr* 35:79–88. <https://doi.org/10.1016/j.autcon.2013.03.002>
7. Nepal MP, Park M, Son B (2006) Effects of schedule pressure on construction performance. *J Constr Eng Manag* 132(2):182–188. [https://doi.org/10.1061/\(asce\)0733-9364\(2006\)132:2\(182\)](https://doi.org/10.1061/(asce)0733-9364(2006)132:2(182))
8. Sambasivan M, Soon YW (2007) Causes and effects of delays in Malaysian construction industry. *Int J Proj Manag* 25(5):517–526

# **Construction Management: Sustainable Construction**

# Optimal Planning of Renewable Energy Integration for Off-grid Residential Buildings in Northern Regions



Don Rukmal Liyanage, Kasun Hewage, and Rehan Sadiq

**Abstract** Buildings consume 40% of the energy and contribute significantly to the world's GHG emissions (Cao et al. in *Energy Build* 128:198–213 [1]; Ürge-Vorsatz et al. in *Renew Sustain Energy Rev* 41:85–98 [2]). Although the majority of the building energy in the world is consumed for electricity, 17% of the global population in 2017 lived without electricity grid connectivity (Das et al. in *Appl Energy* 196:18–33 [3]). Most of this off-grid population relies on fossil fuel combustion to generate electricity, heating, and cooling energy, while many countries such as Canada try to reduce GHG emissions. In Canada, Northern territories comprise one-third of off-grid communities (Canada Energy Regulator in *Market snapshot: overcoming the challenges of powering Canada's off-grid communities* [4]). Approximately two-thirds of the Arctic and Northern communities in Canada rely on diesel (Canada's Arctic and Northern Policy framework [5]), which reduces the energy affordability and security in the residents in Northern regions. Exploiting the local renewable energy sources may reduce the energy operating costs and improve the local economy and energy security. However, it is necessary to consider the sustainability of implementing renewable energy sources. This study aimed to evaluate the environmental and economic sustainability of implementing renewable energy systems in off-grid residential buildings in the Northern region in Canada using a scenario-based assessment. The study conducted a life cycle GHG emission assessment, cost assessment, and discounted payback period analysis. The economic and environmental performances of energy system scenarios were aggregated using eco-efficiency parameters. The life cycle GHG emission assessment indicated that biomass heating systems can

---

D. R. Liyanage (✉) · K. Hewage · R. Sadiq  
School of Engineering, University of British Columbia (Okanagan Campus), 1137 Alumni  
Avenue, Kelowna, BC V1V 1V7, Canada  
e-mail: [liyanagedrd@alumni.ubc.ca](mailto:liyanagedrd@alumni.ubc.ca)

K. Hewage  
e-mail: [kasun.hewage@ubc.ca](mailto:kasun.hewage@ubc.ca)

R. Sadiq  
e-mail: [rehan.sadiq@ubc.ca](mailto:rehan.sadiq@ubc.ca)

reduce GHG emissions by over 75%. Furthermore, implementing wind turbines can improve the GHG emissions savings by up to 96%. The life cycle cost analysis indicated that implementing renewable energy systems can significantly reduce the life cycle cost with acceptable payback periods due to substantial operational cost savings. Furthermore, the energy system that included biomass heating system and micro-wind turbine that supply 80% of the electricity requirement has the lowest life cycle cost for energy. However, the eco-efficiency assessment showed that implementing biomass-based heating systems has the highest viability compared to all the other energy system scenarios due to the higher investment costs of micro-wind turbine installation. The findings of this study will assist community developers, policymakers, and researchers in planning renewable integration in off-grid communities.

**Keywords** Building energy · Renewable energy · Northern region · Life cycle thinking · Sustainability

## 1 Introduction

One of the key goals in the United Nations 2030 Agenda for Sustainable Development is to ensure everyone has affordable access to reliable and sustainable energy sources. However, approximately 1.3 billion people do not have access to modern energy services [6]. In Canada, about 0.2 million people in 280 communities are not connected to the North American electricity grid and natural gas pipe network [4]. Northern territories comprise one-third of Canada's off-grid communities [4]. In addition to the off-grid remote communities, grid-connected rural communities in Canada also suffer from a lack of energy security and higher energy costs. Approximately two-thirds of the Arctic and Northern communities in Canada rely on diesel. Diesel transportation significantly increases the costs and emits GHG emissions [5].

On the other hand, Canada consists of a significant percentage of unpopulated land area (nearly 40% of the landmass [7]). The majority of this unpopulated area belongs to the Northern regions in Canada. Recent publications indicate that the Canadian Northern regions are getting warmer three times the average global warming rate [8]. Therefore, it can be expected that extreme effects of climate change may occur in Northern regions [8]. For example, communities such as Inuit in Northern regions experience climate change effects such as changing patterns of sea ice, loss of snow and ice cover, and thawing permafrost [8]. The communities are trying to adapt their fishing industry to overcome the issues such as lack of ice [8]. The effects of climate change may also disrupt petroleum-based fuel transportation.

On the other hand, the studies that analyzed the impacts on renewable energy technologies by climate change predicted that wind, hydro, and biomass energy sources in the world's Northern hemisphere would be improved due to climate change [9]. Therefore, one of the solutions to improve the affordability and security of energy in Northern regions is to exploit local renewable energy sources further. Overall,

Canada is highly benefitted and relies on renewable energy sources. Hydroelectricity accounts for 55% of the total energy generation capacity in Canada. Hydro energy is considered a dispatchable energy source. Future energy projections in Canada indicate that the energy capacity shared by hydroelectricity will be reduced from 55 to 48% in 2040 compared to 2016 values. The reduction of hydro energy is mainly due to the expected increase of non-hydro renewable energy generation. It is predicted that the wind capacity will be increased from 12 to 27 GW, solar capacity will be increased from 2.3 to 8.6 GW, and the biomass capacity will be increased from 2.7 to 3.5 GW in 2040 compared to 2016 [10]. The increase of the non-hydro energy sources is mainly due to the rise of technical maturity and reduction of costs.

Supplying total energy demand using renewable energy is impossible due to various reasons. The main reason is the intermittent supply of renewable energy technologies such as solar and wind energy. Furthermore, the mismatch between the energy supply and the energy demand is another challenge that needs to be overcome. Therefore, it is necessary to integrate intermittent renewable energy systems with dispatchable energy systems to provide a reliable and steady energy supply for the community. More specifically, these technologies are known as hybrid technologies. Hybrid technologies integrate multiple energy sources, energy storage, and conversion technologies to achieve a common goal [11]. The integration of the technologies can be achieved through a common control framework or physically. Hybrid energy systems are supposed to enhance individual energy generation and conversion technologies [11].

It is necessary to consider sustainability to assess the viability of implementing alternative energy systems. Sustainability is defined as meeting the world's requirements without compromising the ability to meet future generations' needs [12, 13]. Sustainability is held by three main pillars: environment, economy, and society. The environmental sustainability pillar ensures the balance of world's environmental systems and natural resources. The economic sustainability pillar ensures the ability to maintain independence and access to resources and needs. Finally, social sustainability ensures health, security, and human rights while maintaining cultural and regional identity [13]. In addition, it is necessary to consider life cycle thinking on sustainability assessment of energy systems. Life cycle GHG emission assessment is widely used as an environmental performance indicator in energy system performance assessment studies [14–16]. Furthermore, life cycle cost is used in most of the studies to evaluate economic performance [17, 18]. Moreover, eco-efficiency is a combined indicator, which considers both life cycle GHG emission savings and costs savings [19] and is widely used in energy planning studies.

This paper aims to assess the environmental and economic performance of renewable energy systems on off-grid residential buildings in the Northern regions. The study conducted a case study on life cycle GHG emissions, life cycle economic assessment, and eco-efficiency analysis. A single-family residential building in an aboriginal community known as Baker Lake in Nunavut territory was considered.

2 Methodology

The overall methodology consists of building energy demand assessment using energy simulation, determining alternative energy sources based on resource availability, alternative energy system performance assessment, life cycle GHG emission assessment, life cycle cost assessment, eco-efficiency assessment, and payback period assessment. The subsequent sections describe the methods used in each step in the methodology.

2.1 Building Energy Demand Simulation

The study considered the Baker Lake community in Nunavut [20] to assess the economic viability and environmental performance of integrating renewable energy sources for an off-grid single-family detached house. Nunavut is completely dependent on imported petroleum-based fuels to generate electricity and heat. Table 1 shows the details of the selected community.

The software known as HOT2000 was considered to model a residential building. A housing plan provided in the reference [14] was used to model the building. The study considered the maximum overall thermal transmittance and ratio of fenestration and door area to gross wall area based on the national building energy coder 2017 for climate Zone 8 when modeling the building envelope. The details of the building envelop are given in Table 2. The maximum fenestration and door area to gross wall area ratio is 0.2.

The developed building energy model was simulated using HOT2000. The study evaluated the space heating load, domestic hot water heating load, and electrical load for air circulation fans, lights, and appliances.

**Table 1** Details of the selected community

Community	Baker Lake, Nunavut
Type	Aboriginal
Population	1728
Main power source	Diesel
Local grid availability	No
Annual energy demand	17,724 MWh
Heating degree days and climate zone	10235, Zone 8

**Table 2** Overall thermal transmittance for buildings

Building component	Overall thermal transmittance W/(m <sup>2</sup> K)
<i>Above-ground opaque building assembly</i>	
Walls	0.183
Roofs	0.121
Floors	0.142
<i>Other components</i>	
Doors	1.4
All fenestration	1.4
<i>Building assemblies in contact with the ground</i>	
Walls	0.21
Roofs	0.21
Floors	0.21

## 2.2 Alternative Energy System Screening and Scenario Development

The study considered solar heating, solar PV, wind turbine, and biomass heating as alternative energy systems. The average radiation should exceed 3 kWh/m<sup>2</sup>/day to implement a viable solar PV or solar thermal system. In addition, the annual average wind speed must exceed 4.66 m/s for feasible small wind turbine implementation [21]. The selected community has annual average radiation of 2.82 kWh/m<sup>2</sup>/day, lower than the minimum solar PV and thermal energy system installation requirement. The annual average wind speed of Baker Lake is 5.8 m/s and indicates that the implementing wind turbines will be viable [22]. On the other hand, biomass is an abundant renewable energy source in the Northern regions. Therefore, the study considered using wind and biomass as renewable sources that can be used to provide power and heating to the selected single-family residential building.

The study considered the following scenarios given in Table 3 to understand the viability of implementing renewable energy sources. The developed scenarios consist of different power generation sources, space heat generation, and water heating generation.

The required fuel used for biomass and oil heating systems and diesel power generation systems was calculated using the typical seasonal efficiencies of the heating systems and thermal efficiencies power generation systems. The energy contents of each fuel are given in Table 4.

The wind turbines (certified) and technical performance data given in Table 5 were used to evaluate the energy performance of improved cases 2–5.



**Table 3** Details of the scenarios

Scenario	Power generation source 1	Power generation source 2	Space heat generation source	Water heating generation source	Renewable energy coverage
Base case	Diesel	–	Oil	Oil	–
Improved case 1	Diesel	–	Biomass	Biomass	100% space and water heating
Improved case 2	Diesel	Wind	Biomass	Biomass	100% space and water heating using biomass and over 10% electricity coverage using wind turbine
Improved case 3	Diesel	Wind	Biomass	Biomass	100% space and water heating using biomass and over 30% electricity coverage using wind turbine
Improved case 4	Diesel	Wind	Biomass	Biomass	100% space and water heating using biomass and over 80% electricity coverage using wind turbine
Improved case 5	–	Wind	Biomass	Wind	100% space heating using biomass and 100% electricity and water heating coverage using wind turbine

**Table 4** Efficiency and energy content of energy sources

Equipment	Fuel type	Energy content (MJ/L) or (MJ/kg)	Seasonal efficiency (heating)	Thermal efficiency (power generation)
Combo space and DHW heater	Oil	39.21	Space heating: 85% DHW heating: 62%	–
DHW heater	Oil	39.21	55%	–
Biomass space heater	Biomass–wood pellets	18.61	85%	–
Diesel generator	Diesel	18.61	–	0.31

**Table 5** Wind turbine details

Manufacturer	Model	Peak power (kW)	Peak power wind speed (m s <sup>-1</sup> )	Hub height (m)	Swept area (m <sup>2</sup> )	Hub height annual average wind speed (m/s)	Estimated annual energy production (kWh)	Improved scenario
Hi VAWT Technology Corporation	DS3000	1.4	10.50	8.40	10.60	4.98	1280	2
Eveready Diversified Products	Kestrel e400nb	3.0	11.0	12.0	12.0	5.49	3929	3
SD Wind energy	SD6	6.1	17.00	9.00	23.7	5.08	8949	4
Bergey Windpower Company	Excel 10	12.6	16.50	30.0	38.5	6.79	22,300	5

### 2.3 Life Cycle Cost Assessment

The system boundary of the life cycle costing (LCC) consists of the capital cost of the equipment and 20 years of the operational phase of the heating and power systems system. The life cycle costs of the systems are calculated using Eq. 1, which was provided by the life cycle costing manual for the federal energy management program to assess the LCC of building energy and water conservation projects [23].

$$LCC = AQC - RES + FC + RM + OM - RG, \quad (1)$$

where

- LCC is life cycle cost of the heating system
- AQC is present value of the acquisition cost of the system
- FC is present value of the total fuel cost
- RES is present value of residual value of the system
- OM is present value of maintenance cost.

Annualized energy cost in each scenario is calculated using the equations given below [24] (considering heat and electricity together as some scenarios can supply both types of energy).

$$ALCC = I_0 + \sum_{t=1}^n \frac{(I_t + M_t + F_t)(1 + e)^t}{n(1 + e)^t}, \quad (2)$$

**Table 6** Cost parameters

Parameter	Value	References
Discount rate	1.24%	[25]
Facility lifetime	20	[17]
Initial cost of biomass heater	160 (CAD/kW)	[22]
Initial cost of wind turbine	6800 (CAD/kW)	[22]
O&M cost of wind turbine	70 (CAD/kW/year)	[22]
Cost of heating oil	1.10 (CAD/L)	[26]
Cost of diesel	1.79 (CAD/L)	[26]
Cost of biomass	0.33 (CAD/kg)	[27]

**Table 7** Fuel price increase scenarios

Fuel price increase scenario	Annual percentage increase
Fuel price increase scenario 1	0
Fuel price increase scenario 2	1%
Fuel price increase scenario 3	2%
Fuel price increase scenario 4	3%

where

- $I_0$  = initial investment cost
- $I_t$  = investment cost at the year  $t$
- $M_t$  = O&M cost at the year  $t$
- $F_t$  = fuel cost at the year  $t$
- $r$  = discount rate
- $e$  = fuel price increase rate
- $n$  = lifetime of the energy system (years).

Table 6 shows the parameters used to assess the life cycle cost in this study.

The study considered different fuel price escalation scenarios in Table 7. The escalation rates are based on the minimum, average, and maximum electricity price escalation rates in Canada.

## 2.4 Payback Period Analysis

In addition to evaluating the life cycle cost, the study assessed the payback period of implementing renewable energy systems using Eq. 3.

$$\text{DPP} = \ln\left(\frac{1}{1 - \frac{\text{IC} \times r}{E}}\right) / \ln(1 + r), \quad (3)$$

where

- DPP is discounted payback period
- IC is the investment cost
- $E$  is the annual savings
- $r$  is the discount factor.

## 2.5 Eco-efficiency Assessment

The equation shown below was used to evaluate the eco-efficiency of the improved cases compared to the base cases.

$$EEP = \begin{cases} \frac{|LCE|}{IC}; & \text{if } LCE < 0 \text{ AND } LCC \leq 0 \\ -\frac{|LCE|}{LCC}; & \text{if } LCE < 0 \text{ AND } LCC > 0 \\ 0; & \text{if } LCE \geq 0 \end{cases} \quad (4)$$

where

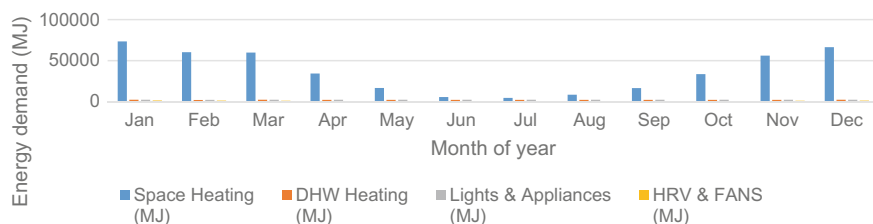
- ICE is the life cycle GHG emissions savings
- IC is the investment cost
- LCC is the life cycle cost savings.

## 3 Results

The results obtained from the energy simulation, life cycle GHG assessment, life cycle cost analysis, and discounted payback analysis are presented in this section.

### 3.1 Energy Simulation Results

Figure 1 shows the heating demand for space and DHW heating and electrical demand for lights, appliances, and air circulation systems. Energy simulation results indicated that the building energy load is predominantly for space and water heating, which is 93%. The electricity loads for lights, appliances, and air circulation account for 7% of the annual energy demand. Furthermore, the results indicate that the selected region requires heat energy during the whole year. The maximum heat and electricity load were observed in December, while the lowest was observed in July.



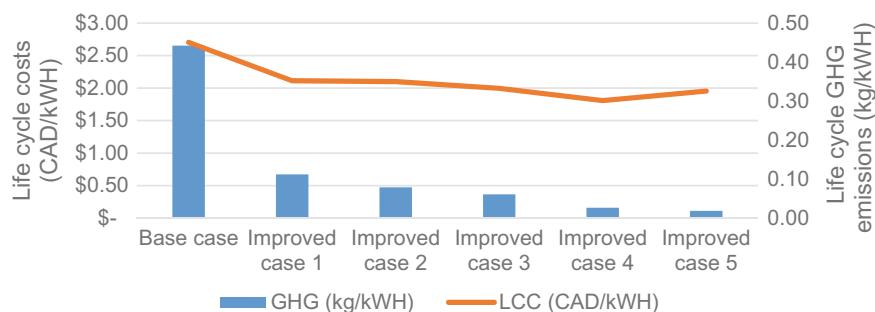
**Fig. 1** Energy demand results

### 3.2 Life Cycle GHG and Cost Analysis

Figure 2 shows the life cycle GHG emissions and costs of each scenario. The life cycle GHG emissions assessment indicates that implementing replacing the space heating system with the biomass energy system can reduce the life cycle GHG emissions by 75%. The improved case 2–4 reduced the life cycle GHG emissions by 82%, 86%, 94%, and 96%.

The figure shows that the base case, which provides energy using diesel and oil, has the highest life cycle energy cost per 1 kWh. Moreover, the percentage energy cost savings in improved cases 1, 2, 3, and 4 were 22%, 22.4%, 26%, 33%, and 27.7%. In addition, the highest cost saving was observed in the improved case 4, which includes biomass to cover 100% of heating demand while covering 80% of the electricity requirement using wind energy.

Figure 3 shows the results obtained from the eco-efficiency assessment. The results show that the eco-efficiency index reduced from improved case 1 to case 5, making improved case 1 the best scenario. This result is mainly due to the higher investment costs of wind power systems. Furthermore, the heating energy requirement is significantly higher than the electricity requirement. Therefore, substantial life cycle GHG emission reduction is caused by biomass energy systems for space and water heating.



**Fig. 2** Life cycle GHG emissions and costs

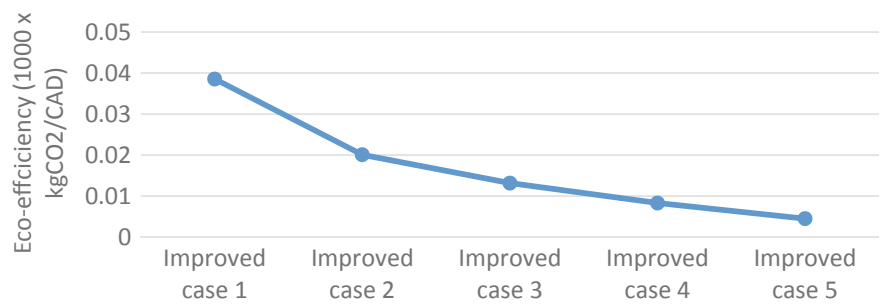


Fig. 3 Eco-efficiency index

3.3 Payback Period Analysis

Figures 4, 5, 6 and 7 show the cumulative cost savings of implementing renewable energy systems. Furthermore, the figures include the investment costs of renewable energy that helps to visualize the payback period of each improved case.

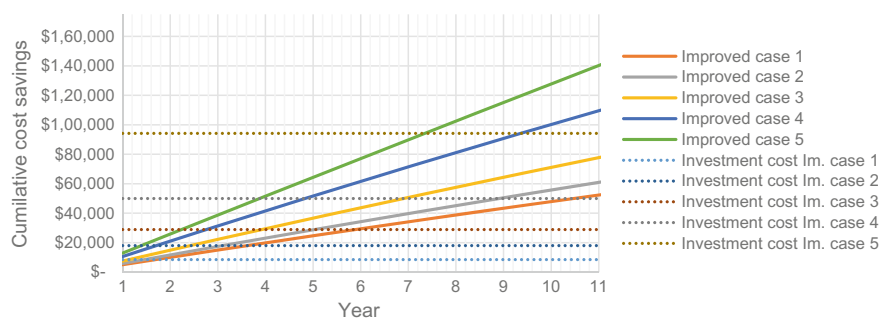


Fig. 4 Cumulative cost—fuel price increase scenario 1

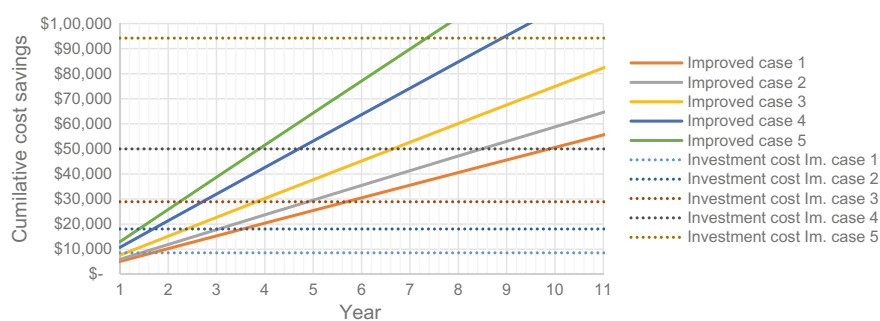
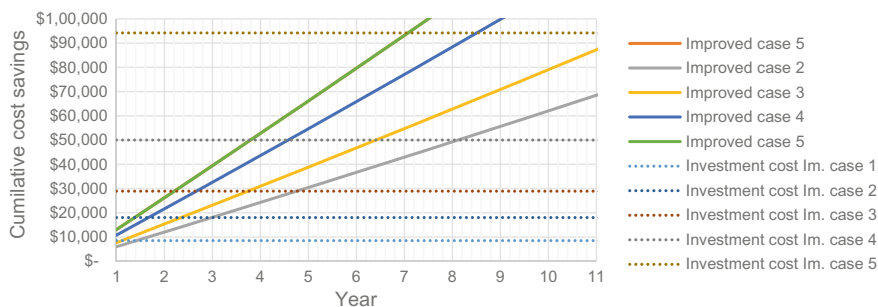
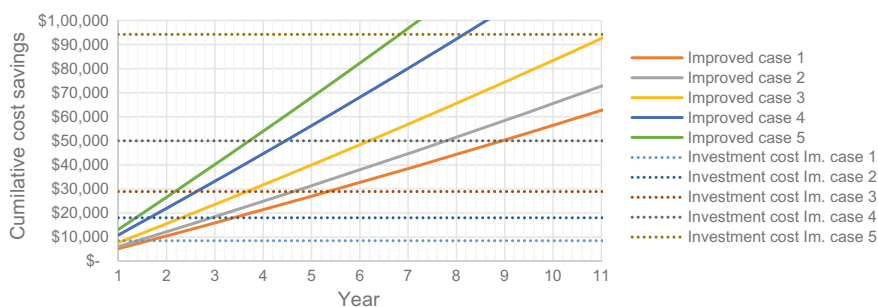


Fig. 5 Cumulative cost—fuel price increase scenario 2



**Fig. 6** Cumulative cost—fuel price increase scenario 3



**Fig. 7** Cumulative cost—fuel price increase scenario 4

Table 8 shows the summarized payback periods of each improved case. The results showed that the payback period is from 2 to 8 years. Improved case 1 has the lowest payback period, and case 5 has the highest. In addition, higher fuel escalation rates cause lower payback periods as expected. However, the reduction of the payback period is not substantially high when the energy price escalation is higher. On the other hand, Figs. 4, 5, 6 and 7 show that the cost saving is significantly high, although the payback period is high in the improved cases with higher renewable energy fractions.

**Table 8** Payback periods

Energy price escalation scenario	Payback period (years)				
	Case 1	Case 2	Case 3	Case 4	Case 5
1	2	4	4	5	8
2	2	4	4	5	8
3	2	3	4	5	8
4	2	3	4	5	7

## 4 Discussion

The building energy simulation results in this study indicated over 90% of the energy demand of the residential building considered in this study is contributed by the building space and water heating. Implementing renewable energy systems can reduce GHG emissions significantly in the off-grid residential buildings located in the Northern communities. Especially, implementing biomass-based heating systems can reduce GHG emissions by 75% due to the significant energy demand for heating in the Northern communities. Furthermore, implementing wind energy systems combined with biomass systems can reduce up to 96% GHG emissions.

The study clearly shows that substantial energy cost savings can be obtained despite the high investment cost associated with renewable energy implementation on single-family residential buildings in Northern regions. The biomass-based heating system has relatively lower investment costs than the wind energy system integration and appeared to be a better alternative than wind energy sources. The eco-efficiency assessment also indicated that the biomass-based heating system strategy has higher combined economic and environmental performance compared to the other scenarios that consider wind energy. This is mainly due to the higher heating requirement and the lower electricity requirement in the Northern regions. This result indicated that more research focus must be established on renewable thermal energy systems such as solar thermal and biomass heating and thermal energy storage when considering the Northern regions. On the other hand, biomass-based energy systems can be considered dispatchable and can provide steady energy output. One of the limitations of this study was not considering power generation through biomass-based combined heating and power (CHP) energy systems. Further studies on biomass-based CHP energy systems and multi-energy systems must be considered in future studies. Especially, lower electrical efficient and cost-effective biomass-based CHP energy systems may be more suitable for Northern communities due to the heating energy-dominated energy demand.

On the other hand, the economic and environmental performance of the selected energy systems can be substantially varied when considering mild climatic regions. Due to the lower heating demand, electricity-generating energy systems such as wind energy may have higher viability if there is sufficient wind speed. However, this study did not consider the intermittent energy supply of wind energy. Although wind energy systems can fulfill the required annual energy, they may not fulfill the short-term energy demand when the electricity demand is high. Large electrical energy storage can be used to store energy when wind energy is available and use it when there is energy demand. However, significant investment costs of electrical storage may reduce the viability of using large electrical storage.

The study indicated that the higher investment cost of renewable energy implementation is one of the main barriers to energy transition in the Northern regions. Policymakers may focus on providing subsidies and loans to implement renewable energy in building scale off-grid energy systems. One of the strategies to reduce the unit investment cost is considering centralized energy systems instead of off-grid



building-level energy systems. However, the energy transmission losses in off-grid single-family buildings are significantly lower as the energy is supplied in a location near the building compared to large community centralized energy systems. In addition, centralized energy systems will only be feasible in concentrated communities due to the high costs and transmission losses when delivering energy to long distances.

Considering the significant advantages of biomass as a renewable energy source, the government of Canada funded many renewable energy projects in the Northern and Arctic regions to reduce diesel use [5]. Many of the funded renewable energy projects are based on developing biomass-based energy systems. Biomass is considered one of the most viable and abundant renewable resources in the Northern regions in Canada. Biomass comes from many sources, such as harvestable trees, harvest residues, and trees killed by disturbances. Biomass energy can be generated using direct combustion, thermochemical conversion, chemical conversion, and biological conversion. It can be considered a renewable energy source and produce significantly less GHG emissions than fossil fuel. However, it is necessary to have a proper forest reformation strategy to consider biomass as a truly renewable energy source [28–30]. Therefore, further feasibility studies must be conducted on biomass-based renewable energy integration accounting for forest reformation and sustainable forest harvesting strategies.

## 5 Conclusion

The study conducted a life cycle thinking-based economic and environmental assessment of integrating renewable energy systems on off-grid residential buildings in the Northern region. The study evaluated the environmental performance of renewable energy systems using life cycle GHG emissions, life cycle cost of energy, discounted payback period, and eco-efficiency parameter. The renewable energy systems considered in this study are micro-wind turbines and biomass heating systems.

The life cycle GHG emission assessment indicated that biomass heating systems could reduce GHG emissions by over 75%. Furthermore, implementing wind turbines can improve the GHG emissions savings up to 96%. The life cycle cost analysis indicated that implementing renewable energy systems can significantly reduce the life cycle cost with acceptable payback periods due to substantial operational cost savings. However, the eco-efficiency analysis indicated that the implementation of the biomass heating system is more viable than the micro-turbine implementation. The result was mainly due to the substantial heating energy demand in the Northern regions. Therefore, wind turbines' significantly high investment cost cannot be justified considering the lower electricity demand.

The study indicated that it is important to focus on renewable thermal energy generation instead of electricity generation when supplying energy for buildings and communities in the Northern regions. However, the applicability of the results

in this study may substantially vary with the climatic regions and the availability of the energy sources. In addition, future studies must consider uncertainties such as global price fluctuations, technology advancements, and geopolitical conditions when evaluating the sustainability of renewable energy systems.

## References

1. Cao X, Dai X, Liu J (2016) Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy Build* 128:198–213. <https://doi.org/10.1016/j.enbuild.2016.06.089>
2. Ürge-Vorsatz D, Cabeza LF, Serrano S, Barreneche C, Petrichenko K (2015) Heating and cooling energy trends and drivers in buildings. *Renew Sustain Energy Rev* 41:85–98. <https://doi.org/10.1016/j.rser.2014.08.039>
3. Das BK, Al-Abdeli YM, Kothapalli G (2017) Optimisation of stand-alone hybrid energy systems supplemented by combustion-based prime movers. *Appl Energy* 196:18–33. <https://doi.org/10.1016/j.apenergy.2017.03.119>
4. Canada Energy Regulator (n.d.) Market snapshot: overcoming the challenges of powering Canada's off-grid communities. <https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/market-snapshots/2018/market-snapshot-overcoming-challenges-powering-canadas-off-grid-communities.html>. Accessed on 8 Dec 2021
5. Canada's Arctic and Northern Policy Framework (n.d.) <https://www.rcaanc-cirnac.gc.ca/eng/1560523306861/1560523330587>. Accessed on 11 Dec 2021
6. Akinyele DO, Rayudu RK (2016) Strategy for developing energy systems for remote communities: insights to best practices and sustainability. *Sustain Energy Technol Assessments* 16:106–127. <https://doi.org/10.1016/j.seta.2016.05.001>
7. About the North (n.d.) <https://www.cannor.gc.ca/eng/1368816431440/1368816444319>. Accessed on 15 Dec 2021
8. Warren FJ (2019) Canada in a changing climate: national issues. Natural Resources Canada, Canada
9. Solaun K, Cerdá E (2019) Climate change impacts on renewable energy generation. A review of quantitative projections. *Renew Sustain Energy Rev* 116:109415. <http://doi.org/10.1016/J.RSER.2019.109415>
10. National Energy Board (2017) Future energy projections in Canada 2017, 32
11. U.S.D. of Energy (2021) Hybrid energy systems: opportunities for coordinated research
12. McGill University (n.d.) What is sustainability?
13. Mateus R, Bragança L (2011) Sustainability assessment and rating of buildings: developing the methodology SBToolPT–H. *Build Environ* 46:1962–1971. <https://doi.org/10.1016/J.BUILDENV.2011.04.023>
14. Feng H, Liyanage DR, Karunathilake H, Sadiq R, Hewage K (2020) BIM-based life cycle environmental performance assessment of single-family houses: renovation and reconstruction strategies for aging building stock in British Columbia. *J Clean Prod* 250:119543. <https://doi.org/10.1016/j.jclepro.2019.119543>
15. Hwang JJ, Kuo JK, Wu W, Chang WR, Lin CH, Wang SE (2013) Lifecycle performance assessment of fuel cell/battery electric vehicles. *Int J Hydrogen Energy* 38:3433–3446. <https://doi.org/10.1016/J.IJHYDENE.2012.12.148>
16. Ristimäki M, Säynäjoki A, Heinonen J, Junnila S (2013) Combining life cycle costing and life cycle assessment for an analysis of a new residential district energy system design. *Energy* 63:168–179. <https://doi.org/10.1016/j.energy.2013.10.030>
17. Karunathilake H, Hewage K, Mérida W, Sadiq R (2019) Renewable energy selection for net-zero energy communities: life cycle based decision making under uncertainty. *Renew Energy* 130:558–573. <https://doi.org/10.1016/j.renene.2018.06.086>

18. Prabatha T, Hewage K, Karunathilake H, Sadiq R (2020) To retrofit or not? Making energy retrofit decisions through life cycle thinking for Canadian residences. *Energy Build* 226:110393. <https://doi.org/10.1016/j.enbuild.2020.110393>
19. Waththage TPHG (2022) A life cycle thinking-based energy retrofits planning approach for existing residential buildings. University of British Columbia. <https://doi.org/10.14288/1.0406300>
20. Royer J (2011) Status of remote/off-grid communities in Canada
21. WindEnergy Systems (n.d.) Natural Resources Canada Ressources naturelles Canada stand-alone a buyer's guide
22. Government of Canada, RETScreen (n.d.) <https://www.nrcan.gc.ca/maps-tools-and-publications/tools/modelling-tools/retscreen/7465>. Accessed on 5 Mar 2022
23. Fuller S, Petersen S (1995) Life-cycle costing manual for the federal energy management program, Gaithersburg
24. Karunathilake H, Hewage K, Brinkerhoff J, Sadiq R (2019) Optimal renewable energy supply choices for net-zero ready buildings: a life cycle thinking approach under uncertainty. *Energy Build* 201:70–89. <https://doi.org/10.1016/j.enbuild.2019.07.030>
25. Bank of Canada (n.d.) Canadian interest rates and monetary policy variables: 10-year lookup. Bank of Canada. [https://www.bankofcanada.ca/rates/interest-rates/canadian-interest-rates/?lookupPage=lookup\\_canadian\\_interest.php&startRange=2010-06-16&rangeType=daily&dFrom=2010-06-16&dTo=2020-06-16&rangeValue=1&rangeWeeklyValue=60&rangeMonthlyValue=60&series%5B%5D=V39078&ByDate\\_frequency=daily&submit\\_button=Submit](https://www.bankofcanada.ca/rates/interest-rates/canadian-interest-rates/?lookupPage=lookup_canadian_interest.php&startRange=2010-06-16&rangeType=daily&dFrom=2010-06-16&dTo=2020-06-16&rangeValue=1&rangeWeeklyValue=60&rangeMonthlyValue=60&series%5B%5D=V39078&ByDate_frequency=daily&submit_button=Submit). Accessed on 5 Mar 2022
26. Prices | Infrastructure (n.d.) <https://www.inf.gov.nt.ca/en/services/fuel-services/prices>. Accessed on 5 Mar 2022
27. Coffin DR (n.d.) Bulletin #7231, What we have learned about heating with wood pellets in Maine. Cooperative Extension Publications. University of Maine Cooperative Extension. <https://extension.umaine.edu/publications/7231e/>. Accessed on 5 Mar 2022
28. González A, Riba JR, Rius A (2016) Combined heat and power design based on environmental and cost criteria. *Energy* 116:922–932. <https://doi.org/10.1016/J.ENERGY.2016.10.025>
29. Kanematsu Y, Oosawa K, Okubo T, Kikuchi Y (2017) Designing the scale of a woody biomass CHP considering local forestry reformation: a case study of Tanegashima, Japan. *Appl Energy* 198:160–172. <https://doi.org/10.1016/J.APENERGY.2017.04.021>
30. Lam HL, Varbanov PS, Klemeš JJ (2011) Regional renewable energy and resource planning. *Appl Energy* 88:545–550. <https://doi.org/10.1016/J.APENERGY.2010.05.019>

# The Impact of Occupancy Pattern on Energy-Efficient Building System Selection: A Case Study of a Living Laboratory in Okanagan (BC)



S. R. Sultana, M. R. Kamal, M. F. A. Khan, M. Kamali, A. Rana,  
M. S. Alam, K. Hewage, and R. Sadiq

**Abstract** The building sector is at the frontline in consuming energy worldwide and a substantial contributor to the global carbon footprint. An energy-efficient building can perform a significant role in lowering the adverse impacts of greenhouse gas emissions. Buildings' physical characteristics and occupants' behavioural patterns intensely influence buildings' energy usage. With wide-scale adoption of the work-from-home concept in the last few years, the hours spent in residential buildings have spiked drastically. Building energy simulations can determine the associated increase in residential energy use. This paper proposes a research methodology that can help evaluate the impact of occupancy time on energy upgrades selection. This paper considers three levels of occupancy patterns, 25%, 50%, and 80% time spent at home and performance of three upgrades: heating, ventilation, and air conditioning system (HVAC), wall insulation (WL), and solar panels (PV), as well as their combinations. The performance is assessed in terms of carbon emissions and energy consumption costs. A real-life case study, a single-family detached home in Okanagan, Canada, demonstrates this methodology. The energy upgrade combinations were ranked through a multi-criteria decision-making method that considered economic and environmental criteria. The results revealed that PV and WL respond well under economic criteria. On the contrary, when greenhouse gas emission is the primary concern, HVAC combined with WL or PV has superior performance. The findings of this study can assist the construction practitioners (such as builders, developers, designers) and homeowners in making informed decisions for energy upgrade selection based on occupancy patterns.

**Keywords** Occupancy pattern · Energy efficient building system · Living lab in Okanagan (BC)

---

S. R. Sultana (✉) · M. R. Kamal · M. Kamali · A. Rana · M. S. Alam · K. Hewage · R. Sadiq  
The University of British Columbia, Okanagan, Canada  
e-mail: [rubaiy01@mail.ubc.ca](mailto:rubaiy01@mail.ubc.ca)

M. F. A. Khan  
United Nations Development Programme, Dhaka, Bangladesh

# 1 Introduction

Buildings are responsible for around 40% of the global energy use, while in Canada the building sector contributes nearly one-fourth of greenhouse gas (GHG) emissions [23]. Residential buildings have become a leading contributor to energy consumption [16]. This energy increase can be attributed to increased home sizes, large-scale adoption of heating, and hot water appliances [5]. As a result, there have been initiatives for exploring efficient ways of energy demand reductions in residential energy consumption considering the limited supply and environmental demerits of non-renewable resources [19]. According to Organization for Economic Development and Cooperation, the residents of Canada are one of the top per capita energy consumers in the world [17]. However, in line with the COP26 conference, Canada has committed to reducing its national GHG emissions to 45% by 2030 [10].

Adoption of energy efficiency measures (EEMs) is an effective way to minimize the energy needs of buildings. It helps to benefit from advanced technologies and access renewable energy sources designed to reduce energy consumption. The notion of EEMs is using energy appliances, equipment, or control systems that contribute to reducing energy use in their operation stage without compromising the occupants' comfort. Proper energy design and management in a building should consider four levers—building envelope (e.g. wall, roof, foundation); heating and cooling systems (e.g. HVAC); integration of renewable energy; and energy use patterns, which influence energy performance [3]. Among these four criteria, the first three have been widely explored, but there is dearth of studies related to impact of energy use patterns. Hence, this paper intends to fill the current research gap by considering a real-life case study building and utilizing calibrated energy model to help make informed decisions. The main objective of this work is to quantify how the energy efficiency of a building can be enhanced by energy upgrades under short and long occupancy times at home. For this purpose, three components including wall (building envelope), heating, ventilation, air conditioning (HVAC), and solar (renewable energy) are considered to experiment with energy upgrades based on occupants' energy use patterns. The goal of the study is to compare three different pictures for future energy upgrades recommendations as a guideline for EEMs.

Findings of this research would impact not only the energy efficiency of buildings but also the operational energy cost. The results can assist developers and homeowners to upgrade their residences based on the most efficient combinations of the aforementioned components. In addition, the study recommendations would benefit the policymakers' planning regulations, strategies, and policies for specific energy consumption standards that would impact on the environment and economy. Finally, the research outcome would guide different investors concerned about the environment and economy to choose standard energy upgrades with diverse occupancy profiles.

## 2 Occupancy Profiles

In recent years, with the advent of COVID-19 pandemic, governments have imposed lockdowns in almost every country worldwide. It restricted citizens' movement and confined them in quarantine for a long time [8]. Working from home has become the norm where people can continue working during or even after the recent pandemic. Many organizations, including governments across the globe, now, encourage people to work from home. It has been reported that almost three in four Chief Financial Officers (CFOs) plan at least 5% of formerly on-site employees to work remotely post-COVID [1]. People might stay home longer for several reasons, such as age, health, or other emergencies. Recent studies found that, during the lockdown, there was an alteration in electricity and hot water consumption among the residents of Canada. For example, the electricity and hot water consumptions increased by 46% and 103% between 9 AM and 5 PM in April 2020 [20]. Research conducted in Serbia also showed that before COVID-19, the average energy consumption in a house was 3414 kWh, and after several restrictions such as hand hygiene, room ventilation increased to 4509 kWh [7]. In Ottawa, Canada, a significant change of 13% in the daily electricity demand (from 16.3% to 29.1%) due to the COVID-19 pandemic was reported [2].

Several attempts have already been made to reduce the energy use in residential houses. A number of studies on multiple energy efficiency combinations in residential buildings were conducted and an optimum set of recommendations for newly constructed buildings have been recommended [12]. The results suggested different types of retrofits or energy-efficient upgrades such as low-flow showerheads, compact fluorescent lamps, upgrade of domestic hot water (DHW), among others. It was shown that the usage of cheap or no-cost appliances can be equivalent to costly energy efficiency works [22]. Occupants' behavioural patterns can also influence residential buildings' energy consumption. Scholars are researching the identification and classification of occupants' behaviour based on time use data [13]. In the case of air conditioning systems, behavioural patterns act as the most influential role in household energy consumption. A study conducted with a questionnaire survey exhibited that the correlation between using heating and ventilation systems and occupants' behaviour [21] changes the pattern of household energy consumption [15]. The heating and cooling system consumes more energy than any other system. According to the office of energy efficiency and renewable energy, the heating and cooling system is accountable for around 50% of a conventional home's energy usage [18]. A building can be turned into more energy efficient by upgrading with appropriate insulation, air sealing, and thermostat settings. This process can reduce GHG emissions and save energy consumption cost.

3 Methodology

This study aims to assess and identify the optimum energy-efficient measures under three different occupancy patterns. A three-stepped methodology is applied that includes data collection and monitoring, data modelling, and data analysis (Fig. 1).

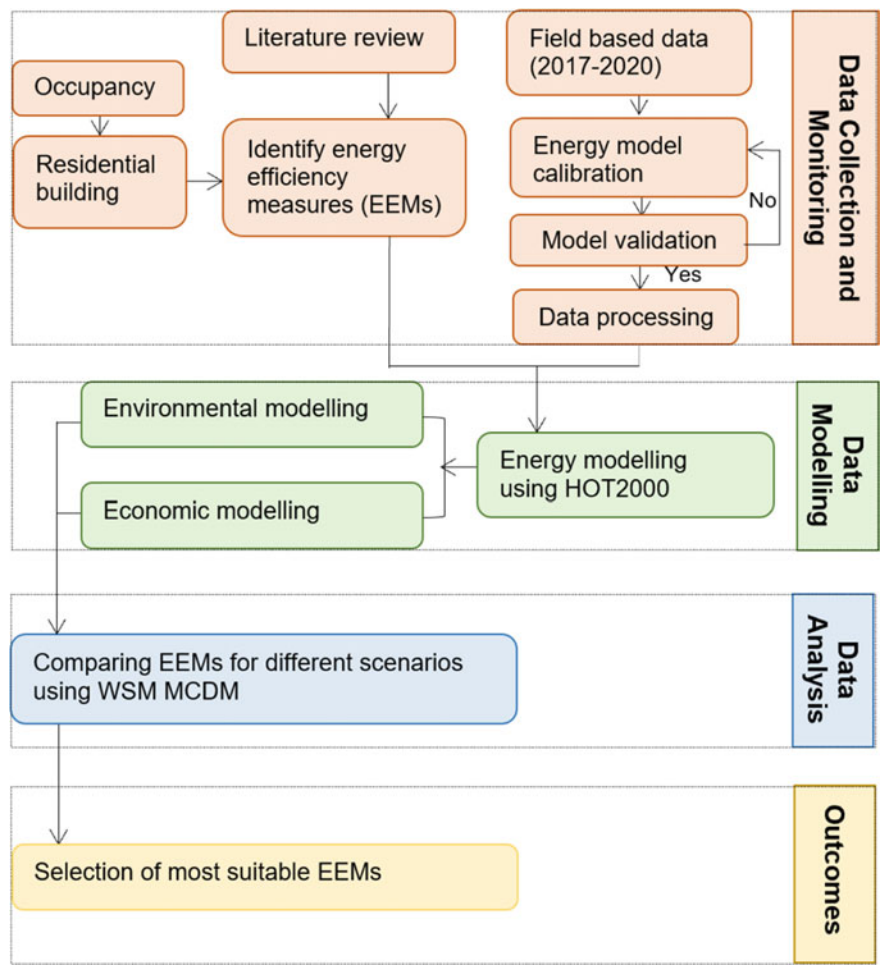


Fig. 1 Research methodology

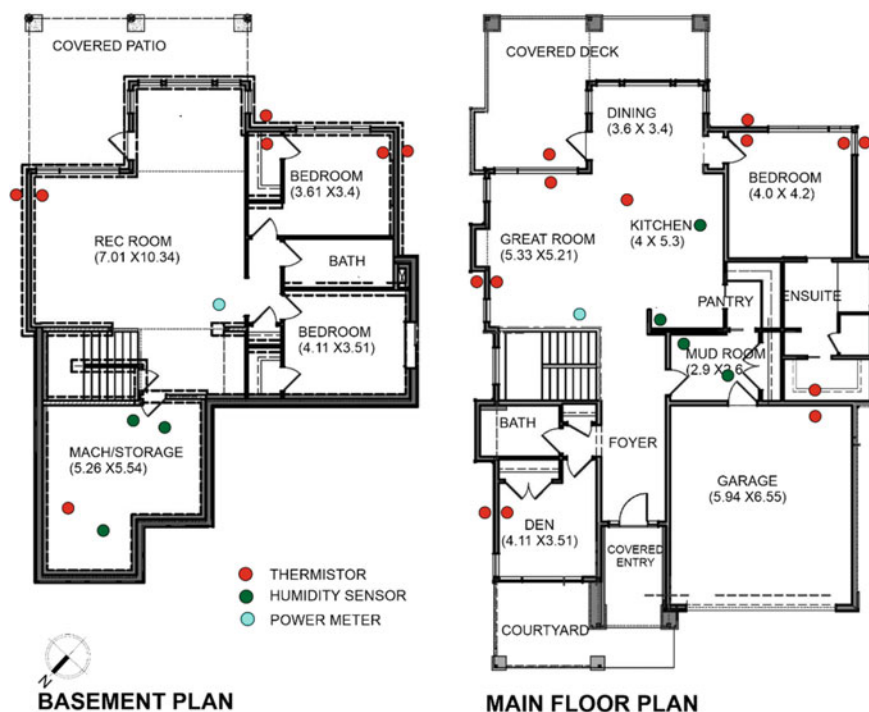


Fig. 2 Floor plans of the case study home and sensor locations

### 3.1 Data Collection and Monitoring

#### 3.1.1 Case Study

A single-family residential building is considered as the base case in this study. It was constructed in 2016 based on the BC building code 2012 [11], which is a prototype of the conventional residential buildings in the Okanagan region. It is a double-story building with a partially submerged basement. The pitched roofed building is northwest–southeast axis oriented. Its floor area covers 291.25 m<sup>2</sup> and contains living space, bedrooms, dining, kitchen, toilets, covered patio, and garage. The garage is located on the west corner of the building. Figure 2 and Appendix 1 represent general information, physical characteristics, and floor plans of the case home.

#### 3.1.2 Occupancy Profile

Two adults and two children were the occupants of the case study home. The adopted modelling in this study is not truly representing the actual occupancy profile. Three occupancy profiles—25%, 50%, and 80%—were chosen based on information



**Table 1** Technical specifications of energy efficiency measures (EEMs)

Energy efficiency measures (EEMs)	Nomenclature	Before upgrades	Energy efficiency measures
Space heating and cooling	HVAC	Energy star rated dual fuel Payne (PG92SCS) AFUE 92.1% 17.58 kWh Natural Gas & Payne (PA14NC) 9.96 kWh 14 SEER Split A/C	5 series (500A11) e Geothermal c/w ECM variable speed blower (Cooling 5.6 COP, Heating 4 COP)
Wall insulation	WL	1/2" drywall, USI 0.28 Batt (Eff. USI 0.32)	9.525 mm EPS Styrofoam and USI 0.28 Batt
Solar photovoltaic system	PV	None	Slope 25' azimuth 15' and 10 panels (16.3 m <sup>2</sup> )

Local developers, Authentech Homes, and RS Means residential cost database (RS Means Company, 2016) are the sources of costs associated with these energy efficiency measures in the study.

acquired from the literature review and data from the field. During the COVID-19 lockdown situation, people have started spending more time at home. The new normal situation introduces a new trend of lifestyle. Work-from-home or any unavoidable situation might encourage people to spend more time indoor in the coming future.

### 3.1.3 Energy Efficiency Measures (EEMs)

This study has chosen three components—space heating and cooling (HVAC), wall insulation (WL), and solar panels (PV) for energy upgrades. As stated earlier, suitable combination of energy efficiency measures (EEMs) would reduce energy usage without hampering the users' comfort. The energy system upgrades are selected based on the literature, neighbourhood developer preference, and local availability of equipment. The chosen EEMs for this research are given in Table 1.

## 3.2 Data Modelling

The study used HOT2000 to construct the base case model. HOT2000 is a tool designed for residential buildings in Canada to simulate provided annual electricity and natural gas consumption under various energy upgrade combinations [11]. The tool created energy models of the case study home addressing general information of its location, number of occupants and their occupancy at home, the specification of materials, and mechanical equipment used in the home. A typical minimum standard energy model household was converted to an advanced standard household through upgrading HVAC, wall insulation and installing solar panels. The energy models evaluated annual electricity and natural gas consumption.

The annual GHG emission is extracted from the energy simulation results of each set of energy upgrade combinations. The GHG emission ( $\text{kgCO}_2\text{eq./GJ}$ ) is considered the core of environmental criteria in this study and generated based on electricity and natural gas consumption results. The emissions were determined based on Eq. (1) using  $2.8 \text{ kgCO}_2\text{eq./GJ}$  for electricity and  $49.87 \text{ kgCO}_2\text{eq./GJ}$  for natural gas using British Columbia energy factors [4].

$$\text{Annual GHG Emissions} = \text{Annual energy consumption} \times \text{Emission factor}. \quad (1)$$

Economic modelling involved performing life cycle cost analysis based on upgrades costs provided from the developers. The extra expenditure needed was CA\$9000 for HVAC, CA\$1000 for wall insulation and CA\$8600 for PV system. Operation costs were determined based on local utility current rates of electricity and natural gas [9]. Difference in base case (S1) and upgraded scenarios (S2–S8) is considered for evaluating the operational cost saving. These savings were used to determine difference in total life cycle costs and to determine associated payback period of each upgrade combination using Eq. (2).

$$\text{Payback Period} = \text{Difference in life cycle costs/Extra expenditure}. \quad (2)$$

### 3.3 Data Analysis

Stakeholders such as developers, contractors, governmental policymakers, and users have different interests and preferences regarding energy upgrades. Much research noted that construction industry practitioners value more the economic criteria [14]. However, governments and building professionals prioritize sustainable building practices [6] specifically on environmental impact. The study opted the environmental and economic criteria as the main decision criteria and developed five scenarios—pro-environmental, pro-economical, neutral, and two intermediates by scoring them considering stakeholders' preferences as listed in Table 2.

Weighted sum method (WSM) as one of the simplest and effective multi-criteria decision methods (MCDM) is chosen to determine the best alternative based on environmental and economic criteria for this study. WSM ranked the energy upgrades combinations under the three occupancies (25%, 50%, and 80% time spent at home). Creation of decision criteria matrix, normalization of the decision matrix, and weighting and ranking are the three significant steps that usually WSM follows.

**Table 2** Weighting scenarios for the energy upgrades’ decision criteria

Scenario	Description	Environmental criteria	Economic criteria
		Carbon emissions	Payback Period
C1	Pro-environmental	1	0
C2	Pro-economical	0	1
C3	Neutral	0.5	0.5
C4	Intermediate	0.75	0.25
C5	Intermediate	0.25	0.75

Very important—1; moderately important—0.75; neutral—0.5; slightly important—0.25; not important—0

## 4 Result and Discussion

The results of energy simulations and annual GHG emission reduction potential due to the use of different energy upgrade combinations are presented in this section. Three energy upgrades have been selected with the potential for environmental and economic impact reduction. A total of eight combinations, denoted by S1–S8, were developed for each occupancy pattern, altogether 24 combinations. Each upgrade combination was separately applied to the base home with the help of the HOT2000 tool. Each upgrade combination delineated electricity consumption (in kWh) and natural gas consumption (in MCF), annual energy consumption (in GJ/Year), payback period (in years), and operational GHG emissions (in kgCO<sub>2</sub>eq/Year). Table 3 presents the key results, and Appendix 2 contains the details of the HOT2000 result.

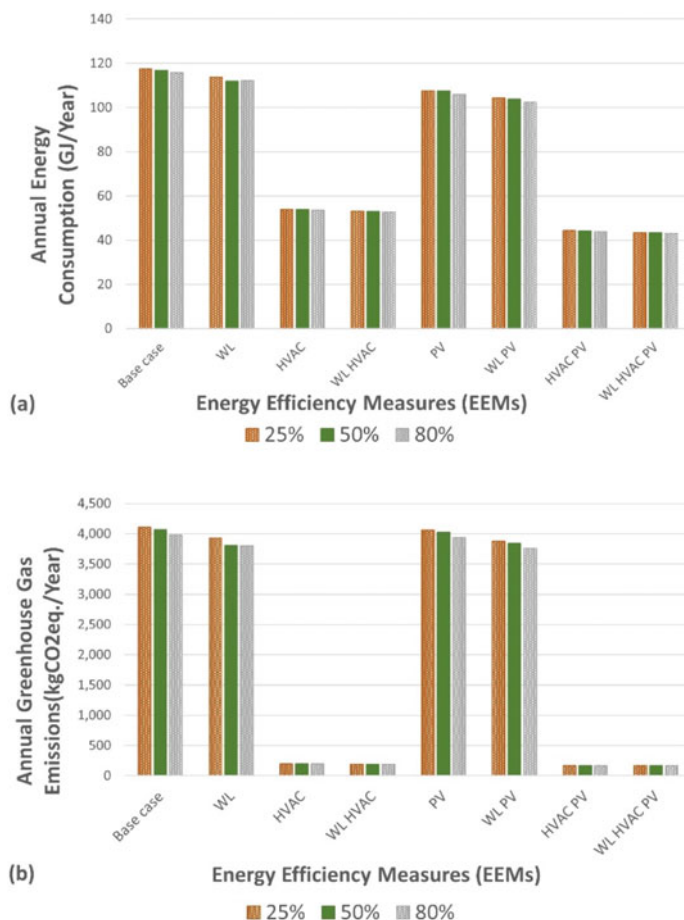
### 4.1 Environmental Performance

The annual energy consumption and GHG emissions for the three occupancy patterns are shown in Fig. 3. From Table 3 and Fig. 3, it is evident that the EEMs and their combinations are beneficial in reducing energy consumption and GHG emission. However, some of the EEMs increase the demand on electricity. From the trends, it is clear that electricity consumption rises with increase in occupancy times. Natural gas consumption, on the other hand, follows the opposite trend.

Energy saving due to different upgrade combinations can range between 43 and 117 GJ in the 25% occupancy time. In the case of electricity consumption, upgrading the HVAC system alone consumes the most electricity, while it shows the reverse trend for natural gas consumption. Moreover, the HVAC system reduces the GHG emission to more than 95% for all the occupancy patterns. Upgraded wall insulation does not affect electricity consumption, although it reduces natural gas consumption over 4%. It is also found that installation of solar panels can decrease electricity

**Table 3** Effect of different EEMs on different occupancy patterns

Occupancy	Scenario	EEEs upgrades	Electricity consumption (kWh)	Natural gas consumption MCF)	Annual operation cost—electricity (CA\$)	Annual operation cost—natural gas (CA\$)
25% at home	S1	Base case	10,343.8	76.1	1513.9	1167.7
	S2	WL	10,317.0	72.6	1509.9	1121.7
	S3	HVAC	14,728.5	0.9	2163.9	178.2
	S4	WL HVAC	14,473.3	0.9	2126.1	178.2
	S5	PV	7691.9	75.7	1120.8	1162.5
	S6	WL PV	7818.6	72.2	1139.6	1116.4
	S7	HVAC PV	12,047.8	0.9	1766.5	178.2
	S8	WL HVAC PV	11,794.8	0.9	1729.0	178.2
50% at home	S1	Base case	10,381.7	75.3	1519.5	1157.2
	S2	WL	10,508.4	70.3	1538.3	1091.4
	S3	HVAC	14,704.7	0.9	2160.4	178.2
	S4	WL HVAC	14,449.8	0.9	2122.6	178.2
	S5	PV	7884.8	75.0	1149.4	1153.2
	S6	WL PV	7873.1	71.5	1147.6	1107.2
	S7	HVAC PV	12,023.6	0.9	1762.9	178.2
	S8	WL HVAC PV	11,771.1	0.9	1725.5	178.2
80% at home	S1	Base case	10,613.5	73.6	1553.9	1134.8
	S2	WL	10,591.5	70.2	1550.6	1090.1
	S3	HVAC	14,617.0	0.9	2147.4	178.2
	S4	WL HVAC	14,365.1	0.9	2110.0	178.2
	S5	PV	7962.0	73.3	1160.8	1130.9
	S6	WL PV	7939.2	69.9	1157.4	1086.1
	S7	HVAC PV	11,937.6	0.9	1750.2	178.2
	S8	WL HVAC PV	11,685.7	0.9	1712.8	178.2



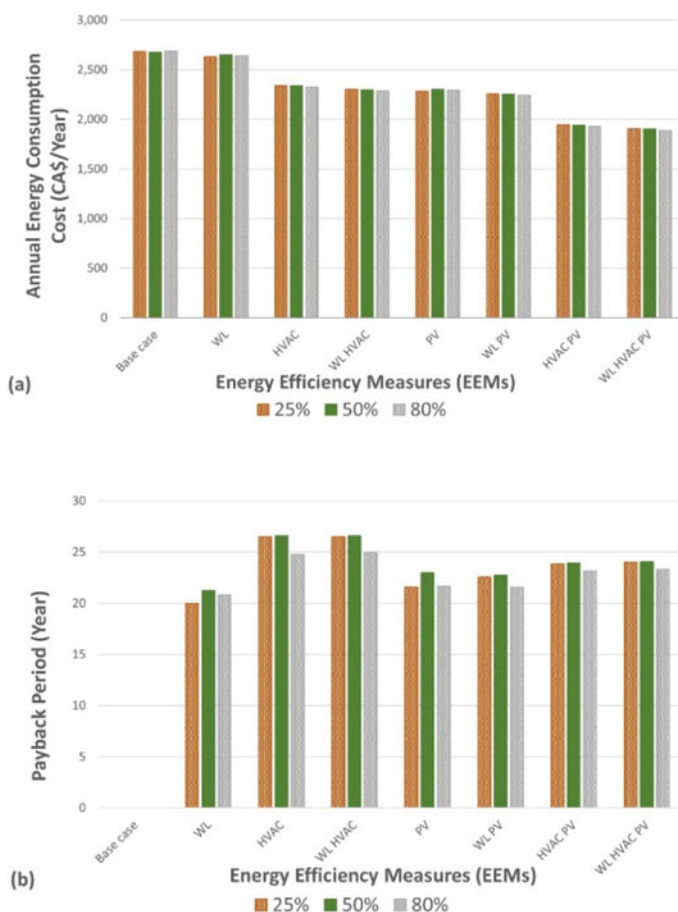
**Fig. 3** Environmental impact. **a** annual energy consumption (GJ/Year) and **b** annual GHG emissions (kg.CO<sub>2</sub>/year)

consumption up to 26%. The combination of upgraded wall insulation and solar panels can reduce the total electricity consumption by 24.5%. Although the combination of wall and HVAC system and the combination of HVAC and solar panels increase the energy consumption by 40% and 16.5%, respectively, these combinations significantly reduce the GHG emissions. Lastly, upgrading the house with all the three EEMs has the highest potential to reduce GHG emission.

## 4.2 Economic Performance

The change in annual energy consumption cost and the corresponding payback period due to the addition of the EEMs was analysed for the different occupancy patterns. Figure 4 illustrates the annual energy cost and the corresponding payback period, respectively.

It is observed that the annual operational cost of electricity follows the trend of electricity consumption pattern. However, the total operational cost does not necessarily follow the trend of total energy consumption. For instance, the HVAC upgrade consumes nearly half of the energy consumed by the PV upgrade, but the former results in more energy bills than the latter for all occupancy patterns. The electricity cost is the lowest for PV upgrades at 25% occupancy. However, for the other two



**Fig. 4** Economic impact. **a** annual energy consumption cost (CA\$/year) and **b** payback period (year)

occupancy patterns WL and PV combination results in the lowest cost of electricity. When the cost of natural gas is taken into consideration, the HVAC upgrade and its combinations result in over 84% cost reduction for all the occupancy patterns. Overall, the combination of all the three EEMs leads to the highest reduction in total energy cost for all occupancy patterns.

When a single energy upgrade is considered, WL can be considered the best as it leads to the lowest payback period. In the case of pairwise energy upgrades, the WL and HVAC combination is the worst combination as it has the highest payback period. The HVAC and PV combination results in the highest decrease in total energy cost. However, this combination has a slightly higher payback period than the WL and PV combination. Therefore, the HVAC and PV combination performs as good as the combination of the three EEMs with respect to energy cost and payback period. Overall, the study found significant differences on the performance of the EEMs themselves, whereas there are minor environmental and economic impacts from the three different occupancy patterns (Figs. 3 and 4).

### **4.3 Multi-criteria Analysis**

This study prioritized the environmental (GHG emission) and economic (payback period) criteria based on the literature and stakeholders' preferences. WSM MCDM was used to evaluate and rank all the 24 upgrades combinations. Such ranking would guide the stakeholders to choose the best combination in the three occupancy patterns (25%, 50%, and 80% time stay at home). Table 4 portrays three top energy upgrade combinations concerning five scenarios, including two extreme scenarios—pro-environment and pro-economic, neutral and two intermediate scenarios. The result illustrates that the combination and rankings are similar in a pro-environmental scenario for different occupancies. WL HVAC PV, HVAC PV, and WL HVAC are ranked first, second, and third, respectively. In pro-economic scenarios, WL ranked first for three different occupancy patterns. The WSM suggested WL, PV, or a combination of WL and PV performs better in the pro-economic scenario. In the neutral scenario, similar combinations and rankings are indicated for 25 and 50% occupancy patterns, while 80% occupancy pattern has a change. Two intermediate scenarios for 25 and 80% occupancy patterns are the same. However, in the 50% occupancy pattern, the two intermediate scenarios are different. In an intermediate case, two combinations carry the same ranking—HVAC PV and WL HVAC PV combinations are in the top-ranked position. One of the significant findings of this study shows that HVAC is a common component for all the scenarios except pro-economic scenarios. All three EEMs combination is suggested quite several times.

**Table 4** Top three energy upgrades combinations under five occupancy pattern scenarios

	Rank	Pro-Environmental C1	Pro-Economic C2	Neutral C3	Intermediate C4	Intermediate C5
25% time stay at home	1	WL HVAC PV	WL	HVAC PV	HVAC PV	HVAC PV
	2	HVAC PV	PV	WL HVAC PV	WL HVAC PV	WL HVAC PV
	3	WL HVAC	WL PV	WL HVAC	WL HVAC	WL HVAC
50% time stay at home	1	WL HVAC PV	WL	HVAC PV	HVAC PV/ WL HVAC PV	HVAC PV
	2	HVAC PV	WL PV	WL HVAC PV	WL HVAC	WL HVAC PV
	3	WL HVAC	PV	WL HVAC	HVAC	WL HVAC
80% time stay at home	1	WL HVAC PV	WL	HVAC PV	HVAC PV	HVAC PV
	2	HVAC PV	WL PV	WL HVAC PV	WL HVAC PV	WL HVAC PV
	3	WL HVAC	PV	HVAC	HVAC	HVAC

## 5 Conclusion

This paper proposed a new methodology to analyse the impacts of occupancy profile variation on energy upgrade selection. The method was applied to a real-life case study residential building. The environmental and economic performances were explored for three occupancy patterns, including 25, 50, and 80% of the occupants' time spent at home concerning eight energy upgrade combinations. In this study, HOT2000 tool was used for energy modelling and WSM MCDM was applied to rank and select the most appropriate EEMs for stakeholders to determine the optimal investment required for GHG emission reduction and a satisfactory payback period. The outcomes of the research advocate opting either HVAC or combined with WL and PV in favour of environmental, neutral, and intermediate criteria. The study suggested that all three EEMs (WL HVAC PV) combinations have the most potential to reduce GHG emissions. However, in the case of the economic criterion, HVAC does not seem to be a good choice; instead, solar panels and wall insulation respond well to the economic criteria.

This study considered the occupancy patterns from 25 to 80% to observe the impacts of GHG emission and financial viability. However, investigating the same criteria addressing extreme scenarios such as 0% and 100% occupancies could give



stakeholders more options to choose from the possible EEMs for their homes. Furthermore, this research used only three energy upgrades that may not offer a broader range of variation in the ranking process of upgrade combinations. In terms of the economic analysis (i.e. payback period), addressing inflation and energy incentives given by governments might help stakeholders to receive a more accurate picture of monetary benefit in the long run by comparing the present time with the future.

**Acknowledgements** The authors are grateful to Wilden Living Lab and Authentech Homes for delivering the information of the case study home and data of the energy upgrades. The authors also like to thank Canada's Natural Sciences and Engineering Research Council (NSERC) and Mitacs through NSERC Alliance (COVID-19) and Accelerate programmes. Finally, support from UBC's Green Construction Research and Training Centre (GCRTC) and Life Cycle Management Laboratory (LCML) is appreciated.

## Appendix 1: Details of the Case Study Home

Component	Building parameters description of the base home	
Occupancy	Occupancy profile	Two working adults Two teenagers
	Occupancy time	25, 50, and 80%
House area and built form	Construction year	2016
	Number of floors	2
	Total area of the household (m <sup>2</sup> )	291.25
	Floor area (m <sup>2</sup> )	249.91
	Wall area (m <sup>2</sup> )	364.57
	Door area (m <sup>2</sup> )	6.60
	Window area (m <sup>2</sup> )	57.91
Foundation and basement	Foundation	8" reinforced concrete, USI 0.26 Batt (Eff. USI 0.32), 3/8" OSB sheathing, 2" × 6" wood studs @ 24" o/c
	Basement slab	4" concrete
Wall system	Exterior wall	1/2" drywall, USI 0.28 Batt (Eff. USI 0.32), 3/8" OSB sheathing, 2" × 6" wood studs @ 24" o/c
	Interior wall	1/2" drywall, 2" × 4" wood studs
Roofing system	Roof	1/2" OSB sheathing, engineered trusses (wood), asphalt
Glazing type	Windows	Tightness: 0.2 L/sm <sup>2</sup> ; vinyl double-glazed windows c/w 180 low-E Air

Appendix 2: HOT2000 Result

Occupancy	Scenario	Upgrades	Environmental Criteria			Economic Criteria				Payback period (Year)
			Electricity consumption (kWh)	Natural gas consumption (MCF)	Total energy consumption (GJ/Year)	GHG emission (kgCO <sub>2</sub> eq./Year)	Annual operation cost electricity (CAS)	Annual operation cost —natural gas (CAS)	Annual energy consumption (CAS/Year)	
25% time stay at home	S1	Base case	10,343.8	76.1	117.5	4108.1	1513.9	1167.7	2681.6	
	S2	WL	10,317.0	72.6	113.7	3923.7	1509.9	1121.7	2631.6	20.0
	S3	HVAC	14,728.5	0.9	54.0	195.8	2163.9	178.2	2342.1	26.5
	S4	WL HVAC	14,473.3	0.9	53.1	193.2	2126.1	178.2	2304.3	26.5
	S5	PV	7691.9	78.7	107.6	4060.3	1120.8	1162.5	2283.2	21.6
	S6	WL PV	7818.6	72.2	104.3	3877.5	1139.6	1116.4	2256.0	22.6
	S7	HVAC PV	12,047.8	0.9	44.3	168.8	1766.5	178.2	1944.7	23.9
	S8	WL HVAC PV	11,794.8	0.9	43.4	166.2	1729.0	178.2	1907.2	24.0
50% time stay at home	S1	Base case	10,381.7	75.3	116.8	4066.4	1519.5	1157.2	2676.7	
	S2	WL	10,508.4	70.3	112.0	3804.6	1538.3	1091.4	2629.7	21.3
	S3	HVAC	14,704.7	0.9	53.9	195.6	2160.4	178.2	2338.6	26.6
	S4	WL HVAC	14,449.8	0.9	53.0	193.0	2122.6	178.2	2300.8	26.6
	S5	PV	7884.8	75.0	107.5	4025.4	1149.4	1153.2	2302.6	23.0

(continued)

(continued)	Occupancy	Scenario	Upgrades	Environmental Criteria			Economic Criteria				Payback period (Year)
				Electricity consumption (kWh)	Natural gas consumption (MCF)	Total energy consumption (GJ/Year)	GHG emission (kgCO <sub>2</sub> eq./Year)	Annual operation cost electricity (CAS)	Annual operation cost —natural gas (CAS)	Annual energy consumption (CAS/Year)	
80% time stay at home		S6	WL PV	7873.1	71.5	103.8	3841.2	1147.6	1107.2	2254.8	22.8
		S7	HVAC PV	12,023.6	0.9	44.2	168.5	1762.9	178.2	1941.1	23.9
		S8	WL HVAC PV	11,771.1	0.9	43.3	166.0	1725.5	178.2	1903.7	24.1
		S1	Base case	10,613.5	73.6	115.9	3979.3	1553.9	1134.8	2688.7	
		S2	WL	10,591.5	70.2	112.2	3800.2	1550.6	1090.1	2640.7	20.8
		S3	HVAC	14,617.0	0.9	53.6	194.7	2147.4	178.2	2325.6	24.8
		S4	WL HVAC	14,365.1	0.9	52.7	192.2	2110.0	178.2	2288.2	25.0
		S5	PV	7962.0	73.3	106.0	3936.8	1160.8	1130.9	2291.7	21.7
(continued)		S6	WL PV	7939.2	69.9	102.3	3757.7	1157.4	1086.1	2243.6	21.6
		S7	HVAC PV	11,937.6	0.9	43.9	167.7	1750.2	178.2	1928.4	23.1
		S8	WL HVAC PV	11,685.7	0.9	43.0	165.1	1712.8	178.2	1891.0	23.3





Appendix 3: Weighted Sum Method (MCMDM) MCMDM Result

Scenario	EEMs Upgrades	25% time stay at home					50% time stay at home					80% time stay at home				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
S1	Base case															
S2	WL	0.3118	0.1211	0.2164	0.2641	0.1688	0.3070	0.1265	0.2167	0.2618	0.1716	0.3111	0.1300	0.2206	0.2658	0.1753
S3	HVAC	0.0156	0.1606	0.0681	0.0518	0.1244	0.0158	0.1582	0.0870	0.0514	0.1226	0.0159	0.1546	0.0853	0.0506	0.1200
S4	WL HVAC	0.0154	0.1606	0.0880	0.0517	0.1243	0.0156	0.1581	0.0869	0.0512	0.1225	0.0157	0.1558	0.0858	0.0507	0.1208
S5	PV	0.3226	0.1308	0.2267	0.2747	0.1788	0.3248	0.1367	0.2307	0.2777	0.1837	0.3223	0.1351	0.2287	0.2755	0.1819
S6	WL PV	0.3081	0.1367	0.2224	0.2652	0.1795	0.3099	0.1353	0.2226	0.2663	0.1789	0.3076	0.1346	0.2211	0.2644	0.1778
S7	HVAC PV	0.0134	0.1447	0.0791	0.0462	0.1119	0.0136	0.1422	0.0779	0.0458	0.1101	0.0137	0.1444	0.0791	0.0464	0.1117
SB	WL HVAC PV	0.0132	0.1455	0.0794	0.0463	0.1124	0.0134	0.1430	0.0782	0.0458	0.1106	0.0135	0.1455	0.0795	0.0465	0.1125

## References

1. A Gartner, Inc (2020) Gartner CFO survey reveals 74% intend to shift some employees to remote work permanently. <https://www.gartner.com/en/newsroom/press-releases/2020-04-03-gartner-cfo-surey-reveals-74-percent-of-organizations-to-shift-some-employees-to-remote-work-permanently2>
2. Abdeen A, Kharvari F, O'Brien W, Gunay B (2021) The impact of the COVID-19 on households' hourly electricity consumption in Canada. *Energy Build* 250:111280
3. Ascione F, Bianco N, Stasio C, de Mauro GM, Vanoli GP (2018) Energy management in hospitals. In: *Comprehensive energy systems*, vols 5–5. Elsevier Inc, pp 827–854
4. BC Ministry of Environment (2014) BC best practices methodology for quantifying greenhouse gas emissions
5. Calero M, Alameda-Hernandez E, Fernández-Serrano M, Ronda A, Martín-Lara MÁ (2018) Energy consumption reduction proposals for thermal systems in residential buildings. *Energy Build* 175:121–130
6. Cousineau C (1999) Breaking through the barriers to sustainable building: insights from building professionals on government initiatives to promote environmentally sound practices
7. Cvetković D, Nešović A, Terzić I (2021) Impact of people's behavior on the energy sustainability of the residential sector in emergency situations caused by COVID-19. *Energy Build* 230:110532
8. Donthu N, Gustafsson A (2020) Effects of COVID-19 on business and research. *J Bus Res* 117:284–289
9. FortisBC (2021) FortisBC new home program—better homes BC 2021. <https://betterhomesbc.ca/rebates/fortisbc-new-home-program/>
10. Govt. of Canada (2021a) Govt. of Canada, 2021. Canada's achievements at COP26, Govt. of Canada, 2021
11. Govt. of Canada (2021b) Govt. of Canada, 2021(b). Tools for industry professionals. Govt. of Canada 2021(b)
12. Griego D, Krarti M, Hernández-Guerrero A (2012) Optimization of energy efficiency and thermal comfort measures for residential buildings in Salamanca, Mexico. *Energy Build* 54:540–549
13. IEEE Power & Energy Society, IEEE International Conference on Service-Oriented Computing and Applications, IEEE PES Innovative Smart Grid Technologies Conference 19–22 Feb 2014, Washington, DC & ISGT 19–22 Feb 2014, Washington, DC, 2014, IEEE PES Innovative Smart Grid Technologies Conference (ISGT), 2014 Washington, DC, USA, 19–22 Feb 2014
14. Kamali M, Hewage K (2017) Development of performance criteria for sustainability evaluation of modular versus conventional construction methods. *J Clean Prod* 142:3592–3606
15. Kashif A, Ploix S, Dugdale J, Le XHB (2013) Simulating the dynamics of occupant behaviour for power management in residential buildings. *Energy Build* 56:85–93
16. Mazzarella L (2015) Energy retrofit of historic and existing buildings. the legislative and regulatory point of view. *Energy Build* 95:23–31
17. OECD (2019) Organization for economic development and cooperation (OECD). Greenhouse gas emissions
18. Office of Energy Efficiency and Renewable Energy (2018) Heating, ventilation, and air conditioning (HVAC). <https://rpse.energy.gov/tech-solutions/hvac>
19. Refahi AH, Talkhabi H (2015) Investigating the effective factors on the reduction of energy consumption in residential buildings with green roofs. *Renew Energy* 80:595–603
20. Rouleau J, Gosselin L (2021) Impacts of the COVID-19 lockdown on energy consumption in a Canadian social housing building. *Appl Energy* 287:116565

21. Santin OG (2011) Behavioural patterns and user profiles related to energy consumption for heating. *Energy Build* 43(10):2662–2672
22. Shandilya A, Hauer M, Streicher W (2020) Optimization of thermal behavior and energy efficiency of a residential house using energy retrofitting in different climates. *Civil Eng Archit* 8(3):335–349
23. Stinson S (2017) Energy efficiency in the buildings sector presentation to the Senate Committee on Energy, The Environment and Natural Resources



# An Approach for Teaching the Design of Net Zero-Energy Buildings



Mokhtar Ahmed

**Abstract** Architecture education must prepare future architects for the challenges of global warming. Therefore, performance-based designs, particularly those with an energy focus, should be understood and appreciated by architecture students. This is particularly essential considering the requirements of many countries to have net zero-energy new buildings within a decade and to retrofit old buildings to that standard within two to three decades. This paper presents an approach to teaching a course on the subject to architecture students in a college that is oriented to form-based designs. The paper discusses the pedagogical challenges and the author's approach to overcoming them. It also identifies the learning outcomes to cover this broad subject from an architectural education point of view. The paper discusses the rationale for the course structure and outlines the course contents in view of the pedagogical challenges. Emphasis is put on the role of an energy modeling software in visualizing the relationship between architectural design decisions and the energy performance of a building. The software is also used to guide students through a final project to retrofit of an existing building to reach a net zero-energy status. A survey of students who took the course one year earlier is presented to get feedback on the pedagogical value of the used approach.

**Keyword** Net zero-energy buildings

## 1 Introduction

Future architects will face more regulations to reduce energy consumption of buildings. These include the ones that aim for net or nearly zero-energy buildings (e.g., [3, 11]). Therefore, the relationship between the architecture design of a building and its expected energy performance must be a mainstream design issue. It must be similar

---

M. Ahmed (✉)

American University of Sharjah, Sharjah, United Arab Emirates

e-mail: [mokhtar@aus.edu](mailto:mokhtar@aus.edu)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_35](https://doi.org/10.1007/978-3-031-34593-7_35)

to having the building structure as a mainstream design issue. Indeed, some architectural programs put emphasis on energy performance by providing multiple required and elective courses about the subject and insist on addressing it progressively in the design studios. Yet, there are many other programs that hardly consider the issue, or they address it in a very limited way. The author teaches in an architectural program that belongs to the later category of programs. To overcome this hurdle, the author tries to make such a technical subject attractive to students who are trained to focus on the formal aspects of design. The hope is to encourage them embrace the subject in their designs.

Through review of some documented work, various approaches were reported to include the general subject of sustainability in architectural curricula. Passe [6] described a pedagogical approach that can be used in a required sustainable design studio but at the graduate level. The approach uses “*a highly structured 10-stage workflow for architectural design equips students with knowledge, tools, and processes to integrate and predict dynamic performances of light, sun, heat, and air movement in their design decisions*”. Simulation tools are used in the process, and the students examine indices such as energy-use intensity (EUI), daylight autonomy as they make design decisions. As this process is integrated in a design studio, the focus is on aspects such as orientation, window size, and massing and how these aspects impact energy consumption. Srivastava [10] describes a process to be used in a studio setup as well. It aims to minimize the EUI of a building. Students go through iterations among the following strategies: “(1) *Applying passive strategies such as orientation, massing, size, location, proportions of building form and locations, orientation, and sizes of transparency (windows, skylights, glazed doorways, etc.) and opacities (walls, roofs), shading systems, and natural ventilation systems. (2) Thermal properties of building envelope (R-values), glazing, and heating, ventilation and air-conditioning (HVAC) systems selections. (3) Any deficit in reaching the target goals was met by calculating the size of a solar array system required to reach the 70% reduction*”. These two examples show the typical way to address the subject in a studio. The tendency is to apply a design solution that seems to work with the architecture design at hand. Vassigh and Spiegelhalter [12] propagated the need to integrate different disciplines from the conceptual design stage and developed an interactive software and a hardcopy textbook to help architecture and building design students designing carbon-neutral buildings. The objective is to encourage students to integrate building systems to improve the building energy performance. This approach is certainly better but work only when students come from various disciplines.

Using energy simulation software is an important tool to link architecture design with the fundamentals of energy performance of building. Several researchers reported their experience with such use. Soebarto [9] reported the teaching of a simulation tool to architecture students at the master's level. The experiences showed that students need to have basic knowledge of building heat transfer, thermal properties and behavior of materials, and basic load calculations before making use of the software. Else, there will be issue with meaningful inputs to the software and in understanding the output results. He also confirms that architecture students do not

consider energy consumption as the most important thing in their design. Elias-Ozkan and Hadia [2] endorsed the need for the students to understand climatology, material properties, building physics, HVAC systems, internal and external gain factors, solar impacts, among other things before being able to use energy simulation tools. They were also teaching to architecture students at the master's level. Schmid [7] reported on several experiments for using energy simulation software with undergraduate architecture students and concluded that *“Building simulation can be a powerful ally in teaching building physics, raising the students’ interest for issues of thermal and acoustical behaviour of buildings. The additional effort of data input required by the adopted software is justified, as students learn to evaluate the effects of their own design choices. However, model complexity and timely effective tutoring are critical variables”*. Hopfe et al. [4] conducted a survey on the use of building performance simulation tools in architecture schools, and some of the conclusions indicated the difficulty of using such tools in studio without prior knowledge of building physics.

The above documented work demonstrates the difficulty of covering the subject in an architectural curriculum. In a studio, the tendency is to incorporate pre-perceived energy-related design solutions with the design at hand. Evidence that these solutions will work requires the use of simulation software. The appropriate use of these software requires good knowledge on the fundamentals of buildings energy performance. There is hardly any time to cover these fundamentals in a studio.

Hoping to overcome these difficulties, this paper documents the author's experience in teaching a course on energy performance of buildings with the intention to reach net zero-energy status. The described course is an elective one for undergraduate architecture students. The program where these students study is oriented toward form-based rather than performance-based designs. Yet, students who sign up for the course have some recognition of the importance of the subject. They hear about it in the news, from alumni, or from offices where they attend the required internship. The paper defines the course objectives and the pedagogical challenges. These include the interest of the students and the nature of the program they are studying in. The paper explains the approach used by the author to cover the subject and shows the sequence of delivering the contents. Finally, it shows the feedback of the students after at least one year of taking the course.

## 2 Learning Objective and Pedagogical Challenges

Because the course is for architecture students, its main objective is to help the student define an energy performance design strategy that suits a particular building project with the hope to reach net zero-energy status. Hence, the strategy should minimize the building requirements for energy and maximize the renewable energy generated

at the project site. The strategy must reflect three aspects for the building project at a hand:

- The climate of the project site (e.g., hot and humid vs. temperate vs. cold).
- The building's typology (e.g., residential apartment vs. shopping center vs. school).
- The building's geometry (e.g., horizontal with a large roof area vs. high-rise tower).

To achieve this objective, the author faced the following challenges.

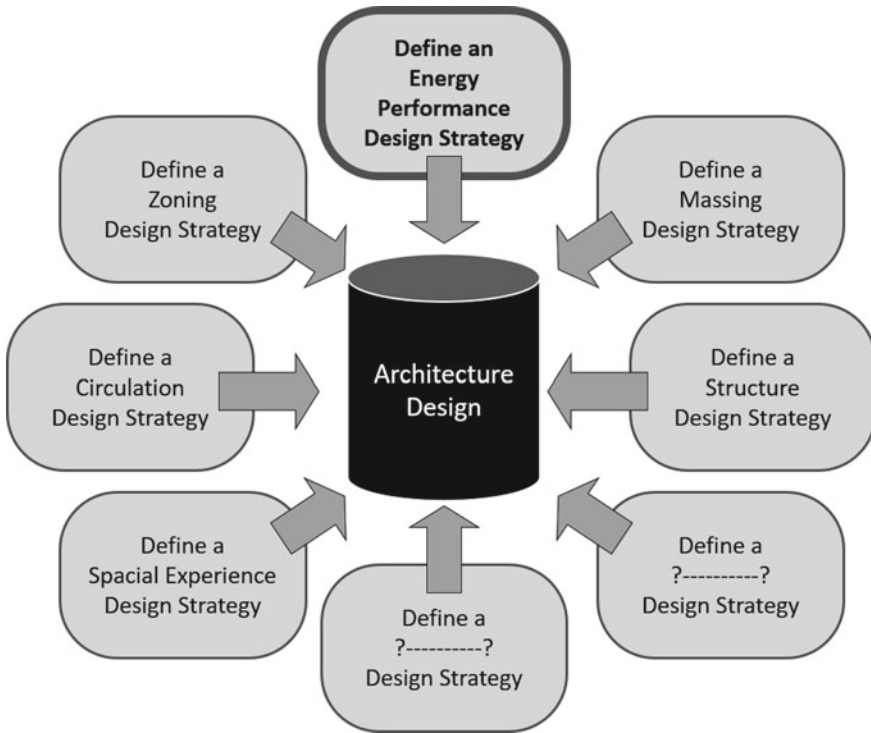
## ***2.1 The Learning Mode***

Architecture design is complicated. It requires the development of various design strategies that address issues like zoning, circulation, special experience, structure, massing, *hopefully* building energy performance, and others. An architecture student needs to learn not only how to develop these various strategies, but also how to integrate them to come up with a piece of architecture that satisfies a wide range of esthetics, economic, social, and other concerns (see Fig. 1). Therefore, and as the above review of some documented work shows, the studio is the best learning mode when the student needs to deal with the difficulty of developing and integrating all these design strategies.

However, the author has a different point of view on this issue. Energy performance of buildings is a complicated enough subject by itself. It makes pedagogical sense to isolate the subject and let the student comprehend its fundamentals before getting into the difficulty of integrating it with other complex subjects in a studio. Therefore, the author selected the lecture mode for teaching the fundamentals of the subject. Yet, the course assignments and the project concentrate on making design decisions and not on doing calculations. Hence, these link to the studio mode of thinking but in a focused subject and without the need to address the many other design strategies.

## ***2.2 The Complexity of the Subject***

Energy performance of buildings depends on many variables, and most of them are interrelated. The student needs to recognize the impact of each variable on the other variables. Therefore, the subject should be dissected to its fundamentals, and gradually, the relationships between the variables are established and are linked to design decisions. The author developed a pedagogical approach with the aim to achieve this. The approach as illustrated in Fig. 2 helps the student learn three foundational subjects. These are (1) the climate data of the project site, (2) the factors

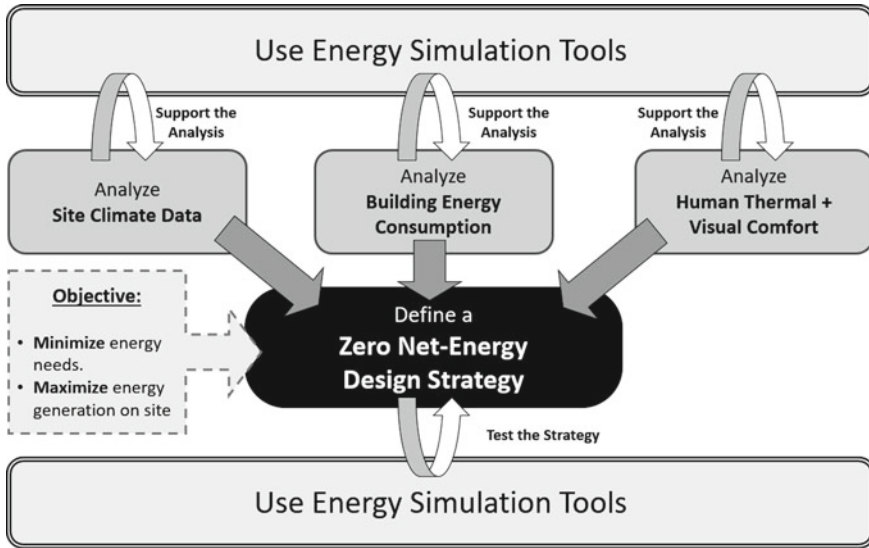


**Fig. 1** Energy performance as one of many design strategies required for an architecture design

impacting human thermal and visual comfort, and (3) the components of a building energy consumption. Each of these is further dissected into different variables as will be shown later.

### 2.3 The Visual Nature of Studying Architecture

Architecture students depend heavily on visuals to understand design. Meanwhile, energy is invisible. We can make sense of it when we solve equations or take measurements in an existing building. Yet, architecture students barely use these tools in their curricula. To overcome this challenge, the author uses an energy simulation tool that can show graphically the results of energy performance calculations. In addition, students hardly do any calculations in the course. Equations are mentioned for the sole purpose of clarifying the relationships between design decisions and some energy performance variables (e.g., the impact of having an air gap on a wall U-value or the impact of using low-E glass on the U-value of a glass). The author uses the values and charts generated by the simulation tool to help students visualize



**Fig. 2** Components of the course covering the three fundamental subjects that help a student define an energy performance design strategy

the impact of making design decisions on the building energy performance. He illustrates how that impact varies by the three aspects of climate, building typology, and building geometry as mentioned above.

## 2.4 Overcoming the Green Wash Approach

Architecture students hear about sustainable design and low-energy performance buildings in different courses and studios. However, when these issues are attempted in a studio, the tendency is to jump to a design solution such as adding louvers, increase natural lighting, and adding more insulation. Typically, there is no analysis that leads to the solution. Changing this approach is a challenge. Therefore, in organizing the course, the author starts by stressing the importance of analyzing the specific project at hand and gradually helps them develop relevant analytical skills. As Fig. 2 shows, each of the three covered subjects depends on such analysis. When a design strategy is defined, its impact is also analyzed to make sure that it makes sense. The simulation software is critical in supporting this approach.

## **2.5 *Form-Oriented Architecture Education***

The main objective of an architecture design is to generate a form. The form reflects the designer ability to integrate various design strategies. Typically, one of these design strategies dominates the form (e.g., the structural strategy to dissipate wind forces is the dominant reason for the building form of Burj Khalifa). One of the challenges is to help architecture students realize the potential for using energy performance strategy as a form-maker. The author tries to overcome this challenge through showing examples of both historic and modern buildings where form was generated as a result of a passive or/and active energy performance strategy (e.g., Bahrain World Trade Center and Masdar Headquarter in Abu Dhabi).

## **3 Learning Outcomes and Course Structure**

Recognizing these challenges and to enable achieving this main objective of the course as mentioned above, the author identified the following course learning outcomes. A student should be able to:

- Analyze a climate and identify the potential positive and negative impact of its components on both the building energy consumption and potential energy generation.
- Identify the factors that impact both human thermal and visual comfort. Recognize the potential of manipulating these factors to achieve such comfort with no or with minimal energy consumption.
- Identify the means through which energy is consumed in a building.
- Identify the means with the largest consumption and those with potential reduction in consumption through better design decisions (particularly architecture-related ones.)
- Use energy simulation tools in analyzing a building energy consumption and in evaluating the impact of any relevant design decision.

To achieve the above course learning outcomes and considering the nature of the students, the author structured the course with the following guidelines in mind:

- Trigger an interest in the subject. This is particularly important given the school focus on the formal aspects and not the performance aspects of design. In addition, students must recognize that what they will learn is related to architecture design decisions and not to engineering decisions that they can ask engineers to do.
- Explain the different components of energy performance independently at the beginning of the course. During the final third portion of the course, these components are integrated together. The learning of the independent components is supported by focused assignments. Meanwhile, their integration is achieved through the final project.

- Cover the components initially with focus on the fundamentals rather than on the design solution. For example, focus on reducing solar gain rather than on using louvers. The author believes this encourages innovation in finding design solutions. Yet, keep linking the components, the assignments, and the project with architecture design decisions, so students do not lose interest in the subject.
- Approach the technical issues at the conceptual level rather than at a detailed calculations level. This fits the nature of architecture design where students start from the bigger picture (conceptual design) and get into more details. This is in contrast with typical engineering courses that usually start with the small details (e.g., equations and calculations of heat transfer) and gradually move to the larger perspectives (e.g., how to make a zero net-energy building).
- Learn the simulation tool progressively and in relation to learning the independent components using the relevant assignments. In the final project, several capabilities of the simulation tool are needed to assist integrating the independent subjects.

## 4 The Learning Modules

### 4.1 Introduction

The main purpose of the course introduction is to stimulate the students' interest in the subject. The documentary "Before the Flood" [8] is presented in class followed by a discussion on the ethical role of architects in addressing the climate change problem. The author presents and analyzes various charts that show the contribution of buildings to the problem. The discussion covers how architecture and planning decisions make important impact on energy use not only for the buildings sector but also for the transportation and for the manufacturing sectors. Based on this, the author emphasizes the importance of performance-based architectural design and contrast that with form-based architectural design.

The focus then moves to the subject of zero net energy. The author explains what it means and the different possible goals such as Zero Energy, Zero Net-Energy, Nearly-Zero Net-Energy, and Nearly-Zero Energy. He relates these goals to different regulatory requirements in some parts of the world.

The author explains the graph in Fig. 2 along with the sequence of learning the various subjects through the course. This graph is shown at the beginning of every lecture as a reminder of a systematic process to design low-energy building. Showing the graph also helps in relating the previous subjects to the new ones.



## ***4.2 Analyzing the Climate of a Site***

This learning module discusses the importance of reading and analyzing the climate of the project's site. This starts by identifying the major components of climate data and how each component may have an impact on the architecture design of the building (e.g., the relationship between cloud cover and the need for shading). The author explains the concept of Typical Meteorological Year (TMY), the reasons for its use, and the worth of the calculated energy consumption when TMY is used by simulation software. He also identifies the possible sources for obtaining TMY for a particular location.

Samples of TMY that covers a hot and dry, a hot and humid, a tropical, and a cold climate are used to compare climate data. Climate Consultant [1] and IESVE [5] are used to visualize and compare the different climate data.

The Psychrometric Chart is explained as part of this learning module. Examples are given to demonstrate its use for making some design decisions (e.g., use of evaporative cooling in dry vs. humid climates). The solar movement in the sky for different locations on earth and for different times of the year is also explained.

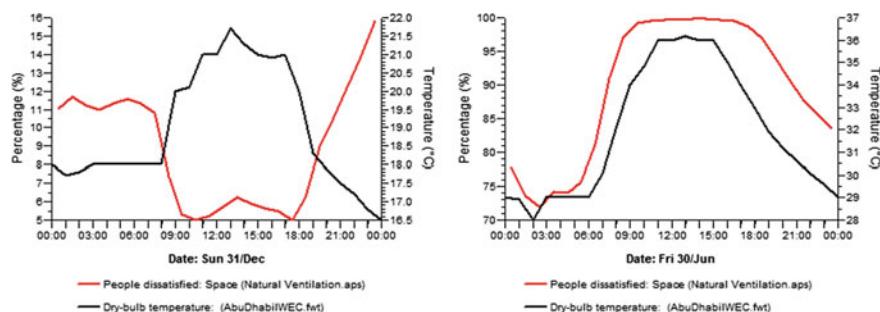
A class exercise is assigned to train on analyzing the climate of the site. The student selects a city of interest and answers the following questions:

- Do you consider cooling or heating or both as the main challenge for the design?
- Will you use shading of windows as a design strategy? If yes, when?
- Will you use greenhouse as a design strategy to heat the building?
- Will you use natural ventilation as a design strategy to cool/ventilate the building?  
If yes, where would you put windows?
- Will you use evaporative cooling as a design strategy to cool the building?
- What is the maximum and minimum angle for the sun in midday?

## ***4.3 Understand Human Thermal and Visual Comfort***

This learning module starts by viewing the human body as a heat generating and a heat exchanging subject. The physics of heat exchange is linked to how the body is interacting with its surrounding environment. The body reaction to extreme environment is described, and from that human thermal comfort is defined. The importance of thermal comfort on human productivity and survival is discussed. Ultimately, the four environmental factors and the two human factors that affect human thermal comfort are deducted.

The author shows examples of passive design strategies that aim to manipulate some of these environmental factors to achieve thermal comfort. For instance, fostering natural ventilation as a mean to manipulate the environmental factor of air speed. This includes clarifying the physics of air movement through pressure difference or temperature difference and how an architect can create these differences through architectural forms.



**Fig. 3** Showing a sample relationship between air temperature and percentage of thermally dissatisfied people in a summer and in a winter day for a room in Abu Dhabi

The various indices that help measure human thermal comfort (e.g., Predicted Percent of Dissatisfied People) are explained. The simulation tool IESVE is used to demonstrate these indices in a non-air-conditioned, naturally ventilated building during different climate seasons for the city of Abu Dhabi (see Fig. 3). A class exercise is assigned where students simulate a naturally ventilated room in cities that belong to three different climates for the same day. The indices of Predicted Mean Vote, Predicted Percent of Dissatisfied People, and Comfort Index are compared for the three cities in relation to the room temperature.

To continue with this module, the author explains the illumination requirements for different human functions. Further discussion on light is covered in a later module.

## 4.4 Understand How Buildings Consume Energy

This module is the largest one in the course. It aims to fragment the subject of energy consumption into easier to understand components. In addition, it aims to link energy consumption to architectural design decisions. The module has several submodules. Each one focuses on a particular aspect of the subject.

### 4.4.1 Identify Categories of Energy Consumption in Buildings

This keystone submodule identifies the five categories of energy consumption in a building. That is HVAC, lighting, domestic hot water (DHW) heaters, equipment that runs the building, and plug-in loads. The submodule uses the simulation tool to clarify how the building function, location, design, and schedule of use result in different distribution for the contribution of each of the five categories to the building energy consumption. The intention is to emphasize that design strategies vary depending

on the nature of the project and that replicating design strategies or design solutions from case studies is not the appropriate approach to deal with energy performance of buildings.

The submodule discusses the typical sources of energy for buildings and the difference between primary energy and site energy. It explains the concept of Energy Utilization Index (EUI) and refers to standard index values for different building typologies in different climate zones.

Finally, a class exercise is assigned where students use the energy simulation tool for an already-modeled multi-story deep-plan hospital. They run the simulation for cities in very hot, temperate, and very cold climates. They analyze the energy consumption for the hospital and compare the results of the contribution of each of the five categories to total building energy consumption. The same exercise is repeated for an already-modeled small villa. The objective is to recognize the impact of the climate, the building typology, and the building geometry on the energy consumption pattern. The students calculate EUI for each case and analyze the meaning of the calculated value.

4.4.2 Identify Factors Impacting HVAC Energy Consumption

Typically, HVAC is one of the major consumers of energy in a building. This submodule helps the students to recognize the factors impacting HVAC energy consumption. The purpose is to link these factors with relevant architecture design decisions. The submodule starts with explaining Fig. 4 and with discussing the role of the architect in both reducing the HVAC load and minimizing the time needed to use the HVAC system.

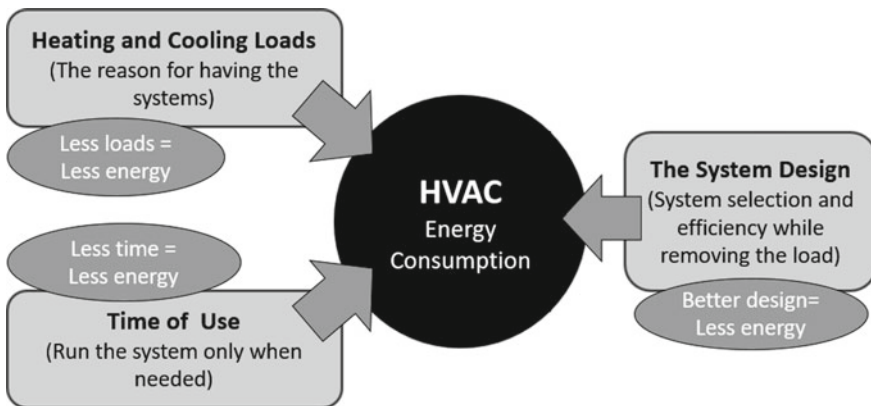
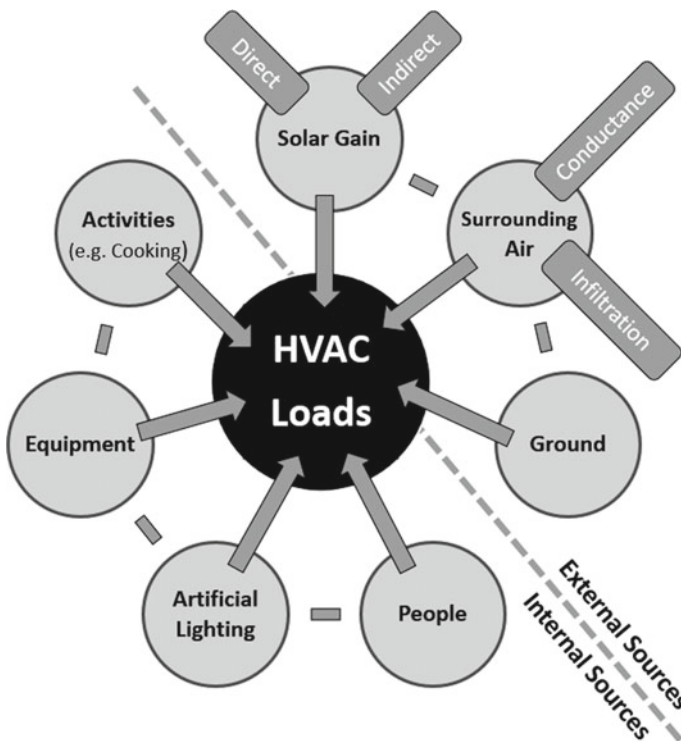


Fig. 4 Factors impacting the energy consumption of the HVAC system

The submodule continues to identify the external and internal sources of heat gain/loss in a building as summarized in Fig. 5. It explains each source in more details and finds the factors impacting the heat gain/loss from that specific source. For each source, the submodule discusses possible bioclimatic design strategies to reduce/increase the heat gain/loss in a building from that source. The energy modeling software is used to visually show the results of these design strategies. For example, the submodule discusses the indirect heat gain from the sun. It shows the factors that impact the quantity of direct solar energy received by a building solid surface. As the solar incident angle is one of these factors, it shows examples of how the architect can manipulate this incident angle to reduce/increase the indirect solar gain from the roof of a building. The simulation tool is used to quantify the impact of this design decision as shown in Fig. 6.

The submodule addresses the issue of minimizing the “time of use” for the HVAC and how it should be linked to the climate of the site. For example, the author discusses a design strategy that supports natural ventilation and shutting down the HVAC system when the outside temperature permits. The simulation tool helps testing the strategy, and the students compare the results of using such strategy on the total energy of the building.



**Fig. 5** Sources for cooling and heating load for an HVAC system

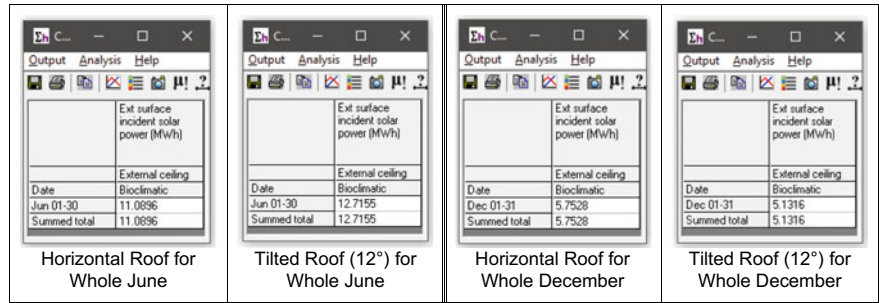


Fig. 6 Impact of tilting a roof on its incident solar energy in summer and winter (for Abu Dhabi)

4.4.3 Understand Conductance Through the Building Envelope

The building envelope is an important part of architecture design. In many buildings, conductance from the envelope is a major contribution to the heat gain/loss in a building. Therefore, it is essential for an architecture student to appropriately design walls, roofs, ground floors, and external glazing to minimize the energy consumption of a building.

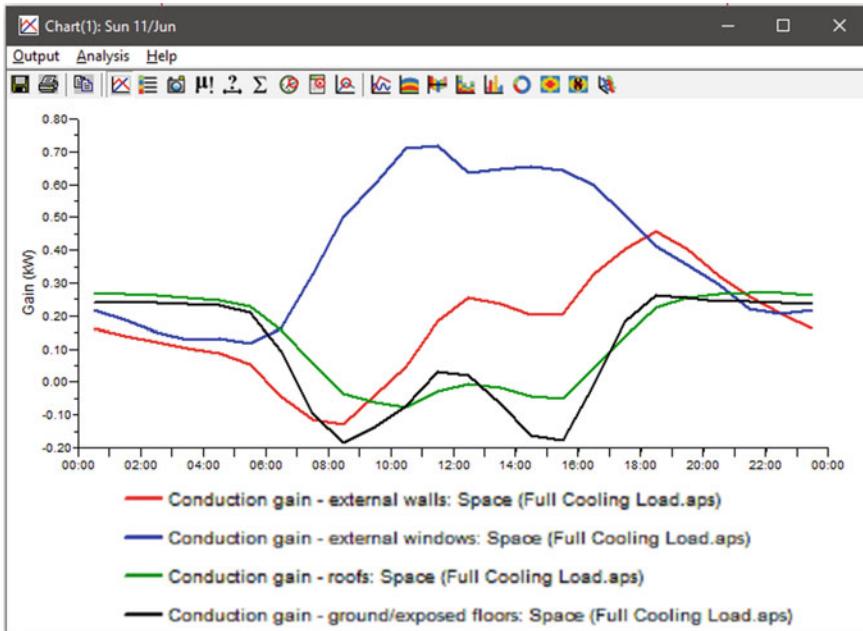
The submodule starts by explaining the heat flow equation through a building envelope and discusses the implication of each of its variables on architecture design decisions. It clarifies the important role of the air films, air gaps, and materials’ emissivity in the flow of heat, and it gives examples of various types of envelopes that utilize these.

A class exercise uses the energy simulation tool to evaluate the impact of changing the envelope design of a room (wall, roof, and glazing) on its external conduction gain/loss. The exercise requires the student to compare the results in different climates and to identify the critical envelope components (see example results in Figs. 7 and 8).

4.4.4 Quantify Solar Energy and Underrated Its Related Technologies

Solar energy plays various roles in buildings energy performance. It impacts cooling load by direct penetration through transparent and translucent components of the envelope and by raising the temperature of all exposed envelope components. It is also an important source for energy to heat DHW and to generate electricity on site. The goal of this submodule is helping the students understand these different roles and determine when, where, and how they should maximize or minimize the capturing of solar energy.

The submodule starts by explaining the nature of the sun, its distance from earth, and its surface temperature. Hence, radiation equation is used to quantify the solar energy received by a perpendicular surface just outside the earth atmosphere. The impact of the different conditions in the earth atmosphere is discussed to explain the



**Fig. 7** Comparing the contribution of the different sources of conduction gain for a particular space to identify the more critical ones

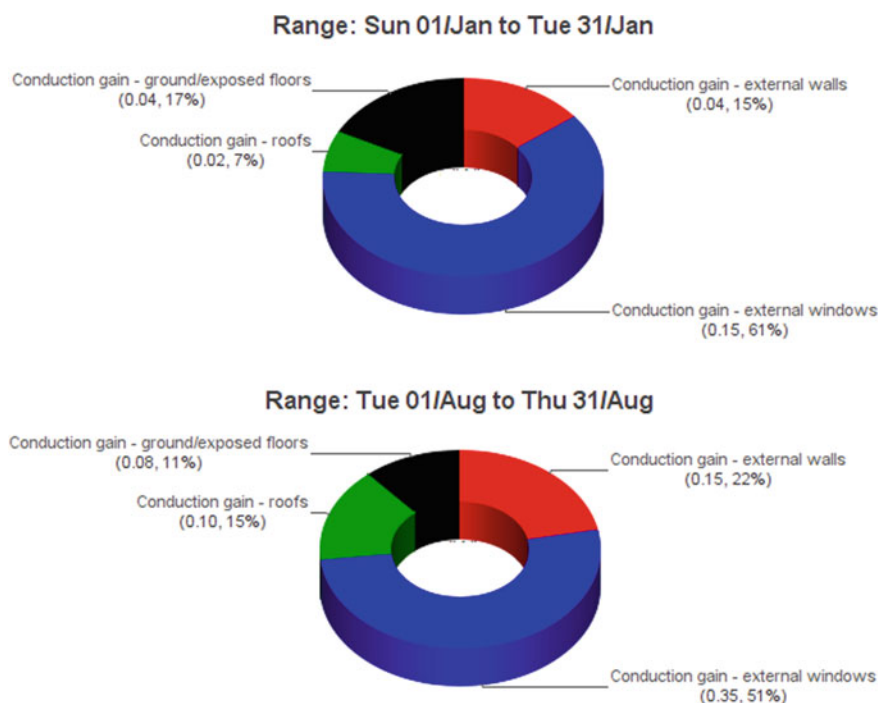
variation of solar insolation throughout the year in various parts on earth. The author explains the azimuth and altitude angles. He discusses the importance of the incident angle and how architecture design decisions effect this angle and hence the received solar energy by a surface.

The submodule explains the relevant properties of glass in terms of its transmittance to shortwave, longwave, and visible radiation. It discussed the criteria to select glass for the different parts of the envelope to accommodate the energy requirements of different buildings in different climates.

The submodule describes the basics of the photovoltaic technology. This covers types of PV cells and types of PV systems (AC vs. DC, on-grid vs. off-grid). It also identifies the different types of solar collectors for DHW.

#### 4.4.5 Understand the Basics of Lighting and Its Control

The aim of this submodule is to explain the requirements for visual comfort in a space and how to achieve that comfort with minimal energy consumption. Hence, it covers the fundamentals of light and light measuring units. It describes the diverse types of lamps and their characteristics including Luminous Efficacy and Lighting



**Fig. 8** Comparing the sources of conduction gain for a whole building during a sample summer day and a sample winter day

Power Density. It provides guidelines for effective utilization of natural lighting and the use of light sensors to reduce energy requirements for lighting.

A class exercise is assigned where the simulation tool is used to test the impact of replacing florescent lamps with LED lamps. The tool is also used to test the impact on light energy consumption when using light sensors with different daylight strategies.

#### **4.4.6 Recognize the Efficiency of Equipment and the Importance of Their Times of Use**

Architecture students do not have the background to select the appropriate mechanical equipment for a project. Therefore, the purpose of this submodule is to help them communicate with mechanical engineers when there is opportunity to do so. They also need the information to put reasonable data while using the simulation tool. Hence, the author provides basic explanation for concepts of Efficiency, Coefficient of Performance, and Energy Efficiency Ratio.

The submodule puts more focus on recognizing the relationship between the schedules of using the various spaces and the energy needs for these spaces. For example, in the case of an office building, it discusses the possibility of raising the

thermostat temperature during the night and turning off the light in non-working hours. It emphasizes with examples the impact of architectural design on allowing such possibilities by appropriately zoning the building spaces with the consideration for the times of using the spaces.

A class exercise is assigned where the simulation tool is used to quantify the energy saving due to changing the thermostat set temperature during weekends, holidays, and outside daily working hours.

## ***4.5 The Final Project***

After covering the fundamentals in the form of separate subjects, the aim of the final project is to integrate the gained knowledge and to recognize the interdependency of these subjects.

The project is structured so the students try to retrofit an existing building, so it can reach a Zero Net-Energy or Nearly-Zero Net-Energy status. The reason for retrofitting—rather than designing a new building—is to focus on the issues related to the energy performance of the building without the complications of integrating it with the numerous other design decisions related to other aspects of design (e.g., zoning, circulation, spacial experience, structure, massing, and others). Yet, the author allows the students to make some unrealistic retrofitting decisions. This includes changing the building orientation, roof and/or walls tilt angles, and glazing areas.

The project avails three pre-modeled buildings. These are a primary school, a small hotel, and a high-rise office. The author choses these building typologies, so they have different geometry and different internal heat gains. The project avails three cities with different climates (Abu Dhabi, Montreal, or Singapore). Each group of students selects a combination of a building typology and a city, so every project is different. The author encourages the students to go through the project in a systematic way following the below steps:

- Analyze the weather data of the selected city.
- Analyze the current energy consumption pattern for the existing building.
- Identify strategies to reach Zero Net-Energy based on the above analysis. The strategies aim to:
  - Reduce cooling/heating load.
  - Maximize energy generation from renewable sources on site.

The students go through iterations of making and testing design decisions till they reach the Zero Net-Energy, Nearly-Zero Net-Energy status, or justify the inability to do so. Each group presents their project to the class providing the rational for their retrofitting design decisions. The author and the other groups discuss these rational and debate the used strategies. The author considers these discussions as the



most important part of the course as it relates together the different and interrelated subjects of building energy performance.

5 Students Feedback

The author conducted an anonymous survey for students who took the course one or two years earlier. The objective is to get their feedback about the course. The purpose for leaving at least a year is to see if the course impacted how the students approach the subject outside the immediate pressure of exams and project. About 18 students out of total 29 who were asked to take the survey have replied. Figure 9 indicates how the students compare their knowledge about the subject to that before taking the course.

The students were also asked about their preference to take the course in a studio format. The majority preferred this format as shown on Fig. 10. This is something that the author needs to reflect upon and may require deeper research with two groups of students. One group takes the subject in a studio and another in the presented course. The level of the students' comprehension of the subject can be examined for each group to know which teaching mode provides better learning. This is particularly

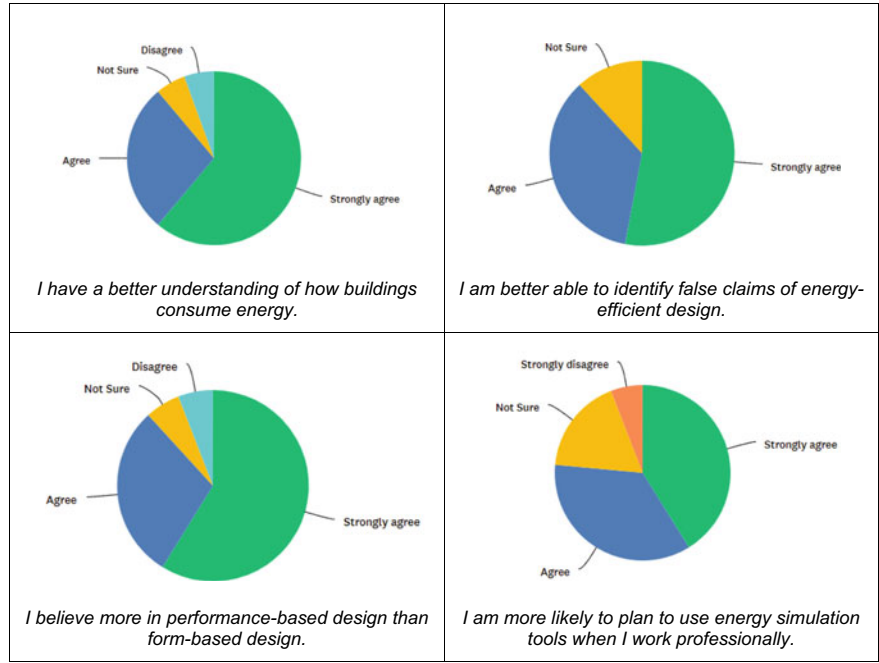
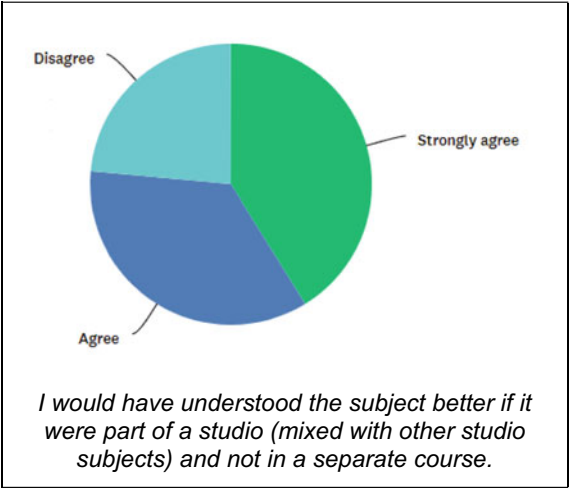


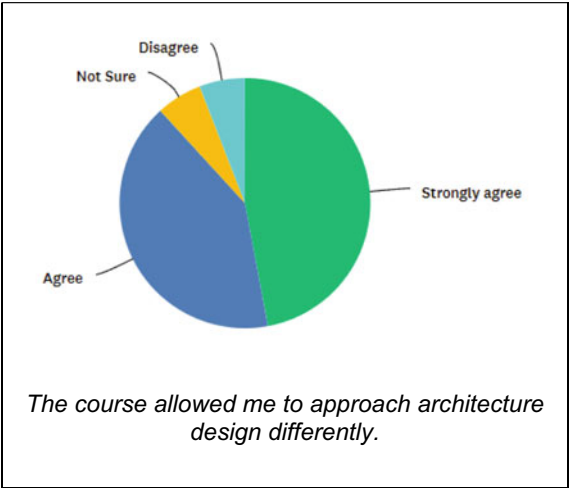
Fig. 9 Shows the students reply in comparison with before taking the course

important considering the students’ answers for the survey question about the course impact on their approach to architecture design. As shown in Fig. 11, the majority agreed or strongly agreed that it did.

**Fig. 10** Students reply to the survey question about the studio format vs. lecture format



**Fig. 11** Students reply to the survey question about the value of the course to their profession



## 6 Summary

This paper presents a course on teaching the fundamentals of the energy performance of buildings, with the aim of reaching net zero-energy status. The paper outlines the pedagogical challenges in addressing this subject, particularly considering the nature of the students taking the course. The course is oriented toward architecture students who are more focused on form-oriented rather than performance-oriented design. The paper discusses the learning outcomes and how the course is structured to achieve these outcomes. The author outlines the guidelines he follows to structure the course modules. These modules are explained with the rationale for their contents. This includes several assignments that use energy simulation software to help students visualize the impact of architecture design decisions on energy performance. In a survey, students who took the course more than a year earlier provided positive feedback on how the course affected their interaction with the subject.

## References

1. Climate Consultant 2020 (2022) Climate Consultant 6.0. Energy design tools. Available from <https://climate-consultant.informer.com/6.0/>. Accessed 18 Mar 2022
2. Elias-Ozkan S, Hadia H (2014) Teaching and learning building performance virtualisation. In: The 7th international conference of the Arab society for computer aided architectural design ASCAAD, Jeddah, Saudi Arabia, 31 March–3 April, pp 323–330
3. European Commission (2022) European climate law. Available from [https://ec.europa.eu/clima/eu-action/european-green-deal/european-climate-law\\_en](https://ec.europa.eu/clima/eu-action/european-green-deal/european-climate-law_en). Accessed 18 Mar 2022
4. Hopfe CJ, Soebarto V, Crawley DB, Rawal R (2017) Understanding the differences of integrating building performance simulation in the architectural education system. In: Presented at the building simulation 2017: the 15th international conference of IBPSA, San Francisco, 7–9 Aug
5. IES 2022 (2022) Virtual environment. Available at <https://www.iesve.com/software/virtual-environment>. Accessed 18 Mar 2022
6. Passe U (2020) A design workflow for integrating performance into architectural education. *Build Cities* 11:565–578. <https://doi.org/10.5334/bc.48>
7. Schmid A (2008) The introduction of building simulation into an architectural faculty: preliminary findings. *J Build Perform Simul* 1:197–208
8. Stevens F (2016) Before the flood. National Geographic Documentary Films, United States
9. Soebarto V (2005) Teaching an energy simulation program in an architecture school: lessons learned. In: Ninth international IBPSA conference, Montréal, Canada, 15–18 Aug, pp 1147–1153
10. Srivastava M (2020) Cooperative learning in design studios: a pedagogy for net-positive performance. *Build Cities* 11:594–609. <https://doi.org/10.5334/bc.45>
11. UAE Government Portal (2022) UAE Net Zero 2050. Available from <https://u.ae/en/information-and-services/environment-and-energy/climate-change/theuaesresponsetoclimatechange/uae-net-zero-2050>. Accessed 18 Mar 2022
12. Vassigh S, Spiegelhalter T (2014) Integrated design pedagogy for energy efficient design: tools for teaching carbon neutral building design. *Energy Proc* 57:2062–2069

# A Framework Approach to Utilize Real-Time Spatial Labor Data from Construction Sites



Hoda Author Abouorban, Khaled Nassar, and Elkhayam Dorra

**Abstract** Real-time monitoring and control of on-site resources is a developing field. In recent years, there has become a growing need for enhanced monitoring systems on construction sites using real-time technologies. Such technologies have been used heavily in fields like IT monitoring and healthcare industries. However, the construction industry is yet to seize the potential of deploying real-time monitoring systems on construction sites. Limited research has been done in the area of applying real-time monitoring, such as RFID tags and visual sensing, for the control of on-site safety performance and labor and earthmoving equipment productivity. Most of the studies delve into analyzing the data collected, with the data only used for verifying the accuracy of the used real-time monitoring technologies used. This study aims to uncover the potential outputs of utilizing real-time data collection and analysis of workers' locations on a construction site. The outputs are achieved by developing a semi-automated framework that studies the real-time data collected. The framework is divided into two stages. First, site data is collected semi-manually, and real-time workers' spatial data is collected using GPS tracking technologies. Second, the collected data is prepared for visual analysis. The proposed framework is then applied to a single case study for a commercial project in New Capital, Cairo, Egypt. Outputs from the case study reaffirm the feasibility of monitoring workers' productivity by identifying their GPS locations in real-time on-site. The framework aims to emphasize the potential benefits gained from using real-time technologies on construction sites. Thus, this research provides stakeholders with a tool that collects and analyzes data from construction sites for improved monitoring and control of workers' performance.

**Keywords** Framework approach • Real-time spatial labor data

---

H. A. Abouorban (✉) · K. Nassar · E. Dorra  
The American University in Cairo, Cairo, Egypt  
e-mail: [hoda94@aucegypt.edu](mailto:hoda94@aucegypt.edu)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_36](https://doi.org/10.1007/978-3-031-34593-7_36)

## 1 Introduction

Real-time monitoring of on-site performance is still in its infancy. However, in recent years, real-time monitoring technologies for construction sites have gained popularity. Most of the research has concentrated on real-time health and safety monitoring [7].

The studies demonstrated that technologies such as RFID tags, Bluetooth, and IoT-based technologies can help improve the performance of site security. Such technologies, according to studies, could be used to track workers and equipment entering a danger zone. Additionally, wireless networks could be used to monitor hazardous construction sites to ensure worker safety [7].

Additionally, the study examined the feasibility of using real-time monitoring to assess worker and equipment productivity. When the location of the resources is used, it can be assumed that they are idle or actively engaged in ongoing activities. The productivity of each location is determined by comparing the amount of time spent in each location to the amount of work completed during that period [4].

According to the literature review, the most effective control system is a fully automated real-time monitoring system. Additionally, research indicates that GPS technologies outperform other methods of real-time monitoring. Despite this, studies indicate a dearth of research on real-time labor monitoring. Future research should focus on this area due to its apparent efficacy and untapped potential.

## 2 Literature Review

### 2.1 Waste Allocation—Lean Construction

A component of optimizing the project's performance measures is identifying and minimizing waste during the construction process, with waste accounting for 30–35% of the project's total cost [3]. Waste identification and minimization are two major Lean principles. According to Thomas et al. [9], optimizing labor workflows, using Lean principles, can reduce waste and improve labor performance. Nikakhtar et al. [6] classified potential construction waste using the Lean approach identifying the idle and waiting time of construction workers as a type of waste. The first Lean principle for reducing waste is to manage the flow of processes. The Lean model developed improved the cycle time of a construction process by 9.22%, demonstrating the value of Lean principles in construction [6]. Wang et al. [10] also used Lean principles and simulation optimization to improve the efficiency of the layout of a combined hospital Emergency Department (ED) in Taiwan. Using Value Stream Mapping (VSM) to optimize staffing, service levels, and patient wait times has shown promising results as there was a 61% improvement in the service level of the department.

## 2.2 *Real-Time Monitoring Systems*

Real-time construction monitoring has grown in popularity as it has made controlling labor and equipment productivity and safety easier, making it easier to implement Lean construction theories. Efforts to fully automate project performance control and calculate performance indicators from indirect parameters have been researched. Navon [4] deployed real-time monitoring systems (RTMSs) to monitor project performance parameters. The study focused on certain project performance indicators as a measure of a project's success, such as labor and earthmoving equipment productivity using the regular location of the resource.

Moreover, one of the most common uses of RTMSs in the construction industry is in health and safety. Nazi Soltanmohammadlou et al. [7] conducted a comparative review of the most prevalent research done in the application of RTMSs in monitoring labor and equipment for better site safety. Despite the importance of RTMSs for enhanced site safety, the study states that research on the topic has been limited to the following: uses safety monitoring of workers, safety monitoring of equipment, safety monitoring of the working environment, RTMS technologies-aided accident prevention, and RTMS technologies-aided behavior-based safety. However, the study recommends further research into RTMSs in other areas of health and safety to fill this research gap. The most recent research by Zhang et al. [12] aimed to fill the research gap identified by Soltanmohammadlou et al. by developing a system that automatically generates alert levels and emergency plans. The early warning system relied heavily on sensing deformations of the surrounding rock masses and the stress state of the support structures, in real-time, to generate an early warning system. The system was successfully implemented in China's Lianghekou Hydropower Station underground cavern construction project [12].

Additionally, on-site productivity is another area where RTMSs have been studied. Navon [4] investigated the use of RTMS technologies in measuring productivity. The research recommends automated labor and equipment monitoring over previously deployed partially automated control methods like RFID, Barcodes, or PDA technologies, which still require manual labor input. The automated system assumes that the presence of labor or equipment near the work area means the resource is productive. Similarly, Navon and Goldschmidt [5] used the location of labor/equipment as an indirect measure of productivity. Worker/equipment presence at a specific time and location indicated their involvement in the current activity. To calculate labor productivity, GPS-located labor was compared to the amount of work completed during that period. Navon and Goldschmidt [5] and Navon [4] concluded that a fully automated monitoring system based on resource location using GPS technologies was the most effective control system. However, studies show that despite its apparent effectiveness, real-time labor monitoring has received little attention. Teizer [8] proposed using simple videoing technology to monitor and track temporary resources. Using a camera or video-based monitoring technology and processing algorithms, this system converts real-time images into real-time data. The study found that time-lapsing can

be used to track resources on-site, document daily workflow, and provide important information about site activities for better decision-making processes. Despite this, there are still major open challenges in the field of video-based sensing, as the construction site is not a laboratory-like environment for using video sensing technologies. Also, significant camera technology expertise is required for algorithm design and testing [8].

Another study by Jiang et al. [2] focused on labor monitoring using GPS or GIS (GIS). The tracking technology used included smartphones, servers, wireless base stations, a display, and an app. The system's analysis could then be used to make accurate decisions, especially when negotiating payments with contractors. It was concluded that a fully automated management system for on-site labor consumption proved to be feasible and effective and was critical to the project's success [2]. Considering the construction industry's digitalization, Calvetti et al. [1] sought to investigate the potential implementation of near real-time monitoring technologies, such as portable and/or wearable devices: radio frequency identification (RFID), ultra-wideband (UWB), global positioning system (GPS), barcodes/QR codes, labels, or tags, and smartphones, for measuring and modeling workers' productivity on the job site. Based on worker movement, Calvetti et al.'s Worker 4.0's systemized framework should help stakeholders focus on improving skills, efficiency, mechanization, and productivity. Although implementing sensing technologies can be difficult and expensive, integrating measurement methodologies and data collection devices could improve its added value. According to the research, future studies should focus on implementing electronic monitoring with a larger workforce sample and evaluating their effectiveness in real-world scenarios [1].

### **2.3 Research Gap**

In previous research, waste management in the construction industry has been identified as a growing field. Many studies focused on minimizing unnecessary motion, queue, interruption, and waiting time of the workforce using Lean construction to maximize the efficiency of the available manpower. This is a tedious, time-consuming, and resource-draining process. Recent research indicates that the field is moving toward using real-time monitoring techniques in construction. These techniques are widely used to track on-site resources' location and movement. Using real-time monitoring techniques in construction safety has received the most research, followed by worker productivity. However, there is a literature gap for implementing real-time monitoring techniques to collect data about the workforce. Also, little research has been done on the use of spatial statistical analysis and correlation of real-time data. As a result, the primary objective of this research is to develop a framework that addresses this gap in the literature.

### 3 Objective

The overall goal of this study is to shed light on the previously unknown potential of using real-time monitoring technologies on construction sites to reduce waste and improve project performance. The research does not quantify the methods for optimizing site performance parameters; rather, it introduces a novel way of depicting areas of time and production waste, which will be referred to as waste going forward.

The primary objective is subdivided into the following sub-objectives:

1. Collecting and analyzing spatial data about workers' locations using real-time technologies to monitor workers' productivity on-site.
2. Analyzing the performance of the construction site statistically.

### 4 Scope of Work

The scope of this research is limited to the study of workers' two-dimensional, easting and northing, spatial behavior. Additionally, this study considers only the following site performance parameters: safety, quality, and productivity. Also, the research examined the use of GPS tracking technologies to monitor the spatial behavior of on-site workers. Finally, the developed framework utilized the workers' smartphones as a monitoring device.

### 5 Framework Development

To fulfill the research objectives, a framework is developed. The framework consists of three stages. Firstly, there is the **data collection** stage where there are three types of data: (1) geographic data about the site, (2) periodic data about the site, and (3) real-time geospatial data about the site. Here, (1) and (2) are gathered manually, whereas (3) is gathered using GPS tracking technologies. Secondly, in the **data preparation** stage, using Python®, the collected data is split and grouped for analysis. Finally, in the **data analysis** stage, the prepared data is then analyzed visually and spatially-temporally using statistical techniques. Python® was also used to apply the algorithms of the techniques to the data.

#### 5.1 Stage 1—Data Collection

The first step in data collection is determining the site's location and zone coordinates (SGD). The framework will then collect performance data regularly (daily, weekly, bi-weekly, etc.). Finally, on-site workers will provide real-time geospatial



**Table 1** Safety incidents log

Date	No	Time	Location	Injured personnel	Title	Description
DD/MM/YY	X <sub>i</sub>	HH:MM XM	A	...	...	...

**Table 2** Inspection requests log

Date	No	Time	Location	Construction element	Dwg. ref	Code
DD/MM/YY	X <sub>i</sub>	HH:MM XM	A	...	...	A/B/C

data (RGSD). After that, the data is stored on a cloud-based server like Microsoft SharePoint (MSP). The data collection methodology is detailed below.

### 5.1.1 Site Geographic Data Collection (SGD)

A single pre-defined coordinate pair (latitude, longitude) for the site was extracted from the project tender file. The coordinates are then determined using Google Earth Pro's GIS software. Also, the WGS84 coordinate system is preferred over the Egypt 1907/Red Belt coordinate system. The Construction Site Layout Plan (CSLP) defines the zones on the job site. Once exported as XML, this data is used to perform the necessary transformations described in data preparation.

### 5.1.2 Site Periodic Data Collection (SPD)

Following the SGD, the SPD is collected for site safety and quality. Manual collection occurs as the site's HSE and QC teams provide periodic safety and inspection records. Each record must refer to the CSLP zones. Finally, the teams must upload a tabulated periodic record of the data to the same MSP.

The HSE team is responsible for documenting site safety incidents, accidents, and near-misses regularly. The records demonstrate the site's safety record. The data shall be documented in a tabulated Excel sheet in the format specified in Table 1.

Simultaneously, the QC team must regularly track both accepted and rejected inspection requests. The data shows the site's quality performance. The more inspection requests accepted, the better the quality of work. A tabulated Excel sheet is shown in the format specified in Table 2.

### 5.1.3 Real-Time Geospatial Data Collection (RGSD)

Finally, the final type of data collected is on-site geospatial data from blue- and white-collar workers. To collect this geospatial data, workers must download a GPS

**Table 3** Sample of prepared SGD

Area code	Ref. no	Easting	Northing	Elevation
A-XX	X	...	...	...

tracking app to their work phones and leave it running continuously while on-site. The workers save their tracks and upload them to the MSP, where the data is transformed and analyzed. The GPS-Logger 7.3.0 version of MyTracks was chosen as the GPS tracking application. It is a free app for iOS and Android smartphones that was recommended by Iowa State University’s Geospatial Technology Training Program. The coordinates are accurate to within 3 m horizontally. The recorded tracks’ precise coordinates can be exported as a gpx file and saved to the MSP.

**5.2 Stage 2—Data Preparation and Cleaning**

Cleansing and preparation of data are necessary because it is collected in a variety of formats, some of which are inefficient to process. This step is automated by utilizing the same programming language as Stage 3—Data Analysis. Each dataset collected during Stage 1—Data Collection must be prepared in such a way that it generates a specific final output for analysis.

**5.2.1 Site Geographic Data Preparation (SGDP)**

Since the SGD is exported as an XML file, the data must be modified. By splitting the KML polygon coordinates, the easting and northing coordinates, as well as the elevation for each area, can be obtained. The final output is shown in Table 3.

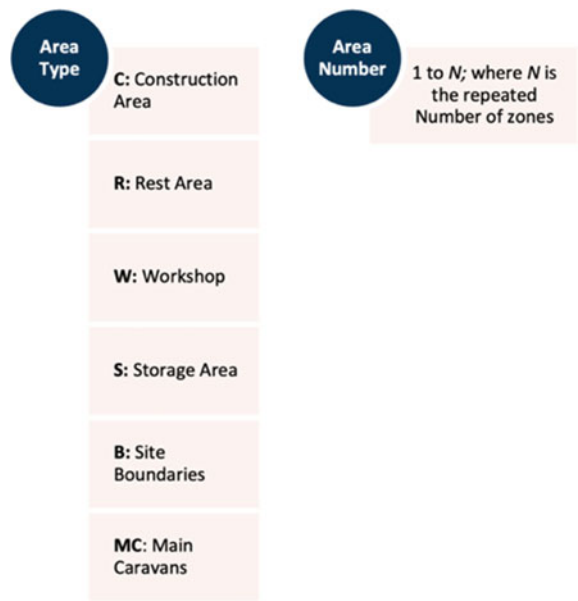
The area type is what determines the area code in the table above. Each area on the site is designated as a working area (WA), a rest area (RA), or a travel path (TP). Classification is necessary, as explained later in this chapter. Rest areas, as implied by their designation, are considered rest areas. Undefined site locations within the site boundaries are referred to as traveling paths. This is accomplished automatically using the area code, and the area is defined in Fig. 1 as WA, RA, or TP.

**5.2.2 Site Periodic Data Preparation (SPDP)**

Typically, the site’s SPD records are lengthy, and not all data is required for the framework analysis. As a result, the data is grouped into a more manageable format for analysis. The safety records gathered on-site are consolidated into the format shown in Table 4, with each incident being referenced by its unique area code.

The dataset referenced above is referred to as Site Safety Data (SSD).

**Fig. 1** Area code composition



**Table 4** Sample of prepared safety data

Area code	No. of safety incidents	Date
A-XX	...	...

**Table 5** Sample of prepared quality data

Area code	No. of accepted inspection requests	No. of rejected inspection requests	Date
A-XX	.	.	.

Additionally, the on-site inspection records are consolidated into the format shown in Table 5, with each record referencing the inspection’s unique area code.

The Site Quality Data (SQD) table above is used to evaluate the project’s execution.

**5.2.3 Real-Time Geospatial Data Preparation (RGSDP)**

Finally, the real-time data contained in the gpx file is converted to a more readable and processable format. The gpx file for each worker is converted to an XML file. The gpx file is divided into columns for each attribute, with information about the employee and their coordinates contained within. The XML data is shown in Table 6.

**Table 6** Sample of prepared real-time data

Employee ID	Employee type	Working activity	Easting	Northing	Distance	Time
A-XX	.	.	.	.	.	.

5.3 Stage 3—Data Analysis

After the data has been transformed, it is analyzed in two dimensions. There are two types of statistical temporal analysis: statistical temporal analysis and spatial-visual analysis. Python® is used to write the algorithmic model.

5.3.1 Analysis Techniques

Numerous mathematical techniques are available for the analysis of spatial-visual and statistical temporal data. Several of these have been selected for inclusion in the algorithmic model. The techniques are:

- 1. Mean
- 2. Standard deviation
- 3. Regression analysis
- 4. Point pattern analysis
- 5. Heatmaps.

5.3.2 Analysis Techniques Implementation and Outputs

The transformed data is then analyzed as described above. Each dataset generates multiple outputs. According to the implementation matrix in Table 7, the various methods are applied to each dataset in Stage 2—Data Preparation and Cleaning. As a result, the analysis is classified by the dataset. There are three types of data analysis: (1) safety data analysis (SDA), (2) quality data analysis (QDA), and (3) productivity data analysis (PDA). This section details the possible outputs for each category.

**Table 7** Framework implementation matrix

	STATISTICAL			
	Heatmaps	Point Pattern Analysis	Regression Analysis	Mean
Site Periodic Data	.		.	.
Real-time Data	•	•		

### a. **Outputs of SDA**

When safety incidents are analyzed on a construction site, the following outcomes are possible:

1. Site safety performance (SSP) using temporal mean.
2. Site safety behavior (SSB) using moving average linear regression.
3. Site safety zones (SSZ) using heatmaps.

- *Mean Analysis*

Provides an average of safety incidents over a specified time period. A construction site with a lower mean number of safety incidents indicates that the site is performing better in terms of safety.

- *Moving Average Linear Regression*

This establishes the relationship between the duration of a project and the number of on-site safety incidents. This relationship can then be used to extrapolate the expected number of safety incidents that may occur over the course of the project's remaining life. A moving average algorithm is used to draw the regression line.

- *Heatmap Analysis*

Visual analysis uses density heatmaps to represent the site's safety risk zones. Heatmaps are created using KDE. Kernel Density Estimation (KDE) estimates the densities of on-site safety incidents. Safety incidents are calculated using the Gaussian Kernel Function and plotted against time  $t$ .

Finally, the KDE density function is used to generate a two-dimensional color contour plot. The contouring of the plot shows the densities of safety incidents that occurred on-site at a given time  $t$ . Zones with a high degree of intensity contouring can be considered high-risk Site zones and vice versa.

### b. **Outputs of QDA**

Analyzing inspection requests on a construction site can result in the following outcomes:

1. Site quality performance (SQP) using temporal mean.
2. Site quality behavior (SQB) using moving average linear regression.
3. Site quality zones (SQZ) using heatmaps.

- *Mean Analysis*

Provides the average number of inspection requests accepted and rejected over a specified time period. A construction site with a higher acceptance rate reflects a higher quality of work than a site with a lower acceptance rate.

- *Moving Average Linear Regression*

This establishes the relationship between the duration of a project and the number of inspection requests accepted and rejected on-site. This relationship can then be

used to extrapolate the expected rate of acceptance or rejection of future requests. A moving average algorithm is used to draw the regression line.

- *Heatmap Analysis*

The Kernel Density Estimate (KDE) calculates the densities of inspection requests that are accepted and rejected on-site. The KDE density function is then used to generate the two-dimensional color contour plot. The plot depicts the densities of accepted and rejected inspection requests that occurred on site at a specified time  $t$  or over a specified period of time  $t$ . Zones of high intensity contouring of accepted inspection requests can be considered as high-quality site zones and vice versa.

### c. **Outputs of PDA**

Multiple analyses could be performed on the workers' real-time GPS transformed coordinates. The findings could help decision-makers understand how workers' spatial behavior affects the project's performance. A worker's temporal spatial coordinates are collected, and the analysis produces the following outputs:

1. Central tendency of workers on site.
2. Spatial randomness and clustering of workers on site.
3. Workers' density on site using heatmaps.
4. Workers' time distribution.

PPA is used to analyze workers' spatial-temporal data and determine their central tendency or dispersion. An event is a worker's location on the premises. The worker's point pattern is defined by his or her location and time. Location also indicates productivity, as explained later in this section.

The point pattern is critical in construction because it provides critical information about personnel behavior based on their locations. Their actions are compared to other site performance indicators to see if their actions affect the project's performance. If a link exists between worker behavior and project performance, it can be classified as positive or negative.

PPA could be performed on a random sample of workers' spatial-temporal data by calculating the measures:

1. *Mean Center*

On-site worker central tendency and dispersion are revealed by the mean center of a dataset of spatial-temporal coordinates. It estimates the location of each point's gravity center. All points are averaged to find the dataset's center. For the mean center, you can use data from one day to one year. In either case, the data points are grouped by time period and the mean center is calculated. This metric represents the central density of workers on-site at any given time.

- *Standard Deviation Ellipse*

Shows the directional effect of dispersion. The standard ellipse calculates the standard distances between the easting and northing axes. The major elliptical axis follows

the direction of major worker dispersion. This is very useful because it shows the worker's dispersion around the site, thus indicating the lowest worker density areas on site and possibly the workers' regular workflow.

## 2. *Spatial Randomness and Clustering of Workers on Site*

Using nearest neighbor distance analysis (NND) can help define spatial randomness or clustering of workers on site. If there are clusters of workers centered around different mean centers, the analysis shows that. First, calculate the mean of NND. The z-score indicates the distribution of workers on a construction site. The greater the value of the z-score, the more significant the pattern.

The G-distance function of the workers' coordinates also indicates their clustering or dispersion pattern across the site. G is calculated according to Eq. (1) [11]

$$G(d) = \frac{\sum(D_{jk} < d)}{n} \quad (1)$$

For on-site workers,  $D_{jk}$  is the shortest distance between adjacent workers that is less than  $d$ . In the bounding box of the workers' coordinates,  $G$  is computed for a cumulative  $d$  between zero and the maximum possible distance between workers. A worker's  $G$ -function  $Gg(d)$  is plotted and compared to the results of  $G(d)$  to determine whether the workers are clustered or dispersed on site.

## 3. *Workers' Density on Site*

It refers to the areas of a job site where workers are concentrated at any given time. Denser areas could be interpreted as more active. These densities could be compared to other site parameters to see if there is a link between worker spatial concentration and site performance. The density is determined using KDE in a similar manner to that explained under the SDA and QDA sections.

- *Heatmap Analysis*

KDE can estimate workers' spatial density over time. For each worker's easting and northing coordinates, the density function can be plotted. Workers' density functions can be visualized using a KDE contour plot (heatmap). The more workers in an area, the higher the probability density, the closer the contour lines in a KDE plot, and the higher the color intensity. The plot and heatmap show the level of activity in various zones. Colors representing high-activity zones have a higher intensity than other colors that may represent low-activity zones.

## 4. *Workers' Time Distribution*

Finally, once the worker concentrations are identified, the time spent on-site can be calculated. The location of workers on-site can reveal whether they are working, traveling, or idle. This is done by determining the worker's bounding zone. If a worker's coordinates are within the working area, they are assumed to be working. The worker is considered idle if he or she is in a resting area; if not, the worker is considered traveling. It is also possible to calculate each worker's percentage of

**Table 8** Output of workers' time distribution analysis

Time category	%
Working	X
Traveling	Y
Idle	Z

time spent working, idle, and traveling. The percentages can be used to calculate on-site worker productivity; higher percentages of working time typically translate into higher productivity. If workers spend most of their time traveling on-site, there may be a layout issue, wasting time that could be spent more productively if the layout was improved.

The above logic is then applied iteratively to all workers', with the result shown in Table 8. The assumption that the worker's location is the sole criterion for categorizing is significant. Unattended workers in a construction zone are therefore undetectable. However, calculating a worker's time on-site using percentages is still valid.

## 6 Case Study

### 6.1 Project Information

The case study presented here is for commercial development in Ain Al Sokhna, Egypt. The project, dubbed New Capital Sportive Village, is part of the Egyptian government's coastal city development strategy. The contract, worth approximately 4 billion Egyptian pounds, was scheduled to be completed in four years. Due to the size of the building areas, the primary project is divided into multiple sub-projects. The framework was implemented on a single sub-project, the hotel area, and data was collected during the construction phase of the chosen sub-project.

### 6.2 Framework Implementation

To begin, a single site coordinate was obtained from the surveying team at the site. The single coordinate was used to locate the site on a geographic information system and to obtain the remaining site coordinates that were required. Secondly, the construction team on-site provided the pre-defined site zones. Thirdly, multiple GPS tracking applications were evaluated to determine which one was the most efficient. Finally, two participants downloaded and used the tracking application for a specified period, and their daily tracks were analyzed.





**Fig. 2** Worker’s daily track displayed in Google Earth Pro

**6.2.1 Data Collection**

The construction site’s geographic data and real-time worker spatial data were collected using the data collection methods described in Sect. 5.1—Stage 1—Data Collection.

Google Earth was used to locate the site geographically and determine the site zones using the single coordinates obtained from the site team, 29.90080957 and 31.68311107. The coordinate was searched for, and the location was obtained.

Once on-site, the engineers let the app track their location for the entire day. After work, the engineers export their gpx tracks and upload them to the framework’s SharePoint. The daily tracks were viewed on Google Earth to verify the workers’ location. Figure 2 shows the workers are correctly located on-site, confirming the tracking app’s accuracy and defining site boundaries.

**6.2.2 Data Preparation and Cleaning**

The data collected could not be directly processed by the analysis algorithms and had to be transformed into a unified format. The site geographic data was manually grouped to create the zone bounding polygon. As shown in Table 9, each set of coordinates corresponded to a specific area of the code.

**Table 9** Site geographic data

Area code	Ref. no	Easting	Northing	Elevation	Category
C01	1	31.673178	29.900439	400	WA

For the site studied, the areas used in the framework are as shown in the table above. These areas are:

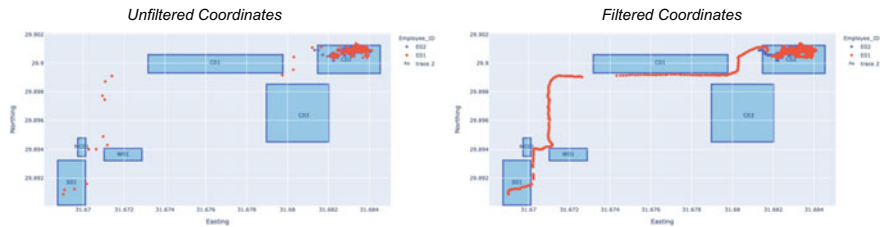
1. Construction Area—C01: The cartouche building.
2. Construction Area—C02: The hotel building.
3. Construction Area—C03: The electronic shooting building
4. Main Storage Area—S01: The project’s warehouse.
5. Main Caravan—MC01: The project’s main caravan where all indirect manpower is located.
6. Workshop—W01: The project’s main rebar carpentry workshop.

For the real-time data, the gpx file was split using Python® splitting to create a data frame. As shown in Table 10, the data frame contained employee spatial and temporal data. Each engineer was assigned a unique ID and a working activity, which in their case is supervision.

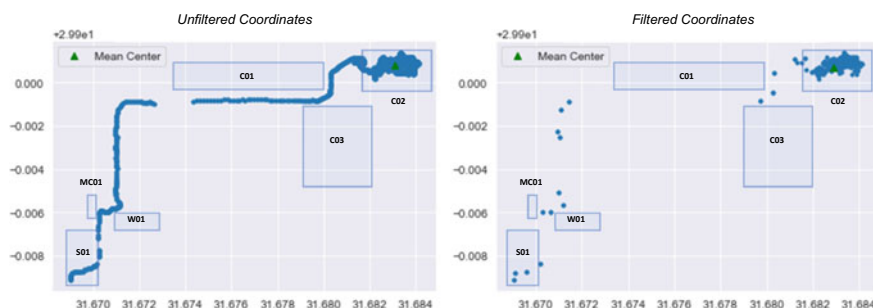
While it is possible to analyze and visualize data without filtering, the process becomes inefficient and time-consuming, and the results are more influenced by noise data points. As a result, noise data points are removed from the visualization and analysis processes using the Ramer-Douglas-Peucker (RDP) algorithm. Thus, the algorithm is used to filter the workers’ coordinates for a more efficient analysis process. Additionally, the data visualization should be enhanced. The scatterplots of the workers after the RDP algorithm was applied to their coordinates are shown in Fig. 3.

**Table 10** Real-time spatial data

Employee ID	Employee type	Working activity	Easting	Northing	Distance	Time	Date
E02	Senior engineer	Supervision	31.682429	29.900498	0.006606	11:31:02	4/9/2019



**Fig. 3** Visualization of workers’ coordinates



**Fig. 4** Mean center of workers

As expected, given that both workers are in charge of supervising construction activities in the hotel area, the majority of their coordinates fall within the boundaries of area C02. Also, the true coordinates of the workers are more visible following RDP analysis, as the points were reduced from 23,869 to 792, compressing the data by 96.7%. Regardless, as will be demonstrated later, the final outputs of the analysis are not significantly affected by the noise filtering, even though the processing time is reduced by 90.4%.

### 1. Central Tendency of Workers on Site

By examining the central tendency of both engineers on-site, the following central measures were obtained:

- *Mean Center*

Before RDP analysis, the mean center is almost identical to the center of C02 (31.682870, 29.900224), indicating that the engineers spent most of their time in C02 during the 10-day period. After applying the RDP algorithm, the mean center was calculated to be (31.68290094, 29.90070515), indicating that the RDP had no effect on the analysis's results. This is shown in Fig. 4.

- *Standard Deviation Ellipse*

Finally, the ellipses in Fig. 5 show the engineers' direction dispersion. It is the same ellipse for both coordinates. Both ellipses' larger axes point west-south and east-north. Due to the site's layout, the main storage area and workshop area are in the far west-south corner, while the C02 is in the far east-north corner. Engineers are expected to move between the corners, causing directional dispersion.

### *Spatial Randomness and Clustering*

Exploring the engineers' spatial randomness and clustering was made easier knowing their central tendencies. For unfiltered data, the NND analysis yielded a p-value of  $1.96588 \times 10^{-6}$  and for filtered data,  $1.40432 \times 10^{-5}$ . Due to the engineers' clustering and lack of complete spatial randomness, both p-values are less than 0.05. The

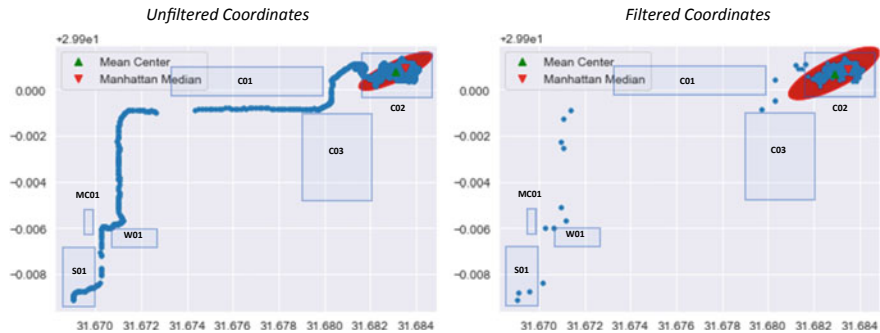


Fig. 5 Standard deviational ellipse of workers

engineers’ clustering was validated using g-distance. XX shows the engineers’ G-function plots, while XX shows their G-function envelopes.

2. Spatial Randomness and Clustering of Workers on Site

The G-plots confirm the engineers’ spatial clustering. The engineer observed coordinates G-plot to the left of the expected worker G-plot on the construction site. The G-distance increased rapidly at shorter distances, indicating a concentration of engineers as shown in Fig. 6.

3. Workers’ Density on Site

Along with clustering the workers, the following techniques were used to determine the workers’ density:

- Heatmap Analysis

The KDE is used to visualize the engineers’ densities using Voronoi analysis. The estimated densities were then used to plot a 2D contoured map of the engineers’ densities using a bandwidth of 0.6. The densities determined the site’s productivity zones. Zones with high intensity-colored contours have more engineers, indicating higher

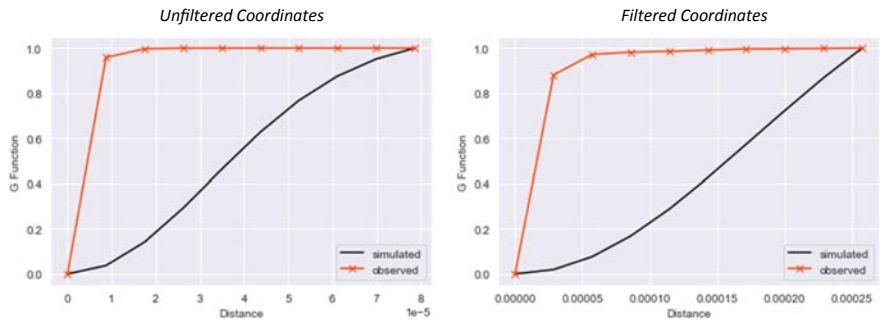


Fig. 6 G-distance plot of workers

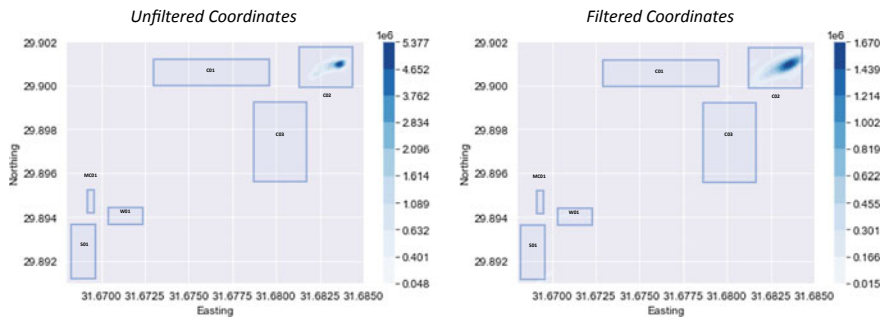


Fig. 7 Heatmaps of workers

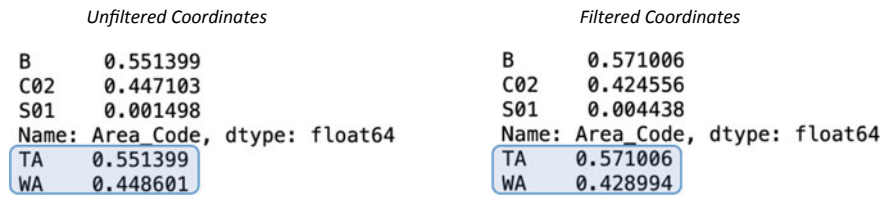


Fig. 8 Workers' time distribution

productivity. Zones with low intensity have medium to low productivity. Figure 7 shows the generated heatmaps for unfiltered and filtered engineer coordinates.

The densities are proportional to the total number of coordinates in both heatmaps. The heatmaps show C02 as the most productive zone on site, as it has the highest intensity contour. This is expected, as the chosen engineers only supervise area C02.

4. Workers' Time Distribution

The last analysis in the case study is the time distribution of the workers. The results of the analysis are shown in Fig. 8.

Both engineers spend 44.9% of their time working and 55.1% traveling using unfiltered coordinates. For both unfiltered and filtered coordinates, the percentage of time spent traveling on-site is higher. This could be due to the calculation's rigidity. In the construction zone, there is no spatial distance beyond which the worker is still considered present. The worker is confirmed to be traveling if located anywhere on site outside of a working or resting area. Engineers' travel time is still longer than ideal for a construction site. This could be due to poor site layout or worker inefficiency. The chosen site appears inefficiently laid out, with the workshop, main caravan, and storage areas distant from each other. This could help decision-makers improve the site's productivity.

## 7 Conclusion and Recommendations

### 7.1 *Research Contributions*

While real-time monitoring has not yet become a standard practice in the construction industry, this research sought to illuminate the potential applications of real-time technologies on construction sites and to fill a knowledge gap by making the following contributions to the body of knowledge:

- GPS monitoring technologies are being used on construction sites.
- Using a Python<sup>®</sup>-based processing algorithm, perform spatial-temporal statistical analysis on data from construction sites.
- Creation of a framework for monitoring and controlling site performance parameters that incorporates GPS monitoring and spatial-temporal analysis.
- Validation of the framework's applicability on a real-world construction site.
- Conducting a study of construction workers' on-site behavior using point pattern analysis.
- Creation of heatmaps and Voronoi diagrams to aid in the interpretation of spatial-temporal data collected from sites.

### 7.2 *Recommendations for Future Research*

Future research recommends that the entire data collection process be fully automated, with data on safety incidents and inspection requests collected in real-time. A mobile phone application could be developed to allow on-site users to input data in real-time, enhancing the data collection stage of the framework. Additionally, when examining workers' spatial behavior, it is possible to take their elevation into account, as this may impact the site's safety, quality, and progress. Additionally, using technologies such as smartwatches, physical parameters such as oxygen saturation, heart rate, and calories burned can be studied. The physical characteristics of workers and their three-dimensional spatial data can be used to ascertain the level of effort exerted by each worker in a particular area on the job.

## References

1. Calvetti D, Meda P, Goncalves MC, Sousa H (2020) Worker 4.0: the future of sensed construction sites. *Buildings (Basel)* 10(10):1
2. Jiang H, Lin P, Qiang M, Fan Q (2015) A labor consumption measurement system based on real-time tracking technology for dam construction site. *Autom Constr* 52:1–15
3. Josephson PE, Saukkoriipi L (2005) Waste in construction projects—need of a changed view. *Fou-väst, report*, p 507

4. Navon R (2005) Automated project performance control of construction projects. *Autom Constr* 14(4):467–476
5. Navon R, Goldschmidt E (2003) Can labor inputs be measured and controlled automatically? *J Constr Eng Manag* 129(4):437–445
6. Nikakhtar A, Hosseini AA, Wong KY, Zavichi A (2015) Application of lean construction principles to reduce construction process waste using computer simulation: a case study. *Int J Services Oper Manag* 20(4):461–480
7. Soltanmohammadlou N, Sadeghi S, Hon CKH, Mokhtarpour-Khanghah F (2019) Real-time locating systems and safety in construction sites: a literature review. *Saf Sci* 117:229–242
8. Teizer J (2015) Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Adv Eng Inform* 29(2):225–238
9. Thomas HR, Horman MJ, Minchin RE, Chen D (2003) Improving labor flow reliability for better productivity as lean construction principle. *J Constr Eng Manag* 129(3):251–261
10. Wang T, Yang T, Yang C, Chan FTS (2015) Lean principles and simulation optimization for emergency department layout design. *Ind Manag Data Syst* 115(4):678–699
11. Yuan Y, Qiang Y, Bin Asad K, Chow TE (2020) Point pattern analysis. In: Wilson JP (ed) *The geographic information science & technology body of knowledge* (1st quarter 2020 edition)
12. Zhang S, Shang C, Fang X, He S, Yu L, Wang C, Yan L (2021) Wireless monitoring-based real-time analysis and early-warning safety system for deep and large underground caverns. *J Performance Constr Facilities* 35(2)

# **Construction Management: Building Information Modeling**



# Robotic Additive Manufacturing Using a Visual Programming BIM Environment



Tayeb Boualam Allah, Walid Anane, and Ivanka Iordanova

**Abstract** In recent years, the application of AM in construction has been increasingly studied. According to scientific literature, the use of robotic AM can help improve efficiency in construction by reducing the cost of completion, minimizing waste, decreasing lead times, and increasing profitability. Building Information Modeling (BIM), a digital technology widely adopted by the industry today, is known to boost construction productivity and quality. However, studies on the use of BIM in the different stages of AM are lacking in the scientific literature, even though there are major challenges to overcome in this area, including the lack of data interoperability between BIM platforms and additive manufacturing systems. In particular, BIM is almost solely used for modeling in the AM process, and other software is needed to process the 3D models and link them to AM robots. This research aims to close the gap between BIM data and AM. To achieve this, attention was paid to data interoperability between BIM platforms and automated building systems (a robotic arm) to preserve BIM model data throughout the manufacturing process. The proposed approach helps prevent the data loss of a 3D BIM model in an additive manufacturing process by using the parametric modeling tool Dynamo. A review of scientific publications, industry best practices, and academic laboratories research was conducted to identify current technologies and practices in the use of BIM in AM. Workflows were developed and tested on technical case studies using an operational framework for integrating AM processes with BIM. The studied cases represent a robotically 3D printout of native BIM models as a practical illustration. Results show that the use of robotic AM in the field of construction can contribute to improving efficiency, reducing the cost of completion, limiting waste, decreasing delays, increasing profitability, and avoiding the need to redo detailed work. It can also contribute to improve construction site safety and provide architects with more freedom of design and functional integration.

**Keywords** Robotic additive manufacturing · Visual programming BIM environment

---

T. B. Allah · W. Anane · I. Iordanova (✉)  
École de technologie supérieure ÉTS, Montreal, Canada  
e-mail: [ivanka.iordanova@ens.etsmtl.ca](mailto:ivanka.iordanova@ens.etsmtl.ca)

## 1 Introduction

Many construction projects today are designed and built using Building Information Modeling (BIM) software. It is a technology that is widely adopted by the industry, and numerous studies have evaluated and communicated the benefits of using BIM throughout the life cycle of a property—from design to operation. Additionally, several recent studies have focused on building automation systems using BIM. Despite its potential benefits for automated construction, its involvement in the various stages of automated construction and the details of its implementation have not been sufficiently discussed. There are major challenges to overcome in this area, including the lack of data interoperability between BIM platforms and building automation systems [4].

Additive manufacturing (AM) is one of the focal points of Industry 4.0 as it enables strong product customization and flexible mass production. To do this, robots are often used. Robot-assisted construction will bring several benefits to the construction industry, such as increased quality, reduced risks, and a faster work cycle. In addition, many developments have been made by industry and academia, covering several distinct applications, for example, automation with masonry machines, installation of steel beams and robotic wooden constructions, and large-scale prefabrication and automation for modular buildings [17].

The research work presented in this report is aimed at construction automation systems using BIM. Its main objective is to propose a process to integrate the AM process with BIM, so that the manufacturing process can be performed from a native BIM environment. To achieve this, attention will be paid to data interoperability between BIM platforms and automated construction systems (robotic arms) and to the preservation of data throughout the manufacturing process.

The paper is structured as follows. After presenting the literature and common practices review on the topic, this article proposes an operational framework for integrating AM into BIM using various digital tools. Then, the adaptation of the framework is validated through several case studies. Finally, a discussion and analysis of the results obtained are presented in the conclusion, and avenues for future development are proposed.

## 2 Literature and Common Practices Review

### *2.1 State of the Art in Additive Manufacturing for Construction*

The application of AM in construction has been increasingly studied in recent years. Large robotic arm and gantry systems have been created to print construction parts using materials based on aggregates, metals, or polymers [13].

Recent research and practices such as Contour Crafting (CC), D-shape, and 3D Concrete Printing (3DCP) have all demonstrated the potential for large-scale processes that adopt AM techniques as an alternative way to construct building components. Conventional building processes share the concept of mold-based shaping with manufacturing. Therefore, AM processes have advantages over conventional building processes, such as customization without additional tools or molds, and the promised freedom of design and functional integration, with simple assembly [11].

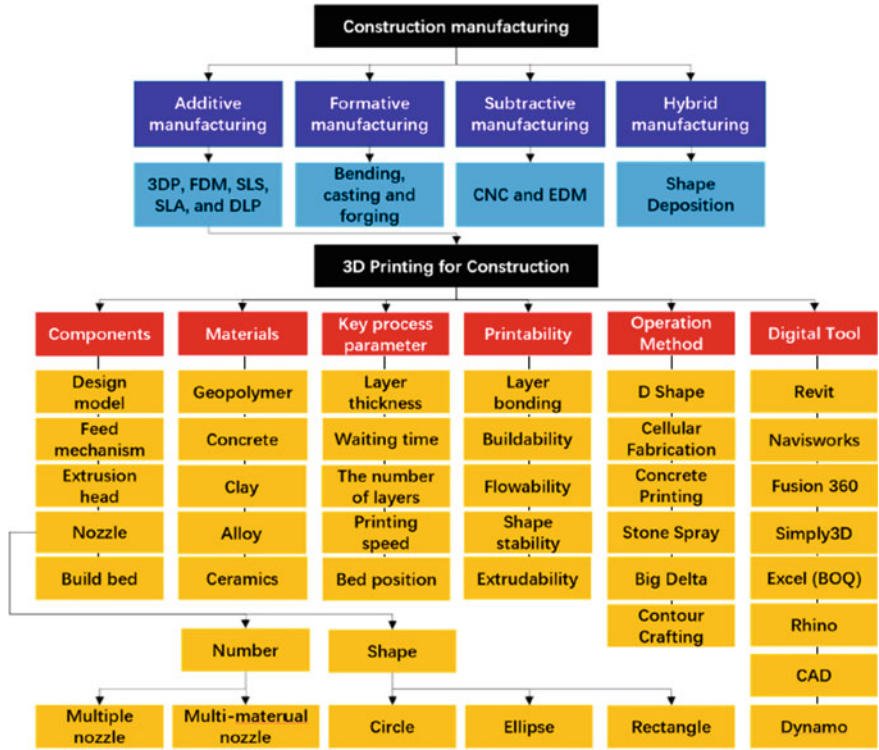
According to Wu et al. [19], 3D printing (3DP) is an emerging technology in the construction industry that has the potential to reduce costs, shorten project timeframes, improve quality, and reduce reliance on labor. It continues to revolutionize manufacturing and production methods by combining digital information with tangible objects.

Figure 1 shows a classification of 3DP systems, digital tools, operation methods, and materials required for 3DP. Among the different digital manufacturing methods, additive, formative, subtractive and hybrid, only additive manufacturing, and more precisely 3D printing in construction, will be developed in detail as it is the object of this paper. The options available for the following aspects are presented in their corresponding columns: components, materials, key process parameter, printability, operation method, and digital tool. It should be noted that in Fig. 1, there is no correspondence between the elements on the same row among the options of the above-mentioned aspects.

## ***2.2 Workflows Supporting Additive Manufacturing in Construction***

Building Information Modeling (BIM) is a holistic construction management approach that covers the full lifecycle of a construction project, including modeling, construction planning, cost estimating, and the management of facilities after delivery. The main feature of a BIM model is that it not only contains geometric information, but also encompasses material, resource, equipment, and manufacturing data. Such a model containing all the required data in a single package would be the ultimate digitization solution for the building and construction industry [18]. BIM has been evaluated by many researchers in recent years and is recognized as a technology capable of improving productivity throughout the building life cycle, from design to operation and maintenance. Current construction projects are primarily designed using BIM platforms such as Autodesk Revit and Bentley Systems MicroStation which are widely adopted in the architecture, engineering, and construction (AEC) industry [4].

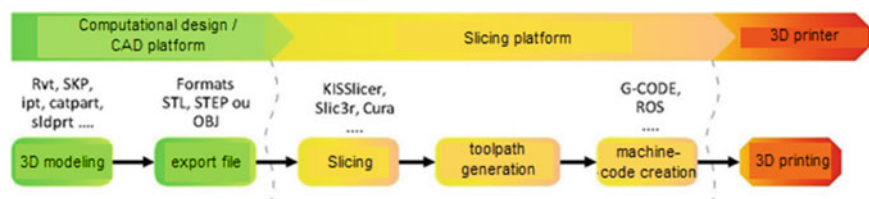
According to Weng et al. [18] and Hager et al. [6], the main processes are common and remain the same for additive manufacturing—they start with digital modeling. The model is then exported to the common 3D data interchange format (STL), which



**Fig. 1** Classification of 3DP systems, digital tools, operating methods, and materials required for 3D printing in construction. [16] Adapted from Sepasgozar et al.. (FDM: fused deposition modeling; SLS: selective laser sintering; SLA: stereolithography apparatus; DLP: digital light processing; CNC: computerized numerical control; EDM: electrical discharge machining; BOQ: bill of quantities)

is finally sliced into several cross sections resulting in a set of 2D contour lines. These layers are then processed to generate the tool path which defines the motion paths the nozzle must follow to produce the desired shape. Finally, the robotic code is generated accordingly before being fed into a 3D printer. In this context, exploiting the potential of BIM has not been fully explored, and its use has been limited to a simple export of the model for printing only. As soon as the model is exported, the parameters associated with it are destroyed, which leaves only a raw model that is difficult to modify according to the printing parameters or for other uses.

According to Weng et al. [18], the current process is not robust as it is vulnerable to data loss and inefficient due to the constant transfer of data from one software to another (Fig. 2). The STL format only uses the three-dimensional description of the surface geometry, that is, it only transmits geometric data without generating other information relating to the properties of the components, for instance, material properties, texture, or color [14]. The STL file format is conceptually simple and easy



**Fig. 2** Conventional 3D printing process

to generate but has issues related to its size and numerical accuracy. It is also not possible to specify material properties, so the fabrication of multi-material structures requires the use of multiple STL files [3]. As a result, much effort and time are involved in the data conversion process, and the quality and accuracy of the final data for practical printing may be affected. As well, the printed result is not always optimal on the first attempt, needing to modify the model according to the difficulties encountered. As such, all the steps related to 3D printing will have to be repeated from the beginning without a possibility of retroaction, which will not be optimal for construction efficiency [2].

### 2.3 *BIM for Digital Additive Manufacturing*

In recent practices, BIM is only used as a modeling software for AM, and other software is needed to process the printing path. Data loss and data non-interoperability may occur due to data transfer problems between different platforms. Since BIM is a widely used approach in the construction industry for designing, managing, and sharing construction information, there is a need for AM technology to be better integrated with BIM. While studies have been carried out to use BIM in additive manufacturing [4, 12], in most cases, the integration of AM into BIM design activities still faces major challenges related to transferring information between BIM tools and building automation systems.

To simplify this time-consuming process, a revised design process flow for 3D printing is proposed by [18]. BIM software is designed to be able to extend its functions, so plugins can be developed to implement slicing functions and send the generated program directly to a 3D printer. Contrary to the traditional process of 3D printing, in this proposed approach, various steps including 3D modeling, clipping, path generation, and robotic code generation are undertaken in the same platform. This workflow consists of developing a script package coupled with Dynamo visual programming tools and integrated with Autodesk Revit 2018 BIM software. Still, the parametric workflow of the script package proposed by [18] is not discussed in detail, and it is only relevant for Cartesian printers and not robotic arms.

To explore the issue of interoperability between BIM and AM, note that Janssen divides parametric BIM modeling into several approaches. Indeed, Janssen [9] classified parametric BIM modeling between embedded and coupled approaches. The embedded approach is to use the functionality of the software. The coupled approach is further subdivided into two: tightly coupled, which brings two systems together through the modeling software's application programming interface (API), and loosely coupled, which involves model exchange through the transfer of file formats. This classification is further explored in the work of [2] where the approaches of using BIM for robotic 3D concrete printing are categorized according to the data flow through the process.

## ***2.4 Visual Programming Environments***

### **Dynamo**

Dynamo is an open-source visual programming editor that allows users to build generative tools to their own specifications. Dynamo offers a robust selection of "nodes" that can be organized into custom "graphs". These terms refer respectively to elementary (lower order) and composite (higher order) operations in a Dynamo workspace, the graph being an arrangement of nodes connected by wires connecting input and output. Nodes tend to be functionally independent. They are specific, basic, and versatile so that their combination avoids predefined results. Although these latest notes suggest that a Dynamo workspace requires somewhat greater interaction than comparable software such as Grasshopper, it exposes advanced functionality while avoiding outright simplicity [10]. For example, to perform a specific function in Grasshopper, only one node may be required, while in Dynamo, a group of nodes may be required to perform that function.

### **Grasshopper**

Grasshopper is a plugin for the Rhinoceros software which is designed by David Rutten. It is primarily known in the construction field as a visual programming language that enables parametric modeling for structural engineering, architecture, and digital manufacturing. The first version, called Explicit History, was released in September 2007, and its popularity has grown ever since. The main reason for its success is that it provides an intuitive way to explore a design without having to learn conventional programming. Users can manipulate their design graphically using so-called components rather than specifying them textually [17]. Grasshopper is a flexible parametric design tool that allows the creation of new algorithms or the modification of existing ones in a graphical scripting environment. It provides freedom and flexibility for designers, including those without programming or scripting knowledge, to apply new rules of behavior and investigate design possibilities in specific problems under different architectural design conditions [11].

## **2.5 Robotics for Additive Manufacturing in Construction**

### **Industrial robots in additive manufacturing**

An industrial robot is defined by the International Organization for Standardization (ISO) as a “general-purpose, reprogrammable, automatically controlled manipulator, programmable along three or more axes, which can be either stationary or mobile for use in industrial automation applications”. Industrial robots are available everywhere, are increasingly inexpensive, and are very versatile. They are also characterized by high precision and dynamics and are easy to control thanks to significant advances in software development. Conventional stationary industrial robots have a limited working space, i.e., a maximum horizontal reach of approx. 3.9–4.7 m, which makes them more suitable for printing objects of a few meters in size and of limited height. Industrial robots were originally designed for stationary use in factories, thus they are generally not suitable for the harsh conditions of construction sites, and this type of manipulator (industrial robot) seems to be less suitable for large-scale printing of buildings. One way to extend the working space of the industrial robot is to mount it on a larger carrier system or make it mobile itself. The original 3.2 m range of the robot can theoretically be extended to an unlimited working space [12].

### **Collaborative robots**

The use of collaborative robots (called cobots) and the collaboration between human robots has increased in various industrial fields in recent years. However, cobots are rarely applied in construction work even though there is a growing need for robotic solutions in this field [5]. Collaborative robotics marks the future stage of convergence between design and construction, an undeniably favorable space where many architects and designers will have the ability to better control not only their work, but also the modes and processes used. Construction sites are often completely different from the traditional industrial sites for which this technology was originally intended [15]. Although these cobots are not huge, they are powerful machines for high payload operations and they can support human workers in strenuous repetitive tasks in the growing field of human–robot collaboration. Since a cobot enables co-manipulation of objects with the human worker, it can provide many benefits, including “force amplification, inertia masking, and guidance via surfaces and paths”. Its installation, programming, and reconfiguration are relatively easy. Compact and lightweight, it has built-in security features. Thus, instead of being used only by experts, it can be used by people with little or no expertise in robotics [1].

### **Computational robotic programs**

Computational robotic programs are ranges of custom components packaged as plugins for parametric design software (Dynamo/Grasshopper) to program industrial robots directly, including a full kinematic simulation of the robot. The generated programs can be run on the robot, without requiring any additional software. Table 1 summarizes the plugins found in the literature depending on the robot manufacturer and the compatible design software.

**Table 1** Summary of robotic programming plugins based on robotic arm manufacturer and appropriate visual programming tools

Visual programming tool	Robotic programming plugin	Robotic arm manufacturer			
		ABB	KUKA	UR	Other
Dynamo	KUKA prc		✗		
	TORO	✗			
	MACHINA	✗		✗	
Grasshopper	KUKA prc		✗		
	HAL	✗	✗	✗	✗
	MACHINA	✗		✗	✗
	ROBOTS	✗	✗	✗	✗
	TACO	✗			

**Robotic simulation**

The simulation of robotic arm actions is a key part of planning the robotic digital manufacturing process. Simulations help to check for errors, such as collisions, out-of-range errors, inaccessible points or singularities, and other problems. Robotic simulation helps to evaluate a feasible toolpath and production plan for 3D printing. He et al. [7] investigates the basic motion and trajectory of the robotic arm to assess the effectiveness of 3D printing. Simulations can be performed in robot simulation and control environments based on visual programming systems such as Grasshopper and Dynamo. These simulations can also be supported in external offline robot simulation and control environments, such as RobotStudio, PowerMill, RoboDK, and Robotmaster, depending on the type of robot used.

**2.6 *Synthesis of the Literature and Common Practices Review***

This literature review provided an overview of common practices in the application of additive manufacturing in construction and more specifically, the application of BIM to additive manufacturing in construction. Indeed, this manufacturing approach in the AEC industry and academia is currently oriented toward discrete products rather than large-scale construction projects and mainly implements AM technology. This may explain the lack of interest in BIM and the increased orientation toward esthetic design processes that do not give much importance to the interoperability of the data used or the collaboration of stakeholders, things a large project would entail. In most instances, BIM is used as a modeling software that exports its model to the manufacturing process and has a unique focus on its geometric shape. No interest is given to the benefit of BIM as a collaborative platform that supports the parametric information management of construction products. However, the growing



interest in automation and emerging construction technologies could change this perception in an industry that is undergoing an abrupt digital shift. In this article, increased importance is given to the design and manufacturing environment of which the product is still discrete but in retrospect is oriented toward extensive projects that prioritize data integration.

### 3 Methodology

This project aims to implement additive digital manufacturing of part of a BIM model directly from the native modeling context (BIM environment). The overall goal of this research is to create, examine, and evaluate an artifact that helps eliminate information loss and ensure data interoperability throughout robotic AM processes in a BIM environment. Indeed, there is a need to better integrate AM technology into the BIM platform to make AM more robust in the construction industry.

As mentioned previously, new workflows must be developed to improve the integration of robotic additive manufacturing from a BIM Revit platform. The specific objectives of this study are twofold: (1) establish an integrated workflow with a streamlined process that enables robotic additive manufacturing of specific geometry from a BIM model in a single BIM environment that preserves printed object information while minimizing the risk of loss of data; (2) verify that this workflow allows for direct interaction between the designer (architect and engineer) of the digital model and the additive manufacturing process in the same native design environment.

To carry out our research project, we adopted a Design Science Research (DSR) methodology—a research paradigm in which a researcher answers questions relevant to human problems through the creation of innovative frameworks, thus contributing new knowledge to the body of scientific evidence. The frameworks designed are both useful and fundamental to understanding this problem. The fundamental principle of design science research is that knowledge and understanding of a design problem and its solution are gained in the construction and application of an artifact [8]. The DSR process consists of six steps: problem identification and motivation, definition of solution objectives, design and development, demonstration, evaluation, and communication; and four possible entry points: problem-centered initiation, goal-centered solution, design and development-centered initiation, and customer/context initiation.

4 Investigation of Possible Workflows for Integrated Additive Robotic Manufacturing Process

As described in the literature, the conventional AM workflow consists of six steps: digital modeling, export, slicing, toolpath generation, robotic code generation, and printing. The overall BIM construction management approach is designed to have the ability to expand these capabilities. In order to streamline the time-consuming conventional process, a new operational framework for integrating AM with BIM is proposed. In this operational framework, the different phases, including 3D modeling, slicing, tool path generation, and robotic code generation, in addition to the simulation phase, are all integrated into the same BIM platform. With built-in default inputs and a standby robotic arm, it will now be possible to combine all these phases into a fully integrated procedure, which will greatly simplify the workflow, from digital design to printed elements. These components will be described in depth below.

4.1 Mapping of Common Digital Tools Needed to Integrate AM into BIM

As seen in the literature, there are several digital tools that can be used to achieve the different phases of the proposed operational framework for AM. Table 2 lists and classifies the different tools in relation to their functionality in the proposed operational framework.

These tools do not all work together, and it is not possible to control and generate codes for all families of robots. A comparison of the characteristics and applicability of these tools is performed to identify their suitability for implementing the proposed AM process. Considering these elements, their compatibility, and their applicability, the mapping of these digital tools on the proposed operational framework for integrating AM into BIM allowed us to draw up several alternatives (workflows). Table 3 lists and ranks the different robotic AM alternatives considering their functionality in the proposed operational framework.

Table 2 Sorting of the digital tools according to their functionality in the mapped workflows

AM phases	Modeling	Slicing	Toolpath generation	Simulation	Robotic programming plugin
Digital tools	<ul style="list-style-type: none"><li>• Revit</li><li>• Rhinoceros</li><li>• Dynamo</li><li>• Grasshopper</li></ul>	<ul style="list-style-type: none"><li>• Dynamo</li><li>• Grasshopper</li></ul>	<ul style="list-style-type: none"><li>• Revit</li><li>• Rhinoceros</li><li>• Dynamo</li><li>• Grasshopper</li></ul>	<ul style="list-style-type: none"><li>• Dynamo</li><li>• Rhinoceros and/or Revit</li><li>• RobotStudio</li><li>• PowerMill</li><li>• RoboDK</li></ul>	<ul style="list-style-type: none"><li>• KUKA prc</li><li>• TORO</li><li>• MACHINA</li><li>• HAL</li><li>• ROBOTS</li><li>• TACO</li></ul>

**Table 3** Alternative workflows of robotic AM according to the compatibility and applicability of digital tools considering their functionality in the proposed operational framework for integrating AM into BIM

Workflow	Environment	Robotic Programming Plugin	Embedded/External Simulator	Alternative Workflows for AM based on the use of digital tools					
				Modeling	Slicing	Toolpath Generation	Embedded / External Simulation	Robotic Code Generation	Robotic Arm for the Printing Process
1	Dynamo	Machina	RobotStudio	Dynamo	Dynamo	Dynamo	RobotStudio	Machina	ABB
2			PowerMill	Dynamo	Dynamo	Dynamo	PowerMill	Machina	ABB / UR / Kuka / etc.
3		Toro	RobotStudio	Dynamo	Dynamo	Dynamo	RobotStudio	Toro	ABB
4			PowerMill	Dynamo	Dynamo	Dynamo	PowerMill	Toro	ABB
5		KUKA pre	Dynamo	Dynamo	Dynamo	Dynamo	Dynamo	KUKA pre	KuKa
6	Grasshopper (Powered by Rhino Inside Revit)	KUKA pre	Rhinoceros And/or Revit	Grasshopper	Grasshopper	Grasshopper	Rhinoceros And/or Revit	KUKA pre	Kuka
7		Hal	Rhinoceros And/or Revit	Grasshopper	Grasshopper	Grasshopper	Rhinoceros And/or Revit	Hal	ABB / UR / Kuka / etc.
8		Machina	RobotStudio	Grasshopper	Grasshopper	Grasshopper	Robot-Studio	Machina	ABB
9			PowerMill	Grasshopper	Grasshopper	Grasshopper	PowerMill	Machina	ABB / UR / Kuka / etc.
10		Taco	Rhinoceros And/or Revit	Grasshopper	Grasshopper	Grasshopper	Rhinoceros And/or Revit	Taco	ABB
11		Robots	Rhinoceros And/or Revit	Grasshopper	Grasshopper	Grasshopper	Rhinoceros And/or Revit	Robots	ABB / UR / Kuka / etc.
12		RoboDK	RoboDK	Grasshopper	Grasshopper	Grasshopper	RoboDK	RoboDK	ABB / UR / Kuka / etc.

4.2 Workflow Analysis

The mapping of digital tools on the operational framework for integrating AM with BIM (Table 3) allowed us to develop 12 possible scenarios, each with its own limitations, functionalities, and characteristics. The differences reside mainly in the families of robots that each combination can command, the approach to the transfer of data between digital tools to carry out a stage of the process (tightly coupled or weakly coupled according to the classification of Anane et al. [2]), and also the simulation platform to be used, internal or external. Table 4 summarizes the possible differences between the various combinations of digital tools for the proposed operational framework.

**Table 4** Possible discrepancies between the various combinations of digital tools for the proposed operational framework

Workflow	Robotic arm manufacturer				Data transfer		Simulator type	
	ABB	UR	KUKA	Other	Tightly coupled	Loosely coupled	Embedded	External
01	X					X		X
02	X	X	X	X		X		X
03	X					X		X
04	X					X		X
05			X		X		X	
06			X		X		X	
07	X	X	X	X	X		X	
08	X					X		X
09	X	X	X	X		X		X
10	X				X		X	
11	X	X	X	X	X		X	
12	X	X	X	X		X		X

**4.3 Workflow Selection Criteria**

The twelve combinations make it possible to apply the AM process from a BIM environment. To determine the most technically, economically, and reliably robust workflow, we have defined the following criteria:

1. Diversity of connected robots. Some combinations only allow one family of robotic arms available on the market to be programmed. Therefore, the optimal combination is one that offers a wider choice.
2. Does not require a license. This criterion particularly relates to the parametric controls of the robot as well as the simulation environment (internal or external) and does not include the modeling platforms because they are all under general license (Revit and Rhinoceros).
3. No external simulator required (tightly coupled approach). The suit must allow the simulation to be performed without using an external simulator, which combines two systems through the modeling software’s application programming interface (API).
4. Data transfer is not required and does not involve model exchange via file format transfer (tightly coupled).
5. No knowledge of robotic language is required. To achieve our goals, the workflow should not require any prior knowledge of robotic languages, which simplifies the designer’s integration into the AM process.
6. Optimization of robot handling.
7. Impact on the quality of the final product. The workflow should not affect the quality of the final product.

8. Lead time. The duration of the process should be as short as possible.
9. Accuracy of the final printed geometry. The precision of the printed object is an important factor in determining the quality of the construction; this criterion is very important in determining the workflow.
10. Does not require manual handling. For the workflow to be robust, it should not support any direct manual manipulation; all action should be occurred through the application programming interface.

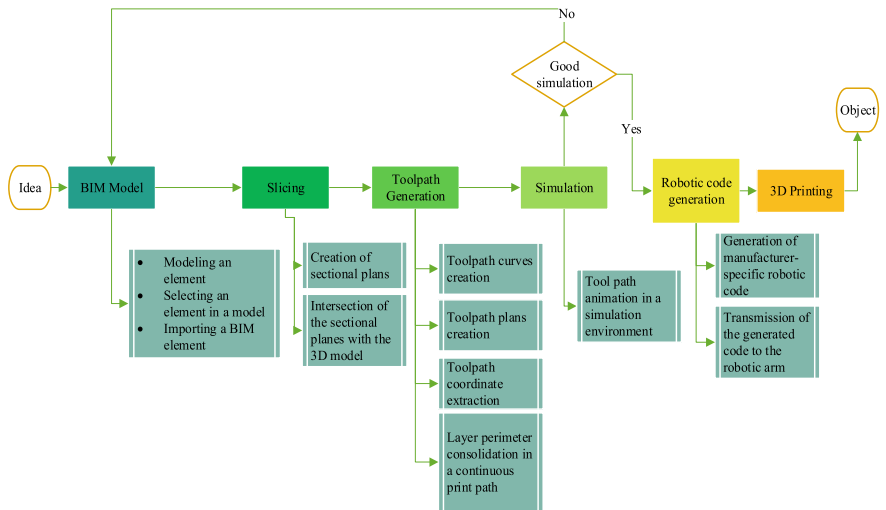
Based on these selection criteria and after analyzing possible scenarios, we have deduced that combination number 11 (Rhino.Inside.Revit/Grasshopper/Robots) is the most suitable. It offers the possibility of programming a multitude of families of robots, the digital tools used do not require licenses, all the phases of the AM process are executed in the same API (thus, no data transfer is required), and no skills in robotic language are needed (simplifying the designer's AM integration into the BIM process) [2]. However, this paper aims to develop a workflow within a BIM environment, without using additional licensed software (such as Rhinoceros). Therefore, below, we will evaluate workflows that do not include Rhinoceros.

#### ***4.4 Workflow Developed According to the Selection Criteria***

To propose an operational framework and workflow for integrating AM with BIM, a set of scripts was developed using Dynamo within Autodesk's BIM Revit software. Dynamo uses visual scripts in the form of nodes and links, where each node or group of nodes performs a specific function. The workflow of the scripting package we developed is shown in Fig. 3. The objective of this work is to implement the presented approach and produce by AM a predefined column as a practical demonstration case study.

### **5 Application of the Selected Integrated Workflow on Case Studies**

Because the workflow determined in the previous section is already being studied in another research project, and to come back to the above-mentioned problem (reasons for which Dynamo was not chosen for AM), we have decided to choose two alternatives that include Dynamo from the scenarios determined for implementation (scenarios 1–5 from Table 4): a first alternative tightly coupled and that the second, loosely coupled. Considering the efficiency of data interoperability, we opted for the fifth workflow which is tightly coupled. As part of the loosely coupled approach, we opted for the first workflow, since the Machina parametric command is free, and the trial version of RobotStudio stays active even beyond its expiration, allowing us to carry out the simulation, even though some advanced functions will be suspended.



**Fig. 3** Flowchart of the developed framework to perform the process BIM robotic AM

**5.1 *Tightly Coupled Workflow: Dynamo/KUKA for Robotic AM***

As mentioned, in this section, we will develop and test scenario number 05 (Dynamo/KUKA|prc/Dynamo) of Table 3. This workflow consists of applying the entire AM process using Dynamo integrated with Autodesk’s BIM Revit software. The practical application will introduce the proposed method and present the tests carried out to evaluate it. The last step of this workflow will be exempt from this test as we do not have a KUKA robotic arm at our disposal.

The implementation of this workflow will take place in five major steps (Fig. 4). The version of Dynamo used is 1.3, available with the 2018 release of Autodesk Revit. The reason a previous version of Dynamo was used is that the available version of KUKA|prc does not actively support new versions of Dynamo.

**5.2 *Loosely Coupled Workflow: Dynamo/Machina for Robotic AM***

Here, we develop and test scenario number 01 of Table 3 (Dynamo/Machina/RobotStudio). This workflow consists of performing the entire AM process using Dynamo integrated with Autodesk Revit BIM software, except for the simulation which will be performed in an external RobotStudio simulation environment. This practical application will also introduce the proposed method and present the tests carried out to evaluate it. The last step of this workflow will involve a 3D printing

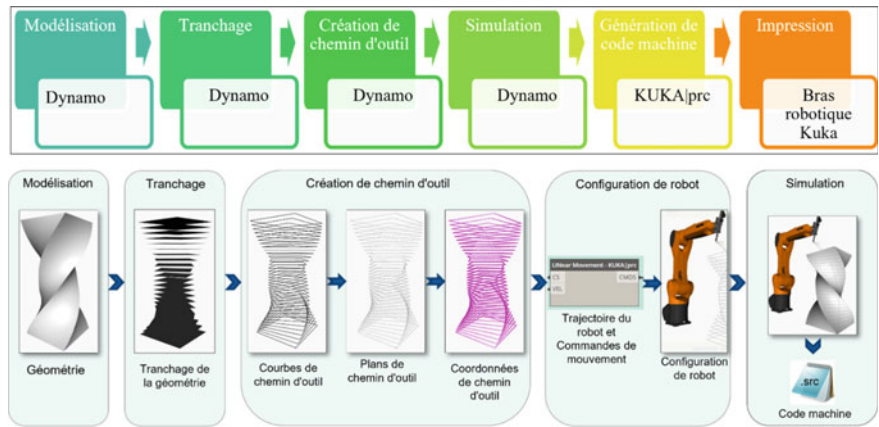
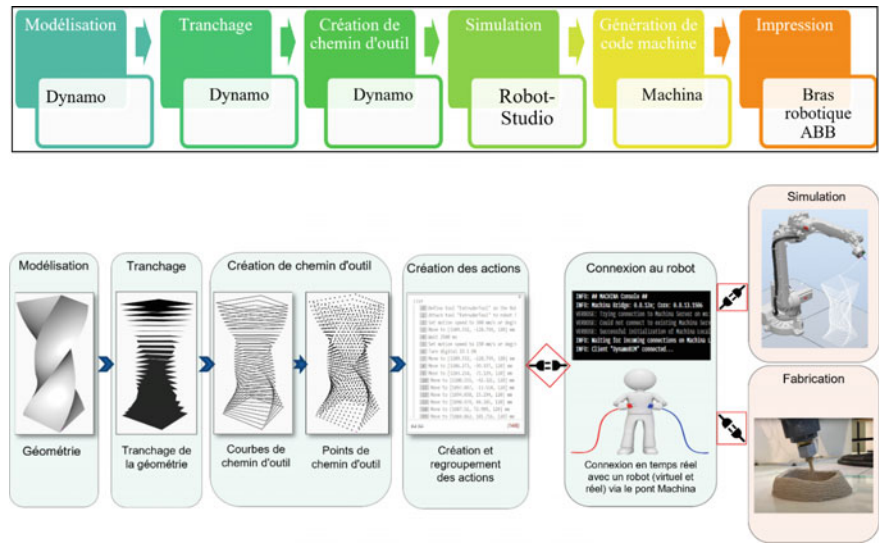


Fig. 4 Workflow of scenario number 05 (Dynamo/KUKA|prc)

in mortar with an ABB-IRB6700 robotic arm. The implementation of this workflow will take place in six major steps (Fig. 4). Dynamo version 2.1 was used, which is available with the 2022 version of Autodesk Revit, and it was the latest version available while working on this research project, and as for Machina for Dynamo, version 0.8.8 was used.



The fact that Dynamo Studio is no longer available for purchase and that it will be removed from the Autodesk AEC Collection in 2022 does not preclude using code developed using only the Dynamo version from Revit. The purpose of using Dynamo Studio in the tightly coupled workflow, Dynamo/KUKA for robotic AM,

was to import the digital twin of the extrusion tool into the robot configuration to visualize and simplify the configuration of the Tool Center Point (TCP). This is because the tools used to import the model/digital twin are only available in Dynamo Studio and not in the Dynamo version of Revit. It is possible to configure the TCP without importing the digital twin of the extrusion tool using only Dynamo from Revit, and it gives the same result, but this requires knowledge of the axes of the robot. It is not a major hindrance and can be solved with some practice.

### ***5.3 Results from the Case Studies***

In common practices, BIM is only used as modeling software for AM, and other software is needed to process the created 3D models to get the print path. Once the model is created, it needs to be exported as an STL file, sliced by a 3D printing slicer (e.g., KISSlicer, Cura) to generate the print path, and converted to machine code for 3D printing. Therefore, a lot of effort and time is involved in the data conversion process, and problems (such as data loss and data interoperability) may occur due to data transfer between different platforms. Also, as soon as the model has been exported, all its parameters are lost and it is in a raw state. In addition, the printed outcome is not always optimal on the first attempt, which requires adjusting the model according to the difficulties encountered during the manufacturing process. Indeed, with such a method, all the steps related to 3D printing will have to be repeated from the beginning without reversibility, which will not only be costly for the construction deadlines, but wasteful in the expenditure of material.

In the proposed AM to BIM integration operational framework, all the steps of a conventional AM process are performed and integrated into the same BIM platform. This preserves all the information of the BIM model as well as its geometric accuracy. In addition, this new approach saves unnecessary time and effort in data transformation. Furthermore, in the case of a modification of the model by virtue of the constraints linked to the manufacturing process, or of a conceptual revision, the steps of the additive manufacturing process consisting in cutting the digital model, producing the printing path, and creating the robotic code will be done automatically without the need to recreate a new script or recreate each step separately. This saves time and improves project performance between the design and construction phases.

## **6 Discussion and Conclusion**

The contribution of this research project is to make it possible to undertake AM processes using a native BIM model. The integration of BIM to AM allows better coordination for more precise, error-free, and waste-free manufacturing, as well as a reduction in construction costs, improvement in manufacturing quality, which meets new environmental standards. Digital fabrication now allows architects and



engineers to integrate complex shapes and spaces into building design directly in a BIM environment.

The use of robotic AM in the field of construction can contribute to improving efficiency, helping to reduce the cost of completion, limiting waste, decreasing delays, increasing profitability, insofar as it provides a high-quality product in terms of finish and prevents the need to redo detailed work in a construction project. The use of robotic additive manufacturing in construction also contributes to improving construction site safety, as the use of robots can help to reduce work accidents caused by human factors. As for the quality of the architectural project, robotic AM helps to improve the product because it gives more freedom to architects and engineers to exhibit their ideas and creativity.

**Acknowledgements** The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for their financial support through Discovery Grant 2020-05641.

## References

1. Afsari K (2018) Applications of collaborative industrial robots in building construction, vol 8
2. Anane W, Iordanova I, Ouellet-Plamondon C (2021) The use of BIM for robotic 3D concrete printing. <https://eamciprodendpoint00.azureedge.net/eamciwesteuprod/production-mcigroup-public/05cc1aabf4894a8d9a1b0654dafce423>
3. Craveiro F, Duarte JP, Bartolo H, Bartolo PJ (2019) Additive manufacturing as an enabling technology for digital construction: a perspective on construction 4.0. *Autom Constr* 103:251–267. <https://doi.org/10.1016/j.autcon.2019.03.011>
4. Davtalab O, Kazemian A, Khoshnevis B (2018) Perspectives on a BIM-integrated software platform for robotic construction through contour crafting. *Autom Constr* 89:13–23. <https://doi.org/10.1016/j.autcon.2018.01.006>
5. Gautam M, Fagerlund H, Greicevci B, Christophe F, Havula J (2020) Collaborative robotics in construction: a test case on screwing gypsum boards on ceiling. In: 2020 5th international conference on green technology and sustainable development (GTSD), pp 88–93. <https://doi.org/10.1109/GTSD50082.2020.9303061>
6. Hager I, Golonka A, Putanowicz R (2016) 3D printing of buildings and building components as the future of sustainable construction? *Proc Eng* 151:292–299. <https://doi.org/10.1016/j.proeng.2016.07.357>
7. He R, Li M, Gan VJL, Ma J (2021) BIM-enabled computerized design and digital fabrication of industrialized buildings: a case study. *J Clean Prod* 278:123505. <https://doi.org/10.1016/j.jclepro.2020.123505>
8. Hevner A, Chatterjee S (2010) Design research in information systems, vol 22. Springer US. <https://doi.org/10.1007/978-1-4419-5653-8>
9. Janssen P (2015) Parametric BIM workflows, vol 10
10. King N, Melenbrink N, Cote N, Fagerström G (2016) BUILD-ing the MASS Lo-Fab Pavilion. In: Reinhardt D, Saunders R, Burry J (eds) *Robotic fabrication in architecture, art and design*. Springer International Publishing, pp 362–373. [https://doi.org/10.1007/978-3-319-26378-6\\_29](https://doi.org/10.1007/978-3-319-26378-6_29)
11. Lim S, Buswell RA, Valentine PJ, Piker D, Austin SA, De Kestelier X (2016) Modelling curved-layered printing paths for fabricating large-scale construction components. *Addit Manuf* 12:216–230. <https://doi.org/10.1016/j.addma.2016.06.004>

12. Mechtcherine V, Nerella VN, Will F, Näther M, Otto J, Krause M (2019) Large-scale digital concrete construction—CONPrint3D concept for on-site, monolithic 3D-printing. *Autom Constr* 107:102933. <https://doi.org/10.1016/j.autcon.2019.102933>
13. Paolini A, Kollmannsberger S, Rank E (2019) Additive manufacturing in construction: a review on processes, applications, and digital planning methods. *Addit Manuf* 30:100894. <https://doi.org/10.1016/j.addma.2019.100894>
14. Sakin M, Kiroglu YC (2017) 3D printing of buildings: construction of the sustainable houses of the future by BIM. *Energy Proc* 134:702–711. <https://doi.org/10.1016/j.egypro.2017.09.562>
15. Schwartz T, Andraos S, Nelson J, Knapp C, Arnold B (2016) Towards on-site collaborative robotics. In: Reinhardt D, Saunders R, Burry J (eds) *Robotic fabrication in architecture, art and design 2016*. Springer International Publishing, pp 388–397. [https://doi.org/10.1007/978-3-319-26378-6\\_31](https://doi.org/10.1007/978-3-319-26378-6_31)
16. Sepasgozar SME, Shi A, Yang L, Shirowzhan S, Edwards DJ (2020) Additive manufacturing applications for industry 4.0: a systematic critical review. *Buildings* 10(12):231. <https://doi.org/10.3390/buildings10120231>
17. Vantyghe G, Ooms T, De Corte W (2021) VoxelPrint: a grasshopper plug-in for voxel-based numerical simulation of concrete printing. *Autom Constr* 122:103469. <https://doi.org/10.1016/j.autcon.2020.103469>
18. Weng Y, Mohamed NAN, Lee BJS, Gan NJH, Li M, Jen Tan M, Li H, Qian S (2021) Extracting BIM information for lattice toolpath planning in digital concrete printing with developed dynamo script: a case study. *J Comput Civ Eng* 35(3):05021001. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000964](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000964)
19. Wu P, Zhao X, Baller JH, Wang X (2018) Developing a conceptual framework to improve the implementation of 3D printing technology in the construction industry. *Archit Sci Rev* 61(3):133–142. <https://doi.org/10.1080/00038628.2018.1450727>

# An Integrated BIM-GIS Dashboard to Improve BIM Coordination



Anouar El Haïte and Conrad Boton

**Abstract** For several years now, the construction sector has been undergoing a digital transformation which is changing working methods. Companies in the construction sector must review and improve their internal processes to be more efficient and productive. Building Information Modelling (BIM) is among the increasingly popular technological approaches being employed in the sector. For construction projects using BIM, BIM coordinators very often require rapid and efficient access to information to respond to the various requests they receive from other project teams (site teams, subcontractors, engineering offices, etc.). For infrastructure projects, on the other hand, BIM is associated with geographic information systems in order to conduct spatial analysis. In the latter context, the present research project aimed to develop and implement a GIS-BIM tool to aid in BIM coordination in the case of a major infrastructure project in Montreal. The choice to develop an automated dashboard integrating GIS-BIM information was made following an internal needs analysis. The tool was mainly developed with Microsoft Power BI, using different data from Excel, Revit, ArcGIS, etc. Following the development of the tool, the dashboard was presented to and tested by various users to collect their opinions and thus confirm whether it responded well to the initial problem.

**Keywords** Integrated BIM-GIS dashboard · BIM coordination

---

A. El Haïte (✉) · C. Boton

LaRTIC—Research Laboratory on Information Technology in Construction, Department of Construction Engineering, École de Technologie Supérieure, Montreal, Canada  
e-mail: [anouar.el-haite.1@ens.etsmtl.ca](mailto:anouar.el-haite.1@ens.etsmtl.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_38](https://doi.org/10.1007/978-3-031-34593-7_38)

623

## 1 Introduction

The study aims to provide BIM coordinators with a new aggregator tool. The tool is based on the convergence between BIM and GIS, through a dashboard representation, where textual, tabular, 2D, 3D, and georeferenced information is centralised in a clear, accessible, and visual manner. The main advantage of developing an aggregator tool is the ability to group several quantities or flows into a single final entity.

The division of construction project teams creates a real organisational divide. This is all the more true in the case of a construction project carried out by a consortium of companies. The organisational structure of the consortium chosen for the major infrastructure project studied in the present paper is built on a combination of five major construction companies in Canada. Given the sheer number of subcontractors, it is difficult for one party to easily know the status of the others. The research focuses on the needs of the BIM coordinator who is not constantly on-site and sometimes struggles to know the precise state of progress of on-site work, due to a lack of immediate communication with the work team.

The project aims to produce an aggregator tool that will collect information automatically, allowing BIM coordinators to free themselves from certain time-consuming information research tasks. The tool is presented as an element allowing to reduce information search times, thus leading to increased productivity on the part of the BIM coordinator. Due to its design, this tool will always be flexible, allowing it to better meet the changing needs of a construction project. The overall objective of this project is therefore to implement such a tool combining GIS and BIM, which generates an automated dashboard by synthesising different bits of information from different data sources.

## 2 Literature Review

### 2.1 *BIM and GIS: Convergence Between Two Complementary Information Systems*

Building Information Modelling (BIM) begins with the creation of a multidimensional digital model, which then makes it possible to document the design of a structure and to simulate its construction and maintenance phases [3, 9]. Information on the built facility is integrated within the BIM model and can be used for different purposes such as visualisation, coordination, prefabrication, facility management, planning, and estimation [9]. Depending on the various analyses issued from the 3D model, the project documentation for the construction of the structure can be generated and distributed from the BIM model.

Narrowly defined, the Geographic Information System (GIS) is a computer system used to capture, manipulate, store, and visualise digital spatial data [8]. More generally, GIS can be considered as a digital system for the acquisition, management,

analysis, modelling, and visualisation of spatial data for planning, administration, and control of the natural environment and socio-economic applications [5]. GIS is systems that can study problems or phenomena at different scales (large, medium, and small scales) according to the established need. On a large scale, GIS may be focused on urban planning or network management, while on a medium and small scale, it can, for example, focus on road engineering studies.

As tools, BIM and GIS apply at different scales. BIM tends to focus on information derived from a building or an infrastructure project, while GIS has broader applications, such as those related to the city, the region, or any territorial study, generally speaking. BIM advantages include the richness of semantics associated with the geometric information throughout the project lifecycle, while GIS covers decision-making based on geovisualisation and geospatial modelling [12]. Taken separately, these two scales can be interesting, depending on the needs of the study to be conducted. However, combining information from BIM with that from GIS can provide an entirely new scope. The merging of these two systems is particularly justified in the case of large infrastructure projects extending over large geographic areas.

Planning infrastructure projects requires taking into account very different scales, ranging from chaining in kilometres, for the general routing of sections of tracks, to centimetres, for details of connection points. Multiscale representations [2] are already well established in GIS, where maps represent a specific level of detail to reduce the complexity of content, thus improving its readability and allowing the reader to focus on the essential information being shown by the layer. Such a representation is also widely used in construction planning, through overall plans (aimed at easily contextualising the scope of the project), site plans (aimed at defining the exact location of the project), and even shop drawings showing the assembly details and exact dimensions of each sub-assembly of the system, etc.

## ***2.2 Dashboard as a Tool for Aggregating Information and Helping with BIM Coordination***

To measure the progress and impact of digital transformation in the construction industry, it is essential to follow several key steps, namely benchmarking, performance measurement (impacts and benefits), as well as maturity and competence at the industry, organisational, and individual levels [1]. Key Performance Indicators are generally indicators used for decision support in organisations and which allow to direct actions aimed at accomplishing a strategy. At the industry level, Key Performance Indicators (KPIs) consist of capital costs, construction time, predictability, faults, accidents, productivity, turnover, and profits. At the organisational level, the KPIs are comprised of working hours spent per project, speed of development, cash flow, etc. [1]. KPIs in BIM coordination are mostly situated at the project level, where they include quality control, on-time completion, and generated costs. Sarkar et al.

[10] carried out a study to further develop the role of KPIs in evaluating the use of BIM as a tool to support equipment management. Based on a questionnaire to which 69 participants responded, 41 main KPIs were identified, and 15 were finally selected by factor analysis.

In information integration, information from heterogeneous sources is combined into a new information set in order to reduce redundancy and uncertainty [6]. Information integration is one of the main challenges tackled by BIM, which archives project data and allows users of information to consult it at any time. BIM coordinators are particularly interested in this notion because they must generally search for information in several very diverse databases, which in some cases can lead to a drop in their productivity at work. Good integration of information within the various construction processes is therefore synonymous with reducing the number of time-consuming actions. This then leads to more effective preventive actions by BIM coordinators and thus makes it possible to maintain high quality assurance and therefore to a reduction of any additional costs that may arise due to a lack of synergies [7].

Dashboard tools allow companies to view accessible and understandable information pulled in from various raw data sources. According to [4], dashboards are cognitive tools that improve the extent of control over the different disparate data that a company generates. Dashboards, generally including using Key Performance Indicators, allow users to visually identify trends, patterns, and even anomalies before they occur. This ability, which ensures constant monitoring of the evolution of a situation, makes it possible, among other things, to prevent certain events before they occur [4]. This tool is particularly useful as it allows to make logical sense of certain pieces of information, which taken individually, may not seem obvious at first glance. To this end, the dashboard uses different visual representations (textual, tabular, graphic, etc.). The dashboard is, in short, an aggregating tool that makes it possible to report a very precise situation, at a given moment, and allows to take coherent decisions according to the problem encountered [11]. For the above reasons, if used correctly, dashboards generate real added value in BIM management due to their ability to synthesise information and make it graphically accessible to different users.

### **3 Methods and Tools**

#### ***3.1 File Creation and Validation***

For the purposes of the present research, a file was created. The objective of the file created was to meet a business need, by presenting different visual representations on a dashboard in order to allow a better understanding of a problem. In the file creation phase, it was considered using the ArcGIS Dashboard application in order to simulate graphical representations. However, the possibilities were limited, because

ArcGIS Dashboard does not offer many of the features available in Power BI, such as the integration of data from very diverse sources, including SQL databases, or the integration of Excel spreadsheets, for example. Thus, Power BI was chosen as the platform to develop this dashboard. ArcGIS Online is used for 2D map visualisation and Tracer for 3D visualisation as data sources, and a simulation Excel spreadsheet file has also been developed. This Excel simulation aimed to contextualise various equipment management issues such as logistics and quality management.

An action-research methodology was followed in creating the file. Action-research consists five main steps, namely identifying a problem, establishing an action plan, carrying out actions in the field by collecting data, analysing this data by evaluating the results, and finally, sharing the research findings. For the present research project, we had to start by understanding the needs of future users of the tool in order to design a suitable tool. As the tool developed, meetings were arranged with BIM managers to present the tool development progress and to gather their opinions and comments.

The validation phase of a developed tool is essential because it involves evaluating whether the tool developed adds value to commonly used methods. It also involves evaluating the characteristics offered by the tool based on various criteria, such as ergonomics, accessibility of handling, visual appearance, and technicality. Since this work employed the research-action method, intermediate validation phases were carried out during the entire development phase of the tool. This consisted in questioning practitioners from different disciplines of the project about in which extend the proposed tool meets their needs. Remote video interviews were carried out, starting with an overall presentation of the tool, followed by a scenario in which the user interviewed alone manipulated the dashboard in Power BI. The results of this validation study were collected by hand and allowed to accurately state whether the study could be transposed to other similar works and, thus, to determine whether future uses and improvements for other types of projects could be foreseen.

## ***3.2 Methods and Tools Used in the Development of the Tool***

### **3.2.1 Power BI Desktop and Power BI Service**

The objective of this research project was to propose an integrated BIM-GIS dashboard integrating information from various sources. Microsoft's Power BI was deemed to be the most suitable software for achieving this, which is why it was chosen for the work.

Power BI's various extensions allow users to transform independent data sources into consistent, visually immersive, and interactive information. There are many types of data that can be imported into and analysed in Power BI, and they range from data in a simple Excel spreadsheet to hybrid cloud-based or web-based data. Connecting to Power BI data sources is relatively simple, as is connection to its

generated dashboards which can be easily viewed, tracked, and shared. Power BI made it possible to configure the dashboard in-depth, including for example, the ArcGIS Maps visualisation. This includes the management of several visualisations that can only be fully configured on the desktop version of Power BI.

In order to validate the overall functionality of the dashboard, it was necessary to share it with several users to retrieve their opinions and comments with respect to various criteria such as general aspects, ergonomics, and accessibility. The Power BI Service extension was used for this because it is an online version of Power BI. The Service extension allows users to access the dashboard only using a Uniform Resource Locator (URL) link; it can also be transferred via Outlook or Microsoft Teams.

### **3.2.2 ArcGIS Pro and Maps for Power BI**

ArcGIS is a geographic information software suite developed by ESRI. The application provides many licences, each with specific uses. Among these, ArcMap and ArcGIS Pro are the two versions most used by professionals. ArcGIS Pro offers more geodata processing and geographic analysis capabilities than ArcMap. ArcGIS was chosen over JMap because JMap does not offer information integration in Power BI, while ArcGIS does offer this functionality. In addition, when used in a GIS-BIM study, ArcGIS Pro has the advantage of allowing the integration of a BIM 360 cloud server in order to integrate various Revit 3D models within the main project in ArcGIS Pro. This functionality is precisely the one that was used in this research project. After creating a georeferenced project scene in ArcGIS Pro, a BIM 360 connector was programmed to import the desired Revit models into the map. The geodetic datum (or coordinate systems) that was used is the NAD 1983 MTM Zone 8, which corresponds to the datum used on a construction project.

Once the coordinate system is defined, the desired Revit model can then be integrated into the active map. The project is thus automatically positioned in planimetry and altimetry. Manual corrections to its planimetric and altimetric positions can then be made if necessary. Once the integration is complete, it is possible to execute many commands that can either extract information (e.g., with the extraction of attribute tables) or process and display transformed information, in particular using geoprocessing tools. As part of this research, the project only focused on integrating multiple Revit models through the BIM 360 connector and extracting information through web feature layer sharing. Sharing a web feature layer consists of transferring the layer made up of features (which represent the elements of the Revit model in our case study) to the ArcGIS Online support and making this layer accessible to other platforms.



### 3.2.3 Proving Ground Tracer for Revit and Power BI

After gathering geographic information using ArcGIS for Power BI and successfully developing the related visualisations, a type of visualisation was missing that would make it possible to visualise the details of the BIM model. Thus, the research project was interested in finding a way to integrate one or more 3D models from Revit. The ArcGIS for Power BI feature did not offer a 3D representation solution, even though the source file used and imported into Power BI had a 3D representation in ArcGIS. The Tracer extension of Proving Ground has been chosen for the 3D representation solution in Power BI. Tracer proposes an extension for Revit. It allows to create interactive 2D and 3D diagrams of BIM objects with Power BI. From Revit, and using the Tracer extension, the user can export the geometry as well as the schematic data of the objects making up the BIM model.

The export in Database File (.db) format allows to generate one or more data tables with a multitude of information intrinsically linked to the BIM model, such as all of the object family information for each element of the 3D model. The Database File (.db) file is readable in Power BI using SQLite3 as an ODBC data source. This is a specific Structure Query Language (SQL) server suitable for the Tracer extension in Power BI.

## 4 Main Results

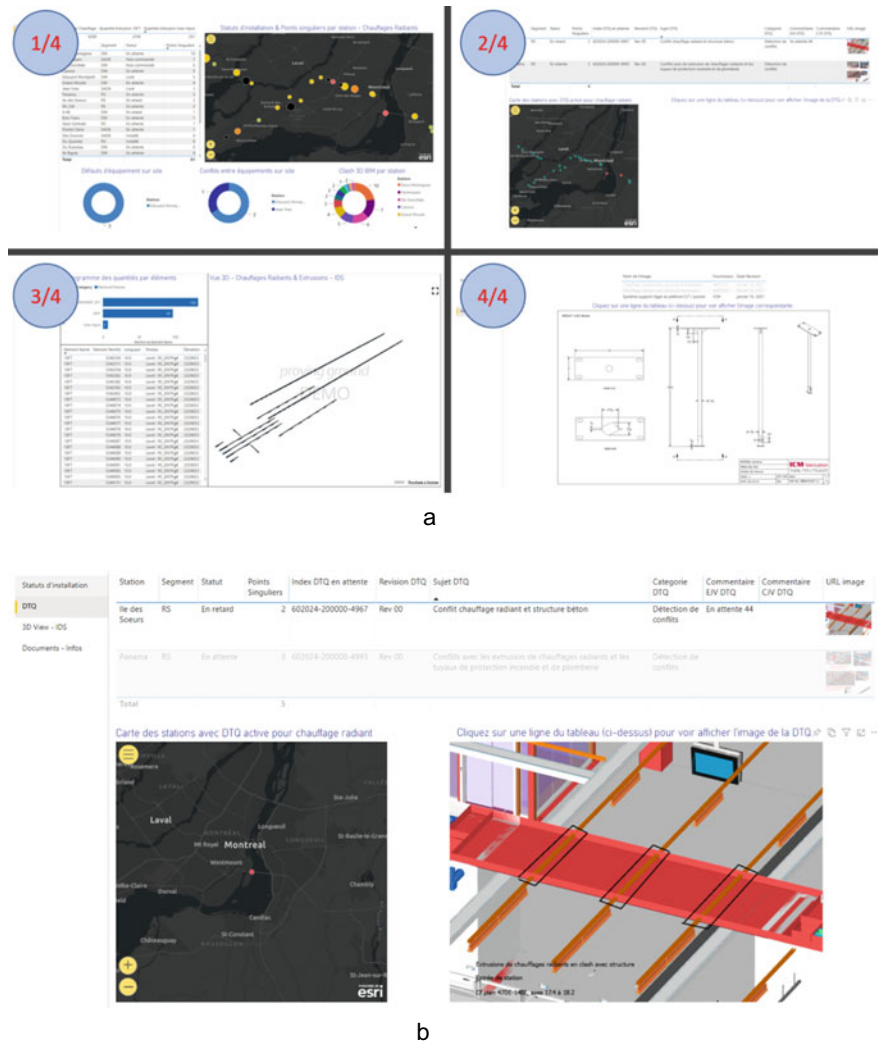
Figure 1 presents an overall overview and screenshots of the developed tool.

Before generating the final visual of the dashboard, we had to design a simulation file for the management of radiant heating equipment. The file was created in Excel and served as the data source for the dashboard in Power BI Desktop.

### 4.1 Development of the Excel Source File

The Excel simulation file, which is one of the dashboard data sources in Power BI, is made up of different sections showing different kinds of information, starting with so-called general information. This information includes the station name, the station's Work Breakdown Structure (WBS), and the segment to which the station belongs. The latitudes and longitudes of each station are also entered in Excel which allows the ArcGIS Maps extension for Power BI to map the locations of the stations on a base map.

Since the dashboard focuses mainly on the representation of radiant heater information, the general information also includes the quantities of heaters and extrusions (10 feet and "user input" for extrusions less than 10 feet) per station.



**Fig. 1** **a** Overview of the dashboard in Power BI Desktop. **b** Screenshot of the dashboard in Power BI Desktop

This research project was also interested in the logistics management of radiant heaters, as the project consortium is responsible for the planning, delivery, and installation of radiant heaters. Thus, five logistics-specific columns were created in Excel. The columns include the heating equipment order dates, expected delivery dates, site reception dates, site installation dates, and finally, the overall status. The global status corresponds to several possible states for a station:

- Not ordered: the station has not yet ordered the heating equipment.

- Pending: the station has placed an order and is awaiting reception on site.
- Delivered: the station has received the equipment, but has not yet installed it.
- Installed: the station has installed the equipment.
- Late: the station is late either on the date of reception or on the date of the installation.

Quality management is an essential component of equipment management in a construction project. For this research project, four quality criteria were adopted:

1. Equipment faults on-site: these may be manufacturing faults attributable to the manufacturer (damaged or incorrectly sized part, for example). This category corresponds to problems identified mainly on-site once the equipment has been delivered.
2. Conflicts between equipment on-site: this category lists the physical conflicts detected on-site between two or more pieces of equipment. The purpose of a BIM coordination is to verify upstream that when equipment is installed on-site, no last-minute conflicts will arise. There are several reasons for potential conflicts between pieces of equipment. For example, it may be due to carelessness on the part of the BIM coordinator when detecting conflicts on the 3D BIM model. It can also be an insufficient level of development (LOD) which does not make it possible to represent all the heating components and therefore to detect all the possible conflicts. It may finally be shifted between the 3D BIM model and the as-built, that is, what was actually built on-site.
3. 3D BIM clash: this category covers all geometric or non-geometric conflicts identified in a BIM coordination software (in the case of the construction project, the consortium uses Autodesk BIM 360 Glue software). It is not necessary to wait for the station construction phase to identify these types of conflicts in the Excel file.
4. Singular points: the term singular points has been used as the sum of equipment faults on-site, conflicts between equipment on-site, and 3D BIM clashes. The summation of these three quantities provides approximate information on the complexity of the operations carried out within the station concerning radiant heaters, even if the station is not yet in the construction phase.

To raise a design or manufacturing problem or initiate a request for information, the consortium has a Design Technical Query (DTQ) system. The purpose of DTQs is to formalise a problem and to redirect it to the person or team affected by the problem. The person affected by the DTQ then has five working days to provide a first solution response to it. It seemed important to include typical DTQ information in the Excel simulation file in the present work, and so the following columns were added:

- Pending index: each DTQ issued has a unique index.
- Revision: indicates the revision number of the DTQ.
- Author: indicates the first and last names of the creator of the DTQ.
- Creation date: indicates the date on which the DTQ was created.
- Elapsed time: indicates the number of days elapsed since the creation of the DTQ.

- Subject: indicates the general subject of the DTQ.
- Category: conflict detection, materials, methods, optimisation, or more details required.
- Engineer's comment: allows engineers to add comments to the DTQ.
- Contractor's comment: allows contractors to add comments to the DTQ.
- Image URL: a direct SharePoint link has been created to retrieve the latest images from the DTQs to illustrate them.

## 4.2 *Building the Power BI File*

Like the breakdown of the Excel simulation file, the dashboard in Power BI is broken down into several pages providing various bits of information. The first is an information page covering installation statuses and special points about the radiant heaters. The page contains six information visualisations:

- A table of the quantities of heaters and extrusions.
- A table of stations with their respective statuses and number of singular points.
- Three donut charts schematically showing the proportions of on-site equipment faults, on-site equipment conflicts, and 3D BIM clash for each station. This decomposition makes it possible to recall that the singular points are the sum of these three quantities.
- A mapping of the infrastructure network using the ArcGIS Maps tool for Power BI. The size of the station points was chosen to be proportional to the number of singular points and to ensure that the colour of the points is a function of the installation status: yellow (pending status), black (non-ordered status), orange (delivered status), green (installed status), and red (overdue status).

When the user hovers their computer mouse over different points on the map, or if they click on the drop-down tab of the ArcGIS Maps for Power BI visual, they can have access to a multitude of information such as the different base maps (synthesised images, satellite images, etc.), the symbology used on the map, the different imported layers, and computer graphics tools.

The second page of the dashboard addresses the topic of DTQ. A table (in the form of a banner) at the top of the page provides essential information on active DTQs for radiant heaters. By clicking on a row of the table, the user automatically sees the image of the active DTQ in.JPG or.PNG format. This is enabled through the use of a Power BI visual that reads URL links that point to a location where an image in.JPG or.PNG format is located. In this case, the consortium mainly uses the Project Document Management Control (PDMC) platform as its database. The purpose of this platform is to centralise the documentation of the construction project. A specific folder for monitoring radiant heater DTQs was created to store images of active and closed DTQs.

To maintain the notion of geographic landmarks on the entire dashboard, a map showing the stations with one or more active DTQ for radiant heaters in red was

produced. In this case, the user can simply click on a red point on the map to instantly know the problem in question at the station concerned.

The third page of the dashboard really introduces the BIM concept of 3D model representation as well as the extraction of the information that a 3D model contains. In the dashboard, the graph on the right side depicts the digital 3D model with the different radiant heating elements (where it is possible to navigate on the 3D view in the same way as on Revit). The visualisation at the left side shows the various information specific to the model such as the element name, element ID, geometric dimensions, level, and elevation. By selecting a bar of the histogram located at the top left of the page, the selected elements can be highlighted in the 3D view and in the Revit data table. It is also possible to focus only on one element of the model, either by clicking on the element in the 3D view or by clicking on the element in the Revit data table.

The Tracer visualisation for Power BI does have several limitations. In particular, it is not possible to integrate several models within the same viewing tab. To represent the entire infrastructure network on the dashboard with the Tracer tool, it would therefore be necessary, for example, to create 26 different pages.

In addition, this research project is interested in GIS-BIM convergence, but unfortunately, the Tracer tool does not allow to integrate a geographic base map, as is the case with ArcGIS Pro. The only visualisation possible with the Tracer tool is that of a single 3D Revit model, on a white background.

Tracer also offers a 2D visualisation tool (two dimensions) on Power BI. Attempts have been made to represent the same as the engineering plans, the arrangement of radiant heaters in the station. Unfortunately, the 2D visual of Tracer does not allow to represent the exact geometry of radiant heating equipment. Indeed, radiant heating equipment can only be represented in 2D by points that seem to represent the centres of the different elements.

The results of the survey show that many potential users find it relevant that the dashboard has a documentary support role, such as PDF extracts, for example. As a result, a fourth and last dashboard page was created to illustrate the various manufacturing plans from subcontractors. The three types of documentation that are offered in the developed dashboard are the workshop plans for radiant heaters, extrusions, and rod supports that attach to the ceiling/beams. These documents are very important for anyone wanting to check certain equipment dimensions or characteristics. The name of the supplier was also entered at the top of these different documents, as was the date of the last revision posted.

## 5 Validation

After the demonstration and handling of the dashboard, the users have been invited to evaluate it. Overall, first impressions were positive that the dashboard appears complex (due to a large number of views per page) while still being accessible and readable. Users found it interesting to see several textual, pictorial, tabular, 3D, and

cartographic representations. Users also underlined the fact that they believed that the use of dashboards, like the one developed in this research project, lent itself well to the different needs they encountered, such as the need to have explicit documentary support during coordination meetings. It was also pointed out that the use of such dashboards would reduce the time spent searching for various bits of information distributed in different databases. These reductions in information search time translate directly into improved work efficiency and improved productivity.

However, some users pointed out that when searching for information on the developed dashboard, they might not be required to use the four pages of the dashboard. Indeed, it is possible that a user, at a given moment, only needs to know certain active DTQ information (represented on the second page of the dashboard), without needing to learn about information of 3D modelling types (represented on the third page of the dashboard).

The main questions from users concerned the GIS component of the dashboard. Users who tried the dashboard found it interesting to add 2D maps to monitor the progress of various issues such as singular points or the distribution of active DTQs of radiant heaters on the infrastructure network. However, they found it unfortunate that for such a study, there were no infographics added to the ArcGIS Maps for Power BI maps. This was not done because our study focuses on equipment management, and therefore it did not make sense to add social- and economic-type reference layers (demography, age, salaries, etc.) in ArcGIS Maps for Power BI. Some users also wondered whether it was possible to add a base map of satellite images (or others) in the Tracer visualisation tool. Unfortunately, this functionality is not allowed by the Tracer tool, which only allows the visualisation of a single model, and only on a white background.

The latter represents an additional drawback posed by using the Tracer visualisation tool for Power BI, which does not appear to be appropriate for infrastructure construction projects. In the case of the construction project under study, there are a total of 26 stations spread over 67 kms across Montreal, and it becomes complicated to present more than 26 dashboard pages just for Power BI's Tracer visualisation tool. Indeed, it would seem that the Tracer visualisation tool for Power BI is more suitable for single-structure construction projects (e.g. in the construction of a single building) or for construction projects where the structures are not too geographically dispersed.

Users also wondered if the dashboard developed in this research project was applicable to equipment management other than radiant heaters. In particular, other general disciplines such as architecture, structure, MEP, and civil engineering were discussed. The answer is that the developed dashboard cannot take into account different pieces of equipment because it can structurally only meet the needs of radiant heaters. To make it multidisciplinary, it would be necessary in particular to review the design of the simulation file created in Excel to generate general table columns, with general criteria, that each discipline can totally or partially fulfil.

## 6 Conclusion

This research project was based on an analysis of the needs, which aimed to probe different profiles of the construction sector on methods used to integrate information into a construction project. Subsequently, it was decided to use the dashboard method to meet the growing needs for information integration. The specificity of the developed dashboard is that it integrates information and visualisations specific to two different information systems, GIS and BIM. The dashboard was developed using an action-research method, including an iterative approach with intermediate validation steps to ensure that the tool developed met the needs initially raised. A final verification phase was set up with the participation of several employees. The purpose of this final phase was to have different users try the dashboard to collect their opinions and opinions about it.

Regarding the limits of the solution proposed in the present research project, it is not possible to transpose the developed dashboard to the management of equipment other than that of the radiant heaters of the construction project. Given that the tool developed was intended to meet specific needs regarding radiant heaters, the structure of the dashboard, as well as the visuals used and the different symbology used, is not fully and perfectly transposable to other management studies.

To allow future improvements to the present research project, it would be interesting to modify and modulate the dashboard developed in Power BI as well as the Excel simulation file to make it possible to generalise the use of the GIS-BIM dashboard to all types of equipment, including equipment used for architectural, structural, civil, MEP, and other disciplines. It would also be interesting to develop a program that makes it possible to represent in 3D visualisations of BIM models on a background of GIS maps Power BI (or another dashboard software). Allowing such features would lighten the density of visualisations in the dashboard and therefore make it more ergonomic and easier to learn for any user.

## References

1. Barlish K, Sullivan K (2012) How to measure the benefits of BIM—a case study approach. *Autom Constr* 24:149–159. <https://doi.org/10.1016/j.autcon.2012.02.008>
2. Borrmann A, Kolbe TH, Donaubaue A, Steuer H, Jubierre JR, Flurl M (2015) Multi-scale geometric-semantic modeling of shield tunnels for GIS and BIM applications. *Comput-Aided Civil Infrastruct Eng* 30(4):263–281. <https://doi.org/10.1111/mice.12090>
3. Botton C, Forgues D (2018) Practices and processes in BIM projects: an exploratory case study. *Adv Civil Eng* 2018:1–12. <https://doi.org/10.1155/2018/7259659>
4. Brath R, Peters M (2004) Dashboard design: why design is important. *DM Rev Online*, 1–4
5. D'Amico F, Calvi A, Schiattarella E, Prete MD, Veraldi V (2020) BIM and GIS Data integration: a novel approach of technical/environmental decision-making process in transport infrastructure design. *Transp Res Procedia* 45:803–810. <https://doi.org/10.1016/j.trpro.2020.02.090>

6. Feng CW, Mustaklem O, Chen YJ (2011) The BIM-based information integration sphere for construction projects. In: Proceedings of the 28th international symposium on automation and robotics in construction, ISARC 2011, pp 156–161. <https://doi.org/10.22260/isarc2011/0025>
7. Forgues D, Botton C, Hittier C (2018) Guide de coordination 3D basée sur des maquettes BIM
8. Karan EP, Irizarry J (2015) Extending BIM interoperability to preconstruction operations using geospatial analyses and semantic web services. *Autom Constr* 53:1–12. <https://doi.org/10.1016/j.autcon.2015.02.012>
9. Sacks R, Eastman C, Lee G, Teicholz P (2018) BIM handbook: a guide to building information modeling for owners, designers, engineers, contractors, and facility managers, 3rd edn. Wiley, Hoboken
10. Sarkar D, Raghavendra HB, Ruparelia M (2015) Role of key performance indicators for evaluating the usage of BIM as tool for facility management of construction projects. *Int J Civ Struct Eng* 5(4):370–378. <https://doi.org/10.6088/ijcser.2014050034>
11. Song K, Pollalis SN, Peña-Mora F (2005) Project dashboard: concurrent visual representation method of project metrics on 3D building models. In: *Computing in civil engineering*, pp 1–12
12. Song Y, Wang X, Tan Y, Wu P, Sutrisna M, Cheng JCP, Hampson K (2017) Trends and opportunities of BIM-GIS integration in the architecture, engineering and construction industry: a review from a spatio-temporal statistical perspective. *ISPRS Int J Geo Inf* 6(12):1–32. <https://doi.org/10.3390/ijgi6120397>



# Overcoming Obstacles in BIM-Based Multidisciplinary Coordination: A Literature Overview



Tabassum Mushtary Meem and Ivanka Iordanova

**Abstract** In spite of the significant amount of research conducted in recent years to improve the efficiency of Building Information Modeling (BIM)-based multidisciplinary coordination processes, unanticipated increases in cost and delays in construction projects still occur. To mitigate factors that hinder the efficiency of construction processes, the literature proposes several solution frameworks. Our work presents an overview of the literature pertaining to these efficiency solution frameworks to identify and address research directions in design conflict resolution. In the existing literature, obstacles having a greater impact on multidisciplinary BIM coordination are sorted into five categories: (i) process, (ii) actor, (iii) task, (iv) context, and (v) team. The implications of each of these categories are considered for separate phases of multidisciplinary coordination. Furthermore, efficiency solution schemes are studied, ranging from shared situational awareness to supervised and hybrid machine learning frameworks. Connections are then drawn between these obstacles, their solutions, and the coordination phases in which they are most applicable. To the best of our knowledge, this is the first work to present a consolidated overview of solution frameworks, and it is our hope that it will help researchers and BIM professionals identify the scope of current research and understand future research directions.

**Keyword** BIM-based multidisciplinary coordination

## 1 Introduction

In recent years, there has been a significant amount of research aimed at making the Building Information Modeling (BIM)-based multidisciplinary coordination process more efficient. However, unanticipated increases in costs and delays in construction projects still occur. Therefore, removing the obstacles for efficient BIM-based

---

T. M. Meem (✉) · I. Iordanova  
École de Technologie Supérieure ÉTS, Montreal, Canada  
e-mail: [tabassum-mushtary.meem.1@ens.etsmtl.ca](mailto:tabassum-mushtary.meem.1@ens.etsmtl.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_39](https://doi.org/10.1007/978-3-031-34593-7_39)

637

multidisciplinary coordination has become an active area of research. To mitigate unforeseen hindrances and expenses in construction processes, the existing literature proposes several solution frameworks to increase the efficiency of multidisciplinary coordination in the industry. This work presents a consolidated overview of these efficient solutions to help researchers and BIM professionals identify research gaps and understand future research directions in the field of BIM-based multidisciplinary coordination by drawing a link between the workflow of multidisciplinary BIM-based coordination, the obstacles affecting it, and possible solutions to these issues.

The objectives of this literature review are: (i) identify the obstacles that affect BIM-based multidisciplinary coordination, (ii) understand the impact these obstacles have over different design coordination phases, (iii) study proposed frameworks for efficient multidisciplinary coordination, and (iv) draw connections between these obstacles, their solutions, and the coordination phases for which they are most applicable.

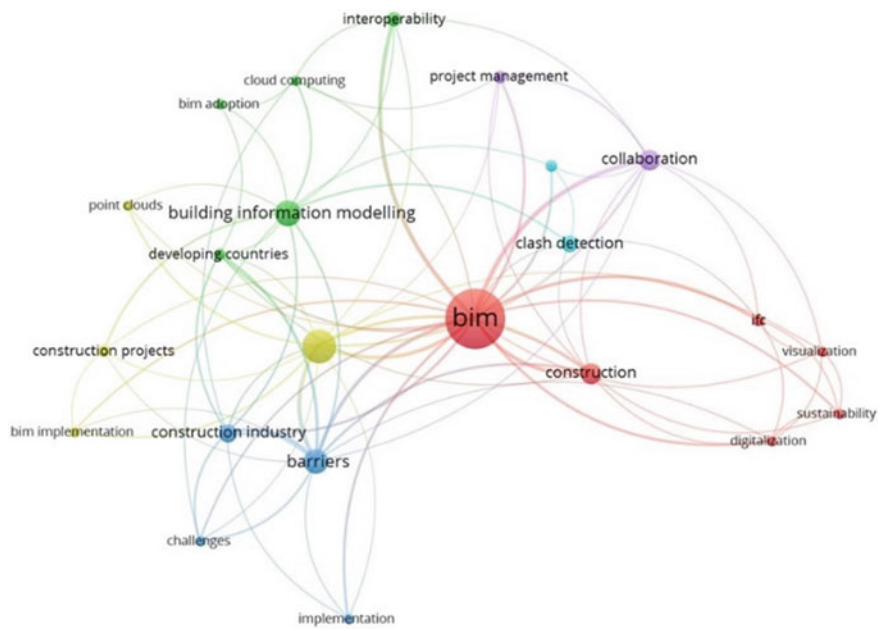
In this study, the terms ‘obstacles’ and ‘barriers’ are used interchangeably. The paper is organized as follows: Sect. 1 presents the literature review methodology, recent literature on factors influencing BIM-based collaboration, and the barriers and bottlenecks inhibiting meaningful multidisciplinary coordination. Different obstacle categories emerging from the studied literature are outlined. These categories are: (i) process, (ii) actor, (iii) task, (iv) context, and (v) team. Sections 2, 3, 4, 5, and 6 discuss the identified categories and the solution schemes that appear in the literature with the potential ways to address these obstacles. The findings from the literature are systemized, and connections are drawn between the obstacles, the coordination phases they affect, and their proposed solutions.

To avoid any potential bias, a ‘mixed-methods systematic review’ was employed which applies quantitative and qualitative methods to thoroughly analyze the available literature. The mixed-methods systematic review for this work was carried out in three stages as illustrated in Table 1. The first two stages of the study involved a keyword search and bibliometric analysis. Stage 1 consisted of a focused keyword search in the Scopus database which produced 1077 pieces of literature. The applied keywords were selected based on their relevance to the topic of study, including Building Information Modeling (BIM), multidisciplinary coordination, real-time collaboration, barriers or obstacles in BIM, clash management, clash relevance, clash filtration, and clash resolution. This approach ensures the comprehensiveness of the investigation by locating associated concepts and word variants (Fig. 1).

In the second stage, 357 pieces of literature were shortlisted based on the publication stage, document type, and source type. For this analysis, English-language published, open-access journal articles, conference papers, and book chapters were selected. Afterward, VOSviewer was used to conduct the bibliometric analysis of the shortlisted literary works. A co-occurrence network map for the author keywords was created to view the topics of research and their interconnection in the field of BIM-based multidisciplinary coordination as shown in Fig. 2. VOSviewer network map uses node circles and lines to illustrate the strength and significance of the interconnection. The size of the circle signifies the frequency of the keyword appearance, whereas the distance between the nodes indicates the level of closeness. The line

**Table 1** Adopted mixed-method systematic literature review procedure

Stages	Evaluation criteria	Number of collected literature
Stage 1	Keyword search based on relevance to the field of study	1077
Stage 2	Published, open access, journal articles, conference papers, book chapters, written in English	357
Stage 3	Studies focused on multidisciplinary collaboration, coordination, and efficiency frameworks	32



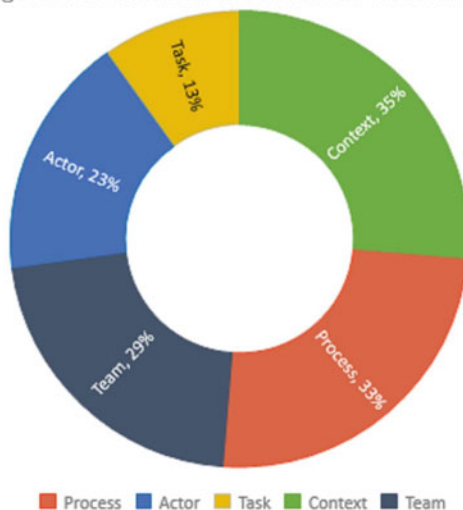
**Fig. 1** Most prominent areas of research shown via author keyword co-occurrence network map

thickness signifies the relation between two keywords and their frequency of co-occurrence. The network map highlights the most frequent topics of research related to multidisciplinary coordination and simultaneously provides insights into the scope of potential research directions. Furthermore, the map shows six distinct groups of keywords appearing as clusters of different colors. BIM in sustainable construction, barriers in BIM, and BIM-based collaboration stand out among these groups.

It can be observed from the map that there has been significant research focused on identifying the barriers or obstacles in BIM-based collaboration and multidisciplinary coordination. However, there is a lack of comprehensive study pertaining to mitigating these obstacles. Qualitative analysis was adopted for the final stage where the relevance of the literature is determined by comparing the contents of all shortlisted works. First, the literary works that directly address the obstacles in

**Fig. 2** Categories of obstacles in BIM-based multidisciplinary collaboration based on [27]

Categories of obstacles in BIM-based collaboration.



BIM-based coordination and collaboration were selected. These shortlisted papers were analyzed carefully to identify and extract information for the next step. Subsequently, they were coded based on the specific BIM coordination or collaboration obstacle they studied to further understand the true extent of investigation done in this field and to identify potential solution frameworks for the said obstacles. This step also helped to recognize any research gaps. The total number of papers shortlisted in Stage 3 of the mixed-methods systematic review was 32.

### 1.1 Categorization of Obstacles

The most influential obstacles in BIM-based multidisciplinary coordination in light of multiple scientific publications are discussed in this section. BIM is regarded as integrative technology with ‘parametric intelligence’ able to escalate efficiency in the construction industry [32]. Investigations show that in recent years BIM-enabled construction coordination is being increasingly adopted in construction projects to drive productivity [18]. However, managing a fruitful BIM-based collaboration network consisting of multidisciplinary participants is challenging despite this being touted as a necessity in BIM-enabled projects [22]. A significant amount of research is aimed at making the BIM-based multidisciplinary coordination process more efficient. Yet the approach toward investigating the obstacles in construction collaboration has been somewhat isolated. This research gap was addressed in recent years, and several researchers recognized that an integrated analysis of the fragmented studies is paramount to identifying the most important challenges and barriers.

One of the contemporary studies that address this research gap is a publication by Oraee et al. [25], which presents a conceptual model capturing the most important challenges to collaboration in BIM-based construction networks. Here, the authors perceive BIM-based construction networks as geographically dispersed stakeholders from various disciplines and organizations performing project tasks. This study adopts the categories from [27], which present the interrelation between attributes exerting influence on BIM-based collaboration networks. This typology was also used as a theoretical lens in [25], where a total of 173 articles on BIM-enabled projects, design coordination, etc., were analyzed to shortlist 26 obstacles under five broad categories that negatively affect BIM-based collaboration networks. These categories are: (i) process, (ii) actor, (iii) task, (iv) context, and (v) team. The authors acknowledge that the obstacles under different categories might be interrelated as well.

### **1.1.1 Process-Related Obstacles**

According to [27], process-related obstacles in multidisciplinary collaboration networks appear in 66% of the sample articles studied. The authors of [25] mention that obstacles related to necessary resources, essential tools, and professional training for BIM-based collaboration are part of this category. Both articles indicate that obstacles regarding relevant technologies, collaboration space, data security, and proper guidelines are the most impactful according to the literature. The articles sub-categorize these obstacles into ‘Tools’ and ‘Resources’, which will be discussed further in this study.

### **1.1.2 Actor-Related Obstacles**

Even though it is relatively underexplored in the context of BIM-based collaboration, actor-related obstacles in multidisciplinary collaboration networks were the focus of 23% of the sample articles studied by [27]. In [25], the authors explain that obstacles related to individual participants such as ‘Knowledge, skills, and ability’ of team members fall within this category. The significance of an individual team member’s skills and abilities was also identified as one of the key attributes influencing social collaboration among BIM actors [32].

### **1.1.3 Task-Related Obstacles**

The demand for the right information at the right time falls under this category along with the task demand, and the task category is acknowledged by 13% of the investigated sample articles [25, 27]. The demand for information will be further discussed in the following sections.

### **1.1.4 Context-Related Obstacles**

Factors related to organizational culture and environment, such as a lack of teamwork mindset, and substantial communication occurring outside BIM, fall under this category of context-related obstacles as highlighted by 35% of the total sample studies [25, 27]. In [26], the authors add a lack of willingness to enforce BIM standards and promote specific BIM collaboration tools as obstacles under the context category. Frameworks that can mitigate the issue of communication occurring outside BIM will be further discussed in this study.

### **1.1.5 Team-Related Obstacles**

As one of the better-investigated categories alongside process and context, the team category is addressed by 29% of the sample articles and includes the obstacles related to the roles and relationships between the team members within this category [25, 27]. The authors also add the factors of BIM managers not having total control and having an unclear understanding of their role as the main obstacles in this category [26].

## **2 Potential Solution Frameworks for Process-Related Obstacles**

This study will address the resources and tools-related obstacles under the process category along with their solution frameworks as, based on the literature, these are the most impactful obstacles.

### **2.1 Resources**

According to the conceptual model presented by the authors of [25], the obstacles related to resources are: (i) lack of a common data environment (CDE) to support collaboration, (ii) lack of guidelines and standards for BIM collaboration, and (iii) data ownership and data privacy concerns.

#### **2.1.1 Potential Framework for Improving CDE to Support Collaboration**

The literature identifies the lack of a CDE able to ensure smooth and efficient collaboration as one of the obstacles. To mitigate this, the smooth and robust integration

of CDE access into BIM authoring tools must be ensured. In [31], a BIM integration framework (BIF) is proposed that provides a steady connection to a central online data platform for each team member regardless of the BIM authoring tools being used by the participants. Instead of transforming the model in IFC format multiple times, the participants can upload files in the native format and communicate and manage clashes via small data packages referred to as topics. This process saves valuable resources such as time, and computing power. Thus, this type of framework could be a potential solution in a BIM-based collaboration.

### **2.1.2 Potential Frameworks for Improving Guidelines and Standards for BIM Collaboration**

Researchers have voiced concerns over the inconsistency of guidelines and standards in BIM-enabled projects and discussed the importance of clear standards for BIM-based multidisciplinary coordination [16, 25]. Khanzode [16] noted the necessity that all disciplines working side by side be fully aware of each other's activities. They also state that the coordination model should include detailed facility functions, as specified by users, to prevent potential reworks. Additionally, the level of development (LOD) of the model needs to be determined before starting the coordination process and receiving the input of all participants. These lessons learned are similar to the industry-developed BIM execution planning guidelines that have been widely followed in recent years, such as the PennState BIM Project Execution Planning Guide and ISO 19650 guidance. Such guidelines can certainly mitigate the process problem, i.e., lack of clear standards.

### **2.1.3 Potential Framework for Improving Data Ownership and Data Privacy Concerns**

Companies and participants often have privacy and security-related concerns over sharing their models on common data environment (CDE) or any cloud-based collaboration platform despite these platforms' tremendous potential to facilitate collaboration [25, 27]. Controlled information exchange and access during such collaboration could resolve this issue. Akponeware and Adamu [3] propose an open work-in-progress framework for greater transparency and acknowledge the need for user-controlled access rights to restrict unauthorized access and promote data security. They also encourage the involvement of a built asset security manager to grant and control such role-based access rights on a need-to-know basis for all relevant participants.

## 2.2 Tools

A significant portion of the tools obstacles is related to failures in technological support [25]. Most of the widely used technologies do not support the functions necessary for efficient BIM collaboration. This sub-section will discuss individual obstacles and some proposed frameworks that could potentially improve the current tools used for BIM-based collaboration.

### 2.2.1 Potential Frameworks for Clash Relevance Prediction

Numerous irrelevant clashes found during the BIM coordination stage increase the delay and cost of construction projects. However, current clash detection tools lack the necessary automation to analyze the clash data, and manual investigation is still needed [29]. Thus, introducing clash relevance prediction in BIM coordination tools has been an active area of research in recent years. Machine learning, more specifically supervised and unsupervised learning, deep learning, and reinforcement learning, can be helpful in clash management [28].

A notable investigation to improve the clash filtration process was conducted by Hu and Castro-Lacouture [13], in which six kinds of supervised machine learning algorithms, such as decision trees, Jrip rules, binary logistic regression, and Bayesian methods, were compared to identify which was the most efficient for clash filtration. This investigation found that Jrip-based rule methods with a prediction accuracy of 80% outperformed the other algorithms. In a similar study, the application of Bayesian statistics, such as naïve Bayesian, the Bayesian network, and Bayesian probit regression for clash relevance prediction, were investigated [12]. The authors identified that the Naïve Bayesian method has average precision, the Bayesian network method is more reliable in predicting irrelevant clashes, and the Bayesian probit regression, which can work with small datasets, has the highest precision when predicting relevant clashes. However, a combination of Bayesian methods through majority rule is more reliable and shows some improvements in the precision of clash classification. Furthermore, clash filtration using a hybrid method of rule-based reasoning and supervised machine learning was attempted which showed that the hybrid method can increase the prediction accuracy by 6–17% compared to individual or ensemble learning classifiers [17].

### 2.2.2 Holistic Clash Detection and Resolution Improvement Frameworks

Design coordination and clash detection are the two most impactful factors influencing design error [33]. According to [14, 15], building components are connected, and they create a network of interdependency which influences the clash impact. These studies represented this relationship via a building component network



centered on hard clash objects which were tested on a real project and showed that irrelevant clashes are reduced by 17%. This network method was also able to group relevant clashes reducing the number of detected clashes significantly. The authors further explored the dependent relationship between building components and their effect on clash resolution by analyzing six types of spatial relations and designed algorithms to query these relations using models in industry foundation classes (IFC) format [14]. After formulating a spatial network for building components based on these relations, the authors tested this method on a real project and identified the use of component-dependent network and graph theory to support clash resolutions and find the globally optimized solutions as the future direction of research.

### **2.2.3 Proposed Framework for Automatic Clash Resolution**

Automatic clash resolution has been investigated in recent years. The dependent nature of building components when investigating clash resolution was investigated by Hu et al. [14, 15], and the individual algorithm's efficacy in automatic clash resolution was investigated by several researchers. One such framework employed a simulated annealing algorithm based on the application programming interface (API) of the BIM authoring tool and proposed a layout modification of the components involved in a clash to reduce the number of detected clashes [11]. Apart from approaches to resolve MEP clashes, contemporary researchers are investigating methods to resolve clashes in steel reinforcement. A two-step genetic algorithm (GA)-based automatic optimization of clash-free reinforcement design was attempted [21]. Applicability of multi-agent reinforcement learning (MARL) [19] and Q-learning in BIM resulting in realistic path-planning for automatic clash-free rebar design have been studied [20].

### **2.2.4 Functional Requirements in BIM Authoring Tools**

To provide appropriate support for BIM-based collaboration, BIM authoring tools need to have some specific functionalities that promote collaboration. In [23], an analysis was conducted of the bottlenecks of BIM-based design coordination and benchmarked popular BIM tools based on their functionalities that ranged from supporting multiple model formats to commenting on the model development. This study pointed out that no currently popular BIM authoring tools support multiple model formats or enable the recording of changes. Solibri emerged as the most compatible BIM tool as this tool supports maximum necessary functionalities.

### **3 Potential Solution Frameworks for Actor-Related Obstacles**

The team members' 'insufficient collaboration knowledge, skills, and abilities' fall under the actor category, which, of all the categories, is the least explored in current literature. This insufficiency can range from a lack of skills regarding BIM authoring tools or coordination platforms to a lack of knowledge about coordination issues faced by team members.

In [24], an attempt to define a taxonomy of design coordination issues was undertaken that would assist team members in gaining a better understanding of such issues as validated by industry professionals. Another important obstacle in mitigating the inadequacy of collaboration knowledge, skills, and abilities is improper documentation of lessons learned from each BIM-based collaboration project. Maintaining an accurate rework log and a log of the VDC team's lessons learned can reduce the chances of missing field-detected issues, as this type of record will provide project participants with the knowledge required to recognize and avoid similar issues in the future projects [4]. Moreover, the literature emphasizes the necessity of skill-developing programs, such as intensive trainings or workshops where experienced professionals can share their expertise to increase the level of understanding and knowledge about BIM [8].

### **4 Potential Solution Frameworks for Task-Related Obstacles**

According to [25], task barriers have the least impact on BIM-based collaboration networks. The subfactor 'demand' within the task category identifies the absence of the right information at the right time. This is a significant barrier to collaboration and results from a low quality of communication [9]. Low quality of communication is the main cause of why team members are unable to acquire the right information at the right time. Thus, introducing ways to promote situational awareness and better communication within the team might help to mitigate this barrier. In [1], a quasi-experiment was conducted with multidisciplinary professionals, and it redefined social BIM (SBIM) and shared situational awareness among remote project participants. This experiment showed that real-time audio-visual collaboration improves communication in BIM-based collaborations.

## **5 Potential Solution Frameworks for Context-Related Obstacles**

The authors of [25] stated that factors related to organizational culture and environment fall under the context category. Specifically, the lack of team working mentality and substantial communication occurring outside BIM had a significant impact on collaboration.

Informal communication outside BIM is a problem that impacts the proper documentation of lessons learned from each project. To prevent team members from using informal communication channels frameworks must be used that enable the members to contact each other over the BIM collaboration platform. The social BIM framework proposed in [1] can prove to be helpful in this case. This study tested four distinct types of collaboration protocols: (i) one-to-one, (ii) one-to-many, (iii) many-to-one, and (iv) many-to-many. The many-to-many protocol offered the maximum amount of shared situational awareness within the team and enabled members to communicate with each other over the BIM platform via text, audio, and video. This protocol has the potential to enable members to avoid informal communication as identified by the participants.

In addition to using such frameworks to promote better formal communication within the team, researchers recommended applying Lean construction strategies to the BIM coordination and identified some steps where Lean problem-solving techniques could be helpful for BIM [30]. The most notable recommendations from this study ranged from flow management and systematic waste analysis during the clash detection phase to facilitating the exchange of lessons learned between projects.

## **6 Potential Solution Frameworks for Team-Related Obstacles**

The team category is one of the better-investigated obstacle categories, along with process and context. Issues related to the roles of BIM collaboration participants and relationships between the team members are included within the team category [25, 27]. The isolated working mentality of team members in BIM-based collaboration in addition to resistance to sharing information is one of this category's most impactful obstacles [25].

## ***6.1 Potential Frameworks for Defining BIM-Specific Roles***

In a regular non-BIM project, the project manager, design manager, and site manager are in charge of coordination. However, in BIM-enabled projects, the BIM coordinator occupies a more strategic position, and this deviates from the traditional multidisciplinary collaboration dynamics. In [5], the authors corroborated this and stated that the level of centrality displayed by the new BIM-based roles such as BIM coordinator or BIM manager is moderate in BIM-enabled projects. They further added that the competition of BIM roles with that of the project managers in terms of leadership can result in ineffective collaboration. Furthermore, non-BIM actors perceive BIM actor roles as focusing more on technical skills than softer skills, while BIM actors believe that their roles drive change by coordination [7]. This implies that even though BIM roles are accepted in the industry, a clear understanding of these roles is still missing.

The impact of BIM-specific roles and their changing definitions to accommodate the needs of the changing technologies and industry has also been explored. A quantitative analysis of the impact and viability of BIM-related jobs and their scope of work showed that BIM-specific jobs supplement the lack of BIM expertise in the traditional project manager. As contemporary project managers absorb more BIM-specific skills, the independent BIM-specific roles will more likely disappear or merge with the project manager's role [10]. Additionally, in Bosch-Sijtsema and Gluch [6], the roles and actions of BIM actors in addition to the changes or disruptions these cause to traditional construction practices were discussed. The investigation stated that BIM-specific actors are challenging and changing the construction standards constantly. These two studies showed clearly that BIM roles and their definitions will change rapidly with time and that the construction industry structure will correspondingly change.

## ***6.2 Potential Framework for Improving the Relationship Between Team Members***

Obstacles pertaining to the relationship between team members, such as lack of trust between team members, are one of the most important obstacles under the team category [25]. The habit of isolated working, which is also known as 'silos' within a collaborative team, is still very much prevalent in the construction industry. It usually stems from project members being reluctant to share their work with other disciplines in the early preliminary stage, and this leads to significantly inefficient BIM-based collaboration [3]. As a potential solution to this issue, researchers proposed an open work-in-progress (OWIP) stage in the common data environment instead of the traditional work-in-progress phase [3]. In the OWIP, all disciplines participating in

the collaboration can have a secure access to the design and provide feedback. Thus, the added security in this proposed framework reduces the designer’s reluctance to share the design with team members, which assists with clash avoidance and promotes efficiency in the coordination phase of the BIM collaboration.

7 Conclusions

This review summarizes current literature that discusses the most impactful obstacles in BIM-based coordination and the proposed BIM efficiency frameworks that can potentially mitigate these obstacles. Furthermore, this work identifies the stage of the BIM-based design coordination most affected by each obstacle category.

This section presents a consolidated view of our findings by connecting the main obstacles identified in the literature with prospective efficiency frameworks. The main steps and workflow in the BIM-based design coordination process in light of the literature are then discussed. Finally, the applicability of the discussed obstacles and appropriate solution frameworks in the steps of BIM-based design coordination is presented.

Figure 3 shows the five obstacle categories: process, actor, task, context, and team and the connections between each obstacle category and the corresponding current BIM efficiency frameworks.

In Fig. 4, the general key steps and workflow of BIM-based design coordination are shown. The diagram is created based on the works of [2, 23]. The key steps of design coordination as identified in the literature are: (i) define coordination strategy, (ii) generate specialty models, (iii) prepare federated model, (iv) perform interference check, (v) analyze detected issues, (vi) publish federated model, (vii) organize coordination meetings, (viii) resolve detected issues, and (ix) update and share 3D

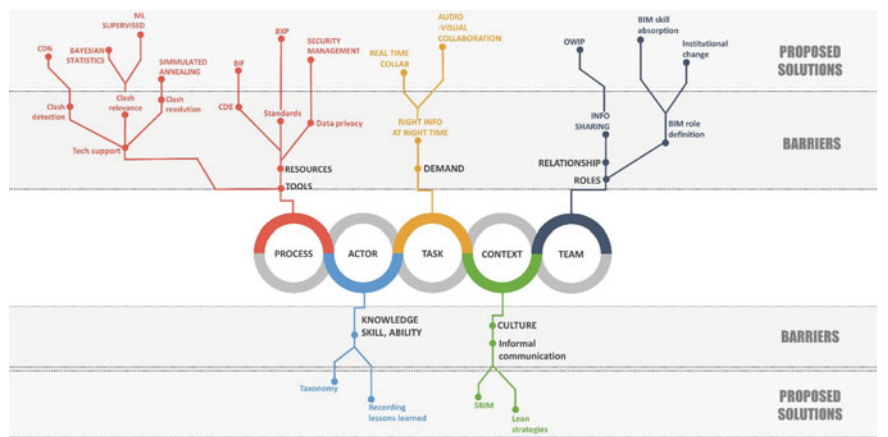
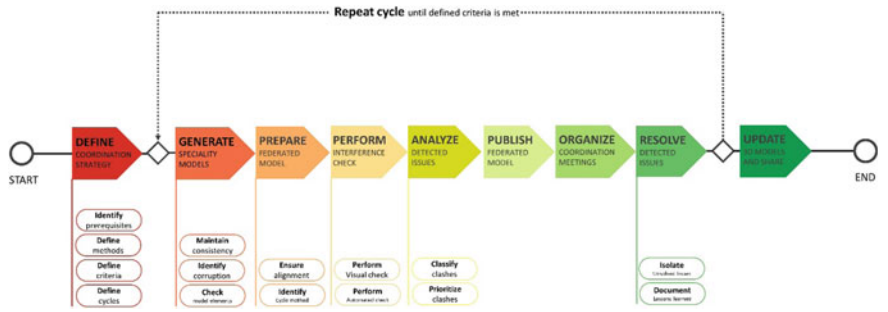


Fig. 3 Obstacles in BIM-based multidisciplinary collaboration. Figure created by the Author

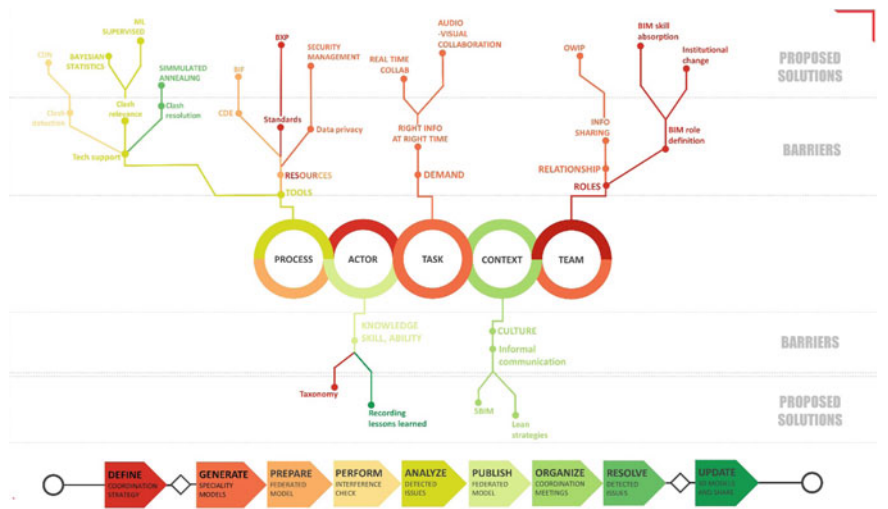


**Fig. 4** Key activities and workflow of BIM-based design coordination. Based on [2, 23]

model. Generally, after the design coordination meetings, the team members from different disciplines agree on how to resolve the detected clashes.

After this stage, the models are edited based on the team’s decision, and this process is repeated until the criteria are satisfied. Based on the initial coordination strategy and BIM execution plan, the key activities of the design coordination phase can change.

In Fig. 5, obstacle categories and the proposed solution frameworks are color-coded according to the different timeline stages of the BIM design coordination phase they affect the most. This diagram visually highlights how the first five stages of BIM-based design coordination are most affected by the obstacles. Process obstacles predominantly affect coordination strategy, model generation, and clash detection stage.



**Fig. 5** Interrelation between BIM collaboration obstacles. Figure created by the Author

Table 2 represents the connection between obstacle categories and the proposed solution frameworks organized by the stages of design coordination they apply to. Actor obstacles affect the very first and last stages of design coordination. Task obstacles mainly have an impact on the individual model generation stage. Context obstacles impact the later stages of coordination when communication between team members is of utmost importance. Finally, team obstacles mostly affect coordination strategy and the individual model generation phase which are also the first steps of the coordination phase. Therefore, it can be stated that the BIM collaboration obstacles that various studies identify mostly affect the initial stages of BIM-based design coordination. If, in the first stages specifically, specialty model generation and federated model creation phases are affected by obstacles, they can hinder the efficiency of coordination as well as the other phases of the construction project.

Based on the analyzed literature, the first two stages of design coordination, which are the defining of coordination strategy and generating of specialty models, have emerged as the stages most affected by the identified obstacles. Team, task, and process are the most impactful obstacles among the five categories. Additionally, these two stages of design coordination have the maximum impact on the efficiency of a BIM-enabled project. Thus, a thorough investigation is needed to resolve the obstacles affecting the first two stages of design coordination. Such investigation will

**Table 2** Interrelation between BIM collaboration obstacles and stages of BIM-based design coordination

Timeline stage of design coordination	Potential solution strategies	Obstacle subcategory	Obstacle category
Define coordination strategy	BIM execution plan for standard maintenance	Resources	Process
	Taxonomy for a better understanding of coordination issues	Knowledge, skills, ability	Actor
	BIM skill absorption by project manager	Roles	Team
	Institutional change to support innovative technology	Roles	Team
Generate specialty models	Security management and role-based access for data privacy	Resources	Process
	Real-time collaboration	Demanding the right information at the right time	Task
	Audio-visual collaboration	Demanding the right information at the right time	Task
	Open work-in-progress (OWIP) to encourage information sharing	Relationship	Team

(continued)

**Table 2** (continued)

Timeline stage of design coordination	Potential solution strategies	Obstacle subcategory	Obstacle category
Prepare federated model	BIM integration framework (BIF) for improved CDE	Resources	Process
Perform interference check	Component-dependent network (CDN) for clash detection	Tools	Process
Analyze detected issues Publish federated model	Supervised machine learning for clash relevance prediction	Tools	Process
	Bayesian statistics for clash relevance prediction	Tools	Process
Publish federated model	Recording lessons learned	Knowledge, skills, ability	Actor
Organize coordination meetings	Social BIM	Informal communication outside the BIM platform	Context
	Applying Lean strategies to BIM process	Informal communication outside the BIM platform	Context
Resolve detected issues	Simulated annealing algorithm to resolve clashes	Tools	Process
Update and share 3D models	Recording lessons learned	Knowledge, skills, ability	Actor

assist the multidisciplinary BIM collaboration team members to avoid coordination conflicts going forward.

**Acknowledgements** The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for its financial support through its CRD program 543867-2019 as well as the industrial partners of the ÉTS Industrial Chair on the Integration of Digital Technology in Construction.

## References

1. Adamu ZA, Emmitt S, Soetanto R (2015) Social BIM: co-creation with shared situational awareness. *J Inf Technol Construct* 20:230–252
2. 351in Model Use Templates Guide (n.d.) BIME initiative. Retrieved January 1, 2022, from <https://bimexcellence.org/resources/300series/351in-model-use-templates-guide/>
3. Akponeware A, Adamu Z (2017) Clash detection or clash avoidance? An investigation into coordination problems in 3D BIM. *Buildings* 7(3):75. <https://doi.org/10.3390/buildings7030075>



4. Alsuhaibani AK (2021) Investigating the causes of missing field fixed issues from BIM-based construction coordination through semi-structured interviews [Application/pdf]. <https://doi.org/10.26153/TSW/14472>
5. Badi S, Diamantidou D (2017) A social network perspective of building information modelling in Greek construction projects. *Archit Eng Des Manage* 13(6):406–422. <https://doi.org/10.1080/17452007.2017.1307167>
6. Bosch-Sijtsema P, Gluch P (2021) Challenging construction project management institutions: the role and agency of BIM actors. *Int J Constr Manag* 21(11):1077–1087. <https://doi.org/10.1080/15623599.2019.1602585>
7. Bosch-Sijtsema PM, Gluch P, Sezer AA (2019) Professional development of the BIM actor role. *Autom Constr* 97:44–51. <https://doi.org/10.1016/j.autcon.2018.10.024>
8. Evans M, Farrell P (2020) Barriers to integrating building information modelling (BIM) and lean construction practices on construction mega-projects: a Delphi study. *Benchmarking Int J* 28(2):652–669. <https://doi.org/10.1108/BIJ-04-2020-0169>
9. Hosseini MR, Banihashemi S, Chileshe N, Namzadi MO, Udaea C, Rameezdeen R, McCuen T (2016) BIM adoption within Australian small and medium-sized enterprises (SMEs): an innovation diffusion model. *Construct Econ Build* 16(3):71–86. <https://doi.org/10.5130/AJCEB.v16i3.5159>
10. Hosseini MR, Martek I, Papadonikolaki E, Sheikhhoshkar M, Banihashemi S, Arashpour M (2018) Viability of the BIM manager enduring as a distinct role: association rule mining of job advertisements. *J Constr Eng Manag* 144(9):04018085. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001542](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001542)
11. Hsu H-C, Wu I-C (2019) Employing simulated annealing algorithms to automatically resolve MEP clashes in building information modeling models. In: 36th international symposium on automation and robotics in construction, Banff, AB, Canada. <https://doi.org/10.22260/ISARC2019/0106>
12. Hu Y, Castro-Lacouture D (2018) Clash relevance prediction in BIM-based design coordination using Bayesian statistics. *Construct Res Cong* 2018:649–658. <https://doi.org/10.1061/9780784481271.063>
13. Hu Y, Castro-Lacouture D (2019) Clash relevance prediction based on machine learning. *J Comput Civ Eng* 33(2):04018060. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000810](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000810)
14. Hu Y, Castro-Lacouture D, Eastman CM (2019) Holistic clash resolution improvement using spatial networks. *Comput Civil Eng* 2019:473–481. <https://doi.org/10.1061/9780784482421.060>
15. Hu Y, Castro-Lacouture D, Eastman CM (2019) Holistic clash detection improvement using a component dependent network in BIM projects. *Autom Constr* 105:102832. <https://doi.org/10.1016/j.autcon.2019.102832>
16. Khanzode A (n.d.) Benefits and lessons learned of implementing building virtual design and construction (VDC) technologies for coordination of mechanical, electrical, and plumbing (MEP) systems on a large healthcare project, p 19
17. Lin WY, Huang Y-H (2019) Filtering of irrelevant clashes detected by BIM software using a hybrid method of rule-based reasoning and supervised machine learning. *Appl Sci* 9(24):5324. <https://doi.org/10.3390/app9245324>
18. Liu H, Skibniewski MJ, Ju Q, Li J, Jiang H (2021) BIM-enabled construction innovation through collaboration: a mixed-methods systematic review. *Eng Constr Archit Manag* 28(6):1541–1560. <https://doi.org/10.1108/ECAM-03-2020-0181>
19. Liu J, Liu P, Feng L, Wu W, Lan H (2019) Automated clash resolution of rebar design in RC joints using multi-agent reinforcement learning and BIM. In: 36th international symposium on automation and robotics in construction, Banff, AB, Canada. <https://doi.org/10.22260/ISA RC2019/0123>
20. Liu J, Liu P, Feng L, Wu W, Li D, Chen YF (2020) Automated clash resolution for reinforcement steel design in concrete frames via Q-learning and building information modeling. *Autom Constr* 112:103062. <https://doi.org/10.1016/j.autcon.2019.103062>

21. Mangal M, Li M, Gan VJL, Cheng JCP (2021) Automated clash-free optimization of steel reinforcement in RC frame structures using building information modeling and two-stage genetic algorithm. *Autom Constr* 126:103676. <https://doi.org/10.1016/j.autcon.2021.103676>
22. Matthews J, Love PED, Mewburn J, Stobaus C, Ramanayaka C (2018) Building information modelling in construction: insights from collaboration and change management perspectives. *Prod Plann Control* 29(3):202–216. <https://doi.org/10.1080/09537287.2017.1407005>
23. Mehrbod S, Staub-French S, Bai Y (n.d.) Analysis of bottlenecks in BIM-based building design coordination process and benchmarking state of the art BIM tools, p 10
24. Mehrbod S, Staub-French S, Mahyar N, Tory M (n.d.) Beyond the clash: investigating BIM-based building design coordination issue representation and resolution, p 25
25. Oraee M, Hosseini MR, Edwards DJ, Li H, Papadonikolaki E, Cao D (2019) Collaboration barriers in BIM-based construction networks: a conceptual model. *Int J Project Manage* 37(6):839–854. <https://doi.org/10.1016/j.ijproman.2019.05.004>
26. Oraee M, Hosseini MR, Edwards D, Papadonikolaki E (2021) Collaboration in BIM-based construction networks: a qualitative model of influential factors. *Engineering, construction and architectural management*, ahead-of-print (ahead-of-print). <https://doi.org/10.1108/ECAM-10-2020-0865>
27. Oraee M, Hosseini MR, Papadonikolaki E, Palliyaguru R, Arashpour M (2017) Collaboration in BIM-based construction networks: a bibliometric-qualitative literature review. *Int J Project Manage* 35(7):1288–1301. <https://doi.org/10.1016/j.ijproman.2017.07.001>
28. Pan Y, Zhang L (2021) Roles of artificial intelligence in construction engineering and management: a critical review and future trends. *Autom Constr* 122:103517. <https://doi.org/10.1016/j.autcon.2020.103517>
29. Pärn EA, Edwards DJ, Sing MCP (2018) Origins and probabilities of MEP and structural design clashes within a federated BIM model. *Autom Constr* 85:209–219. <https://doi.org/10.1016/j.autcon.2017.09.010>
30. Pedo B, Tezel A, Koskela L, Whitelock-Wainwright A, Lenagan D, Nguyen QA (2021) Lean contributions to BIM processes: the case of clash management in highways design, pp 116–125. <https://doi.org/10.24928/2021/0164>
31. Preidel C, Borrmann A, Oberender C, Tretheway M (n.d.) Seamless integration of common data environment access into BIM authoring applications: the BIM integration framework, p 9
32. Raja Mohd Noor RNH, Che Ibrahim CKI, Belayutham S (2021) The nexus of key attributes influencing the social collaboration among BIM actors: a review of construction literature. *Int J Construct Manage* 1–11. <https://doi.org/10.1080/15623599.2021.1946902>
33. Wong JKW, Zhou JX, Chan APC (2018) Exploring the linkages between the adoption of BIM and design error reduction. *Int J Sustain Dev Plan* 13(01):108–120. <https://doi.org/10.2495/SDP-V13-N1-108-120>

# Adopting Ecolabels in the Construction Industry via Blockchain



Dilusha Kankanamge and Rajeev Ruparathna

**Abstract** Green procurement has been a growing area of interest in public sector purchasing. The above has sparked an interest in ecolabels that provide verified environmental performance information. Green procurement and ecolabels are less observed in the construction sector due to data challenges, transparency, and implementation resources. Blockchain is a promising technology that can aid in ecolabel data verification. Combined with Building Information Modeling (BIM), blockchain can be used to ensure the authenticity of construction project data and the validity of sustainability claims. A comprehensive literature review has revealed that no previous research has used BIM nor blockchain for green procurement in the construction sector. This paper proposes a methodological framework for integrating entrusted ecolabels through blockchain for the green procurement of water supply infrastructure. BIM was used as the platform for integrating project data. The outcomes of this study will improve the transparency of green procurement while promoting cutting-edge technology in the construction sector. More importantly, the outcomes of this research support addressing data credibility and accuracy in eco-conscious decision-making.

**Keywords** Ecolabels · Environmental product declarations · Building information modeling · Blockchain · Sustainability · Procurement

## 1 Introduction

According to [48], approximately 1.9 billion people around the world lack access to clean water. Also, clean water and sanitation have been identified as a Sustainable Development Goal by the United Nations [46]. Even the developed countries such as Canada has invested \$1.83 billion in water supply infrastructure within the period of 2016–2021 [18]. Therefore, the importance of water supply infrastructure has

---

D. Kankanamge · R. Ruparathna (✉)

Department of Civil and Environmental Engineering, University of Windsor, Windsor, Canada

e-mail: [rajeev.ruparathna@uwindsor.ca](mailto:rajeev.ruparathna@uwindsor.ca)

© Canadian Society for Civil Engineering 2023

655

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_40](https://doi.org/10.1007/978-3-031-34593-7_40)

never been more important in terms of research and development than ever before. Green procurement is a potential strategy to cater to the demand for water supply infrastructure in an eco-friendly manner [37]. Green procurement is an eco-conscious substitution to traditional low-cost project selection that incorporates environmental considerations into the procurement of goods, works, and services [20, 35, 42]. Since this is a highly data-intensive process, the wider adaptation is hindered by data and implementation challenges.

Ecolabels are an incentive for green procurement. Environmental Product Declaration (EPD) is a type III ecolabel that contains quantified environmental impacts based on pre-set indices [15]. EPD is developed through a life cycle assessment (LCA) [21]. The fidelity of an EPD heavily depends on the accuracy of the life cycle inventory. Ensuring the accuracy of a life cycle inventory is a challenge since less accurate data is available on a majority of backend processes.

Blockchain is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business network [6]. Blockchain can be a potential solution to enhance the accuracy of the life cycle inventory data. Blockchain allows multiple users in a supply chain to distribute the encrypted data via secure logging. It is a decentralized database that works on a network (e.g., Internet) [30]. Each new transaction (change) forms a new block with the updates if the transaction is validated, and the new blockchain is shared among all the users. This concept ensures the credibility and transparency of data, and it can be shared among all stakeholders instantaneously. Even though the blockchain concept is popular as the underlying technology in cryptocurrency, other industries, including the construction sector, have not acknowledged the full potential blockchain [1].

In order to incorporate EPDs in green procurement, a common data platform is required. Building Information Modeling (BIM) is an information repository that can serve as a common data platform for green procurement [38]. External data such as EPDs can be linked with the BIM environment to conduct eco-conscious evaluation [12]. Blockchain can be used for gathering entrusted, encrypted, up-to-date EPD data instantaneously.

According to [3] construction industry is ranked at the second-lowest level to adapt to information technology. However, [36] have emphasized the potential uses of blockchain in the construction sector, such as asset management, file sharing for document management, and construction supply chain management. Kiu et al. [24], Li et al. [26], and Shojaei [40] have also investigated the application of blockchain in the construction sector. Even though scholars have conducted reviews of blockchain adaptation for the construction sector, implementation frameworks for specific applications have not been researched.

This research aims to support eco-conscious decision-making in water supply infrastructure projects. The objective of this research is to propose a conceptual framework architecture to integrate EPDs via blockchain to aid the BIM-based green procurement. The proposed approach will provide updated, transparent, and verified environmental impacts of infrastructure components. This study will advocate BIM and blockchain adaptation in the infrastructure construction industry.

2 Literature Review

2.1 Eco-conscious Purchasing in the Infrastructure Sector

Due to its magnitude, complexity, and various characteristics, construction of water supply, transportation, irrigation, and utility infrastructure requires close scrutiny [11, 22]. The delivery of water supply infrastructure projects is within the mandate of federal, provincial, and municipal governments and international funding agencies [2, 17, 49]. However, environmental considerations of water supply infrastructure projects have been overlooked. Several previous studies have been conducted by previous researchers to enhance eco-conscious decision-making in the construction sector. Table 1 presents key focus areas of published literature in eco-conscious decision-making within 2019–2021.

According to Table 1, several researchers have developed innovative evaluation methods to assess environmental performance for eco-conscious decision-making [25, 33]. However, due to intensive data requirements, the infrastructure sector will not be able to apply the proposed methods without a common data platform. Several researchers have identified performance indicators to compare various suppliers, contractors, or products [4, 7, 23, 33]. A comprehensive review of the above research has revealed that there is no consistency in the evaluation criteria. There are several studies that have identified the challenges pertaining to implementing eco-conscious project evaluation. Foo et al. [13] and Lindblad and Gustavsson [28] have identified that financial capabilities of the manufacturers, intra-organizational integration, transparency, and stakeholder influences are some of the challenges impacting eco-conscious decision-making. In order to address these gaps in the literature, ecolabels can be utilized as a standard set of performance indicators and BIM as a platform to conduct the data-driven evaluation.

**Table 1** Published literature on eco-conscious decision-making in the construction sector

	Innovative evaluation methods	Performance indicator identification	Implementation challenges
Kesidou and Sovacool [23]		✓	
Laosirihongthong et al. [25]	✓		
Pamućar et al. [33]	✓	✓	
Bjerkkan et al. [4]		✓	
Cui et al. [7]		✓	
Lindblad and Guerrero [27]	✓		✓
D’Amico et al. [8]	✓		
Foo et al. [13]			✓

## 2.2 *Ecolabels*

Ecolabels inform customers of the product's environmental impacts in the form of a descriptive label [15]. There are three types of ecolabels, specified by the International Standards Association (ISO):

**Type I:** Type I ecolabel is standardized via ISO 14024, and it is a multi-criteria ecolabel program evaluated by an independent third-party organization [16]. It is also a self-styled environmental symbol, claim, or statement based on the full life cycle considerations [45].

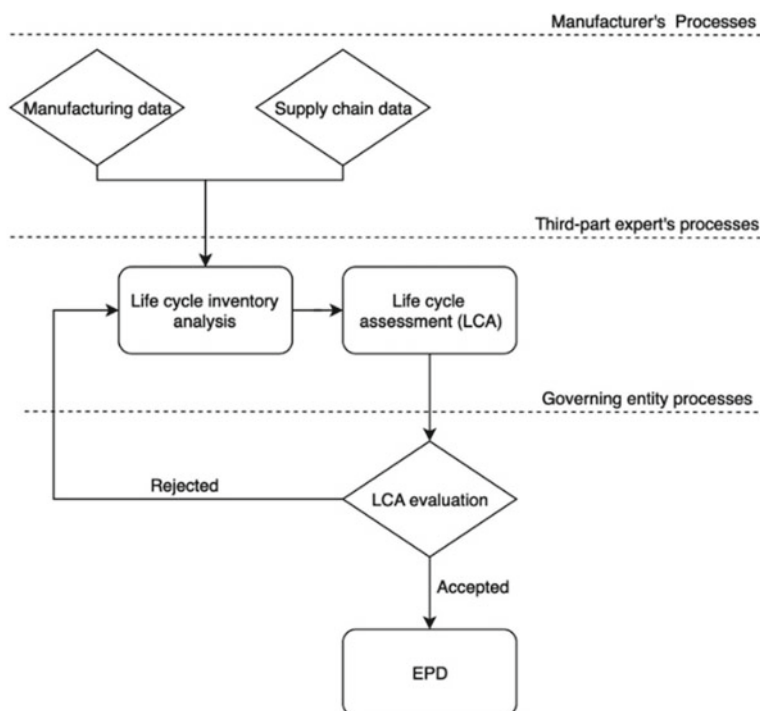
**Type II:** Type II ecolabel is a self-declared claim made by the manufacturers following ISO 14021, considering the selected aspects of the product's life cycle. Type II ecolabels are presented without impartial third-party auditing or verification [31].

**Type III:** Environmental Product Declaration (EPD) uses pre-set indices based on ISO 14025 or EN 15,804 to provide quantified environmental information about a product. An impartial organization verifies EPD before disclosing it to the public [15]. Given that EPDs contain quantified data, a consumer may evaluate a product's environmental performance with its competitors. EPD is an incentive for green procurement of the construction sector [29]. Figure 1 depicts the process for developing an EPD.

LCA and the approval for the LCA are the two main components in EPD development. This process involves extensive data sharing. In conducting the LCA, a life cycle inventory shall be conducted based on the data provided by a manufacturer [19]. The accuracy of life cycle inventory data should be ensured. Moreover, this data can vary with time as a result of changes to production methods, transportation modes and distances, and the type of raw material used. An established EPD does not regularly update by incorporating the time-varied parameters. Therefore, innovative measures such as blockchain can be utilized to verify the data used in conducting the LCA and automatically incorporate life cycle inventory changes into an EPD.

## 2.3 *Blockchain and Its Applications Infrastructure*

The blockchain concept was developed in 1991 and is widely practiced in the financial sector due to its speed, transparency, and encryption. Blockchain is defined as a decentralized electronic form of a database of blocks connected via chains [43]. In a blockchain, the longest chain is identified as the main chain (also referred active chain), generated from the genesis block, and each block is connected with different information and encryption with specific authorization. The blockchain has four elements ledger, cryptography, consensus, and business logic. The ledger element is replicated and shared among all the participants in that blockchain domain [6]. Ledger contains the history of transaction not necessarily in terms of money, while



**Fig. 1** EPD processes

cryptography ensures the integrity of the ledger and the identity of the participants in the blockchain. Consensus is the decentralized protocol and business logic element that contains the logic embedded in the ledger, and it is executed together with the transactions. For each transaction, a new block is created and chained after verification, typically known as Proof of Work (POW), and once it is chained, it cannot be broken. Furthermore, each stakeholder in the blockchain has a private key and public key in order to make a transaction (change) in the blockchain [30]. Once the transaction is approved, the entire chain is duplicated among every stakeholder involved [38]. The blockchain can be public, private, or hybrid based on the developer's requirement, and therefore, the application of blockchain has been extended into currency, health care, and supply chain. Similarly, the infrastructure sector can benefit from the blockchain concept in terms of eco-conscious decision-making via EPDs. Even though there were not any direct literature to support the possibility of using blockchain for eco-conscious decision-making via EPDs, Ølnes and Jansen [32], Kiu et al. [24], Li et al. [26], Rodrigo et al. [36], and Shojaei [40] stated that blockchain is evolving into a support infrastructure sector to aid secure documentation and provide entrusted data. Moreover, blockchain is positioned to positively influence future digital innovations, including in the public sector's e-governance.

The blockchain concept is developed for secure and transparent data sharing via duplication of its ledger between stakeholders. Infrastructure projects involve multiple stakeholders varying from suppliers to governmental authorities [9, 10], where the authenticity of the functional data has always been a major concern. Blockchain has the potential to address this challenge. In applying blockchain as a solution for EPDs' data challenges, all the life cycle inventory, including supply chain data, can be encrypted and shared block-wise to the LCA practitioner, the authority that approves the EPD, and the consumer. The authority will approve or reject the EPD and upload that data into the blockchain so that each party will have access to the most updated entrusted data. Since the supply chain is linked to the blockchain, sudden changes shall be shared instantaneously among the other stakeholders to provide the most updated EPD. Once the entrusted data through blockchain is gathered, a common data platform (e.g., BIM) is needed to evaluate the construction project. Even though the technology to link blockchain data into BIM is yet to be fully established Turk and Kline [44] has proposed architecture of a system using blockchain for BIM transactions.

## ***2.4 BIM Adaptation for Eco-conscious Decision-Making***

BIM is a computer-aided collaborative platform where stakeholders share a large amount of data and conduct analysis to improve a construction projects' time, cost, sustainability, safety, etc. BIM provides the common data platform to make eco-conscious decisions by using EPDs for water supply infrastructure projects. Previous researchers have exploited BIM potentials to conduct various analyses. Sloot et al. (2019) proposed a 4D simulation-based bidder evaluation method to select the best proposal. Ren et al. (2021) suggested a methodology for extracting information from BIM using an ontological knowledge base, permitting more effective financial management decision-making. The costs are extracted from a database and are imported to the BIM environment to conduct an evaluation. Ren et al. (2012) proposed a BIM-integrated framework for material procurement and supplier performance evaluation. They proposed ten evaluation criteria with performance indicators that are coded in a BIM tool to calculate a weighted score for each supplier. Salehabadi and Ruparathna [39] have developed a framework to conduct an environmental evaluation for single-family detached houses in the BIM environment using identified KPIs. Also, Patel and Ruparathna [34] have developed a BIM-based tool to conduct a life cycle sustainability evaluation for road infrastructure. Therefore, based on the literature, BIM has the potential to conduct an evaluation based on EPDs for eco-conscious decision-making even though it has not been researched by previous researchers.

BIM has the potential to incorporate customized parameters to the object elements in the model and conduct assessments. Moreover, BIM can be used to automate the evaluation and utilize third-party data for successful decision-making [12]. KPIs from



an EPD can be considered for bidder evaluation. Out of many multi-attribute decision-making methods (MCDM), TOPSIS method is a popular method for supply chain management and logistics engineering [47]. TOPSIS can provide the most stable performance results even with varying weights [47]. Since the significance of each environmental impact is in a gray area, the TOPSIS method best fits eco-conscious proposal evaluation due to its minimum influence from the weights toward the final output.

EPDs can be utilized in conducting an environmental evaluation to enhance eco-conscious purchasing. However, as mentioned in Sect. 2.2, EPDs shall be improved in terms of transparency and frequent updates to the quantitative data. Even with entrusted EPD data, the manual evaluation is complex due to the magnitude of water supply projects. Hence, blockchain-based EPDs, embedded BIM environments may provide the solutions to solve the above dilemmas. Several studies have proposed frameworks for environmental evaluations; however, EPDs are yet to be used as a parameter for environmental evaluations for infrastructure projects. Similarly, blockchain is yet to be used to share trusted data among the stakeholders in a construction project. This gap in the literature has influenced the research team to propose a framework to adopt blockchain-based entrusted ecolabels into BIM environment to enhance eco-conscious purchasing.

3 Conceptual Framework

This study focuses on water supply infrastructure projects and therefore considers the EPDs of PVC pipes and fittings as a case study for the framework. In order to develop a conceptual framework, life cycle processes in PVC pipes and fittings are identified. Based on the published literature, generalized life cycle processes and relevant stakeholders were identified (Table 2).

Based on the life cycle processes and the stakeholders identified, a conceptual framework architecture was designed to obtain a reliable and updated EPD via blockchain. The most updated EPD containing KPI values before the verification by

**Table 2** Pipes and fittings life cycle and EPD process including the stakeholders. Adapted from [5, 14, 16]

Process	Stakeholder
Raw material extraction	Manufacturer
Manufacturing	
Transportation	
Operation (usage)	
End life disposal	
Conduct EPD	Expert
Verify EPD	Governing authority
Purchasing	Purchaser

the governing authority is denoted by uEPD in this framework. And once the uEPD is approved, it will become the EPD. In certain scenarios, uEPD may differ from the EPD based on the verification time lag. The framework consists of two processes: obtaining the EPD and process 2 being the evaluation of multiple bidders within the BIM platform. Figure 2 presents the architecture of the conceptual framework.

### ***Process 1: EPD through blockchain***

In this process, the quantities and qualities of each process of the life cycle of PVC pipes and fitting are shared among each stakeholder via blockchain. The initial data provided by the manufacturer is included in the Genesis block, and each modification that occurs within the life cycle is added to the blockchain. After the POW, a new block will be created and chained with the updates. An enhanced view of a block is presented in Fig. 3.

The initial data provided in the Genesis block acts as the benchmark for the EPD. The *KPI values (uEPD)*, *approval status*, and the *EPD* appear as *Null* in the Genesis block. Any single transaction (change) in any of the *Life cycle process data*, *uEPD*, *approval status*, and *EPD* are added to the *logic element* in the block. In order to override to existing data, each stakeholder has a private and public key; as an example, if the manufacturer's logistics department change the mode of transportation to trucks from a rail, they add that change to the *Logic* using their private key and verify that change using their public key in order to confirm the authenticity. The signing using their keys is called the Hash, and it is a mathematical problem that must be solved to approve the transaction, which is the POW. Each block contains the Hash of the previous block, and therefore, any data tampering requires changing all the Hash in the succeeding blocks, which is impossible [30]. The new *Life cycle process data*, *uEPD*, *approval status*, and *EPD* are updated based on the logic provided and included to the blockchain as a new block. Furthermore, each change contains its timestamp to ensure the last update of data.

The expert obtains the last updated *Life cycle process data* and provides the uEPD and its KPI values by using their private and public keys. The governing authority verifies the last uEPD based on the timestamp and changes the approval status using their private and public keys to publish the EPD. However, to maintain consistency, the KPIs shall be predefined. Based on the EPD standards, ISO 14025 and EN 15,804 KPIs vary. For the purposes of this case study, KPIs depicted in Table 3 are used.

Since the purchaser has the access the uEPD as well as the EPDs, the evaluation of manufacturers can be conducted. Even though the expert may take time in providing the uEPD or delays in approving an EPD and will not cause any delays to decision-making due to the decentralized ledger with all the previous transactions. Therefore, the purchaser may still be able to make their decision based on previously approved EPD and the most updated KPI values in the uEPD.

The analysis is conducted by using a blockchain-based EPD database. This database contains the manufacturer name and pipe diameters varying from 20 to 1000 mm and the respective environmental KPI (Table 3) values per unit pipe length.

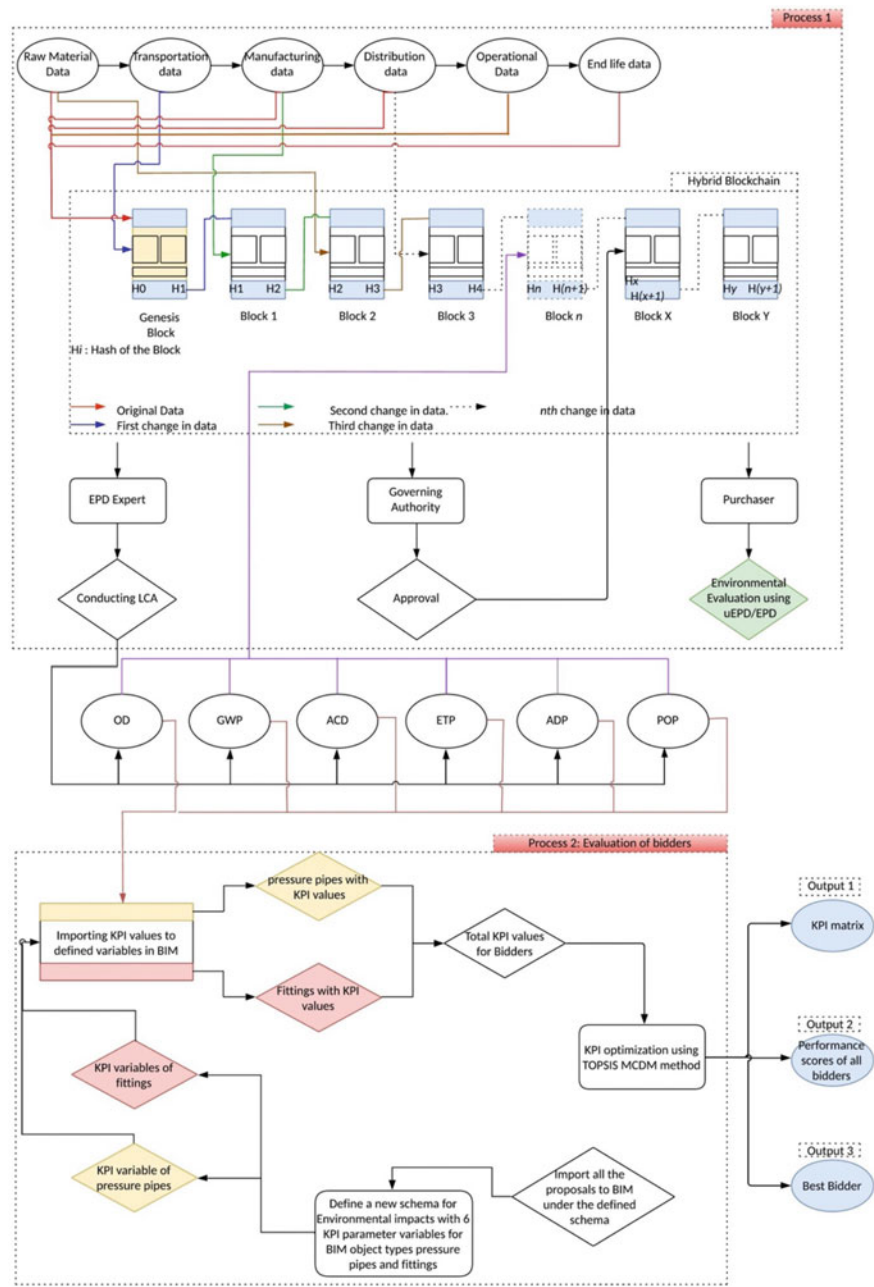


Fig. 2 Conceptual framework

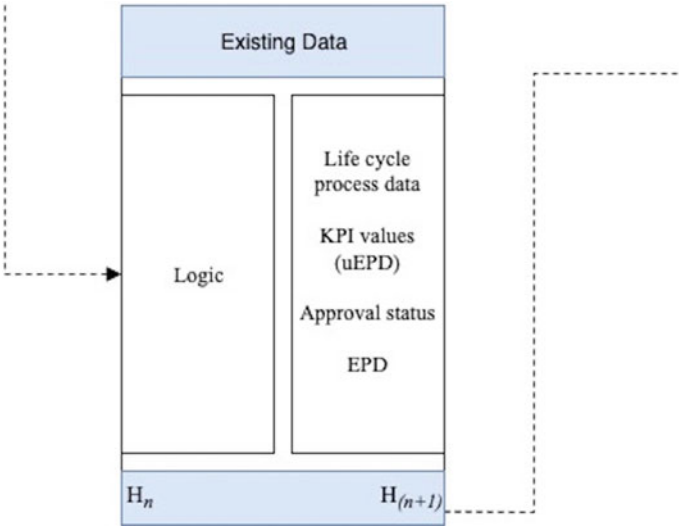


Fig. 3 Enhanced view of block  $n$

Table 3 Key performance indicators

KPI	Unit
Ozone depletion (OD)	kg CFC-11 eq
Global warming potential (GWP)	kg CO <sub>2</sub> eq
Acidification (ACD)	mol H + eq
Eutrophication (ETP)	PO <sub>4</sub> eq/kg N eq
Abiotic depletion potential (ADP)	MJ
Photochemical oxidation potential (POP)	kg C <sub>2</sub> H <sub>4</sub> eq

Process 2: Evaluation of bids

The first step in the bid evaluation is to link KPI values to system elements in the BIM model. Here, an extended schema is defined as PIPELINES\_PROC, to add environmental KPIs. Bidders are required to submit the project design in a BIM open-source file format. Bidder defines information such as manufacturer name, pipe length, and diameter. Environmental performance values are added by using the EPD database. Through the Javascript code, respective EPD data is imported to each bidder’s proposal by looping through EPD database and filtering via manufacturer name and diameter. Environmental KPI scores are calculated based on the actual quantities in the model (Fig. 4).

The  $7 \times N$  matrix, where  $N$  is the number of the bidders, is generated through the script as shown in Table 5 to select the best bidder. The TOPSIS MCDM algorithm is used in the BIM platform to automate the evaluation.

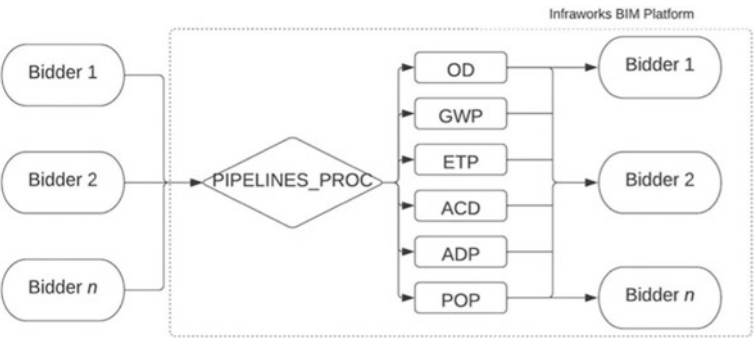


Fig. 4 Exporting bidder proposals to BIM

Table 5 KPI matrix

Environmental KPIs						
	$W_1$	$W_2$	$W_3$	$W_4$	$W_5$	$W_6$
	OD	GWP	ACD	ETP	ADP	POP
Bidder 1	$V_{11}$	$V_{12}$	$V_{13}$	$V_{14}$	$V_{15}$	$V_{16}$
Bidder 2	$V_{21}$	$V_{22}$	$V_{23}$	$V_{24}$	$V_{25}$	$V_{26}$
Bidder n	$V_{n1}$	$V_{n2}$	$V_{n3}$	$V_{n4}$	$V_{n5}$	$V_{n6}$

\* Footnote  $i, j$  vary from 1 to  $n$  for  $V_{ij}$ : KPI values;  $i$  vary from 1 to 6 for  $W_i$ : user input weightage as a percentage

Step 1: Normalization

All the values under each KPI ( $V_{ij}$ ) will be normalized ( $N_{ij}$ ) using the following equation:

$$N_{ij} = \frac{V_{ij}}{\sqrt{\sum_{i,j=1}^n V_{ij}^2}}$$

Step 2: Weighted normalization

Weighted normalization values will be denoted in  $X_{ij}$  ( $i, j$  vary from 1 to  $n$ ).  $W_i$  is the weightage.

$$X_{ij} = N_{ij} \times W_i$$

Step 3: The positive and negative ideal solution

In deciding ideal positive and negative solutions, all the KPIs are considered as non-beneficial values. Therefore, the positive ideal solution is the minimum

weighted normalized value, and the negative ideal solution is the maximum weighted normalized value.

$$X_i^+ = \text{Min} X_{ij}^+$$

$$X_i^- = \text{Max } X_{ij}^-$$

**Step 4: Positive and negative Euclidean distance**

Positive ( $D_i^+$ ) and negative ( $D_i^-$ ) Euclidean distances are calculated as shown below.

$$D_i^+ = \sqrt{\sum_{j=1}^n (X_{ij} - X_i^+)^2}$$
$$D_i^- = \sqrt{\sum_{i,j=1}^n (X_{ij} - X_i^-)^2}$$

**Step 5: Performance score**

Performance score ( $S_i$ ) is calculated for all the alternatives as follows. The final weighted normalized decision-making matrix is shown in Table 6.

$$S_i = \frac{D_i^-}{D_i^- + D_i^+}$$

The BIM-based green procurement evaluation plugin will provide the combined KPI matrix of all proposals, the calculated performance score for each proposal, and the proposal with the lowest environmental impacts as deliverables.

**Table 6** Weighted normalized decision-making matrix

Environmental KPIs									
	$W_1$	$W_2$	$W_3$	$W_4$	$W_5$	$W_6$			
	OD	GWP	ACD	ETP	ADP	POP	$D_i^-$	$D_i^+$	$S_i$
Bidder 1	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$D_1^-$	$D_1^+$	$S_1$
Bidder 2	$X_{21}$	$X_{22}$	$X_{23}$	$X_{24}$	$X_{25}$	$X_{26}$	$D_2^-$	$D_2^+$	$S_2$
Bidder $n$	$X_{n1}$	$X_{n2}$	$X_{n3}$	$X_{n4}$	$X_{n5}$	$X_{n6}$	$D_n^-$	$D_n^+$	$S_n$
$X_i^+$	$X_1^+$	$X_2^+$	$X_3^+$	$X_4^+$	$X_5^+$	$X_6^+$			
$X_i^-$	$X_1^-$	$X_2^-$	$X_3^-$	$X_4^-$	$X_5^-$	$X_6^-$			

\* Footnote ~  $i, j$  vary from 1 to  $n$ ;  $X_{ij}$ : weighted normalized values;  $D_i$ : Euclidean distance;  $W_i$ : user input weightage as a percentage;  $S_i$ : performance score

## 4 Discussion

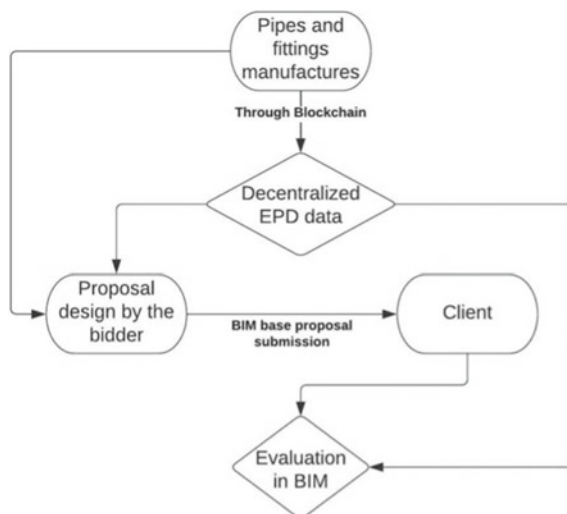
This research developed a conceptual framework to obtain the most updated and reliable EPD to aid the eco-conscious evaluation of water supply project proposals. Furthermore, through this framework, the client will have the opportunity to look into the backend of the supply chain and manufacturing information of the product before making a decision. This framework will deliver the following benefits to the construction industry as well as to the manufacturers.

- Considering the magnitude of investments for water supply infrastructure projects, implementing this framework will significantly lower the potential environmental impacts of the construction industry. Furthermore, since the EPD is conducted based on life cycling thinking, the overall environmental impacts can be minimized by adopting this framework.
- Since blockchain contains a decentralized database of all the transactions, the purchaser will have the ability to monitor and track the life cycle processes of a product. Hence, once the bidders are evaluated, and the contract is awarded, the purchaser will be able to track the purchase order to ensure that the product is delivered according to the expected quality and environmental performance.
- The logistics information of the purchase order will also be shared with the purchaser. Therefore, once the order is placed, the purchaser will be able to track the status of the order and incorporate that data into the project scheduling process to make an accurate and precise project schedule.
- The proposed approach will aid contractors around the world to securely bid for potential infrastructure projects. In fact, the data transparency is enhanced via blockchain; therefore, the purchaser will have no hesitation in the bidder's data. This will allow the manufacturers to expand their customer base on a global scale due to the blockchains' decentralized data sharing.
- Lack of resources and experts and transparency are identified as a barrier to green procurement. Since the evaluation and data sharing are automated, the requirement of expert involvement can be minimized. Furthermore, real-time data can enhance the accuracy of the evaluation process.

### 4.1 *Implementation of the Proposed Framework*

The manufacturers will develop a blockchain for their supply chain as well to transfer Life cycle process data to other stakeholders. The manufacturer will appoint a control officer for each process in the product life cycle to monitor the changes happening in their respective process with a public and private key to update the information in the blockchain. Manufacturers can share the blockchain with their prospective buyers to emphasize their data credibility. The expert who develops the EPD and the approval authority will be provided with timeframes to consider the updates to life cycle process in developing the EPD and approving it. The client will instruct

**Fig. 5** Implementation roadmap



all the manufacturers and experts to follow a single standard (e.g., ISO 14025) in developing the EPD. Lastly, the client will gather the entrusted EPD and conduct the evaluation for eco-conscious decision-making.

In implementing the proposed framework, the client should define the following in the Instructions to Bidder:

- The pipe system model should be developed to Level of Development (LOD) 3.
- The pipe system model shall be submitted as a non-proprietary open-source file format or IMX open-source file format to ensure interoperability. The non-proprietary open-source file has to be updated in order to incorporate the PIPELINES schema class with the environmental KPIs.
- The bidders should provide data for all the attributes in PIPELINES schema class defined in the open sources file format hierarchy (e.g., manufacturer name, pipe diameter, material, etc.)
- The pipe system model shall be developed and exported to open-source file format under WGS84 coordinate system.

Figure 5 depicts the overall industrial implementation summary of the proposed framework.

## 4.2 Limitation of the Proposed Framework

Even though the proposed framework improves the eco-conscious decision-making for water supply infrastructure projects, implementation will be a challenge due to the knowledge and resource limitations of stakeholders. Furthermore, on a technical



horizon, the blockchain process of hashing, which is solving the encryptions for validations of each block, requires a large amount lot of energy which is estimated to be 250–500 MW (enough to power a major city to a small country) [6]. ICT solution providers such as IBM offer platforms to implement blockchain solutions for industries such as supply chain, food, oil, and gas [41]. However, since the blockchain solution providers are limited, there may be delays for manufacturers in developing countries to adopt blockchain. Lastly, the legislation for blockchain and BIM-based applications is pivotal. Therefore, the contract administration between the stakeholders can create disputes. Another limitation of the proposed framework is the delays that may occur in developing and approving EPDs. However, the development of interoperability among different tools such as Civil 3D and OpenLCA can resolve this concern with automated assessments.

## 5 Conclusions and Recommendations

This study was conducted to propose a conceptual framework for eco-conscious decision-making for infrastructure projects. In order to mitigate the data risks, blockchain was proposed to obtain updated manufacturing data to develop EPDs and aid EPD approvals through a transparent, decentralized database. The above data was linked with BIM to conduct the eco-conscious proposal evaluation for green procurement. The proposed architecture was designed for water supply infrastructure projects.

With the use of blockchain and automated bidder evaluation, this framework breaks the transparency barrier in sustainable infrastructure procurement through to the entrusted data delivered through blockchain. Having access to real-time manufacturing data, the assumption that the life cycle processes of the product stay constant is no longer required. Lastly, due to the automation of the evaluation in the BIM platform and blockchain, the need for extra manpower and experts can be minimized.

The authors are currently developing process 2 (BIM-based proposal evaluation) of the proposed framework. Future research can focus on developing a global decentralized blockchain-based EPD database that can contain entrusted EPD data for all construction material. To aid this, interoperability between the LCA, BIM, and blockchain should be enhanced. Interdisciplinary research (e.g., civil engineering and computer science) is required in this domain for technology development and demonstration.

## References

1. Abou Jaoude J, George Saade R (2019) Blockchain applications—usage in different domains. *IEEE Access* 7:45360–45381. <https://doi.org/10.1109/ACCESS.2019.2902501>

2. ADB (2021) Knowledge | Asian Development Bank. (online) Available at: <https://www.adb.org/what-we-do/knowledge>. Accessed 16 Apr 2021
3. Agarwal R, Chandrasekaran S, Sridhar M (2016) Imagining construction's digital future. (online) Available at: <https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future2/28>. Accessed 5 Mar 2022
4. Bjerkan KY, Karlsson H, Sondell RS, Damman S, Meland S (2019) Governance in maritime passenger transport: green public procurement of ferry services. *World Electr Veh J* 10(4). <https://doi.org/10.3390/wevj10040074>
5. del Borghi A (2013) LCA and communication: environmental product declaration. *Int J Life Cycle Assess.* <https://doi.org/10.1007/s11367-012-0513-9>
6. Cachin C, Androulaki E, de Caro A, Osborne M, Schubert S, Sorniotti A, Vukolic M, Weigold T (2016) Blockchain, cryptography, and consensus. IBM Research, Zurich
7. Cui C, Sun C, Liu Y, Jiang X, Chen Q (2020) Determining critical risk factors affecting public-private partnership waste-to-energy incineration projects in China. *Energy Sci Eng* 8(4):1181–1193 (online). <https://doi.org/10.1002/ese3.577>
8. D'Amico F, Calvi A, Schiattarella E, di Prete M, Veraldi V (2020) BIM and GIS data integration: a novel approach of technical/environmental decision-making process in transport infrastructure design. *Transp Res Procedia* 45:803–810. <https://doi.org/10.1016/J.TRPRO.2020.02.090>
9. Dodanwala TC, Santoso DS (2021) The mediating role of job stress on the relationship between job satisfaction facets and turnover intention of the construction professionals. *Eng Constr Archit Manag.* <https://doi.org/10.1108/ECAM-12-2020-1048>
10. Dodanwala TC, Shrestha P (2021) Work–family conflict and job satisfaction among construction professionals: the mediating role of emotional exhaustion. *On the Horizon* 29(2):62–75. <https://doi.org/10.1108/OTH-11-2020-0042>
11. Dodanwala TC, Shrestha P, Santoso DS (2021) Role conflict related job stress among construction project professionals: the moderating role of age and organization tenure. *Construct Econ Build* 21(4):21–37. <https://doi.org/10.5130/AJCEB.v21i4.7609>
12. Farnsworth CB, Beveridge S, Miller KR, Christofferson JP (2015) Application, advantages, and methods associated with using BIM in commercial construction. *Int J Constr Educ Res* 11(3):218–236. <https://doi.org/10.1080/15578771.2013.865683>
13. Foo MY, Kanapathy K, Zailani S, Shaharudin MR (2019) Green purchasing capabilities, practices and institutional pressure. *Manage Environ Q Int J* 30(5):1171–1189. <https://doi.org/10.1108/MEQ-07-2018-0133>
14. FPI (2021) Environmental product declaration
15. Gallastegui IG (2002) The use of eco-labels: a review of the literature. *Eur Environ* 12(6):316–331. <https://doi.org/10.1002/EET.304>
16. Government of Canada (2021a) Environmental labelling programs. (online) Office of Consumer Affairs. Available at: <https://ic.gc.ca/eic/site/oca-bc.nsf/eng/ca02742.html>. Accessed 2 Feb 2022
17. Government of Canada (2021b) Infrastructure Canada—investing in Canada plan. (online) Infrastructure Canada Bilateral Agreements. Available at: <https://www.infrastructure.gc.ca/pt-sp/index-eng.html>. Accessed 7 Nov 2021
18. Government of Canada (2021c) Investing in water and wastewater infrastructure. (online) Available at: <https://www.sac-isc.gc.ca/eng/1525346895916/1525346915212>. Accessed 6 Mar 2022
19. Guinée JB, Heijungs R, Huppes G, Zamagni A, Masoni P, Buonamici R, Ekvall T, Rydberg T (2011) Life cycle assessment: past, present, and future. *Environ Sci Technol* 45(1):90–96. <https://doi.org/10.1021/ES101316V>
20. Hollos D, Blome C, Foerstl K (2012) Does sustainable supplier co-operation affect performance? Examining implications for the triple bottom line. *Int J Prod Res* 50(11):2968–2986. <https://doi.org/10.1080/00207543.2011.582184>
21. ISO (2006) ISO 14025: Environmental labels and declarations. (online) Standards. Available at: <https://www.iso.org/standard/38131.html>. Accessed 26 Feb 2022

22. Kankanamge DH, Santoso DS (2021) Examining and comparing the facility management services of two universities: Asia and Europe perspectives. *Int J Edu Econ Dev* 12(3):267–293. <https://doi.org/10.1504/IJEED.2021.115601>
23. Kesidou S, Sovacool BK (2019) Supply chain integration for low-carbon buildings: a critical interdisciplinary review. *Renew Sustain Energy Rev*. <https://doi.org/10.1016/j.rser.2019.109274>
24. Kiu MS, Chia FC, Wong PF (2020) Exploring the potentials of blockchain application in construction industry: a systematic review. *Int J Constr Manag*. <https://doi.org/10.1080/15623599.2020.1833436>
25. Laosirihongthong T, Samaranayake P, Nagalingam S (2019) A holistic approach to supplier evaluation and order allocation towards sustainable procurement. *Benchmarking* 26(8):2543–2573. <https://doi.org/10.1108/BIJ-11-2018-0360>
26. Li J, Greenwood D, Kassem M (2019) Blockchain in the built environment and construction industry: a systematic review, conceptual models and practical use cases. *Autom Constr* 102:288–307. <https://doi.org/10.1016/J.AUTCON.2019.02.005>
27. Lindblad H, Guerrero JR (2020) Client's role in promoting BIM implementation and innovation in construction. *Construct Manage Econ* 38(5):468–482 (online). <https://doi.org/10.1080/01446193.2020.1716989>
28. Lindblad H, Gustavsson TK (2018) Project managers as involuntary policy implementers? The case of implementing BIM. In: *Proceeding of the 34th annual ARCOM conference*, pp 465–474
29. Manzini R, Noci G, Ostinelli M, Pizzurno E (2006) Assessing environmental product declaration opportunities: a reference framework. *Bus Strateg Environ* 15(2):118–134. <https://doi.org/10.1002/BSE.453>
30. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. *Decentralized business review*. (online) Available at: [www.bitcoin.org](http://www.bitcoin.org). Accessed 16 Feb 2022
31. OECD (1997) Eco-labelling: actual effects of selected programme. (online) Paris. Available at: [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=ocde/gd\(97\)105](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=ocde/gd(97)105). Accessed 2 Feb 2022
32. Ølnes S, Jansen A (2017) Blockchain technology as a support infrastructure in e-government. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 10428 LNCS, pp 215–227. [https://doi.org/10.1007/978-3-319-64677-0\\_18](https://doi.org/10.1007/978-3-319-64677-0_18)
33. Pamučar D, Puška A, Stević Ž, Čirović G (2021) A new intelligent MCDM model for HCW management: the integrated BWM–MABAC model based on D numbers. *Exp Syst Appl* 175:114862 (online). <https://doi.org/10.1016/j.eswa.2021.114862>
34. Patel K, Ruparathna R (2021) Life cycle sustainability assessment of road infrastructure: a building information modeling-(BIM) based approach. <https://doi.org/10.1080/15623599.2021.2017113>
35. Reuter C, Foerstl K, Hartmann E, Blome C (2010) Sustainable global supplier management: the role of dynamic capabilities in achieving competitive advantage. *J Supply Chain Manag* 46(2):45–63. <https://doi.org/10.1111/J.1745-493X.2010.03189.X>
36. Rodrigo MNN, Senaratne S, Weinand R (2020) Blockchain technology: is it hype or real in the construction industry? *J Ind Inf Integr*. (online) <https://doi.org/10.1016/j.jii.2020.100125>
37. Ruparathna RJ (2013) *Emergy based procurement framework to improve sustainability performance in construction*. (online) University of British Columbia. <https://doi.org/10.14288/1.0074139>
38. Sacks R, Eastman C, Lee G, Teicholz P (2018) *BIM handbook: a guide to building information modelling for owners, designers, engineers, contractors and managers*, 3rd edn. Wiley, New Jersey
39. Salehabadi ZM, Ruparathna R (2022) User-centric sustainability assessment of single family detached homes (SFDH): a BIM-based methodological framework. *J Build Eng* 50. <https://doi.org/10.1016/J.JOBE.2022.104139>
40. Shojaei A (2019) Exploring applications of blockchain technology in the construction industry. In: *10th international structural engineering and construction conference*. ISEC Press. <https://doi.org/10.14455/ISEC.RES.2019.78>

41. Singh AJ, Cuomo J, Gaur N (2019) Blockchain for business
42. Srivastava SK (2007) Green supply-chain management: a state-of-the-art literature review. *Int J Manag Rev* 9(1):53–80. <https://doi.org/10.1111/J.1468-2370.2007.00202.X>
43. Tasatanattakool P, Techapanupreeda C (2018) Blockchain: challenges and applications. In: International conference on information networking, 2018-January, pp 473–475. <https://doi.org/10.1109/ICOIN.2018.8343163>
44. Turk Ž, Klinc R (2017) Potentials of blockchain technology for construction management. *Procedia Eng* 196:638–645. <https://doi.org/10.1016/J.PROENG.2017.08.052>
45. UNEP (2022) Eco-labelling-UN environment programme. (online) Eco-labelling. Available at: <https://www.unep.org/explore-topics/resource-efficiency/what-we-do/responsible-industry/eco-labelling>. Accessed 27 Feb 2022
46. United Nations (2015) The 17 goals | Sustainable Development. (online) Available at: <https://sdgs.un.org/goals>. Accessed 4 June 2021
47. Velasquez M, Hester PT (2013) An analysis of multi-criteria decision making methods. *Int J Oper Res* 10(2):56–66
48. World Bank (2021) Infrastructure overview. (online) Available at: <https://www.worldbank.org/en/topic/infrastructure/overview>. Accessed 16 Apr 2021
49. World Bank Group (2021) Infrastructure | Data. (online) Data. Available at: <https://data.worldbank.org/topic/infrastructure>. Accessed 22 Feb 2021

# Role of Electronic Document Management Systems in the Design Change Management Process



Oussama Ghnaya, Hamidreza Pourzareei, Louis Rivest, and Conrad Boton

**Abstract** Design changes are an unavoidable reality in the construction industry. After construction documents are released, changes may be made to them by different stakeholders for various reasons. Design changes must be well managed in accordance with a design change management (DCM) process to avoid time and cost overruns. Furthermore, the right tools must be used to facilitate the flow of documents and ensure that stakeholders have the most recent documents. Electronic document management systems (EDMSs) have been introduced to facilitate the storing, sharing, and tracking of documents in electronic format. This research aims to report on the real use of an EDMS in the construction industry as part of a DCM process. Six semi-structured interviews were conducted with a construction project director to document the DCM processes applied in two ongoing construction projects, one of which involved the use of an EDMS. The data collected served to compare the functionalities of the different tools used during the DCM processes. The main results show that traditional tools (e.g., PDF reader and email application) and the EDMSs adopted share some functionalities. However, the EDMS offers significant advantages in terms of collaboration and traceability.

**Keywords** Electronic document management · Design change management process

## 1 Introduction

The construction industry has always been characterized by its fragmented nature. Various stakeholders with different backgrounds collaborate to deliver a project that responds to the client's requirements. Documents, which can take the form of contracts, instructions, bids, or drawings, play a crucial role in the transfer of knowledge between stakeholders during the design and construction phases [17]. However,

---

O. Ghnaya (✉) · H. Pourzareei · L. Rivest · C. Boton  
École de Technologie Supérieure, Montreal, Canada  
e-mail: [oussama.ghnaya.1@ens.etsmtl.ca](mailto:oussama.ghnaya.1@ens.etsmtl.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_41](https://doi.org/10.1007/978-3-031-34593-7_41)

they are almost inevitably subjected to changes in the construction industry [6]. Changes can be necessary to rectify errors or recommended to make improvements or adapt elements when requested by the client [10]. If changes are not managed properly, they can lead to time and cost overruns. It therefore seems necessary to adopt a design change management (DCM) process to define participants' tasks and the flow of information. Changes have traditionally been transmitted by mail or email to the designated recipient on site with instructions to add, delete, or remove pages from the original document [15]. This does not guarantee that the right person has the right version of a document at the right moment. As a result, one mistake in a single document can lead to costly execution errors.

Today, electronic document management systems (EDMSs) are used in the construction industry to overcome document management issues in a paperless environment. EDMSs can serve to securely index, edit, store, and retrieve documents. In addition, using an EDMS can significantly facilitate DCM [15]. This paper focuses on presenting the role EDMSs play in real-world DCM processes in the construction industry. Two DCM processes, one of which involves the use of a commercially available EDMS (SmartUse), are documented and compared from the standpoint of tool functionality. The article is organized into six sections. Section 2 presents the DCM processes that are proposed in the literature as well as a generic one used to compare the two DCM processes studied. Section 3 reviews the EDMS used and its functionalities. Section 4 presents the methodology adopted to collect and analyze the data describing the two DCM processes in question. Section 5 shows the main results, focusing on the similarities and differences between the functionalities of the tools used in the DCM processes studied. Finally, Sect. 6 provides a discussion of the results and concludes the work.

## 2 Design Change Management Process

Design changes are a common occurrence in construction projects and can be made at any stage of a project [6]. A design change is defined as any change to the scope of work as described in the contract agreements [18]. It may apply to architectural, structural, plumbing, drainage, site work, or other aspects of construction. Design changes can have significant negative impacts on costs and scheduling [13]. Furthermore, [4] considers design changes to be one of the major causes of project failure. Changes can become a major source of contract disputes if they are not managed in accordance with a structured change management process [6]. Some research studies in the literature propose different DCM processes to help control and guide design changes. Table 1 sets out five DCM processes proposed in the literature [6, 8, 12, 14, 18].

It is worth mentioning that it is quite difficult to establish a standard DCM process due to the particularity of each construction project. However, we can notice from Table 1 that there are some similarities between the processes proposed and that

**Table 1** Design change management processes from the literature

Author	Year	Step 1	Step 2	Step 3	Step 4	Step 5
Ibbs et al. [8]	2001	Promote a balanced change culture	Recognize change	Evaluate change	Implement change	Continuously improve from lessons learned
Sun et al. [18]	2006	Start-up	Identification and evaluation	Approval	Implementation and review	
Motawa et al. [14]	2007	Start-up	Identify and evaluate	Approval and propagation	Post change	
Mejl��nder-Larsen [12]	2007	Identification	Filtration	Evaluation	Approval	Implementation
Hao et al. [6]	2008	Identify change	Evaluate and propose change	Approve change	Implement change	Analyze change

some activities remain applicable despite the differences between projects. A DCM process generally includes four stages:

- **Identification:** In this stage, the need for a change is identified. A change can be issued by the client to alter one or more specifications mentioned in the contract, by professionals due to technical issues, or by contractors due to execution problems.
- **Evaluation and proposal:** The proposed change's impact on the project's schedule and budget is evaluated. Evaluation is usually done by a change board that includes the affected stakeholders. In some cases, the client is also included when the change affects the scope of the project. During this stage, various solutions are proposed and evaluated to resolve the issue raised.
- **Approval:** The change board approves one solution to be implemented based on the evaluation of the solutions proposed and the feedback of the different departments impacted.
- **Implementation:** In this stage, the approved solution is implemented, meaning that the impacted documents (drawings, contracts, bids, etc.) are modified and the stakeholders concerned are notified to execute the change on site.

In addition to these four stages, some authors propose a fifth (initial) stage for DCM planning [14, 18] that includes team building, role clarification, and agreement on change management methods and procedures. The goal of this extra step is to improve the project team's ability to respond to change, effectively manage change, and create contingency preparations for any required changes. Moreover, [6, 8] propose a review stage at the end to conduct an assessment, take remedial steps, compare the results to the initial goals, and apply lessons learned.

All the above stages require the accurate and timely analysis of project conditions and timely communication between project stakeholders. Appropriate tools are therefore needed to support the impacted professionals in completing these tasks [18].

### 3 Electronic Document Management System

An EDMS is a system that can securely coordinate and handle operations such as the storage, retrieval, processing, printing, routing, and distribution of electronic and paper documents so that they may be used effectively by authorized personnel when needed [2]. Ahmad et al. [17] consider an EDMS to be a software application that collects paper and electronic documents for secure storage, retrieval, and archiving. Interestingly, [15] do not consider an EDMS to be a single technology, but rather a collection of information technology for scanning, indexing, editing, processing, storing, and retrieving documents to suit an engineer's needs relative to the information in the documents.

An EDMS helps to reduce a project's documentation costs and improve its quality. According to Hung et al. [7], construction firms implement EDMSs for two key



purposes: ubiquitous document access and quick and easy data sharing and collaboration. In addition, EDMs provide secure access to documents, reduce the amount of time required for project engineers to complete document-centered processes, and facilitate change management [15].

A variety of research studies have proposed and defined EDM functionalities [1, 9, 11, 16]. Their common basic functionalities are:

- **Document storage and retrieval:** The repository should serve as the central storage location for all a company's documents, with users able to access documents via search and retrieval or by browsing [9].
- **Foldering:** With virtual folders, a single classification level can be shown in multiple folders throughout the system. Searching for a document is made easier when documents are stored in separate folders according to their context [9].
- **Document relations:** In the context of document management, the term "relations" refers to the relationships that exist between documents. Relationships are frequently established using database relations, which makes relation administration easier. For example, changes made in some documents can be applied automatically in all linked documents depending on the type of document relations [9].
- **Integration with desktop applications:** Users can save documents to EDMs directly in the desktop applications in which they were created, like Microsoft Office [1, 11].
- **Document viewing:** Most EDMs have viewing capabilities to be able to quickly read a document directly in the system. This allows users to see documents without having to open them in the application in which they were created [9].
- **Document review:** Users can add comments to a document without actually changing the document itself [16].
- **Document comparison:** Users can compare two related files, for example, two different versions of a document, and record the differences using a comparison function [9].
- **User access management:** Individual users can have different access rights to documents. User access rights can be defined in a variety of ways and be specific to, for example, a user or user group or a document or document group [9].
- **Check-in/check-out:** The check-in and check-out features regulate who is editing a document and when and ensure that a document is being updated by only one person at a time [1].
- **Workflow:** The workflow functionality routes documents from one user to another in a controlled fashion [16].
- **Version control:** When a document is modified, the system must be able to track the modifications made to it. This is accomplished by assigning the document a version number [1].
- **Document numbering:** Document numbering refers to the automatically numbering of documents according to predefined numbering standards. Numbering can follow a sequential numbering, a predetermined numbering sequence, or a combination of both [9].

- **Auditing:** Auditing, in combination with version control, keeps track of who made changes to a document and when. Authorized users can use the auditing tool to see what modifications have been made to a document since it was initially created [1].

## 4 Methodology

This article aims to report on the real use of an EDMS in the construction industry during a DCM process. Six semi-structured interviews were conducted with a project director from a Quebec construction firm to document and validate the DCM process used in two ongoing construction projects. For confidentiality reasons, the projects studied are referred to as “Project A” and “Project B”. An EDMS is used in Project B. In Project A, documents are managed in a more traditional way.

First, the Business Process Modeling Notation 2.0 (BPMN 2.0) format was chosen to model the DCM processes. BPMN is a graphical representation of the logic of business process steps [3]. This format was created specifically to organize the sequence of procedures and messages that pass between participants in various activities. In addition, project-related documents including emails, contracts, drawings, project descriptions, change requests, and change orders have been analyzed to understand the flow of documents during the DCM process.

Next, the activities identified in Project A and Project B were classified into the four phases that constitute a generic DCM process: identification, evaluation and proposal, approval, and implementation. Then, the different functionalities of the tools used were associated with their relevant activities.

Last, the DCM processes used in Project A and Project B were compared from the standpoint of tool functionality. It is important to mention that this study concerns the tools used to manage released documents after the design phase. For this reason, the study of CAD tool functionalities is outside the scope of this study.

## 5 Results

In this section, the results obtained from the data collected through our semi-structured interviews are presented. We describe the two projects studied and the types of documents used and then provide a detailed comparison of the phases of their DCM processes and the functionalities of the tools used.

## ***5.1 Presentation of the Projects Studied***

### **5.1.1 Project A**

Project A is a private project that is 98% completed. Our industrial partner is the project manager on this project. As with other construction projects, a variety of stakeholders are involved in the project. The main ones are the client, the project manager, professionals (including the engineering department), the general contractor, and subcontractors. It is important to mention that a change can be requested by any stakeholder at any stage. There is no formal DCM process to follow for this project. However, we have tried to document through the interviews the way changes are handled.

### **5.1.2 Project B**

Project B is a private project that is 90% completed. Our industrial partner is the construction manager on this project. Other project stakeholders include the client, professionals, the general contractor, and subcontractors. While traditional document management tools are used for this project, an EDMS has also been adopted to facilitate sharing the various project-related documents. The EDMS is managed by the construction manager, who takes charge of organizing the documents into different folders. The other stakeholders are granted access to the documents in line with their assigned roles.

## ***5.2 Comparative Analysis of the DCM Processes Used***

In this section, we compare the different phases of the DCM processes used in the two projects studied. It is important to mention that various documents are exchanged between stakeholders during a DCM process. The documents used in Project A and Project B are:

- **Request for Information (RFI):** “A formal written procedure initiated by the contractor seeking additional information or clarification for issues related to design, construction, and other contract documents” [5].
- **Design Change Request (DCR):** “A list of changes from engineering” [12].
- **Change Directive (CD):** The instruction proposed by the professionals to implement a change.
- **Contract Modification (CM):** The proposed modifications to the price and details in the contract.

- **Amendment:** The official/approved changes to the content of an agreement.

Our comparison aims to highlight the role of the tools’ functionalities in each activity. For this reason, the activities have been categorized into the four generic DCM process phases identified from the literature: identification, evaluation and proposal, approval, and implementation. It is important to mention that the activities our industrial partner performs in each phase may differ slightly from what is mentioned in the literature.

5.2.1 Identification Phase

Project A

In this phase, the need for a change is identified. It can be raised by the client to modify the scope of the project or elements of the contract. It can also be identified by the general contractor or a subcontractor in the case of a site issue or drawing mistake. In the latter case, an RFI is usually issued to the professionals to get the clarifications needed. An Excel sheet that contains details about the creation of RFIs and their associated answers ensures that RFIs are followed up on. In some cases, an RFI can be promoted to a DCR. The email application is the main tool used for communication between stakeholders.

Project B

As in Project A, a change can be requested by the client to alter the scope of the project or by the professionals in the case of design errors. The general contractor can also issue an RFI to the professionals to obtain clarification about issues related to the construction drawings. In the latter case, there is usually a back and forth between the general contractor, the subcontractors, and the professionals to clarify the issues. In some cases, an RFI can be promoted to a DCR. An Excel sheet that indicates the date of creation of RFIs, the answers, and the departments concerned is used to track and respond to RFIs. Documents are shared in the EDMS, and the review and comment features allow stakeholders to annotate documents and drawings directly in the system. Interestingly, the email application is still used for communication between stakeholders.

Comparison

Table 2 shows the similarity between the activities involved in Project A and Project B that were considered during the identification phase. All the activities of project A were shared with project B and shown in pink cells. In terms of tool functionality, we observe that the EDMS used in Project B has some functionalities in common with other tools. For example, both the EDMS and the PDF Reader are used to annotate and view documents, and both the EDMS and the email application are used to share documents.

**Table 2** Activities and tool functionalities considered in the identification phase activities

Activities	Tools			
	Desktop Applications	PDF Viewer	Email Application	EDMS
Create/revise an RFI	Create text document; Convert document to PDF	Annotate document		Upload document; Move document to a folder; Link documents; Review document
Share an RFI			Share document	Share document
Update the table of the following RFI	Create and modify sheets			
Respond to an RFI	Create text document; Convert document to PDF	View document; Annotate document	Share document	View document; Review document
Evaluate the answer to an RFI		View document; Review document		View document; Review document
Evaluate the need for a DCR		View document; Review document		View document; Review document
Create a DCR	Create text document; Convert document to PDF			
Share a DCR			Share document	Share document

: Project A     : Project B     : Projects A & B

5.2.2 Evaluation and Proposal Phase

Project A

In this phase, the professionals concerned issue CDs to resolve the issues raised in DCRs, and impacted drawings are updated. The CDs are then evaluated by the engineering departments. A CD can be accepted, rejected, or revised. Once a CD is accepted, it is communicated to the general contractor, the client, and the project manager by email with all the updated documents. CD follow-up is done in an Excel sheet.

Project B

In this project, the DCR created is evaluated by the professionals concerned. If the change requested impacts the cost of the project or the details of contracts, multiple CDs are created to propose solutions and evaluated by the professionals to determine their potential impacts before a single CD is approved. The drawings and documents impacted are updated and shared via the EDMS. An Excel sheet is also used to keep track of the CDs created.

Comparison

As in the previous phase, Table 3 shows that there are some similarities between the two projects. We can see that Project B’s evaluation activities are done entirely in the EDMS. Users do not need to access the native applications to annotate documents and leave comments in them. In addition, a comparison feature allows users to compare two drawings either by putting them side-by-side or by overlaying them. Despite these features, traditional practices are also used in Project B. In some cases, documents are reviewed and annotated in the PDF viewer or by hand. In addition, the email

**Table 3** Activities and tool functionalities considered in the evaluation and proposal phase

Activities	Tools			
	Desktop Applications	PDF Viewer	Email Application	EDMS
Evaluate a DCR		Annotate document		Annotate document; Compare document; View document
Create/revise a CD	Create text document; Convert document to PDF			Upload document; Create new revision
Share a CD			Share document	Share document
Evaluate a CD		View document; Review document		View document; Review document
Update the table of the following CD	Create and modify sheets			

: Project A     : Project B     : Projects A & B

application is still used to share documents. It seems that the EDMS has not yet replaced the traditional tools.

5.2.3 Approval Phase

Project A

Once a CD has been issued, the general contractor and subcontractors provide a CM that sets out the working procedure, time frame, and estimated cost to execute the change. The CM is then evaluated and recommended to the client by the professionals and the project manager. The client has the final say whether to accept, reject, or revise the CM. Again, communication and negotiation are done via email, and an Excel sheet is used to ensure CM follow-up.

Project B

The approved CD is communicated to stakeholders. The general contractor and the subcontractors evaluate the CD and put forward bids to execute the change. There is also a back and forth between the professionals, the general contractor, and subcontractors in this stage. The bids are accepted, rejected, or revised by the client. Once the stakeholders agree on how to implement the change, the construction manager creates an amendment to modify the content of the initial agreement. Finally, the amendment is approved by the client.

Comparison

In the approval phase, there are some differences between the two projects in terms of the types of documents used. For example, bids are used in Project B to communicate the new proposed cost estimate, whereas a CM is used in Project A to set out the working procedure, time frame, and estimated cost to execute a change. The differences are mainly due to the nature of the contracts. During this phase, approved documents are signed. In a paperless environment, e-signatures are used in place of traditional signatures (Table 4).

**Table 4** Activities and tool functionalities considered in the approval phase

Activities	Tools			
	Desktop Applications	PDF Viewer	Email Application	EDMS
Create/revise a CM	Create text document; Convert document to PDF			
Share a CM			Send document	
Evaluate a CM		View document; Review document		
Update the table of the following bids	Edit text			
Create/revise bids	Create text document; Convert document to PDF			Create new revision
Share bids				Share document
Evaluate bids				Add a stamp
Update the table of the following CD	Create and modify sheets			
Update the table of the following amendment	Create and modify sheets			

: Project A : Project B : Projects A & B

5.2.4 Implementation Phase

Project A

Once the CM is approved, the updated documents are shared with the contractor and the subcontractors, who execute the change.

Project B

The updated documents are shared with the contractor and subcontractors to execute the change. In this case, using the EDMS ensures that the right version of the documents is available for the right people. The construction manager and the professionals follow up on the execution of changes to ensure compliance with the final document (Table 5).

Comparison

There are no major differences between Project A and Project B in this stage.

**Table 5** Activities and tool functionalities considered in the implementation phase

Activities	Tools			
	Desktop Applications	PDF Viewer	Email Application	EDMS
Share final document			Share document	Share document
Execute change		View document		View document

: Project A : Project B : Projects A & B

## 6 Discussion and Conclusion

The research presented in this paper helps to understand how design changes are managed in the construction industry and underlines the roles that are played by an EDMS during a DCM process.

First and foremost, the results show that the two projects studied rely heavily on desktop applications. Microsoft Word is the main tool used to create and modify documents including RFIs, DCRs, and CDs. Microsoft Excel is used to track the status of documents, with a separate sheet used for each type of document. The email application is the main tool used for communication between stakeholders and to share documents. It is worth mentioning that the telephone is also used in some cases to informally request changes. This type of change request is difficult to track.

The results also show that the EDMS used in Project B has some functionalities in common with the PDF viewer and the email application, such as document sharing, viewing, and review. However, it is the only tool to offer a secure repository where documents can be easily shared and kept track of. In addition, its version control functionality makes quick work of tracking modifications. Users can thus be sure they are working on the most recent version of documents. Despite these advantages, it seems that the EDMS has not yet replaced the traditional way of sharing and tracking documents.

Since this research presents a qualitative comparison of two DCM processes, it was difficult to reveal the real impact of adopting an EDMS, especially in terms of the time frame and cost of a project. The BPMN models generated could be used to simulate the time it takes to execute a design change. However, it was hard to determine the true duration of each activity.

Last, it is worth mentioning that the processes documented helped our industrial partner to gain an overall view of its DCM processes. In addition, even though process improvement was not part of the scope of this research, the documentation of the processes will undoubtedly help our industrial partner make significant improvements to their processes.

## References

1. Adam A (2008) Implementing electronic document and record management systems. Auerbach Publications
2. Asılı H, Tanrıover OO (2014) Comparison of document management systems by meta modeling and workforce centric tuning measures. *Int J Comput Sci Eng Inf Technol* 4(1):57–67. <https://doi.org/10.5121/ijcseit.2014.4106>
3. Chinosi M, Trombetta A (2012) BPMN: an introduction to the standard. *Comput Stand Interf* 34(1):124–134
4. Hallock B (2006) Managing change vs. administering the change order process. *Nielsen-Wurster Communiqué* 1(6)
5. Hanna AS, Tadt EJ, Whited GC (2012) Request for information: benchmarks and metrics for major highway projects. *J Constr Eng Manag* 138(12):1347–1352. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000554](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000554)



6. Hao Q, Shen W, Neelamkavil J, Thomas JR (2008) Change management in construction projects. NRC Publications Archive
7. Hung Kao C, Rzu Liu S (2013) Development of a document management system for private cloud environment. Elsevier. <https://doi.org/10.1016/j.sbspro.2013.02.071>
8. Ibbs CW, Wong CK, Kwak YH (2001) Project change management system. J Manag Eng 17(3):159–165. [https://doi.org/10.1061/\(ASCE\)0742-597X\(2001\)17:3\(159\)](https://doi.org/10.1061/(ASCE)0742-597X(2001)17:3(159))
9. Janne Y (2003) Document management as a part of product lifecycle management. <https://lut.pub.lut.fi/bitstream/handle/10024/34889/nbnfi-fe20031571.pdf?sequence=1>
10. Jarratt TAW, Eckert CM, Caldwell NHM, Clarkson PJ (2011) Engineering change: an overview and perspective on the literature. Res Eng Design 22(2):103–124. <https://doi.org/10.1007/s00163-010-0097-y>
11. Kuasmanen K (2019) Evaluation of an electronic document management system implementation success from an end user perspective in an industrial company. [https://www.utupub.fi/bitstream/handle/10024/147388/Kuosmanen\\_Katariina\\_Opinnayte.pdf?sequence=1](https://www.utupub.fi/bitstream/handle/10024/147388/Kuosmanen_Katariina_Opinnayte.pdf?sequence=1)
12. Mejl  nder-Larsen    (2017) Using a change control system and building information modelling to manage change in design. Architect Eng Des Manage 13(1):39–51. <https://doi.org/10.1080/17452007.2016.1220360>
13. Moayeri V, Moselhi O, Zhu Z (2017) Design change management using BIM-based visualization model. Int J Archit Eng Construct 6(1) Article 1. <https://doi.org/10.7492/IJAEC.2017.001>
14. Motawa IA, Anumba CJ, Lee S, Pe  a-Mora F (2007) An integrated system for change management in construction. Autom Constr 16(3):368–377. <https://doi.org/10.1016/j.autcon.2006.07.005>
15. Jadid MN, Idrees M (2005) Electronic document management system (EDMS) in civil engineering projects. In: 33rd Annual conference of the Canadian society of civil engineering
16. Raynes, M. (2002). Document management: Is the time now right? *Emerald*, 303–308.
17. Ahmad HS, Bazlamit IM, Ayoush MD (2017) Investigation of document management systems in small size construction companies in Jordan. ScienceDirect. <https://doi.org/10.1016/j.proeng.2017.03.101>
18. Sun M, Fleming A, Senaratne S, Motawa I, Yeoh ML (2006) A change management toolkit for construction projects. Archit Eng Des Manage 2(4):261–271. <https://doi.org/10.1080/17452007.2006.9684621>

# An Interactive Decision-Support Tool to Improve Construction Cost Management with Building Information Modeling



Nicolas Strange, Daniel Forgues, and Conrad Boton

**Abstract** Cost overruns in construction projects are currently one of the biggest problems in the industry. Factors causing these cost overruns have been identified in order to improve current management practices. Building Information Modeling (BIM), and in particular the fifth dimension relating to cost management (5D BIM), offers techniques and processes to improve the quality, performance, and efficiency of projects. Current BIM 5D practices are evolving and many software programs are being developed in this direction. However, due to the multiplicity of the proposed solutions and the wide range of the possibilities and associated functionalities, it is challenging for the practitioners to select the right tool adapted to their particular requirements. Previous research works have attempt to solve the issue by proposing some comparison of the existing 5D BIM tools. While very interesting, such comparisons are not sufficient since they do not provide the user with the level of interactivity necessary to adapt the choice to their particular business context. The research project presented in this paper proposes an interactive decision-support tool for the effective choice of 5D BIM solutions. The proposed tool is based on the characteristics and functions of the existing software and uses a personalized weighting by the user for each of the identified cost management requirements. Thus, the proposed tool makes it possible to inform users about 5D BIM software according to their specific requirements and to provide them with informed software selections. The results were evaluated and validated by experts in construction cost management.

**Keywords** Building information modeling · Cost management · 5D BIM · Interactive decision support

---

N. Strange · D. Forgues · C. Boton (✉)  
Department of Construction Engineering, École de Technologie Supérieure, Montreal, QC,  
Canada  
e-mail: [conrad.boton@etsmtl.ca](mailto:conrad.boton@etsmtl.ca)

N. Strange  
e-mail: [nicolas.strange.1@ens.etsmtl.ca](mailto:nicolas.strange.1@ens.etsmtl.ca)

# 1 Introduction

Cost overruns are among the main challenges observed in construction projects and the construction industry is constantly evolving in technology in an attempt to improve cost management, reduce cost overruns, save time in running estimates, and better control costs. Many studies have been dedicated to identifying the problems facing the overall cost management of projects [9, 12] and the Project Management Institute has established a breakdown of cost management into different activities with the aim of facilitating the resolution of cost overruns, as follows: cost management planning, cost estimation, budget determination, cost control, life cycle cost analysis, and claim management [19, 30].

In recent years, advances in Building Information Modeling (BIM) have made it possible to considerably improve these different aspects based on digital models [31]. The BIM-based cost management practices are generally referred to as 5D BIM. The advantages of the 5D BIM are numerous and well documented in the scientific literature [26, 29], and many technological solutions are commercially available in the industry. However, due to the fragmentation of the practices in the industry and the specificities of the existing software, it is very challenging for the practitioners to select the right software adapted to their particular needs and context. While some recent research works have proposed a comparison of existing solutions [37], these tools are constantly evolving and such a comparison needs to be constantly updated. Moreover, the most recent comparisons lack the interactive aspect, necessary to allow each user to adapt the comparison criterion to his particular needs.

The research work presented in this article aims at proposing an interactive decision-support tool in the selection of 5D BIM software for variable business contexts. This tool aims to help estimators in selecting the BIM 5D software that best suit their practices and needs, in order to improve their cost management. The article is organized into four main sections. Section 2 proposes a literature review, including research works related to the use of 5D BIM for construction cost management, the challenges related to its application, and the need for a decision-support tool. Section 3 presents the methodology, encompassing the steps of the development and the validation of the proposed tool. Section 4 presents the structure of the tool, its operation, and the results of the validation process. Section 5 concludes the work, including its limitations and identifies the future work.

## 2 Related Works

### 2.1 *Cost Management in the Construction Industry*

Cost management is defined as “the cost estimation of all activities and effort necessary to deliver the project” [16]. However, cost management encompasses much more, including the “application of management accounting concepts, methods of

data collection, analysis and presentation in order to provide the information needed to plan, monitor and control costs” [36]. Thanks to the work in three studies [4, 30, 32], it is possible to categorize the activities of cost management in construction into cost estimation, cost budgeting, cost control, life cycle cost analysis, and claims.

Cost estimation can be described as “a process that allows stakeholders to determine monetary resources required for a project’s completion” [37]. The budget determination process basically consists of a summation of the activities and work packages’ cost estimates executed earlier [37]. It can be done either “once, early in a project’s life-cycle, or multiple times at predetermined milestones and serve as the baseline against which the project’s performance can be controlled” [30]. The critical aspect of cost control is the analysis of the consumption of project funds in relation to the actual work being accomplished with the aim of maintaining the baseline previously established throughout the whole duration of a project [30]. Life cycle costs’ analysis evaluates the costs involved in a project, from the early stages till the end of the life of the project [7, 38]. It is usually performed early in a project’s life cycle [13] in order to help in evaluating the investment options and easing the choice of design criterion [38]. Construction claims refer to the claims resulting from changes in the scope of the works, delays and productivity factors, or deficient contract documents [23].

## ***2.2 Uses of 5D BIM for Cost Management***

5D BIM offers the possibility of linking cost information with digital models to enable better cost management [22, 33, 37]. 5D BIM models have great potential since they have the ability to contain elements, or assemblies of objects, with associated costs directly through the BIM software, or through an internal or external database [34]. Mitchell [28] argues that the objectives of 5D BIM for design and construction are to provide a transparent framework in order to make the best decisions, in terms of quality, especially in terms of quantity extraction, tendering, variation evaluation, change orders, and progress payments [26, 28, 32]. Thus, 5D BIM is a project visualization and monitoring tool, allowing stakeholders to control, in a collaborative environment, the flow of information, to effectively manage planning and estimation aspects by improving and facilitating control of costs and planning [1, 34].

Cost planning is the basis of overall cost management in a construction project [28]. The quantity extractions, retrieved from 3D models, assembled with the corresponding cost information make it possible to target the available budget on the most important characteristics of the building design [28]. In addition, Stanley and Thurnell show through a study that cost planning, with the quantities extracted using 3D BIM modeling software, is easier and faster than when carried out by estimators from 2D plans [34].

Quantification based on digital models becomes more and more automated as the models develop [32]. Thus, the extraction of quantities makes it possible to improve the good management of projects, in particular on the production of estimates, cost planning, measurements, preparation of bills of materials and tender documents, and finally cost control and proper preparation of evaluations and payments [2, 26, 34]. The study presented by Stanley and Thurnell shows the importance of precise quantities' takeoff based on 5D BIM in making budget estimates [34].

Estimating costs is a very important part of project decision-making. By extracting quantities from digital models, 5D BIM software can produce project cost estimates with varying degrees of precision [22, 37]. 5D BIM enables to quickly create more accurate cost estimates and to explore different project options before and during construction based on variations that are likely to occur. It also helps to identify factors that have different advantages or options in order to select the most interesting proposals [34].

Performing cost monitoring and control allows project cash flow to be forecast more quickly and accurately than at the start of the project life cycle [20, 37]. Thanks to 5D BIM technologies, it is possible to establish a cost at a given point in the project, by isolating the elements completed on the model and using the associated cost information. Thus, it is possible to compare the current calculated budget with that initially planned.

Life cycle costs can be defined as the sum of the present value of all expected costs from the construction phase to the end of the life of the facility [7]. A more recent research presented the development of a solution of life cycle cost analysis (LCCA) using 5D BIM by adding a life cycle cost calculation model structure to an innovative 5D BIM model associated with a spreadsheet file [18]. The main benefit of this process is that it links cost planning, quantity extraction, and life cycle cost calculations in an integrated environment [18]. However, while some 5D BIM software, specific to quantity extraction, provides appropriate costs for performing an LCCA [17], some authors claim that the BIM-based tools and technologies do not yet have the necessary capacities to adapt to the conditions of an effective LCCA [10, 35].

The capabilities of BIM allow the management and the pricing of construction claims, while reducing the number of claims issued during a project [37]. In addition, 5D BIM offers new technological opportunities for document management and visualization [14]. A study by El Hawary and Nassar shows that the use of new BIM technologies has a very positive effect on the reduction of claims in construction projects [15]. New working methods have been devised to find all flaws in logic and planning activities that will not be respected, based on the combination of a claims matrix and a 5D BIM model [25].

### ***2.3 The Need for a Decision-Support Tool for the 5D BIM Software Selection***

The scientific literature shows that there are many challenges to implementing 5D BIM in the construction industry, primarily regarding the details of the design, standardization, and skills required [26]. BIM induces significant changes in the management and the delivery of construction project [5], and its implementation of BIM within companies is also a challenge due to the associated training load from the start, the investment, and the time required to assimilate new processes [35]. In addition, the high costs of this investment are a hindrance, especially if the company plans to use it only for extracting quantities [35].

It is important to note that all trades do not use BIM the same way. Architecture firms are among the most active stakeholders when it comes to BIM adoption [27]. A study presented by Ding et al. [8] shows that motivation plays an important role in BIM adoption by architects and the implications of BIM in the architecture design process have been thoroughly discussed by Marcos [24]. Beyond the design itself, however, architects “may add as much data as any element within the entire architectural fabric could need to be thoroughly defined, properly speaking, the model is a three-dimensional data model (3DD model)” [24]. The model created can then be used for further uses, including estimating and managing costs (5D). Unlike architects who merely use BIM authoring tools, contractors tend to “utilize a broader range of BIM solutions” [3]. A challenging issue regarding contractors’ BIM needs is related to implementing both “preconstruction BIM” and “site BIM”. In their study on the strategies for managing 5D BIM adoption, Chan et al. [6] explored, through a case study, contractors’ perspectives regarding the adoption of 5D BIM. The results suggest that contractors have a passive and conservative attitude, with a relatively low openness to change, as it appears to be difficult for them to perceive the possible benefits [6]. Based on various studies, it can be reasonably argued that contractors have a passive and conservative attitude [6] and they prefer to outsource the 3D modeling task, but prefer to keep BIM-based cost management activities in-house [3], Fountain and [11, 21].

Anyway, whether architect or contractors, due to the multiplicity of the proposed solutions and the wide range of the possibilities and associated functionalities, it is challenging for the practitioners to select the right tool adapted to their particular requirements. Previous research works have attempted to solve the issue by proposing some comparison of the existing 5D BIM tools. The most advanced work on the subject is proposed by [37] who proposed a neutral framework to help practitioners in choosing the appropriate 5D BIM solutions. The proposed framework is particularly and is an important milestone toward the resolution of the issue. Unfortunately, the proposed framework did not come in the form of an interactive tool and did not provide the practitioners with the capability of weighting the criterion according to their specific context and needs. In addition, the proposal does not include a

more visual graph, which is very useful for synthesizing the results and making them more accessible to practitioners. Thus, there is still the need for an interactive decision-support tool in the selection of 5D BIM software for variable business contexts.

### **3 Research Methodology**

The work proposed in this project is a continuation of the projects already carried out on the development of a tool to support 5D BIM software selection decision. Vigneault et al. [37] carried out part of his research work on the representation of a table summarizing the main criteria useful for managing the costs of a construction project, as well as a list of 5D BIM software currently available on the market. This information has served as the starting point for our methodological approach. First, the software list and the criteria have been updated. Then, the decision-support tool has been developed and evaluated with practitioners from the industry.

#### ***3.1 Updating the Software List and the Criteria***

The software selected deals with cost management for all phases of a construction project, such as design, construction, layout, and management. The objective is to identify all the specifics and characteristics of these new technologies, in order to derive full descriptions.

The list of criteria is based on scientific literature, including the Project Management Book of Knowledge (PMBok), the Royal Institute of Chartered Surveyors, and the Australian Cost Management Manual, to classify the criteria according to five cost management categories: cost management planning (criteria for sharing information and management tools), cost estimation (criteria for assessing the quality of models, information extraction capacities, and interoperability), budget analysis (criteria for budget comparisons, budget visualization, and Business Intelligence), cost control (cash flow, monitoring of cost trends, and Business Intelligence), as well as complaints and pricing management (criteria related to modifications, changes, payment requests). These categories are then structured into several sub-categories, then into criteria.

#### ***3.2 Development of the Decision-Support Tool***

The research was conducted in such a way that the board evolves and becomes interactive for users. This tool was produced on an Excel file, so that it is possible to include intelligent calculation formulations. The tool is built iteratively, so that all

users can use it freely, while getting the desired information about these 5D software. The interactivity of the tool is measured thanks to the weighting system put in place, allowing the user to judge the criteria and thus have results that are specific to him. Its development within an Excel-type document offers an iterative and re-executable aspect of all parts of the tool.

### ***3.3 Evaluation of the Developed Tool***

Some experts in the field of cost management in the construction industry were contacted, such as a chief estimator, director of estimation, and BIM coordinator and manager. The goal is to have a broad horizon of trades and specialties in order to have as much feedback and comments as possible and to improve the tool and make it as usable as possible. The questions put to the experts focused on compliance with the requirements relating to cost management, the completeness of the list of BIM 5D software, their assessment of the general appearance of the tool, and finally the relevance of the comparison and presentation of results.

## **4 Main Result: An Interactive Decision-Support Tool to Support 5D Software Selection**

In this section, we present the results obtained, through the structure and operation of the tool developed.

### ***4.1 Structure of the Proposed Tool***

The tool is made up of four parts corresponding to four Excel spreadsheet tabs, to guide the user in handling it (Fig. 1).

The first tab concerns the presentation and gives the first steps to follow to use the tool correctly. The second tab corresponds to the presentation of the structural table of criteria according to the categorization carried out and proposes a definition for each of them in order to have a homogeneous understanding. The user is required to fill in the weighting column for each of the criteria in this tab in order to subsequently observe the results. The third tab represents the main complete table of the tool



**Fig. 1** Four tabs of the interface (*in French*)



comprising the list of criteria, 5D software, and the transcription of the weighting performed in the previous step. This table displays all the detailed results for each software based on the user’s choices. The fourth tab details all the results in the form of graphs, allowing you to better visualize the results obtained from the software, by displaying only totals and sub-totals. The user therefore has access to a summary of the results in a precise and visual manner.

4.2 Operation of the Tool

The purpose of this tool is to interact with the user and give them results that are unique to them. To do this, a weighting system has been put in place. When filling out the tool, a score is assigned to each software for each criterion. The interactivity of the tool developed in this project comes from the weighting and therefore on the user’s participation to obtain results. Thus, the tool offers a weighting system, chosen by the user according to his preferences and the importance he places on each of the criteria. The weighting is free, and each user can fill in what he wants.

The criteria weighting interface is in the form of a table showing all the criteria identified, grouped by category and sub-category. A definition is proposed for each criterion to ensure a common understanding and to avoid any ambiguity or confusion on the part of the user (Fig. 2).

When the weighting is completed by the user, the table is automatically filled to give the results of the information. The table is constructed in such a way that the weighting will multiply with the software score for each of the criteria and a synthesis for the different categories corresponding to the 5D BIM uses is provided (Fig. 3). The results of the tool appear when the user has finished filling in the criteria weighting. They come in the form of detailed notes, sub-totals, totals, rankings, and finally graphs to give a visual aspect to the user. The detailed notes allow the user

Catégorie	Sous-catégorie	Critères	Définition	Pondération
Analyse budgétaire	Comparaison budgétaire	Comparaison des coûts de plusieurs éléments de même type	Capacité du logiciel à comparer lui-même les coûts de plusieurs éléments de même nature, afin de proposer à l'utilisateur de "meilleures" solutions.	
		Comparaison des quantités relevées manuellement et automatiquement	Capacité du logiciel à permettre une comparaison manuelle des quantités relevées grâce aux logiciels de conception et à souligner les différences.	
	Visualisation budgétaire	Rapports sur les coûts modifiables	Capacité du logiciel à sortir des rapports sur les coûts du projet en fonction de ce que l'utilisateur a choisi d'estimer, avec la possibilité pour l'utilisateur de modifier ces rapports par la suite.	
	Business intelligence	Solutions budgétaires proposées par le logiciel avec la base de données	Capacité du logiciel à comparer les solutions budgétaires obtenues après avoir comparé les bases de données et les choix de l'utilisateur, et à proposer ainsi une solution "meilleure" que l'utilisateur peut, ou non, accepter.	
Contrôle des coûts	Flux de trésorerie	Analyse des flux de trésorerie	Capacité du logiciel à analyser en temps réel les flux de trésorerie, permettant à l'utilisateur de contrôler ses coûts.	
		Prévision des flux de trésorerie	Capacité du logiciel à générer un flux de trésorerie à partir de modèles 4D	
	Évolution des coûts et suivi	Évolution des coûts tout au long du cycle de vie du projet	Capacité du logiciel à suivre l'évolution des coûts durant toute la durée du cycle de vie du projet.	
		Valeur des activités complétées	Capacité du logiciel à calculer et mettre à jour les coûts de toutes les activités complétées lors de la construction.	
		Évaluation de la performance des coûts	Capacité du logiciel à évaluer la performance des coûts, c'est-à-dire la différence (ou le ratio) entre le budget planifié et les coûts réels du projet.	
		Rapports personnalisables de suivi des coûts	Capacité du logiciel à réaliser le suivi des coûts tout au long du projet, sur les quelques aspects évoqués dans cette catégorie. Ces rapports sont personnalisables et peuvent être modifiés par l'utilisateur.	
Système d'alerte de dépassement des coûts		Capacité du logiciel à prévenir l'utilisateur lorsque l'estimation des coûts du projet a dépassé les objectifs de ce dernier.		
➤ Présentation <b>Pondération des critères</b> Tableau pondéré Résultats graphiques ➤				

Fig. 2 User criteria weighting interface (in French)

	Assemble	Beel Manager	BIM4YOU	CATO Suite	CostX	Cubicox Suite	Cubit	Destini Estimator	Ideate BI
	Assemble Systems	Beel Manager	BRZ France	Causeway	Exactal	Glodon	Buildsoft	Beek Technology	Ideate Sor
Sous-total "Plan de gestion des coûts"	11	4	12	11	11,5	8,5	4	9	6,5
Sous-total "Estimation des coûts"	34	44	55	44	38	28	32,5	26	5
Sous-total "Budgetisation des coûts"	4	10	7	7	10	10	10	11	4
Sous-total "Contrôle des coûts"	3,5	17	22	20	8	15	2	20	2
Sous-total "Gestion des réclamations et des tarifications"	7	7		7	8	0	0	7	0
TOTAL GÉNÉRAL	59,5	82	104	89	75,5	61,5	48,5	73	17,2

Sous-total "Plan de gestion des coûts"	44,00%	16,00%	48,00%	44,00%	46,00%	34,00%	16,00%	36,00%	26,00%
Sous-total "Estimation des coûts"	34,00%	44,00%	55,00%	44,00%	38,00%	28,00%	32,50%	26,00%	5,00%
Sous-total "Budgetisation des coûts"	13,33%	33,33%	23,33%	23,33%	33,33%	33,33%	33,33%	36,67%	13,33%
Sous-total "Contrôle des coûts"	6,36%	30,91%	40,00%	36,36%	14,55%	27,27%	3,64%	36,36%	3,64%
Sous-total "Gestion des réclamations et des tarifications"	23,33%	23,33%	25,67%	23,33%	26,67%	0,00%	0,00%	23,33%	0,00%
TOTAL GÉNÉRAL PONDERÉ	24,79%	34,17%	43,33%	37,08%	31,40%	25,63%	20,21%	30,42%	7,29%

Fig. 3 Example of results by sub-category for a specific weighting (in table format) (in French)

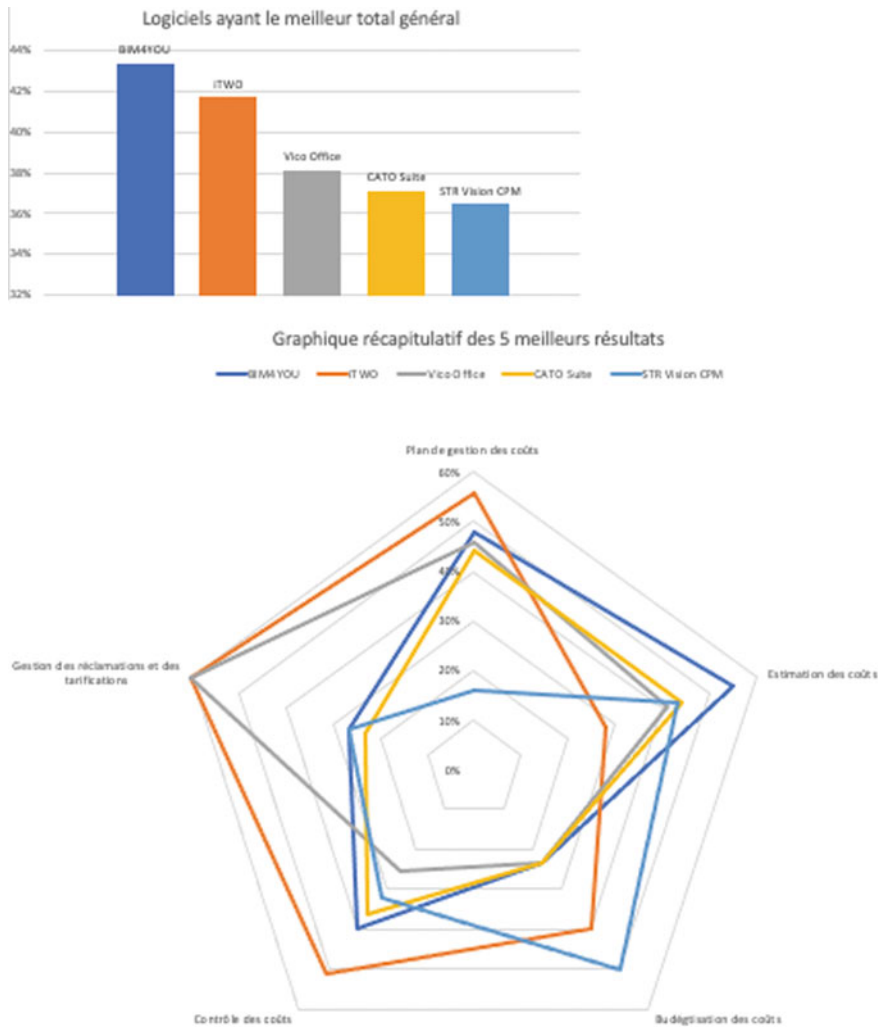
to observe with precision the data of each software. These scores are then added together and give a sub-total. Each of these sub-totals corresponds to a category of criteria. The averages of the sub-totals of all software are then calculated, and the implementation of a visual indicator allows the user to have a global overview of the software sub-totals.

The tool also offers a graphical presentation of the results, so that users can better visualize the results. It takes the form of a bar graph and radar, generated after the criteria have been weighted. The bar chart automatically ranks software according to the total score obtained for all criteria. While this helps to get a general view of software capacity, it is generally necessary to have results for different sub-categories. Radar corrects this lack and allows the user to visualize the strengths and weaknesses of top software according to the different cost management uses (Fig. 4).

5 Conclusion and Future Work

The article presented an interactive decision-support tool in the selection of 5D BIM software for variable business contexts. The aim is to help the practitioners from the construction industry to select the right tools adapted to their particular requirements. The validation carried out with practitioners in the field shows an interest in the tool and the relevance of the criteria and the weighting principle. It also shows that the tool can be improved and made accessible to the entire industry.

To achieve this, it is essential to switch from its current form of Excel workbook to an online tool allowing wider use, but also a permanent update of the list of tools and criteria, directly by the community. Future work will focus on the development of an online version capable of reaching a large audience and a larger scale evaluation of the proposed tool.



**Fig. 4** Example of results by sub-category for a specific weighting (in graphical format) (*in French*)

**References**

1. Agostinelli S, Cinquepalmi F, Ruperto F (2019) 5D BIM: tools and methods for digital project construction management. *Build Inf Modell (BIM) Des Construct Oper III* 1:205–215
2. Aibinu A, Venkatesh S (2014) “Status of BIM adoption and the BIM experience of cost consultants in Australia. *J Prof Issues Eng Edu Pract* 140(3)
3. Becerik-Gerber B, Rice S (2010) The perceived value of building information modeling in the U.S. building industry. *J Inf Technol Construct* 15(February):185–201
4. Boton C (2013) Conception de vues métiers dans les collecticiels orientés service: Vers des multi-vues adaptées pour la simulation collaborative 4D/nD de la construction. Université de Lorraine

5. Boton C, Forgues D (2018) Practices and processes in BIM projects: an exploratory case study. *Adv Civil Eng* 2018:1–12
6. Chan IYS, Liu AMM, Chen B (2018) Management strategies for 5D-BIM adoption in Hong Kong. In: *Proceedings of the 21st international symposium on advancement of construction management and real estate*, 2016, (209889), pp 1023–1039
7. Chang SE, Shinozuka M (1996) Life-cycle cost analysis with natural hazard risk. *J Infrastruct Syst* 2(3):118–126
8. Ding Z, Zuo J, Wu J, Wang JY (2015) Key factors for the BIM adoption by architects: a China study. *Eng Constr Archit Manag* 22(6):732–748
9. Doloi H (2012) Cost overruns and failure in project management: understanding the roles of key stakeholders in construction projects. *J Constr Eng Manag* 139(3):267–279
10. Eastman C, Teicholz P, Sacks R, Liston K (2011) *BIM handbook: a guide to building information modeling for owners, managers, designers, engineers and contractors*
11. Fountain J, Langar S (2018) Building information modeling (BIM) outsourcing among general contractors. *Autom Construct* 95(February 2017):107–117
12. Frimpong Y, Oluwoye J, Crawford L (2003) Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study. *Int J Project Manage* 21(5):321–326
13. Fuller S (2006) Life-cycle cost analysis (LCCA)
14. Gibbs D-J, Emmitt S, Ruikar K, Lord W (2013) An investigation into whether building information modelling (BIM) can assist with construction delay claims. *Int J 3-D Inf Model* 2(1):45–52
15. El Hawary AN, Nassar AH (2015) The effect of building information modeling BIM on construction claims. *Int J Sci Technol Res* 4(8):25–33
16. Herszon L (2017) *The complexity of projects : an adaptive model to incorporate complexity dimensions into the cost estimation process*. University of Huddersfield
17. Kehily D, McAuley B, Hore A (2013) Leveraging whole life cycle costs when utilising building information modelling technologies. *Int J 3-D Inf Model* 1(4):40–49
18. Kehily D, Underwood J (2017) Embedding life cycle costing in 5D BIM. *J Inf Technol Construct (ITcon)* 22(22):145–167
19. Kerzner HR (2013) *Project management: a systems approach to planning, scheduling, and controlling*. Wiley
20. Kim H, Grobler F (2013) Preparing a construction cash flow analysis using building information modeling (BIM) technology. *KICEM J Construct Eng Project Manage* 704:1–9
21. Ku K, Taiebat M (2011) BIM experiences and expectations: the constructors' perspective. *Int J Constr Educ Res* 7(3):175–197
22. Lee XS, Tsong W, Khamidi MF (2016) 5D building information modelling—a practicability review. In: *The 4th international building control conference 2016 (IBCC 2016)*, pp 2–7
23. Levin P (1998) *Construction contract claims changes & dispute resolution*. ASCE Press
24. Marcos CL (2017) BIM implications in the design process and project-based learning: comprehensive integration of BIM in architecture. *WIT Trans Built Environ* 169:101–110
25. Marzouk M, Azab S, Metawie M (2018) BIM-based approach for optimizing life cycle costs of sustainable buildings. *J Clean Prod* 188:217–226
26. Mayouf M, Boyd D (2019) Activity-based vs. object-based scheduling in construction: a phenomenological study of the potential shift of construction scheduling process. In: *CIB world building congress 2019, Hong Kong SAR, China*, pp 1121–1131
27. McGraw Hill Construction (2014) *The business value of BIM in Australia and New Zealand : SmartMarket report* managing editor. SmartMarket report
28. Mitchell D (2012) 5D BIM: creating cost certainty and better buildings. In: *RICS Cobra conference*, pp 1–10
29. Olusola BS, Srinath P, Damilola E, Esther AT (2019) An investigation into BIM-based detailed cost estimating and drivers to the adoption of BIM in quantity surveying practices. *J Financ Manage Property Construct* 25(1):61–81

30. Project Management Institute (2017) A guide to the project management body of knowledge. Project Management Institute Inc
31. Sacks R, Eastman C, Lee G, Teicholz P (2018) BIM handbook: a guide to building information modeling for owners, designers, engineers, contractors, and facility managers. Wiley, Hoboken
32. Smith D, Lovegrove S, Muse A, Pan DDZ, Sawhney A, Watkins P, Whisson G, Seah Kwee Yong T (2015) BIM for cost managers: requirements from the BIM model. RICS guidance note
33. Smith P (2014) BIM & the 5D project cost manager. *Proc Soc Behav Sci* 119:475–484
34. Stanley R, Thurnell D (2014) The benefits of, and barriers to, implementation of 5D BIM for quantity surveying in New Zealand. *Australas J Construct Econ Build* 14(1):105–117
35. Thurairajah N, Goucher D (2013) Advantages and challenges of using BIM: a cost consultant's perspective. In: 49th ASC annual international conference proceedings
36. Verbeeten FHM (2011) Public sector cost management practices in The Netherlands. *Int J Public Sect Manag* 24(6):492–506
37. Vigneault M-A, Botton C, Chong H-Y, Cooper-Cooke B (2020) An innovative framework of 5D BIM solutions for construction cost management: a systematic review. *Arch Comput Methods Eng* 27(4):1013–1030
38. Woodward DG (1997) Life cycle costing—theory, information acquisition and application. *Int J Project Manage* 15(6):335–344

# **Construction Management: Water Resources**

# Machine Learning Approximation for Rapid Prediction of High-Dimensional Storm Surge and Wave Responses



Saeed Saviz Naeini and Reda Snaiki

**Abstract** Storm surge and waves are responsible for a substantial portion of the tropical and extratropical cyclones-induced damage in coastal areas of the USA and Canada. High-fidelity, numerical models can provide accurate simulation results of the water elevation, where a hydrodynamic model (e.g., ADCIRC) is coupled with a wave model (e.g., SWAN). However, they are computationally expensive, hence cannot be employed as part of an early warning system for urban flooding hazards or implemented in probabilistic tropical and extratropical cyclones' risk assessment. In this study, an alternative and efficient approach is proposed based on hybrid machine learning approaches. First, a dimensionality reduction technique based on deep autoencoder is developed to encode the spatial information in a reduced state space. Then, a machine learning-based model is developed in the latent space to predict the maximum surge and significant wave height. The latent space is then decompressed back to the original high-dimensional space using the decoder. The high-fidelity data are retrieved from the North Atlantic Comprehensive Coastal Study (NACCS), released by the US Army Corps of Engineers. Due to its high efficiency and accuracy, the proposed methodology can be employed to analyze the impact of input uncertainties on the simulation results. Four machine learning algorithms are used to predict the maximum surge and significant wave height including artificial neural network (ANN), support vector regression (SVR), gradient boosting regression (GBR), and random forest regression (RFR). The coupled autoencoder-ANN model for the prediction of the storm surge (significant wave height) outperformed all other algorithms with a coefficient of determination  $R^2$  of 0.953 (0.921) for the testing set. In addition, the comparison between deep autoencoder and the widely used principal component analysis (PCA) technique indicated the superior performance of the former since it is able to accurately capture the inherent nonlinearities within the data.

---

S. Saviz Naeini · R. Snaiki (✉)

École de Technologie Supérieure, Université du Québec, Quebec City, Québec, Canada

e-mail: [reda.snaiki@etsmtl.ca](mailto:reda.snaiki@etsmtl.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022, Lecture Notes in Civil Engineering* 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_43](https://doi.org/10.1007/978-3-031-34593-7_43)

**Keywords** Storm surge · Significant wave height · Machine learning · Deep autoencoder · Principal component analysis

## 1 Introduction

Throughout the last several years, natural disasters have dramatically and frequently exerted a far-reaching impact on communities and caused widespread damages [11]. Despite the recent developments in damage mitigation of tropical and extratropical cyclones, recent events (e.g., Hurricane Irma 2017; Harvey 2017; and Dorian 2019) have shown that coastal regions are still highly vulnerable to their effects [10]. The east coast of the USA and Canada has been specifically impacted by several tropical and extratropical storms. This region might be subjected to an increasing storm activity due to the effects of climate change and sea level rise (SLR) [14]. Hence, accurate and efficient models for the prediction of storm surge and waves are of great importance since they can be employed as part of an early warning system for urban flooding hazards or implemented in probabilistic tropical and extratropical cyclones' risk assessment [6].

Physics-based high-fidelity numerical models such as ADvanced CIRCulation (ADCIRC, [7] for the prediction of storm surge and Simulating WAVes Nearshore (SWAN [4]) for modeling the waves have been widely implemented in several engineering applications. However, the high computational cost of these models prevents their implementation in real-time forecasting and probabilistic hazard assessment platforms [6]. Therefore, data-driven models have been alternatively proposed to efficiently predict storm surge [5]. Several surrogate models have been already developed to simulate storm surge (e.g., feedforward neural network, support vector regression, and kriging models). However, most of these studies were applied to limited number of geographical locations in which the model training would be feasible. Therefore, though the prediction has brought impressive accuracy over few selected points, those models cannot be applied to an extensive coastal region that has numerous geographic locations [1, 13]. Considering the importance of storm surge and wave predictions over larger geographical areas, some studies employed the principal component analysis (PCA) technique to reduce the dimensionality related to the output space. For instance, [5] developed a kriging-based model and coupled it with PCA to predict storm surge. Other studies have also attempted to couple PCA with several other data-driven techniques (e.g., [2]). However, the PCA technique is essentially adapted for linear problems, hence might not accurately capture the inherent nonlinearities within the data.

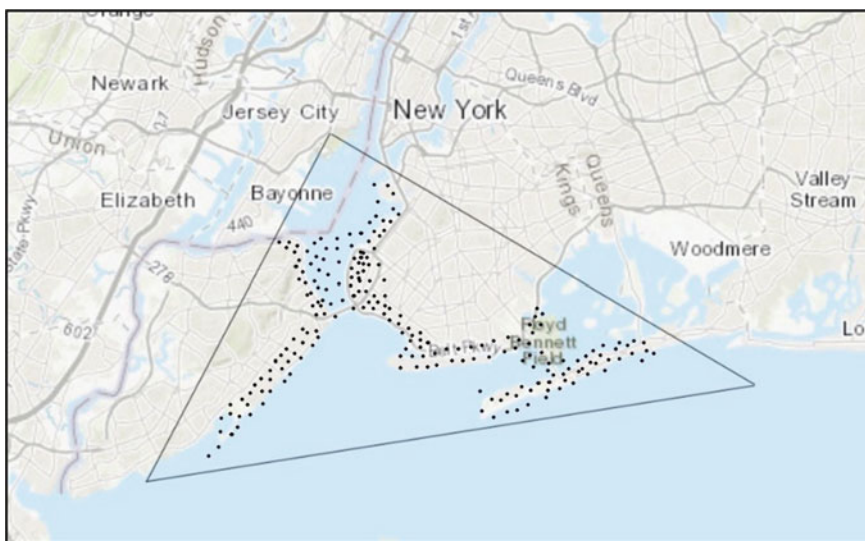
In this study, an advanced dimensionality reduction approach, namely deep autoencoder, is developed to encode the characteristics of geographical points in the latent space, and then, several machine learning algorithms are used in the latent space to predict the maximum surge and significant wave height. Moreover, a decoder is used to decompress the latent space to the original space. To highlight the superior performance of deep autoencoder, it will be compared with the



PCA technique. Four data-driven algorithms are used with the dimensionality reduction techniques, namely feedforward artificial neural network (ANN), support vector regression (SVR), random forest regression (RFR), and gradient boosting regression (GBR).

## 2 Data

The training/testing data are retrieved from the North Atlantic Comprehensive Coastal Study (NACCS) which was conducted by the US Army Corps of Engineers. This database contains the simulation results corresponding to 1050 synthetic tropical cyclones and historical extratropical cyclones using numerical models including ADvanced CIRCulation (ADCIRC, [7], WAVE Model (WAM), and Steady State Spectral WAVE [8]. In this study, the aforementioned database is used for the prediction of storm surge and waves (through the significant wave height) for both tropical and extratropical cyclones. The data went through a data processing stage which resulted in 1031 synthetic storms. The model inputs consist of six storm parameters, namely latitude, longitude, heading direction, central pressure, radius of maximum winds, and translation speed. The outputs consist of the peak storm surge and significant wave height at 159 geographic locations New York City as indicated in Fig. 1.



**Fig. 1** Geographical distribution of the save points in New York City in which storm surge and significant wave height will be predicted

### 3 Model Development

The general framework of the proposed predictive model is elaborated in Fig. 2. The model inputs consist of six storms' parameters and the outputs are either the storm surge or significant wave height (or both) at the 159 save points located in NYC (Fig. 1). Considering the high dimension of the output space, a deep autoencoder method was applied to the dataset in order to reduce its high dimension. The encoder part of the deep autoencoder takes the storm surge/significant wave height at the 159 save points as inputs and has additional five hidden layers. An additional layer which represents the latent space has two neurons. The decoder has a similar architecture (but inversed) as the encoder (8–16–32–64–128–159). The deep autoencoder model was implemented using Keras with a TensorFlow backend. Four different data-driven models were trained in the latent space. The data were also split into three sets containing the training set (80% of the data), the testing set (10% of the data), and the validation set (10% of the data). A brief presentation of the data-driven models will be introduced in the following sections.

#### 3.1 *Artificial Neural Network*

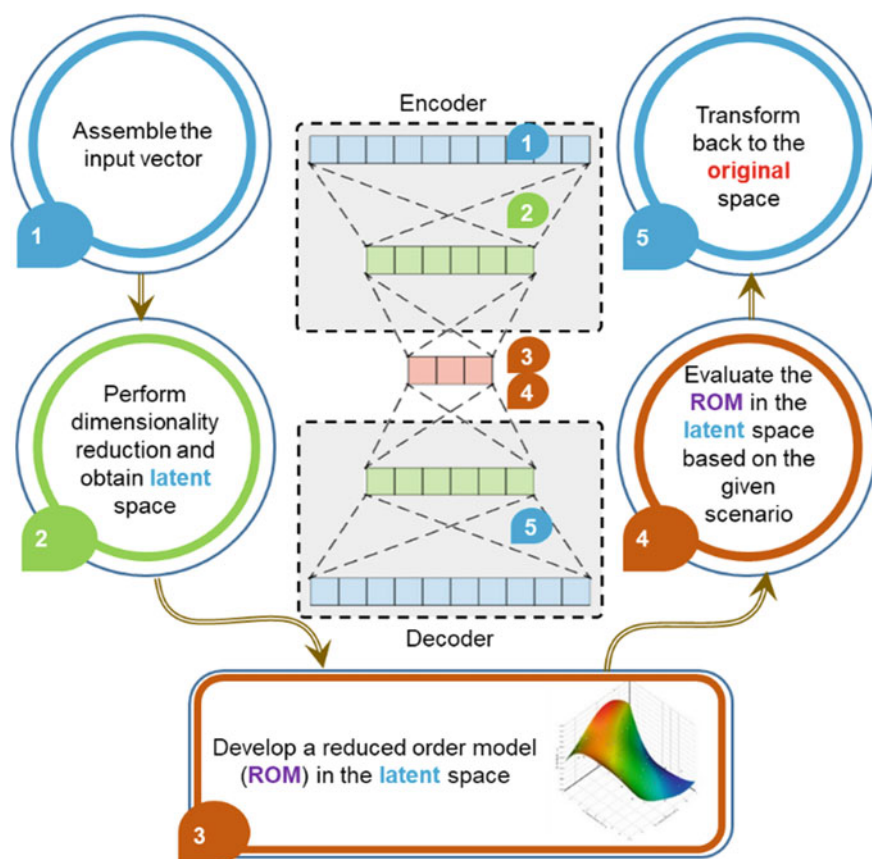
Artificial neural network (ANN) is one of the widely used algorithms for building predictive models. It contains several nodes or neurons which are connected to each other. The typical ANN architecture has an input layer, several hidden layers, and an output layer [3, 13]. In this study, the final ANN architecture was obtained based on a trial-and-error approach (five hidden layers, each having 64 neurons with a ReLU activation function).

#### 3.2 *Support Vector Regression*

Support vector regression (SVR) is a non-parametric statistical algorithm which tries to find a functional relationship between the input matrix and output (response) matrices based on kernel functions. Kernels are linear or nonlinear functions which shape the hyperplane and decision boundary. Several kernels can be employed depending on the problem, including linear, polynomial, and sigmoid functions.

#### 3.3 *Gradient Boosting Regression*

Gradient boosting regression (GBR) is a machine learning technique that is implemented for regression analysis problems. This ensemble algorithm creates at first



**Fig. 2** Schematic of the proposed hybrid data-driven model

several weak models which are typically decision trees and then combine each weak model to build a comprehensive model that has improved accuracy [15].

### 3.4 Random Forest Regression

Random forest regression (RFR) is a supervised learning algorithm that is based on the ensemble learning technique. This algorithm creates several decision trees in the training process and calculates the mean of the classes as the prediction of the whole decision tree. Then, it selects the model which performs the best [9].

## 4 Model Performance

In this study, the coefficient of determination  $R^2$  and the mean squared error (MSE) are used to evaluate the performance of the predictive models. The coefficient of determination  $R^2$  evaluates the scatter of the data points around the fitted regression line and is defined as:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (1)$$

where RSS is the sum of squared residuals, and TSS is the total sum of squares.  $R^2$  value is between 0 and 1, where 0 represents a poor predictive model and 1 represents a robust predictive model.

The MSE measures the average of the squares of the errors and is given as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

where  $Y_i$  is the vector of observed values,  $\hat{Y}_i$  is the vector of predicted values, and  $n$  is number of predicted data points.

## 5 Simulation Results

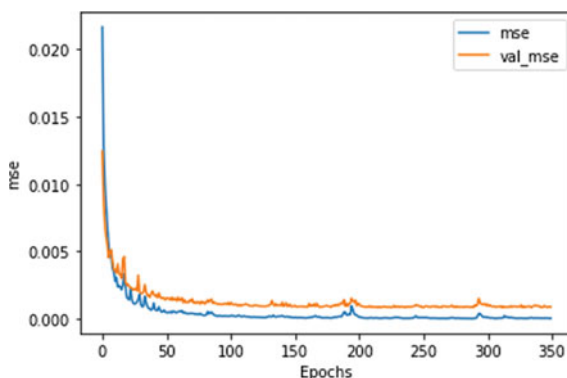
### 5.1 Deep Autoencoder-Based Models

As noted before, 80% of the data were selected as training set, 10% as testing set, and 10% as validation set. In addition, the Bayesian optimization technique was employed to find the best model hyperparameters. Specifically, Bayesian optimization identifies the minimum (or maximum) of an objective function and builds a probability model of the objective function which allows for the convenient selection of the best hyperparameters for the model [12].

The obtained  $R^2$  score for the ANN-based model predicting storm surge was 0.988 for the training, 0.953 for testing, and 0.927 for validation. The MSE results are indicated in Fig. 3 where 350 epochs have been selected. Similarly, the  $R^2$  score for the ANN model predicting the significant height was 0.964 for the training set, 0.921 for the testing set, and 0.917 for the validation set. The obtained results indicate that a good prediction performance has been achieved with the hybrid deep autoencoder-based ANN model for both storm surge and significant wave height.

Figure 4 shows the scatter plot of the predicted storm surge and their corresponding real values (a similar plot can be generated for the significant wave height) using deep

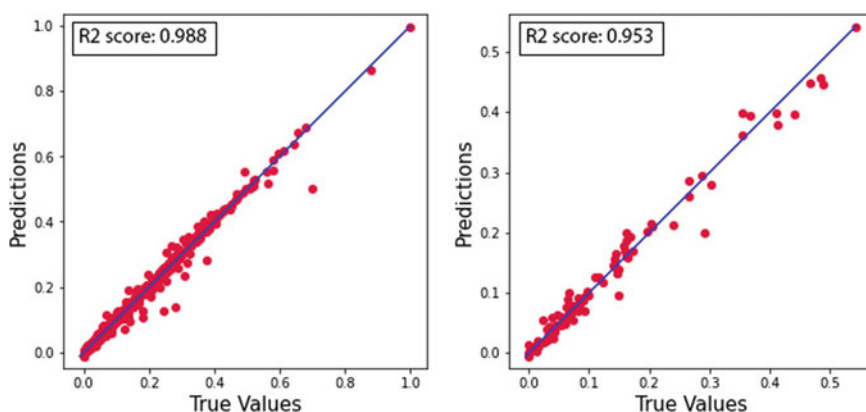
**Fig. 3** MSE results for the ANN model trained over the deep autoencoder-generated latent space



autoencoder as a dimensionality reduction technique. The corresponding  $R^2$  score is also indicated in the plots.

The simulation results of the predicted storm surge, in terms of the  $R^2$  score, corresponding to the selected machine learning models (i.e., SVR, GBR and FRF) combined with deep autoencoder are summarized in Table 1. Based on the  $R^2$  scores, it is clear that the deep autoencoder-based ANN model (followed by RFR model) has outperformed all other machine learning techniques in predicting storm surge values. In addition, the SVR algorithm has the lowest accuracy for the training, testing, and validation sets.

The simulation results of the significant wave height, in terms of the  $R^2$  score, corresponding to the selected machine learning models (i.e., SVR, GBR, and FRF) combined with deep autoencoder are summarized in Table 2. Similarly, the ANN model outperformed all other machine learning algorithms followed by RFR, then GBR and SVR.



**Fig. 4** Scatter plot of the simulated storm surge data and their corresponding real values using the hybrid deep autoencoder-based ANN model for both training (left) and testing sets (right)

**Table 1** Comparison of the simulation results of the predicted storm surge using the selected machine learning algorithms trained in the autoencoder-based latent space

Algorithms	Train $R^2$ score	Test $R^2$ score	Validation $R^2$ score
ANN	0.988	0.953	0.927
SVR	0.937	0.919	0.718
GBR	0.965	0.934	0.846
RFR	0.984	0.928	0.864

**Table 2** Comparison of the simulation results of the predicted significant wave height using the selected machine learning algorithms trained in the autoencoder-based latent space

Algorithms	Train $R^2$ score	Test $R^2$ score	Validation $R^2$ score
ANN	0.964	0.921	0.917
SVR	0.934	0.852	0.736
GBR	0.936	0.870	0.749
RFR	0.980	0.905	0.782

5.2 *PCA-Based Models*

To assess the performance of the deep autoencoder, it will be compared with the widely employed using principal component analysis (PCA) technique. For fair comparison, the same architecture and hyperparameters of the trained machine learning algorithms are used for the PCA-based models. Two principal components were required to explain 95% of the total variance in the model for the storm surge prediction. The simulation results corresponding to the storm surge prediction are summarized in Table 3.

**Table 3** Comparison of the simulation results of the predicted storm surge using the selected machine learning algorithms trained in the PCA-based latent space

Algorithms	Train $R^2$ score	Test $R^2$ score	Validation $R^2$ score
ANN	0.993	0.932	0.929
SVR	0.966	0.856	0.763
GBR	0.944	0.858	0.777
RF	0.970	0.842	0.790

In general, it can be concluded that, although not pronounced, the deep autoencoder results are better than those generated using the PCA technique. Similar results are obtained for the significant wave height prediction as indicated in Table 4.

It should be noted that with the increasing number of save points, the involved nonlinearities will be more pronounced; therefore, it is expected that the deep autoencoder-based model will perform even better than the PCA-based model.

**Table 4** Summary of the performance of each algorithm with the implementation of PCA for the prediction of the significant wave height

Algorithms	Train $R^2$ score	Test $R^2$ score	Validation $R^2$ score
ANN	0.945	0.903	0.902
SVR	0.831	0.812	0.778
GBR	0.969	0.885	0.846
RF	0.980	0.870	0.845

## 6 Conclusion

In this study, several machine learning models, including ANN, SVR, GBR, and RFR, were used to predict the storm surge and significant wave height at an extended geographical region. The necessary training/testing data were retrieved from the NACCS database. Due to the high dimensionality of the output space, a dimensionality reduction technique, namely deep autoencoder, was developed to encode the spatial information in a reduced state space. The simulation results indicated that the deep autoencoder-based ANN model outperformed all other machine learning techniques and the coefficient of determination ( $R^2$ ) for the testing set in predicting the storm surge was 0.953, and the  $R^2$  score of the coupled autoencoder-ANN model for the prediction of the significant wave height was 0.921 for the testing set. In addition, the SVR algorithm had the lowest accuracy for the training, testing, and validation sets. Furthermore, the comparison between deep autoencoder and the widely used PCA technique indicated the superior performance of the former since it is able to accurately capture the inherent nonlinearities within the data. Due to its efficiency and accuracy, the proposed model can be utilized as part of an early warning system for urban flooding hazards or implemented in probabilistic tropical and extratropical cyclones' risk assessment.

## References

1. Al Kajbaf A, Bensi M (2020) Application of surrogate models in estimation of storm surge: a comparative assessment. *Appl Soft Comput* 91:106184
2. Bass B, Bedient P (2018) Surrogate modeling of joint flood risk across coastal watersheds. *J Hydrol* 558:159–173
3. Hashemi MR, Spaulding ML, Shaw A, Farhadi H, Lewis M (2016) An efficient artificial intelligence model for prediction of tropical storm surge. *Nat Hazards* 82(1):471–491
4. Booij N, Holthuijsen LH, Ris RC (1997) The SWAN wave model for shallow water. In: *Coastal engineering 1996*, pp 668–676
5. Jia G, Taflanidis AA (2013) Kriging metamodeling for approximation of high-dimensional wave and surge responses in real-time storm/hurricane risk assessment. *Comput Methods Appl Mech Eng* 261:24–38
6. Lee JW, Irish JL, Bensi MT, Marcy DC (2021) Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. *Coast Eng* 170:104024

7. Luetlich RA, Westerink JJ, Scheffner NW (1992) ADCIRC: an advanced three-dimensional circulation model for shelves, coasts, and estuaries. Report 1, theory and methodology of ADCIRC-2DDI and ADCIRC-3DL
8. Smith JM, Sherlock AR, Resio DT (2001) STWAVE: steady-state spectral wave model user's manual for STWAVE, version 3.0. Engineer Research and Development Center Vicksburg MS Coastal and Hydraulicslab.
9. Smith PF, Ganesh S, Liu P (2013) A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J Neurosci Methods* 220(1):85–91
10. Snaiki R, Wu T, Whittaker AS, Atkinson JF (2020) Hurricane wind and storm surge effects on coastal bridges under a changing climate. *Transp Res Rec* 2674(6):23–32
11. Wang Z, Ye X (2018) Social media analytics for natural disaster management. *Int J Geogr Inf Sci* 32(1):49–72
12. Wu J, Toscano-Palmerin S, Frazier PI, Wilson AG (2020) Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In: *Uncertainty in artificial intelligence*. PMLR, pp 788–798
13. Wu YC, Feng JW (2018) Development and application of artificial neural network. *Wireless Pers Commun* 102(2):1645–1656
14. Zhang K, Douglas BC, Leatherman SP (2000) Twentieth-century storm activity along the US east coast. *J Clim* 13(10):1748–1761
15. Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. *Transp Res Part C Emerg Technol* 58:308–324



# Predicting Pumps-as-Turbine Characteristics with the Use of Machine Learning Applications



Alex Brisbois and Rebecca Dziedzic

**Abstract** The water–energy–carbon nexus is crucial to the consideration of renewable energies and sustainable developments when analyzing the intensive requirements of water services. Installing a turbine in key areas of water distribution networks to recover the wasted energy proves to be beneficial to counter waste such as leakages in pipes and pressure reductions. Previous studies have proposed models to predict characteristics of pumps as turbines (PaT) based on pump best efficiency point. These models apply statistical methods to determine the turbine characteristics, but accuracy in these models is still questionable. Data science and machine learning applications have become increasingly popular to show correlations in datasets and offer deeper insight in predicting the relationships of a PaT in pump and turbine modes. The goal of this study is to apply machine learning algorithms to model characteristic curves of PaTs. A database of over 140 PaT experimental records was compiled to compare machine learning results to that of the predicted values from the established models. Considering the limited number of features and data points, regression models were developed, including elastic-net, ridge regression, support vector regression, etc. The predicted results were evaluated based on the coefficient of determination ( $R^2$ ) as well as error according to the root mean squared error (RMSE) and mean absolute deviation (MAD). The results show the proposed method to be more reliable with smaller margins of error in predicting PaT characteristics than previous studies.

**Keywords** Pumps • Turbine characteristics

---

A. Brisbois (✉) • R. Dziedzic  
Concordia University, Quebec, Canada  
e-mail: [alex.brisbois@mail.concordia.ca](mailto:alex.brisbois@mail.concordia.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_44](https://doi.org/10.1007/978-3-031-34593-7_44)

711

## 1 Introduction

As the world starts to recognize the need for sustainable practices regarding the way humans consume energy, a significant amount of effort should be focused toward minimizing waste of energy. Municipalities need energy for the services they provided to their citizens, and one of the more important services is to provide water. Providing water is extremely energy intensive considering that the water needs to be extracted, filtered, then distributed. A significant amount of energy is also associated to the disposal of water once we are done with it. McNabola mentions that the water sector consumes an estimated 305–60% of expenses of municipalities is associated to water industry [8]. With this consideration of how energy intensive the water sector is, analyzing and reducing the wasted energy in the water distribution network are gaining popularity.

Water distribution networks distribute the water within a municipality and these systems can differ significantly all according to its geographical characteristics. Within the system, if the pressure is too high and needs to be reduced, a pressure reducing valve (PRV) is usually inserted to control the pressure. These PRVs dissipate the pressure into through friction, and therefore, its energy potential is wasted. At these locations, micro-hydro turbines can replace these PRVs to not only reduce the pressure in the system but to also recover some of the wasted energy. The only negative aspect to this approach is that the micro-hydro turbines can be expensive. An alternative to these micro-hydro turbines would be pumps-as-turbines (PaTs), which is the practice of inserted a pump in a reverse orientation and to couple it with a generator so that it acts as a turbine. The only drawback to this method is that the manufacturers do not provide the characteristic attributes of the pumps in reverse mode and therefore require expensive testing of the PaTs to understand the capability of the PaT. Previous researchers have proposed equations to relate the PaTs' characteristics in pump mode to turbine mode but with varying capabilities in terms of accuracy.

Recently, data science and machine learning practices have proven to show a wide range of applications in terms of predicting numerical attributes. Considering that relating the pump-mode characteristics to the turbine-mode characteristics of a PaT is purely geometrical, an assumption that correlations between the two can be made. Machine learning libraries are extensive in their lists for regressions, and comparing them to one another can be made instantaneously. Rossi has applied this methodology using neural networks and has shown great potential in terms of accuracy against the previous research [15]. A step further in terms of a wider comparison of regressions and parameter optimization can potentially be superior in choosing a PaT for energy recovery in water distribution networks.

The goal of this study is to commission a library of machine learning algorithms to libraries of PaTs at its BEP and characteristic curves to expose their accuracies in predicting their turbine-mode characteristics. Furthermore, both libraries will be considered in their dimensioned and dimensionless natures to assess the difference in accuracy and correlation.

## 2 Literature Review

### 2.1 Energy Recovery Potential of PaTs

Several studies have been conducted to assess the potential energy harvest in water distribution networks around the world with a particular popularity in Europe. Mitrovic pursued a study showing the potential energy capture associated to the drinking water, wastewater, and irrigation networks in multiple northeastern European countries [9]. The study specifically analyzed the potential energy capture locations in the networks as well as the potential energy capture from the provided databases which showed a potential recovery of 17,944 kW for a total extrapolated value of 548.2 GWh.

Algieri conducted a study of potential energy capture and an analysis on the costs of PaTs against micro-hydro turbines [2]. His study specifically dealt with the energy capture of the irrigation system in the peninsula of Calabria in Southern Italy demonstrating a potential recovery of 4193.2 kW for a total of 21,140.6 MWh. The study also showed that the mean cost per kilowatt of a PaT is 532 (Euro/kW) against 1926 (Euro/kW) for a micro-hydro turbine.

PaTs having a low cost and the same harness capability of a micro-hydro turbine have rapid payback periods. Stefanizzi researched the application of a PaT to replace the location of a PRV and summarized all the costs associated to the installation of a PaT as well as the surplus of funds from recovering energy at that location [19]. At a conservative outlook, the payback period was estimated to be 1.93 years with a reduction of CO<sub>2</sub> emissions at an equivalency of 1.3 ton CO<sub>2</sub>/km per automobile.

### 2.2 Parameters and Dimensionless Parameters

PaTs have specific parameters to measure its capabilities in a system: the impeller diameter, the rotational speed of the impeller, the flow rate of water, the pressure inside the PaT represented as the hydraulic head, the power output of the PaT, and the efficiency associated to the PaT. The specific speed is also a defining parameter since it considers the flow, head as well as the rotational speed to be represented as a unique characteristic which is an identifier regardless of typology and impeller diameter. The specific speed of a PaT in pump and turbine mode can be calculated by using Eq. (1) from Pérez-Sánchez et al. [13].

$$N_s = \frac{N \sqrt{Q_{\text{BEP}}}}{H_{\text{BEP}}^{3/4}} \quad (1)$$

The dimensionless parameters are more specific characteristics of these established parameters but normalized with the consideration of the impeller diameter to better unify the results. The dimensionless parameters for flow, head, and efficiency

are represented by the Greek letters phi, psi, and eta, respectively, as defined by Pérez-Sánchez et al. [13].

$$\Phi = \frac{Q[\text{m}^3/\text{s}]}{\omega[\text{rad/s}] \cdot (D[\text{m}])^3} \quad (2)$$

$$\Psi = \frac{g[\text{m/s}^2] \cdot H[\text{m}]}{(\omega[\text{rad/s}])^2 \cdot (D[\text{m}])^2} \quad (3)$$

$$\Lambda = \frac{P[\text{W}]}{\rho[\text{kg/m}^3] \cdot (\omega[\text{rad/s}])^3 \cdot (D[\text{m}])^5} = \Phi \Psi \eta \quad (4)$$

where  $\eta$  is the efficiency. Rossi uses the dimensionless parameters of the PaT to unify the results so that they can be compared universally. The study explains that the analysis in a non-dimensional orientation allows the results to be applied in a broader consideration for applications and for future studies [16].

### 2.3 Best Efficiency Point

Research has yielded equations which have been derived both theoretically and statistically to predict the capability of a pump acting as a turbine with the assumption that the specific speed of turbine mode is equal to the product of the pump specific speed and its efficiency. This implies that the PaT in turbine mode is required to work harder to achieve the same efficiency as in its pump mode.

$$N_{\text{st}} = N_{\text{sp}} \eta_{\text{p}} \quad (5)$$

Stepanoff was the first to derive the theoretical bridge between pump and turbine characteristics, dependent on the PaTs' best efficiency points [20]. The relationship shown in Eq. (5) was also theoretically derived by Stepanoff. Sharma developed his own theoretical equation which follows the template proposed by Stepanoff, but with the assumption that the turbine-mode specific speed is equal to the pump-mode specific speed multiplied by the root of the efficiency in pump mode [18]. As more information became available between a PaTs' performance from pump mode to turbine mode, Alatorre-Frenk formulated his set of equations based on statistical correlations by curve-fitting experimental data [1]. More recent equations have been developed by Yang which include the theoretical relationship between pump and turbine modes but also supported by statistical regressions [21]. Barbarelli developed equations solely based on statistical regression only dependent on the pump-mode specific speed of the PaT [3]. Some limitations to the studies using statistical regression to formulate the equations are the amount of data used to formulate the equations. For Alatorre-Frenk and Barbarelli, they used limited amount of data points

**Table 1** BEP equations of PaTs from literature

Author	Flow	Head
Stepanoff [20]	$\frac{Q_t}{Q_p} = \frac{1}{\sqrt{\eta_p}}$	$\frac{H_t}{H_p} = \frac{1}{\eta_p}$
Sharma [18]	$\frac{Q_t}{Q_p} = \frac{1}{\eta_p^{0.8}}$	$\frac{H_t}{H_p} = \frac{1}{\eta_p^{1.2}}$
Alatorre-Frenk et al. [1]	$\frac{Q_t}{Q_p} = \frac{0.85\eta_p^5+0.385}{2\eta_p^{9.5}+0.205}$	$\frac{H_t}{H_p} = \frac{1}{0.85\eta_p^5+0.385}$
Yang et al. [21]	$\frac{Q_t}{Q_p} = \frac{1.2}{\eta_p^{0.55}}$	$\frac{H_t}{H_p} = \frac{1.2}{\eta_p^{1.1}}$
Barbarelli et al. [3]	$\frac{Q_t}{Q_p} = 0.00029N_{sp}^2 - 0.02771N_{sp} + 2.01648$	$\frac{H_t}{H_p} = -3 \times 10^{-5}N_{sp}^3 + 4.4 \times 10^{-3}N_{sp}^2 - 0.20882N_{sp} + 4.6493$

in specific operating ranges, furthermore, they used PaTs of mixed typologies in their datasets which could hinder the accuracy of the results (Table 1).

2.4 Characteristic Curves

An essential piece of information needed for the consideration of pumps or even turbines is their characteristic curves. These curves with respect to flow show the capability that the pump or turbine can operate for its hydraulic head and efficiency, or power. Derakhshan and Nourbakhsh developed equations to predict the characteristic and power curve of a PaT based on the ratio of the flow at its best efficiency point [5]. In turn, the head curve is represented as head over its head best efficiency point as well as for the power. Rossi developed his own set of characteristic curve equations but with respect to the PaTs dimensionless parameters. It is important to note that the dimensionless parameters and dimensioned parameters divided by their best efficiency point would yield the same result since they are ratios with respect to its correlating best efficiency point. A recent study from Pérez-Sánchez developed these characteristic parameters but with respect to its dimensioned characteristics [13]. Once again, these characteristic curve equations are based on a limited number of data points and can also be subjected to error if multiple different typologies are used in the analysis (Table 2).

2.5 Machine Learning

Similarly, to how popular optimization equations were in the scientific community in the nineteenth century [6], machine learning is gaining traction in popularity and the range of applicability is growing significantly. Numerical regressions paired with an automatic optimization process can show correlations between attributes

**Table 2** Characteristic curve equations of PaTs

Author	Variable	Equation
Derakhshan and Nourbakhsh [5]	$\frac{H_t}{H_{t,BEP}}$	$1.0283\left(\frac{Q_t}{Q_{t,BEP}}\right)^2 - 0.5468\left(\frac{Q_t}{Q_{t,BEP}}\right) + 0.5314$
	$\frac{P_t}{P_{t,BEP}}$	$-0.3092\left(\frac{Q_t}{Q_{t,BEP}}\right)^3 + 2.1472\left(\frac{Q_t}{Q_{t,BEP}}\right)^2 - 0.8865\left(\frac{Q_t}{Q_{t,BEP}}\right) + 0.0452$
Rossi et al. [17]	$\frac{\psi}{\psi_{t,BEP}}$	$0.2394\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^2 + 0.769\left(\frac{\varphi}{\varphi_{t,BEP}}\right)$
	$\frac{\eta}{\eta_{t,BEP}}$	$-1.9788\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^6 + 9.0636\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^5 - 13.148\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^4 + 3.8527\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^3 + 4.5614\left(\frac{\varphi}{\varphi_{t,BEP}}\right)^2 - 1.3769\left(\frac{\varphi}{\varphi_{t,BEP}}\right)$
Pérez-Sánchez et al. [13]	$\frac{H_t}{H_{t,BEP}}$	$0.406\left(\frac{Q_t}{Q_{t,BEP}}\right)^2 + 0.621\left(\frac{Q_t}{Q_{t,BEP}}\right)$
	$\frac{\eta_t}{\eta_{t,BEP}}$	$-1.219\left(\frac{Q_t}{Q_{t,BEP}}\right)^4 + 6.95\left(\frac{Q_t}{Q_{t,BEP}}\right)^3 - 14.578\left(\frac{Q_t}{Q_{t,BEP}}\right)^2 + 13.231\left(\frac{Q_t}{Q_{t,BEP}}\right) - 3.383$

and volume of data with the potential of yielding accurate predictions of an isolated variable. Close to no research has been done to demonstrate these correlations using machine learning except for Rossi who used neural networks to find a link between turbine and pump modes [15]. The research is based on a library of 36 PaTs and shows strong correlations between pump-mode and turbine-mode attributes.

The machine learning library, SciKit learn, offers an extensive list of machine learning regression algorithms with opportunities to increase the applicability of the regression by hyperparameter tuning [11]. Some of these regressions include elastic-net, ridge regression, support vector regression, Bayesian ridge regression, and ARDR. In addition to the SciKit learn library, XGBoost offers regression capabilities and is a popular machine learning algorithm [4]. A study applying machine learning to a geometric subject was done by Niu who compared results from a support vector regression against an artificial neural network for the behavior of machine diesel engine [10]. The study shows the difference between a machine learning algorithm and a neural network and how they react differently to the same provided data by comparing the application of the characteristics of an engine to both a support vector machine algorithm and a neural network. Another interesting article by Rodriguez uses an artificial neural network to investigate the connection in predicting the useful life of blades found within steam turbines [14]. These studies loosely show that correlation can exist between predictions of geometric subjects with the use of machine learning.

### 3 Research and Methods

#### 3.1 Data Gathering and Exploratory Data Analysis

Prior to any machine learning modeling or optimization, a library of attributes needs to be compiled and the quality of the data needs to be addressed. The availability of experimental data on PaTs is quite scarce but not nonexistent. Considering the popularity of PaT technology in Europe, an organization known as REDAWN has numerous studies and experimental results pertaining to the behavior of PaTs. One of REDAWN's members, Pérez-Sánchez compiled an extensive list of PaT behavior at its BEP and made it publicly available [12]. This catalogue contained only the turbine-mode BEP characteristics but also provided the reference for which it was extracted. Since the current study is to relate pump-mode data to turbine-mode behavior, the pump-mode data needed to be extracted to make the information worthwhile. Another limitation concerns the specifics on the data, it is obligatory that the provided turbine-mode data contained the impeller diameter, rotational speed, flow at the best efficiency point, head at the best efficiency point, and the correlating efficiency. Furthermore, the reciprocated attributes in pump mode needed to be within the study in which it was extracted. Approximately half of the data were available to compose an initial library. The rest of the library was compiled from individual studies which would contain either one source or several sources of data for the present study. The final compiled library contained 180 data points.

Regarding the characteristic curves, similarly to the BEP library, key pieces of information of the PaT characteristic operating in turbine mode and pump mode are needed. Specifically, the impeller diameter, rotational speed, the ratio of flow over its flow BEP, the ratio of head over its head BEP, and either the ratio of the power or efficiency over its corresponding BEP. Once again, REDAWN has made the data it has publicly available [7]. The catalogues' information was accessed through an app created by Fecarotta to display the characteristic curves of PaTs. This database contained 38 PaTs with the characteristic curve information of both turbine and pump mode.

As the information for both the BEP and characteristic curves was compiled, addressing some key assumptions of the data and nature of the study had to be done. First, the typology had to be considered, most of the information provided were applied to the end suction own bearing (ESOB) in which is the most popular type commissioned so far. Therefore, the data points relating only to the ESOB typology were extracted.

At this point, the exploratory data analysis methodology was applied to the datasets to assess the quality of data in terms of consistency and outliers. This process includes checking that all attributes of each data point contained information, checking the maximum and minimum values, and exploring the correlation of each attribute in relation. The amount of data points relating to the BEP have been reduced to 145 data points and from 36 to 21 PaTs for the characteristic curves for a total of 196 data points mainly from removing inconsistent data, unrelated typologies, and data

with higher-than-normal flow rates. The last step before checking the correlations was to calculate the specific speeds using Eq. (1) and to convert all the dimensioned parameters to their corresponding dimensionless parameters using Eqs. (2)–(4). With the complete data sets ready to undergo machine learning applications, addressing the correlations between its attributes offers some wisdom as to how the results may produce.

### 3.2 Machine Learning Models

The number of regression models that SciKit Learn offers is quite extensive and ranges considerably in applicability. Since the number of studies related to turbine technology and machine learning is quite limited, applying a library of varying models is appropriate to conclude which should be studied further and optimized. Popular models like Bayesian Ridge or Decision Tree regressions were selected as well as niche models like Orthogonal Matching Pursuit and Theil Sen regressions were included in the library. To further add some diversity to the library, the XGBRegressor from XGBoost were also included in the model library. The final library contained 24 regression models to commission as an initial analysis and can be viewed in Table 3. All the models come from the SciKit Learn library except for the XGBRegressor which is from the XGBoost Library.

For each attribute to be predicted, the model in which produces the best coefficient of determination will be selected for that respective attribute and will undergo hyperparameter tuning. This capability is also supplied by the SciKit Learn library and can easily be implemented in a process to optimize the machine learning algorithms.

**Table 3** List of regression algorithms

XGBRegressor (XGBoost)*	HuberRegressor*	ElasticNetCV	NuSVR
RandomForestRegression	PassiveAggressiveRegression*	GammaRegressor	LinearSVR
DecisionTreeRegression	OrthogonalMatchingPursuit*	Poisson	KernalRidge
LinearRegression	LassoLars*	Gamma	
Ridge	ElasticNet*	InverseGaussian	
Lars	ARDRegression*	SVR-rbf	
TheilSenRegressor*	BayesianRidge*	SVR-lin	

\* Regressions used for the characteristic curve predictions



### 3.3 *Evaluation Metrics*

As seen in many modern articles for PaT prediction equations, the coefficient of determination ( $R^2$ ) is the most popular metric relating to the performance of the equation in anticipating turbine-mode characteristics.  $R^2$  is widely popular in statistics which compares the trend of the data against its mean value within the dataset. The values can range in value between 0 and 1 where 0 represents no correlation and 1 represents complete correlation. If the value is 0, this would then result in the model representing a capability of representing the data just as much as the mean value of the dataset. However, a negative value is possible, and this would result in negative correlation; in these circumstances, it would be more useful to use the mean value to represent the  $R^2$  for the data.

The root mean squared error (RMSE) is a metric widely used and represents the error associated to predicted and actual results in a normalized manner. This metric squares the error of the predicted result against the actual results in a dataset, which determines the mean value followed by its square root to present a value of error to represent the error of the dataset.

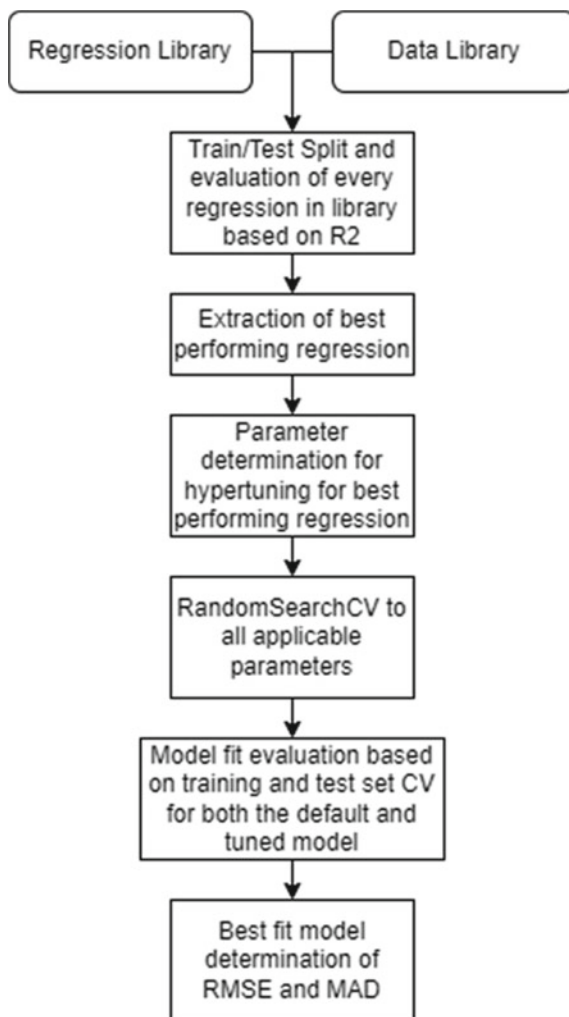
The last key metric used in this analysis is the mean absolute deviation (MAD). Knowing what the deviation of the data produced gives perspective as to how spread out the data is and possibly can show reasoning as to why the model performs better since the values are closer together in value. This can also be used to explain if the model performs poorly with the data spanning over large differences in values in relation to its mean.

These three metrics used together will give an in-depth perspective on the performance of the models against data. The  $R^2$  will help to understand how well the model fits the data, the RMSE will show us how much error is associated to the model, and the MAD will give perspective as to how spread out the data is with respect to the dataset.

### 3.4 *Code Layout*

All the activities regarding the application of the machine learning models were done in a Python programming environment and each library followed a similar methodology. Both libraries relating to predicting the BEP and characteristic curves underwent the same process for analysis and determination for prediction models. The process for each predicting attribute begins with an initial iteration of fitting and scoring through the default parameters of each model, and the best performing model was extracted based on the coefficient of determination. The RMSE and MAD of the model are also presented and stored for reference later on. The summarized process can be seen in Fig. 1 where the data library is representing either the BEP or characteristic curve datasets. Some models required different train and test sizes for optimum results.

**Fig. 1** Code layout for BEP and characteristic curve prediction process



With the best performing model, the code then evaluates the model with hyper-parameter tuning. Considering that each model has unique parameters with specific values, this process to initialize which parameters to hyper tune and to define the ranges are done manually. The function used for this process is the “Randomized-SearchCV” function which randomly picks values with the given ranges and scores the model with those attributes within a defined number of iterations based on cross-validation. Once the function finishes all its iterations and defines the optimum parameters, the model is retrained with the optimum parameters to procure a new  $R^2$  score. Furthermore, a cross-validation analysis is also performed based on  $R^2$  for the default model and the optimized model to check the quality of fit of the model to the data. This quality check addresses if the model is over fitting or underfitting the data. This

is an important step in the analysis considering the test set of the library which can be within 20–30 data points, and inconsistencies can have a big impact on the results.

After the model has been optimized with its best parameters and the quality of the model has been checked, a visual representation of the performance of the model is created. This step graphs the data and shows how well the model performed by plotting the actual results by the predicted results. A perfect model would show up all on the blue line producing a line like  $x = y$ . The RMSE and MAD are then recalculated and shown for the optimized model.

Once all the attributes have been predicted, all the results with respect to the model in its default nature and optimized orientation are presented with their optimized parameters as well as the  $R^2$ , RMSE, and MAD are compiled into a data frame. This data frame can then be exported to a csv file to further present with other software's.

It should also be noted that for the process of the characteristic curves, the model library was reduced from 24 to 8 models to reduce the operating time required for the code to process the data.

Attributes found within the BEP data library are the PaTs' BEP flow, head, efficiency, specific speed for both pump and turbine modes, as well as the impeller diameter, rotational speed, and specific speed. The same attributes are used in the characteristic curve dataset with the edition of the curve values with respect to its BEP for the flow, head, and efficiency (Q/QBEP, H/HBEP, and E/EBEP). The regression library models are mentioned in Table 3 corresponding for both the BEP and characteristic curve predictions.

## 4 Results

### 4.1 BEP Results

A correlation matrix was created for both the dimensioned and dimensionless datasets for the BEP of the PaTs. This gives a preview of how well the information for each attribute correlates to each of the other attributes in the dataset. The correlation matrix for both the dimensioned and dimensionless of the PaTs BEP libraries can be seen in Figs. 2 and 3, respectively. The dimensioned library showed strong correlations between its pump and turbine-mode attributes such as the rotational speed, specific speeds, flow, head, power, and efficiency. The dimensionless library shows a similar trend with stronger correlations when looking at PHI, PSI, and Lambda against each other in both turbine and pump modes.

From these datasets, the specific speed, flow, head, and efficiency of the PaT in turbine mode were then predicted for both the dimensioned and dimensionless datasets. The results for the best predicting models for both the dimensioned and dimensionless libraries can be seen in Tables 4 and 5. The reduction in  $R^2$  scores from the default to optimized parameters of the model is a consequence of the hyperparameter tuning and fitting the model better to the data. It should also be noted

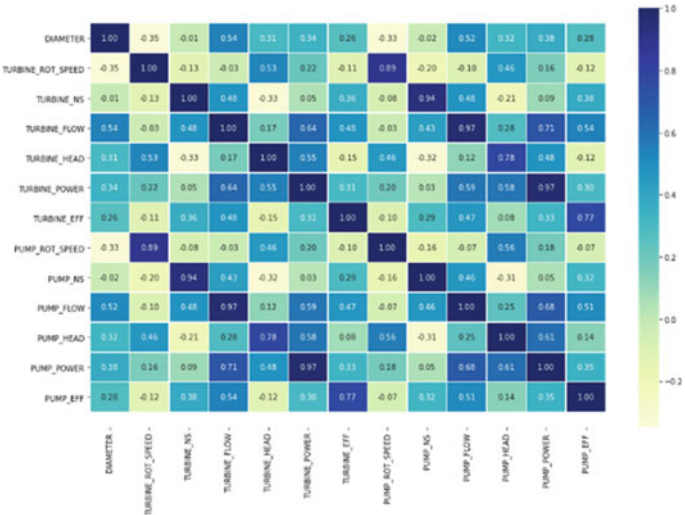


Fig. 2 Dimensioned dataset correlation matrix for the BEP

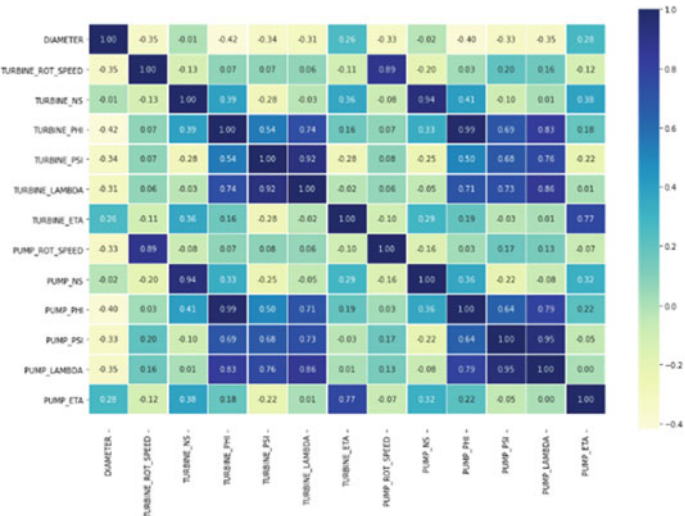


Fig. 3 Dimensionless dataset correlation matrix for the BEP

that the scales of the dimensioned parameters and the dimensionless parameters are different. Models applied to the dimensioned dataset performed better than the dimensionless when looking at the coefficient of determination. However, the prediction of the specific speed performed significantly better on the dimensionless dataset than the dimensioned.

**Table 4** Results of predicted turbine mode attributes dimensioned

Attribute	Best model	Default $R^2$	Optimized $R^2$	RMSE	MAD
Ns	XGBRegressor	0.7908	0.684891	1.675993	1.700475
Flow	HuberRegressor	0.9728	0.972127	9.720491	16.95505
Head	ElasticNet	0.9549	0.931871	7.666108	7.13368
Efficiency	OrthogonalMatchingPursuit	0.81467	0.81466	0.050519	0.063368

**Table 5** Results of predicted turbine mode attributes dimensionless

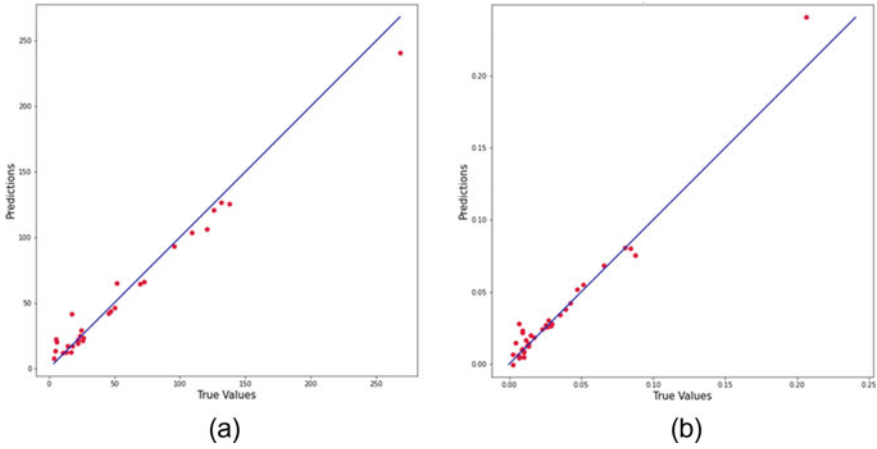
Attribute	Best Model	Default $R^2$	Optimized $R^2$	RMSE	MAD
Ns	HuberRegressor	0.8921	0.884719	1.249263	2.25259
Phi	ARDR	0.9749	0.95257	0.008099	0.013051
psi	HuberRegressor	0.8787	0.877729	0.068265	0.030723
Efficiency	OrthogonalMatchingPursuit	0.8147	0.81466	0.050519	0.063368

Figures 4a, b compare the models predicted results against its actual values for the flow BEP in both its dimensioned and dimensionless orientations, respectively. These two graphs show a near linear relationship to each other. Figures 5a, b show the results of the predictions against the actual values for the head predictions, and Fig. 6a, b show the results for the efficiencies in the same manner. The head results still show a linear trend, but the efficiencies are a little more sporadic in terms of the predicted against actual values. The results of the efficiency predictions for both the dimensionless and dimensioned datasets are identical, and this is a result of the model “Orthogonal Matching Pursuit” being applied to both datasets. In addition, this model has no parameters which can be tuned, therefore given the nature that the conversion and fitting process relating to the efficiency is equivalent, resulting in a model with the same performance for both libraries.

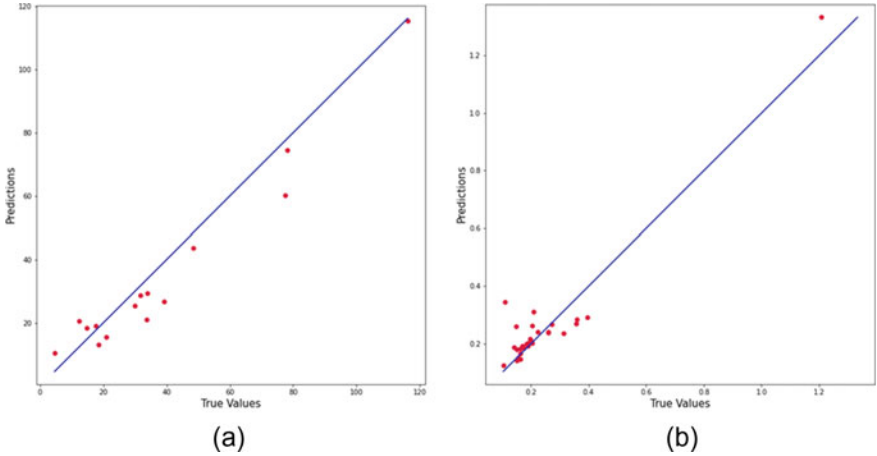
Different train and test splits were used for the models which can be seen in the amount of data in the figures of the actual against predicted data. These splits were between 10 and 30% for the test set with a corresponding interval of 90–70% for the training set. For instance, Fig. 5a shows a test set of 10%, while the data shown in Fig. 4b have a test set of 30%.

## 4.2 Characteristic Curve Results

Like the BEP process, a correlation matrix was created for both the dimensioned and dimensionless datasets for the characteristic curves of the PaTs as seen in Figs. 7 and 8, respectively. The results for the predictions of the characteristic curves of the dimensioned and dimensionless libraries are presented in Tables 6 and 7, respectively.

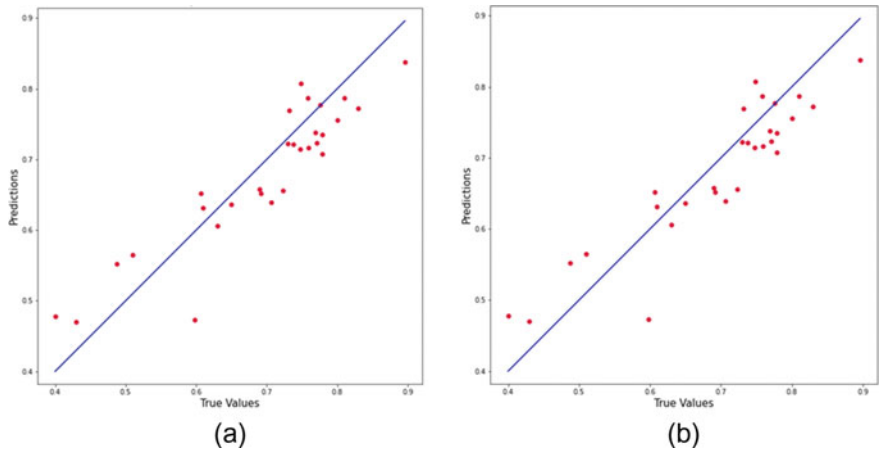


**Fig. 4** Actual versus predicted flow, **a** dimensioned and **b** dimensionless

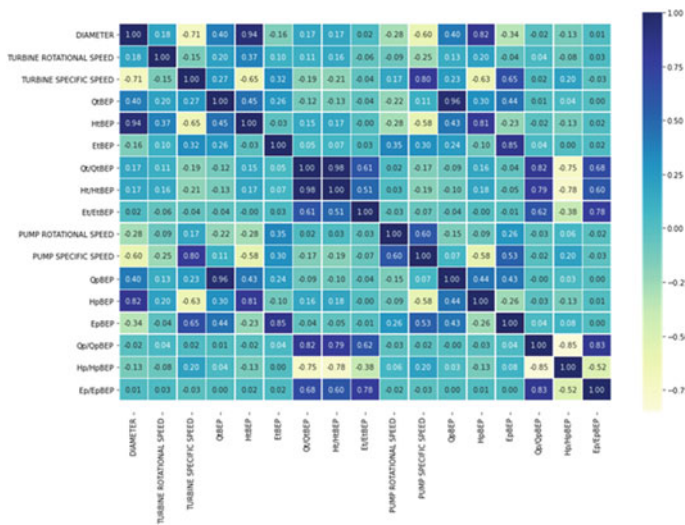


**Fig. 5** Actual versus predicted head, **a** dimensioned and **b** dimensionless

All models performed the best with the XGBRegressor with very similar performances for both libraries. The dimensionless library performs better by a small margin when considering the score of the coefficient of determination of the efficiency curves. For both the dimensioned and dimensionless parameters, the head curve performed with significant accuracy for the coefficient of determination with scores of 0.9972 and 0.9967, respectively.



**Fig. 6** Actual versus predicted efficiency, **a** dimensioned and **b** dimensionless



**Fig. 7** Dimensioned dataset correlation matrix for the characteristic curve

The visual comparison of the actual and predicted values for the head curve values can be consulted in Fig. 9a, b for the dimensioned and dimensionless parameters, respectively. A very strong correlation can be seen since the results produce a linear relationship between the actual and predicted results. The efficiency curve values can also be seen in Fig. 10a, b in the same manner, respectively. These values show more reliability for the predictions for higher efficiency values.

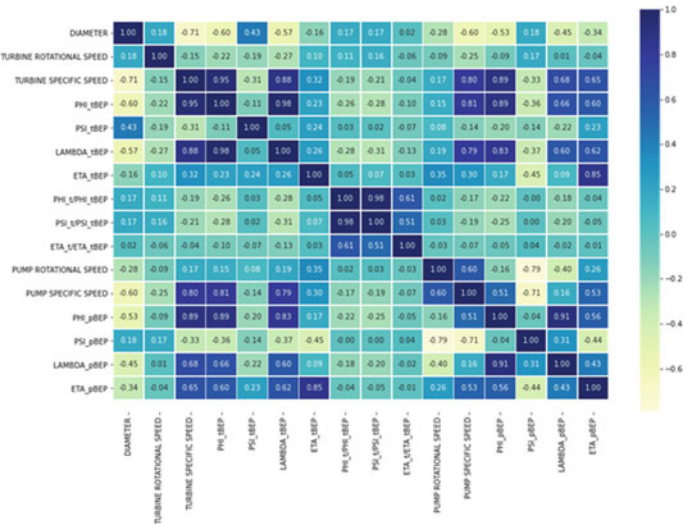


Fig. 8 Dimensionless dataset correlation matrix for the characteristic curve

Table 6 Results of predicted turbine mode characteristic curves—dimensioned

Curve	Model	Default $R^2$	Optimized $R^2$	RMSE	MAD
Head	XGBRegressor	0.9936	0.9972	0.0186	0.2369
Efficiency	XGBRegressor	0.9077	0.9089	0.0539	0.0394

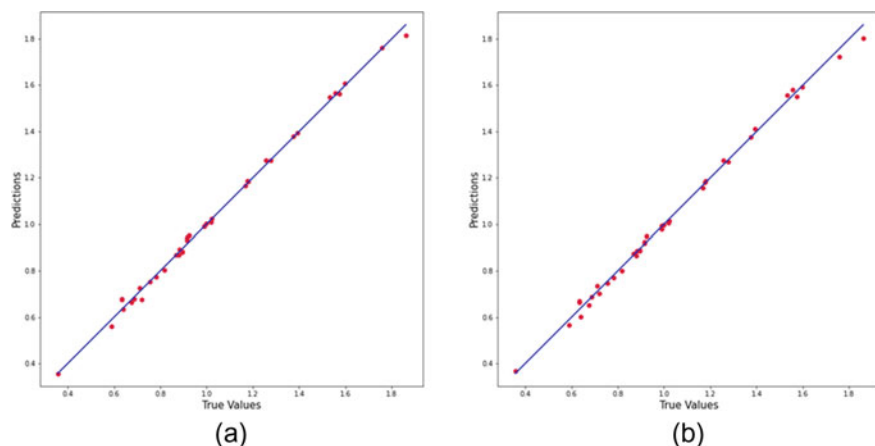
Table 7 Results of predicted turbine mode characteristic curves—dimensionless

Curve	Model	Default $R^2$	Optimized $R^2$	RMSE	MAD
Psi	XGBRegressor	0.9879	0.9967	0.0200	0.2385
Eta (efficiency)	XGBRegressor	0.9170	0.9134	0.0526	0.0400

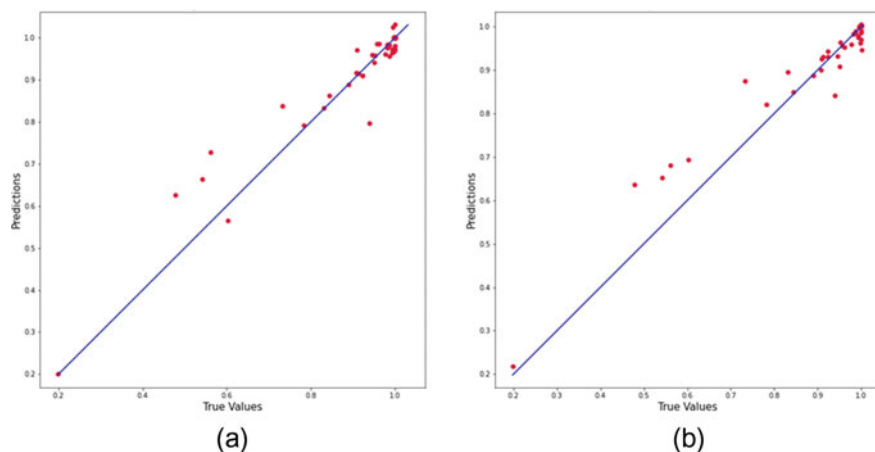
4.3 Comparisons to Established Equations

The results from this study were compared to the performance of equations from previous research. For the BEP comparison, 20 random data points were extracted from the dimensioned dataset and were used to calculate the predicted head and flow at its BEP. This considers that the test set of the machine learning models had were not the same for each attribute which ranged between 15 and 30 data points. The performance of the BEP of the head as can be seen in Table 8 demonstrates that the current model performed better than all the other studies with a coefficient of determination of 0.9318 followed by the model proposed by Sharma with a score of 0.8267. When looking at Barbarelli’s score, a considerably low score was determined, and this could be based on how the equation was formulated. The study uses only 4





**Fig. 9** Actual versus predicted head curve, **a** dimensioned and **b** dimensionless



**Fig. 10** Actual versus predicted efficiency curve, **a** dimensioned and **b** dimensionless

PaTs to form the equation in which the specific speeds range from 14 to 45. Most of the specific speeds of the 20 random points were under 10, and thus, the equation is being used outside of the parameters in which it was created. This can be a reason as to why it performs poorly.

The predicted flow scores seen in Table 7 show that once again the current study performed the best among the five other methods. The current study yielded a score of 0.9721, while Yang's equation scored a value of 0.9651 as the next highest. These scores are very comparable since they are less than 1% in deference.

As for the characteristic curves, the results of the other studies mentioned can be seen in Table 9. The performance of the head curve predictions of the current

**Table 8** BEP comparison of equations from literature

Method	$R^2$ head	RMSE head	$R^2$ flow	RMSE flow
Current study	0.9318	7.6661	0.9721	9.7204
[20]	0.7984	8.8332	0.9145	16.1634
[18]	0.8267	13.5815	0.9271	17.9042
[1]	0.7500	16.3587	0.8187	17.0522
[21]	0.7441	15.7382	0.9651	19.9461
[3]	-6.8072	32.2522	0.7391	23.3874

study scored very high with a value of 0.9972 for the coefficient of determination with the equation proposed by Pérez-Sánchez with a value of 0.9829. These are very comparable results and show almost complete certainty when predicting the head curve values.

The results of the predicted efficiency curve values also scored highly in the current study with a coefficient of determination of 0.9089 with Rossi's equation as the runner up with a score of 0.8685. As these scores are also comparable, there is still a margin of difference when looking at the scores.

When comparing both the predictions of the current study against established research, the current study scored highest for both the prediction of the BEP and the characteristic curves. These results can be justified by the RMSE of the methods being the lowest compared to all the other studies for both the head and flow predictions. The same thing can be said about the RMSE of the characteristic curves, when compared to the other studies, the RMSE is the lowest.

It can be presumed that the performance of the current study is a result of the amount of information compiled to formulate the predictions. The studies mentioned are based on datasets ranging from 4 to 30 PaTs for the BEP compared to 140 for the current study. With a significantly larger database to formulate regression-based correlations, it is consistent that a larger dataset would yield stronger results. The larger the dataset, the opportunity of the correlations being reliable also increase.

**Table 9** Characteristic curve comparison of equations from literature

Method	$R^2$ head curve	RMSE head curve	$R^2$ efficiency curve	RMSE efficiency curve
Current study	0.9972	0.0186	0.9089	0.0539
[5]	0.5450	0.2394	0.7870	0.0870
[17]	0.9546	0.0757	0.8685	0.0684
[13]	0.9829	0.0465	0.7773	0.0890

## 5 Conclusion

The goal of the study was to compare the performance of regression-based machine learning models with support of both a dimensioned and dimensionless library of PaT behavior. Previous authors have used both orientations to formulate equations using statistical regressions and correlations based on experimental data. With a significantly larger library of PaT behavior data from experimental results, a stronger and more reliable model to predict this behavior is possible. Compiling an even larger library of data points to input in a machine learning process would be potentially effective in creating more reliable and accurate models. With the results from this paper, the opportunity of creating a model with near complete accuracy is possible and would be of extreme use to predict the behavior of PaTs. With this information, designers can get a good idea of the requirements of a PaT in turbine mode, based on the pump mode characteristics, thus resulting in savings from having to run the PaT in a lab to get the capabilities in specific scenarios.

**Acknowledgements** Funding for this study has been provided by Concordia University.

## References

1. Alatorre-Frenk C, MI A, Karin A (1994) cost minimisation in micro-hydro systems using pumps-as-turbines
2. Algieri A, Zema DA, Nicotra A, Zimbone SM (2020) Potential energy exploitation in collective irrigation systems using pumps as turbines: a case study in Calabria (Southern Italy). *J Clean Prod* 257. <https://doi.org/10.1016/j.jclepro.2020.120538>
3. Barbarelli S, Amelio M, Florio G (2017) Experimental activity at test rig validating correlations to select pumps running as turbines in microhydro plants. *Energy Convers Manage* 149:781–797. <https://doi.org/10.1016/j.enconman.2017.03.013>
4. Chen T, Guestrin C (2016) {XGBoost}: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 785–794. <https://doi.org/10.1145/2939672.2939785>
5. Derakhshan S, Nourbakhsh A (2008) Experimental study of characteristic curves of centrifugal pumps working as turbines in different specific speeds. *Exp Thermal Fluid Sci* 32(3):800–807. <https://doi.org/10.1016/j.expthermflusci.2007.10.004>
6. Esposito WR (2008) Hamilton–Jacobi–Bellman Equation. *Encycl Optim* 1473–1476. [https://doi.org/10.1007/978-0-387-74759-0\\_258](https://doi.org/10.1007/978-0-387-74759-0_258)
7. Fecarotta O (2021). RPS. <https://doi.org/10.5281/ZENODO.4973447>
8. McNabola A, Coughlan P, Corcoran L, Power C, Williams AP, Harris I, Gallagher J, Styles D (2014) Energy recovery in the water industry using micro-hydropower: an opportunity to improve sustainability. *Water Policy* 16(1):168–183. <https://doi.org/10.2166/wp.2013.164>
9. Mitrovic D, Chacón MC, García AM, Morillo JG, Díaz JAR, Ramos HM, Adeyeye K, Carravetta A, McNabola A (2021) Multi-country scale assessment of available energy recovery potential using micro-hydropower in drinking, pressurised irrigation and wastewater networks, covering part of the eu. *Water (Switzerland)* 13(7). <https://doi.org/10.3390/w13070899>
10. Niu X, Yang C, Wang H, Wang Y (2017) Investigation of ANN and SVM based on limited samples for performance and emissions prediction of a CRDI-assisted marine diesel engine. *Appl Therm Eng* 111:1353–1364. <https://doi.org/10.1016/j.applthermaleng.2016.10.042>

11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. *JMLR* 12:2825–2830
12. Pérez-Sánchez M, Javier Sánchez-Romero F, Ramos HM, Amparo López-Jiménez P (n.d.) Improved planning of energy recovery in water systems using a new analytic approach to pat performance curves. [www.mdpi.com/journal/water](http://www.mdpi.com/journal/water)
13. Pérez-Sánchez M, Sánchez-Romero FJ, Ramos HM, López-Jiménez PA (2020) Improved planning of energy recovery in water systems using a new analytic approach to PAT performance curves. *Water* (Switzerland), 12(2). <https://doi.org/10.3390/w12020468>
14. Rodríguez JA, El Hamzaoui Y, Hernández JA, García JC, Flores JE, Tejada AL (2013) The use of artificial neural network (ANN) for modeling the useful life of the failure assessment in blades of steam turbines. *Eng Fail Anal* 35:562–575. <https://doi.org/10.1016/j.engfailanal.2013.05.002>
15. Rossi M, Renzi M (2018) A general methodology for performance prediction of pumps-as-turbines using artificial neural networks. *Renewable Energy* 128:265–274. <https://doi.org/10.1016/j.renene.2018.05.060>
16. Rossi M, Righetti M, Renzi M (2016) Pump-as-turbine for energy recovery applications: the case study of an aqueduct. *Energy Procedia* 101:1207–1214. <https://doi.org/10.1016/j.egypro.2016.11.163>
17. Rossi M, Nigro A, Renzi M (2019) Experimental and numerical assessment of a methodology for performance prediction of Pumps-as-Turbines (PaTs) operating in off-design conditions. *Appl Energy* 248:555–566. <https://doi.org/10.1016/j.apenergy.2019.04.123>
18. Sharma KR (1985) Small hydroelectric project-use of centrifugal pumps as turbines. In: Kirloskar, Electric Co. Bangalore, India
19. Stefanizzi M, Capurso T, Balacco G, Binetti M, Camporeale SM, Torresi M (2020) Selection, control and techno-economic feasibility of pumps as turbines in water distribution networks. *Renewable Energy* 162:1292–1306. <https://doi.org/10.1016/j.renene.2020.08.108>
20. Stepanoff AJ (1957) Centrifugal and axial flow pumps. Theory, design, and application
21. Yang SS, Derakhshan S, Kong FY (2012) Theoretical, numerical and experimental prediction of pump as turbine performance. *Renewable Energy* 48:507–513. <https://doi.org/10.1016/j.renene.2012.06.002>

# Large Eddy Simulation of Near-Bed Flow Over Bottom Roughness in Open Channel



Bowen Xu and S. Samuel Li

**Abstract** Channel-bed roughness has the influence on near-bed flow, with implications in hydraulic engineering design. Previous studies suggest that small-scale structures of turbulence are an important element for near-bed flow. However, very little research has dealt with turbulence structures of flow in the vicinity of roughness elements at the channel bed. The reasons for this include the high costs and technical difficulties to obtain fine resolution measurements of near-bed flow from natural and laboratory channels. Therefore, Computational Fluid Dynamics (CFD) models become an attractive tool for the investigation of near-bed flows. The purpose of this paper is to use Large Eddy Simulation (LES) as a mathematical model to predict eddy motions, including small-scale motions near the bed and around roughness elements. In this study, LES runs are performed under conditions matching a series of laboratory experiments. The LES runs use OpenFOAM as a numerical solver. The simulations numerically solve one-phase incompressible flow using the finite volume method and yield snapshots of instantaneous velocity and pressure fields. The simulations use the rigid lid approximation at the upper boundary of the model channel and apply cyclic conditions between its upstream and downstream lateral open boundaries. A significant challenge in LES is to resolve small-scale motions of dynamic importance. This paper handles the challenge by resolving the viscous sublayer. This treatment permits the use of no-slip condition, which is realistic. The ensemble averages of instantaneous velocities are validated using experimental data. The LES results are further analyzed to reveal eddy motions and turbulent flow patterns. The results provide details of turbulence structures near the bed with the presence of roughness elements.

**Keywords** Near-bed flow • Bottom roughness • Open channel

---

B. Xu • S. Samuel Li (✉)

Department of Building, Civil and Environmental Engineering, Concordia University, 1455 de Maisonneuve Boulevard West, Montreal, QC H3G 1M8, Canada  
e-mail: [sam.li@concordia.ca](mailto:sam.li@concordia.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_45](https://doi.org/10.1007/978-3-031-34593-7_45)

731

# 1 Introduction

The physical condition of an alluvial river, including the morphological condition of the riverbed, can change due to anthropogenic activities and/or natural processes. An assessment of the changes is important from the perspective of hydraulic engineering and water management. Key tasks for assessing these changes include conducting bed topography surveys and taking measurements of bed roughness. Such data prove invaluable for habitat simulations. The data is useful for habitat simulations [7].

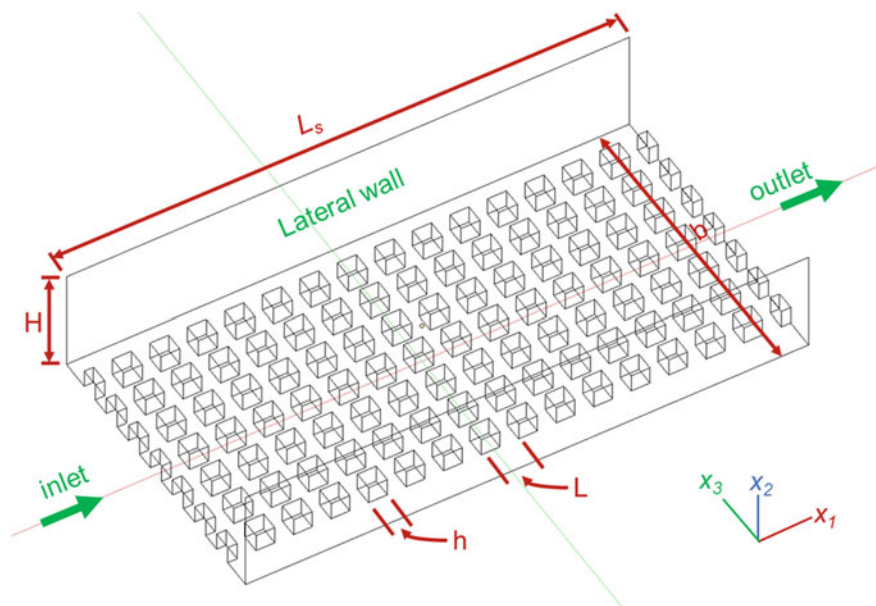
Sturm [13] describes bedforms as an irregular feature at the alluvial riverbed, with dimensions much larger than the size of bed sediments. Ripples, dunes, and antidunes are examples of bedforms, each with its own physical origin. Ripples result from the expansion of any significant discontinuity on the bed because of deformation. Dunes usually associate with the greatest scale turbulent eddies, having a height on the order of the flow depth. Alternate scour and deposition areas appear in the flow direction, causing the bedforms to evolve into a rather stable shape. Since the size of the biggest eddies depends on depth, the wavelength of dunes must relate to the depth as well [15]. At near unity Froude number, standing surface waves generate antidunes [4]. Both bedforms and roughness elements (REs) at the bed produce whirlpools and turbulence as the near-bed flow interacts with the REs. The characteristics of turbulence are complicated, especially in the cavities between REs [12].

Yager et al. [14] suggested that large sediment grains bear a significant portion of the total shear stress and thereby reduce the stress available to move finer sediments in a mountain river. Meanwhile, the discharge of a river may increase with decreasing bed roughness when the flow cascades or the channel transforms to bedrock morphology [2].

Li and Li [6] reported flume experiments of flow over three types of roughness, namely d-type, intermediate-type, and k-type. The flume channel simplifies an irregular channel bed using uniform REs. These artificial REs were arranged in rows and columns as shown in Fig. 1. The experiments produced measurements of flow velocity from an acoustic Doppler velocimetry. It was concluded that the velocity fluctuated slightly in the main flow; the standard deviation of the velocity distribution varied from 5 to 7% among three types of REs. Their results showed that the turbulence intensity was the highest near the channel bed, and turbulence shear was the highest at the crest level of REs.

Zhang and Li [16] carried out the analysis of flow over uniform REs through 2D numerical simulations. They concluded that the pitch-to-roughness height ratio dictated the fluctuation of turbulence intensity and turbulence stress distribution, along with the number, location, and shape of eddies in the cavity.

It is well known that computations using the Reynolds-averaged Navier–Stokes (RANS) equations cannot realistically capture turbulent fluctuations expected in the vicinity of bed roughness elements. The fluctuations have important implications with respect to natural channel stability. The purpose of this study is to investigate the fluctuations by means of Large Eddy Simulation (LES). LES has the capacity of predicting transient turbulent structures. Specifically, this paper aims to build a



**Fig. 1** Three-dimensional view of model channel geometry, for RANS and LES computations

LES model that matches the laboratory experimental setup of Li and Li [6] and that extends the experimental results. The experimental data are utilized to validate the ensemble average of instantaneous velocities from LES. The LES model results are examined to highlight evolving eddy motions and turbulent flow patterns. The findings reveal detailed turbulence structures near the REs.

## 2 Methods

The computational model channel and bottom roughness in this study (Fig. 1) match the laboratory channel of Li and Li [6]. The computational channel has a cyclic inlet and outlet (Fig. 1). The water surface is treated as a slippery surface. The channel bottom, lateral walls, and REs are a non-slippery surface.

The flow depth is  $H = 0.088$  m, the length of the channel is  $L_s = 0.5715$  m, and the width of the channel is  $b = 0.309$  m. The dimension of each cubic roughness element is  $h = 0.01905$  m (or 3/4 inches), and the pitch-to-roughness height ratio is  $L/h = 2$ .

The continuity and momentum equations that govern the flow are solved using the CFD software OpenFOAM version 9, under the GNU-GPL 3 license. The partial differential equations are discretized on 3D unstructured mesh of polyhedral cells for

finite volume solutions. The flow solver is built within a robust, implicit, pressure–velocity, iterative solution framework. Domain decomposition parallelism is built into the core of OpenFOAM and is implemented at a low level. This allows the solver to be designed without any “parallel-specific” coding; in this work, the total number of computational cells that cover the model channel was divided into 32 parts using the built-in “scotch” method in OpenFOAM. The computations were performed on a 32-core CPU of the High-Performance Computing (HPC) Facility at Concordia University. A random-access memory of 512 GB was allocated for the computation. Details of the CFD model are given below.

## 2.1 Hybrid System of Reynolds-Averaged Navier–Stokes Equations and Large Eddy Simulation

This study implements a hybrid system of Reynold-Averaged Navier–Stokes (RANS) Equation computations and Large Eddy Simulation (LES). The RANS model ensures that the quality of mesh is sufficient to resolve at least 80% of the total turbulent kinetic energy (TKE) in LES. It also provides initial fields for LES, by mapping from the RANS model. The LES resolves eddies of large sizes that carry most of the turbulence energy and produce instantaneous flow velocities for postprocessing.

The RANS simulation uses the  $k$ - $\omega$  shear stress transport (SST) turbulence model to close the model equations. It uses the  $k$ - $\omega$  model in the region of boundary layer and shifts to the  $k$ - $\varepsilon$  model in the free shear flow. In the  $k$ - $\omega$  and  $k$ - $\varepsilon$  models, the Boussinesq hypothesis links the turbulent stresses to the mean velocity gradients by means of a turbulence viscosity  $\nu_t$  [9].

The RANS simulation provides steady-state solutions, using the SIMPLE algorithm, specifically for a incompressible Newtonian fluid. The pressure field and velocity field are coupled without considering gravity [1]; this solver is called simpleFoam in OpenFOAM. The steady-state Navier–Stokes equations are

$$\nabla \cdot \mathbf{U} = 0 \quad (1)$$

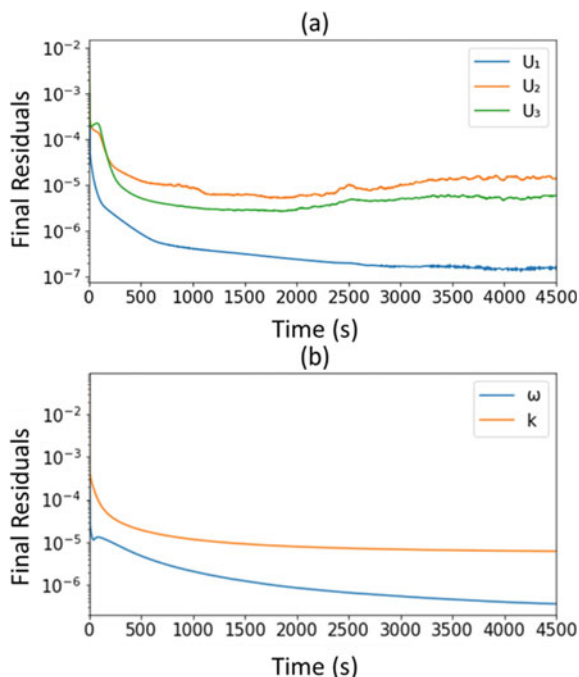
$$\nabla \cdot (\mathbf{U}\mathbf{U}) - \nabla \cdot (\nu \nabla \mathbf{U}) = -\nabla p / \rho \quad (2)$$

where  $\mathbf{U}$  is the velocity vector;  $\nu$  is the dynamic viscosity;  $p$  is the pressure divided by water density;  $\rho$  is the fluid density. The RANS simulation allowed 4500 iterations and reached a steady state. In Fig. 2, the residuals over iterations are plotted, showing the final fluctuations of the residuals over iterations.

After validating the mesh quality by the RANS simulation, the velocity and pressure fields are mapped to the LES model as initial conditions. The LES proceeds for 60 s of model time. The LES is a transient simulation. It uses a filter to retain large-scale flow in the model domain and a dynamic subgrid model for turbulence closure.



**Fig. 2** Convergence of RANS simulation: **a** Final residuals of velocity; **b** final residuals of  $\omega$  and  $k$

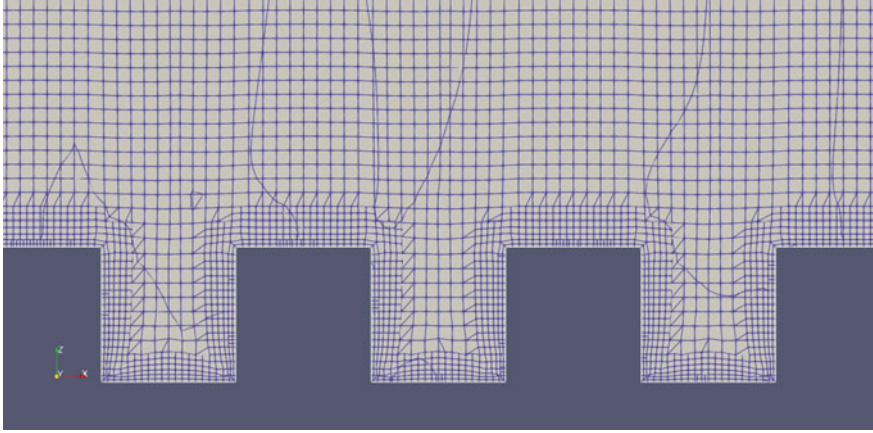


The LES uses the pressure–implicit with splitting of operators (PISO) algorithm for the transient state to couple the pressure field with the velocity field [1]. In OpenFOAM, the solver is called `pimpleFoam`, which merges the PISO and SIMPLE algorithms. The outer corrector is set to 1 in the model to select PISO. The time step of LES is set as 0.0005 s, and the output variables are saved every 20 time steps (0.01 s or 100 Hz).

## 2.2 Mesh

Two mesh generators in OpenFOAM, namely `blockMesh` and `snappyHexMesh`, are employed to create the unstructured mesh with hexahedral cells for RANS and LES models. At first, `blockMesh` is utilized to generate a structured hexahedral base mesh, with cubes of 2 mm, covering the whole domain. This step also specifies patches of an inlet, outlet, water surface, lateral walls, and bottom of the model channel. Then, the module of `snappyHexMesh` is used to add REs, along with refinement for regions of interest and prismatic layers on a solid surface.

As illustrated in Fig. 3, the refinement level for near-wall regions is set to 1, where the cubes are 1 mm in dimension. This reduces mesh distortion and improves simulation accuracy. Two layers are added to the surfaces of REs and the channel bed. The thickness of the first cell is 0.5 mm, and the growth ratio is 1.2. In the same



**Fig. 3** Longitudinal cross-section at the model channel centerline, showing mesh configuration and near-wall refinement

manner, two layers are added to the channel lateral walls. The final mesh consists of 4,078,835 cells, of which 3,967,954 are hexahedral cells.

### 2.3 Boundary and Initial Conditions

The flow computations use boundary conditions identical to those in Li and Li's [6] experiments. In both RANS and LES computations, the incompressible Newtonian fluid has a constant temperature of 18 °C, a kinematic viscosity of  $\nu = 1.05 \times 10^{-6}$  m<sup>2</sup>/s, a turbulence intensity of  $I = 5\%$  (for RANS model only), and a cyclic velocity between the inlet and outlet. More details of boundary and initial conditions are listed in Table 1.

In Table 1, the Reynold number ( $R = UR_h/\nu$ ) is evaluated as 29,095, which is greater than 20,000 and thus the flow is fully turbulent. The Froude number ( $Fr = U/\sqrt{gH}$ ) is 0.58, which means subcritical flow. The initial values of turbulent parameters,  $k$  and  $\omega$ , are estimated as

$$k = 1.5(I|U|)^2 \quad (3)$$

$$\omega = \frac{k^{0.5}}{C_\mu^{0.25} l_m} \quad (4)$$

where  $I$  is the turbulence intensity;  $U$  is the mean velocity;  $C_\mu$  (equal to 0.09) is a turbulence model constant; and  $l_m$  is a mixing length, given by  $l_m = 0.07D_H$ , where  $D_H$  is the hydraulic diameter.

**Table 1** Boundary conditions for RANS and LES computations

Elements	$Q^a$ (m <sup>3</sup> /s)	$p^a$ (Pa)	$k$ (m <sup>2</sup> /s <sup>2</sup> )	$\Omega$ (s <sup>-1</sup> )	$\nu_t$ (m <sup>2</sup> /s)
Internal regions	0	0	13.36	1083.43	0
Inlet	14.72 (Cyclic)	Pressure gradient linked to the velocity	13.36	1083.43	0
Outlet	14.72 (Cyclic)	Total pressure with the value of 0	Zero gradient	Zero gradient	0
Water surface	Slip	Slip	Slip	Slip	Slip
Lateral walls, channel bottom and REs	0	0	Wall function <sup>b</sup>	Wall function <sup>b</sup>	Wall function <sup>b</sup>

<sup>a</sup>LES model uses only the flowrate and pressure, and their initial internal field values are mapped from RANS model

<sup>b</sup>In the RANS model, the turbulent parameters  $\varepsilon$ ,  $\omega$ , and  $\nu_t$  are determined by the wall functions on the non-slippery surfaces [3]

## 2.4 Residuals

For a conserved variable, its residual is the difference between the expected and observed values in a control volume. The scaled residuals are used in OpenFOAM to judge whether this imbalance is acceptable or not. The equation tolerances, namely the absolute and relative quantities, are specified for each conserved variable in the configuration. If the initial residual meets either of the values, the equations are deemed as solved. Table 2 summarizes the linear-solver (abbreviated as *solver* in the table) control and tolerances configured in RANS and LES models.

**Table 2** Linear-solver control and tolerance

Variables <sup>a</sup>	RANS	LES
$P$	<ul style="list-style-type: none"> <li>• Solver: PCG</li> <li>• Preconditioner: DIC</li> <li>• Tolerance: <math>10^{-10}</math></li> <li>• Relative tolerance: 0.05</li> </ul>	<ul style="list-style-type: none"> <li>• Solver: GAMG</li> <li>• Smoother: DICGaussSeidel</li> <li>• Tolerance: <math>10^{-6}</math></li> <li>• Relative tolerance: 0.1</li> </ul>
$U$ , $k$ , and $\omega$	<ul style="list-style-type: none"> <li>• Solver: smoothSolver</li> <li>• Smoother: symGaussSeidel</li> <li>• Tolerance: <math>10^{-10}</math></li> <li>• Relative tolerance: 0.05</li> </ul>	<ul style="list-style-type: none"> <li>• Solver: smoothSolver</li> <li>• Smoother: symGaussSeidel</li> <li>• Tolerance: <math>10^{-5}</math></li> <li>• Relative tolerance: 0.1</li> </ul>

<sup>a</sup>The turbulent parameters  $k$  and  $\omega$  are only for RANS model

## 2.5 Dynamic Subgrid Scale (SGS) Model

Due to energy cascade, in reality, turbulent eddies are unstable, which burst into smaller and smaller eddies. In the end, the molecular viscosity dissipates the smallest eddies into heat. This study employs the LES model to investigate turbulence characteristics in the domain covered by an unstructured mesh. Eddies smaller than the cell cannot be resolved by the mesh, but their velocities are reserved at the cell centroids, which contribute to the accumulation of turbulence kinetic energy (TKE). Therefore, a Dynamic Subgrid Scale (SGS) model is applied to remove the smallest resolved eddies, whose sizes are just larger than the corresponding cells [11]. To achieve this goal, an extra viscous stress tensor is added to the Navier–Stokes equations as

$$\frac{\partial(\rho U_i)}{\partial t} + \frac{\partial}{\partial x_j}(\rho U_i U_j) = -\frac{\partial P}{\partial x_i} + \frac{\partial}{\partial x_j}(\tau_{ij} + \tau_{sgs}) \quad (5)$$

where the subscript  $i$  ( $= 1, 2, 3$ ) refers to the three orthogonal directions,  $\tau_{ij}$  is the viscous stress tensor, and  $\tau_{sgs}$  is the extra SGS stress tensor. To calculate  $\tau_{sgs}$ , the SGS kinematic viscosity,  $\nu_{sgs}$ , becomes the target; their relationship is

$$\tau_{sgs} = 2\rho\nu_{sgs}S_{ij}^* - \frac{2}{3}\rho k_{sgs}\delta_{ij}, \text{ where } S_{ij}^* = \frac{1}{2}\left(\frac{\partial \tilde{U}_i}{\partial x_j} + \frac{\partial \tilde{U}_j}{\partial x_i} - \frac{1}{3}\frac{\partial \tilde{U}_k}{\partial x_k}\delta_{ij}\right) \quad (6)$$

The calculation of  $\nu_{sgs}$  accounts for the eddy size and assumes isotropic turbulence, which makes it a scalar. Given that the study focuses on the distribution of eddies in the entire model domain, including near-wall regions ( $y^+ < 5$ ), the Van Driest Damping Dynamics Smagorinsky model is used to calculate the SGS kinematic viscosity. The function of  $\nu_{sgs}$  is given by

$$\nu_{sgs} = l_0^2 \sqrt{2S_{ij}S_{ij}} \quad (7)$$

where  $l_0$  is a length scale;  $S_{ij}$  is the magnitude of strain rate tensor. The length scale (or the size of eddy) is calculated as

$$l_0 = \left[ \kappa y \left( 1 - \exp \exp \left( \frac{y^+}{A^+} \right) \right), C_s \Delta \right] \quad (8)$$

where  $\kappa$  is the von Kármán constant (equal to 0.41);  $A^+$  is the van Driest constant (equal to 26),  $C_s$  is the Smagorinsky model constant (equal to 0.158);  $y$  and  $y^+$  are, respectively, the dimensional and non-dimensional distances to the wall.

## 2.6 Time Scheme

For the RANS model, the simulation is in steady state, i.e., the temporal derivatives are zero. For the LES, a backward time scheme (an implicit and second-order accuracy) is chosen. The study of temporal accuracy in OpenFOAM by Lee [5] concludes that the backward scheme not only improves the accuracy, compared to the first-order Euler scheme, but also keeps the efficiency at a relatively high level. It requires the field values at previous time steps, given by

$$\frac{\partial}{\partial t}(\phi) = \frac{1}{\Delta t} \left( \frac{3}{2}\phi - 2\phi^o + \frac{1}{2}\phi^{oo} \right) \quad (9)$$

where  $t$  is the time;  $\phi$ ,  $\phi^o$ , and  $\phi^{oo}$  are, respectively, the field values at actual, previous, and previous–previous time levels. At the beginning of a simulation, computations for the first two time steps use the first-order Euler scheme.

## 3 Results and Discussion

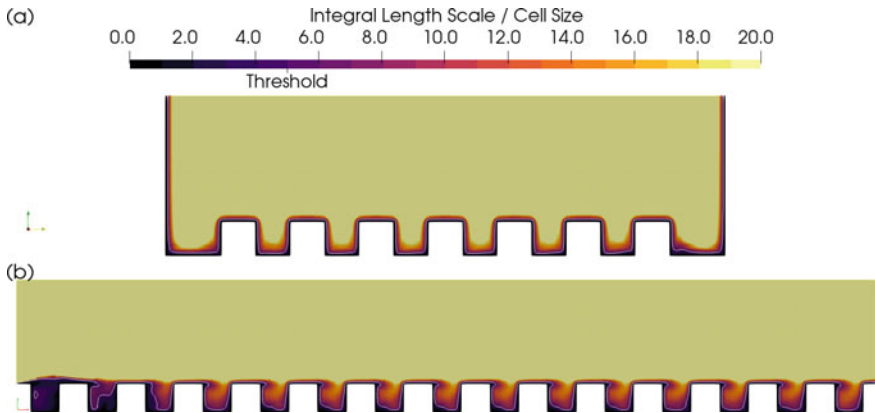
The postprocessing has been achieved via the OpenFOAM's built-in utilities and ParaView, a multi-platform software for the data analysis and visualization. This section presents the validation of the mesh quality, explains the choice of the investigated time interval, compares the results to the ones of Li and Li [6], and illustrates the distributions of the turbulence intensity and Reynolds shear stress in the regions of interest.

### 3.1 Mesh Quality and Resolved TKE

To validate the mesh quality, the integral length scale calculated based on the results of RANS model is used to determine if the mesh is sufficient to resolve 80% of TKE in LES simulation. The integral length scale,  $l_0$ , refers to the size of an eddy with the average TKE of all the eddies in the LES simulation. Maele and Merci [8] presents the function of  $l_0$  estimated from the steady RANS simulation, which gives:

$$l_0 = \frac{k^{3/2}}{\varepsilon} = \frac{k^{1/2}}{C_\mu \omega} \quad (10)$$

Given that the LES simulation is considered valid if at least 80% of TKE is solved [8], we need to ensure that  $l_0/\Delta$ , where  $\Delta = \text{Cell Volume}^{1/3}$ , is superior to 5. As illustrated in Fig. 4, the white contour represents the threshold of 5; the majority of unresolved TKE exists in the near-wall region, especially the cavities in



**Fig. 4** **a** Transverse cross-section, through the middle of a cube column, and **b** longitudinal cross-section at the channel centerline, showing contours of  $l_0/\Delta$  values from RANS computations

the longitudinal direction. Since the area under the threshold is very small, the mesh quality is considered as satisfied.

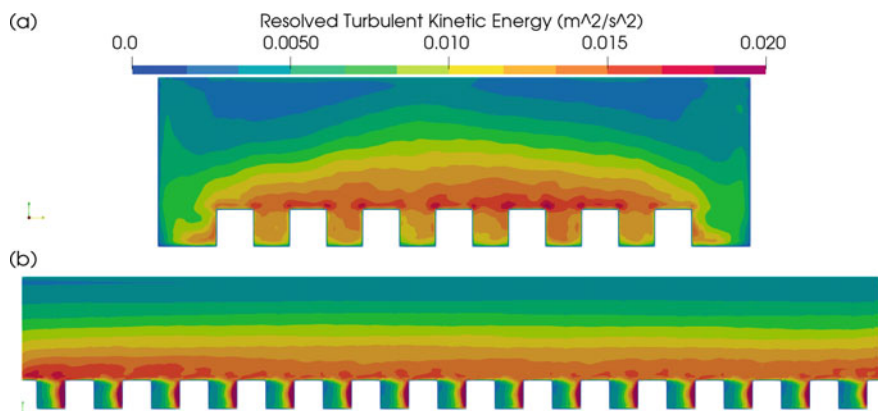
After validating the mesh, the LES computations were conducted, and we can first extract the resolved TKE; the function to calculate the resolved TKE, presented by Pope [11], gives:

$$k_{\text{res}} = \frac{1}{2} \left( \overline{u_{ii}'^2} \right) \quad (11)$$

Figure 5 shows the distribution of  $k_{\text{res}}$  in a chosen transverse and longitudinal cross-sections. The model time of interest is from 10 to 60 s or a duration of 50 s. The values of  $\overline{u_{ii}'^2}$  are calculated based on data of 5000 samples. In the transverse plane (Fig. 5a), the distribution is slightly asymmetrical and shows a convex pattern. The high TKE concentrates on the lateral edges of REs, especially the REs closer to the channel centerline. Figure 5b plots values of resolved TKE in the longitudinal cross-section at the channel centerline. Note that the resolved TKE intensifies the windward face of the REs. In the free-flow region, the distribution shows strip patterns.

### 3.2 Turbulent Statistics

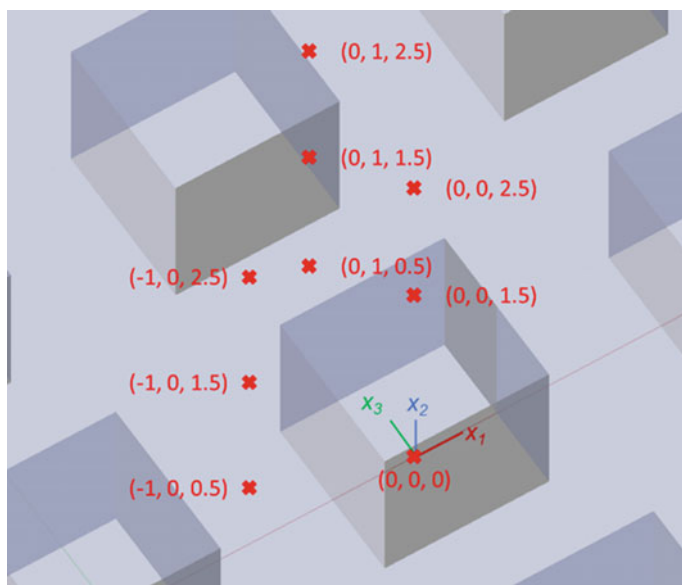
To confirm if the LES model has reached a statistically steady state, it is necessary to examine velocity fluctuations around a constant mean value. Eight probes (Fig. 6) are placed in the LES model to extract velocities over the simulation period of 0–60 s. The origin of coordinates is at the center of a RE's lower surface, and the probes are placed around it. This RE is the ninth unit on the centerline from the inlet. The coordinates are normalized by the RE dimension as  $(x_1/h, x_2/h, x_3/h)$ . Figure 7 shows



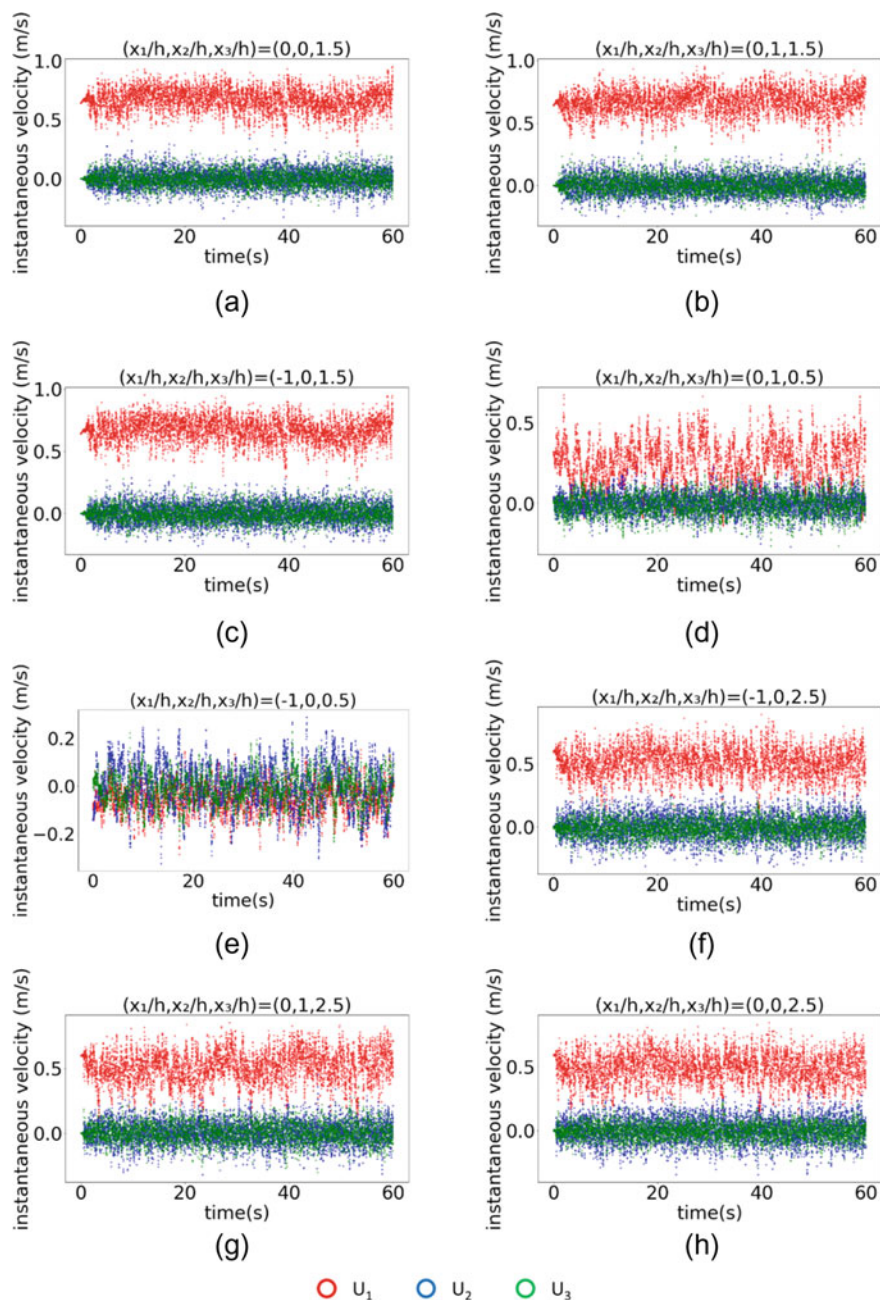
**Fig. 5** **a** Transverse cross-section, through the middle of a cube column, and **b** longitudinal cross-section at the channel centerline, showing contours of  $k_{res}$  values from LES computations

the instantaneous velocity components at the locations of the eight probes. There are fluctuations in the velocity components.

Starting from the initial conditions mapped from the steady RANS model or from well-developed flow, the LES quickly reaches the statistically steady state, as can be seen from the fluctuation patterns.



**Fig. 6** Eight selected locations for the extractions of time series of instantaneous velocity components from LES computations



**Fig. 7** Time series of instantaneous velocity components in three orthogonal directions: **a–h** for the eight locations shown in Fig. 6

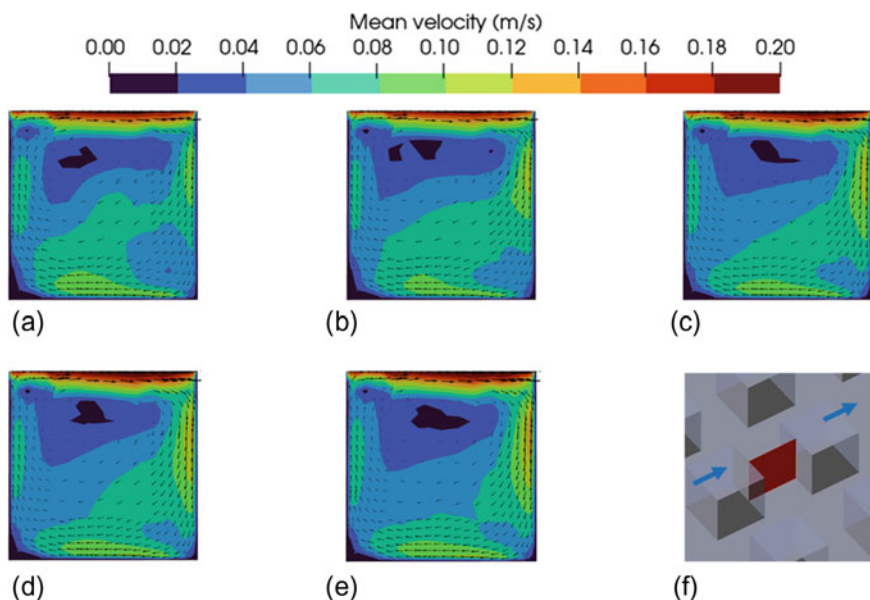


On the basis of the length of the model channel and the specified inlet velocity, the time for a fluid particle to pass through the model channel is estimated to be 1 s, called flow-through time. The LES results are statistically reliable, as they cover 50 flow-through times (from 10 to 60 s of model time).

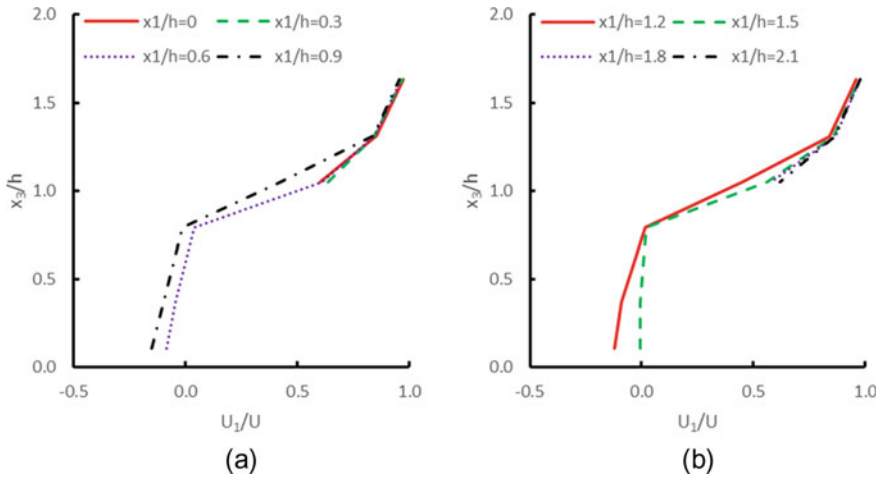
### 3.3 Velocity Field and Its Temporal Stability

The purpose of this section is to evaluate the temporal evolutions of velocity in a cavity. The time-averaged velocities over five increasing durations, namely 2, 5, 10, 20, and 50 s, are illustrated in Fig. 8a–e. Figure 8f exhibits the cross-section for the extraction of LES results. This cross-section, at the channel centerline, is the ninth cavity from the inlet. Patterns of the time-averaged velocities feature a single recirculation in the whole cavity. This is reported from the experiments of Li and Li [6] and LES of Zhang and Li [16].

A fully converged flow is a necessity for the investigation of temporal evaluations of fluctuating velocities. The mean velocity field for the 50 s duration (Fig. 8e) provides a benchmark. The differences between the average flow fields over the other four durations and this benchmark are 26.14%, 18.76%, 8.93%, and 6.88%, respectively (Fig. 8a–e). These results indicate that measurements of velocities for



**Fig. 8** Longitudinal cross-section at the channel centerline, showing the magnitude of  $(U_1^2 + U_3^2)^{0.5}$ , from LES computations, for the ninth cavity from the inlet. The duration is: **a** 2 s; **b** 5 s; **c** 10 s; **d** 20 s; **e** 50 s. There are 50 samples in 1 s duration. Panel **f** shows the cross-section in cavity



**Fig. 9** Vertical profiles of the velocity component  $U_1$ , normalized by the cross sectionally averaged velocity at the inlet  $U$ , at eight selected  $(x_1/h, x_2/h)$  locations: **a** (0, 0); **b** (0.3, 0); **c** (0.6, 0); **d** (0.9, 0); **e** (1.2, 0); **f** (1.5, 0); **g** (1.8, 0); **h** (2.1, 0)

a duration of 2 s may provide qualitative flow patterns, but they are insufficient to eliminate uncertainties.

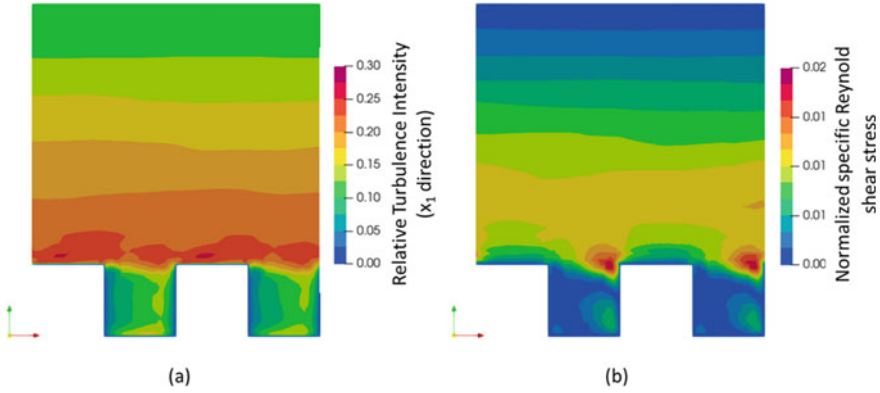
The distributions of time-averaged velocities are compared to experimental results [6]. To overcome potential uncertainties, the duration for average velocity is chosen as 60 s (5000 samples). Figure 9 illustrates vertical profiles of streamwise velocity at selected locations at the channel centerline.

The velocity is normalized by the mean inlet velocity ( $U = 0.541$  m/s), and the height is normalized by the RE dimension,  $h$ . The coordinates  $(x_1, x_2, x_3)$  are the same as described in Sect. 3.1.

In Fig. 9, the vertical coordinates of the data points,  $x_3/h$ , for each of the locations are 0.11, 0.37, 0.79, 1.05, 1.31, and 1.63. Most of the data points are close to the measurements of Li and Li [6] and show the same shapes, although there are some discrepancies for the location  $x_1/h = 1.5$ . The measurements show a gradual decrease of velocity toward the bottom, but the LES results give an almost constant velocity. A plausible reason is that eddies in this near-bottom region are not resolved by the mesh but are simply removed by the SGS model.

### 3.4 Turbulence Intensity and Reynolds Shear Stress

The LES computations produce the velocity field with instantaneous values,  $U_i$ , in three directions. The mean-flow velocity field can be calculated over a selected duration. The selected duration is 50 s with 5000 samples. This section aims to use the velocity output to quantify turbulence intensity and specific Reynolds shear stress.



**Fig. 10** Longitudinal cross-section at the channel centerline, showing contours of: **a** Turbulence intensity and **b** specific Reynold shear stress

The turbulence intensity in the  $x_i$ -direction is given by

$$I_i = \frac{\sqrt{u'_i u'_i}}{U} \quad (12)$$

where  $u'_i$  is the turbulence fluctuation;  $U$  is the mean velocity.

The specific Reynolds shear stress,  $\tau_{ij} = -\overline{u'_i u'_j}$ , is normalized by  $U^2$ . Figure 10 exhibits the turbulence intensity in the  $x_1$ -direction and the normalized specific Reynolds shear stress in the  $x_1 x_2$ -plane. The longitudinal plane is at the channel centerline and cuts through the ninth and tenth roughness elements (REs) from the inlet.

Figure 10a shows that the intensity increases from the water surface to the channel bed, except the cavities, where the values are lower. The highest intensity appears near the top faces of REs, which is approximately 25%. In the cavity, the relatively high value, around 15%, occurs near the windward face of RE cube and the cavity bottom, while near the leeward face of RE cube, the intensity is relatively low, around 5%.

The specific Reynolds shear stress (Fig. 10b) increases with increasing depth below the water surface. The distribution of values is more or less linearly in the upper half of the free-flow region; in the lower half, the distribution is rather irregular because of the REs. The highest Reynolds shear stress appears near the RE's upper edge facing the water flow. In the cavity, the values are relatively low and close to zero in a large part of the area.

## 4 Conclusion

The effects of roughness elements at the channel bed on turbulent flow are numerically investigated in this study. The scope of the work includes Reynolds-averaged Navier–Stokes (RANS) Equation simulation, Large Eddy Simulation (LES), and a comparison of LES results with experimental data.

The conclusions have been reached: A hybrid numerical method incorporating RANS and LES computations allows cross-validation of the models. The method improves the accuracy and efficiency of computations. The ensemble average flow field from the LES results is in reasonable agreement with the experimental data. The quality of LES mesh can be validated through an integral length scale derived from a RANS simulation. The resolved TKE in LES aligns with the expectation; the highest TKE appears near the top and windward faces of roughness elements. An analysis of temporal evolutions of the flow field shows that a short duration of model time (2 and 5 s in this study) is insufficient to remove uncertainties in the ensemble average flow field. It is recommended to use a longer duration (more samples) to determine time-averaged velocity field. This numerical method discussed in this paper is reliable, as confirmed by the comparison with experimental data. Some discrepancies exist in the LES results but can be improved by resolving more eddies with a revised mesh. In the region above the roughness elements, the turbulence exhibits roughly the same behavior as in free flow and is subjected to little influence from the elements.

**Acknowledgements** This study received financial support from NSERC through Discovery grants held by S. S. Li.

## References

1. Ferziger JH, Peric M (2004) Computational methods for fluid dynamics. Rashtriya Printers, Delhi, India
2. Galia T, Hradecký J (2014) Morphological patterns of headwater streams based in Flysch Bedrock: examples from the outer Western Carpathians. CATENA 119:174–183
3. Kalitzin G, Medic G, Iaccarino G, Durbin P (2005) Near-wall behavior of RANS turbulence models and implications for wall functions. J Comput Phys 204:265–291
4. Kennedy JF (1963) The mechanics of dunes and antidunes in erodible-bed channels. J Fluid Mech 16(4):521–544
5. Lee SB (2017) A study on temporal accuracy of OpenFOAM. Int J Naval Archit Ocean Eng 9(4):429–438
6. Li J, Li SS (2020) Near-bed velocity and shear stress of open-channel flow over surface roughness. Environ Fluid Mech 20:293–320
7. Linnansaari T, Monk W, Baird D, Curry R (2013) Review of approaches and methods to assess environmental flows across Canada and internationally. Canadian Science Advisory Secretariat, Ottawa, Canada
8. Maele KV, Merci B (2008) Application of RANS and LES field simulations to predict the critical ventilation velocity in longitudinally ventilated horizontal tunnels. Fire Saf J 43(8):598–609
9. Menter FR (1994) Two-equation Eddy-viscosity turbulence models for engineering applications. AIAA J 32(8):1598–1605

10. Mignot E, Brevis W (2020) Coherent turbulent structures within open-channel lateral cavities. *J Hydraul Eng ASCE* 146(2):04019066
11. Pope SB (2000) *Turbulent flows*. Cambridge University Press, Cambridge
12. Radecki-Pawlik A, Pagliara S, Hradecký J (2018) *Open channel hydraulics, river hydraulic structures and fluvial geomorphology*. Taylor & Francis, Boca Raton, FL, USA
13. Sturm TW (2001) *Open channel hydraulics*. McGraw-Hill, New York
14. Yager E, Kirchner J, Dietrich W (2007) Calculating bed load transport in steep boulder bed channels. *Water Resources Res* 43:W07418
15. Yalin M (1971) On the formation of dunes and meanders. In: *Proceedings of the 14th congress of the international association for hydraulic research*. IAHR, Paris, France, pp C101–C108
16. Zhang Z, Li SS (2020) Large Eddy simulation of near-bed flow and turbulence over roughness elements in the shallow open-channel. *Water* 12(10):2701

# Real-Time Water Distribution System Calibration Using Genetic Algorithm



Ziyuan Cai, Rebecca Dziedzic, and S. Samuel Li

**Abstract** This paper aims to calibrate, in real-time, a hydraulic model by combining an implicit technique with a genetic algorithm (GA). The approach uses demand multipliers as the calibration parameter. The hydraulic model yields an optimal solution through minimizing a nonlinear objective function between observed and simulated values of nodal pressures and pipe flow rates at a five-minute time interval, subject to a set of explicit bound constraints for the demand multipliers and implicit nonlinear hydraulic constraints. The hydraulic modeling software EPANET is applied to solve hydraulic constraints, including the equations of mass conservation for each junction and energy conservation for each pipe, and to retrieve the simulated values required in the objective function after assigning demand multipliers to nodes in the hydraulic model. The simulated values are computed under the same boundary conditions, including tank levels and pump status, as the measurements are collected. The approach is applied to the L-Town hypothetical network as a case study, yielding not only simulated values in good agreement with measurements but also consistent variation trends over time between simulated values and measurements. The case study shows that demand multipliers expressed in real numbers lead to more accurate simulated values, closer to the measurements, than expressed in a string of bits. Further efforts will focus on leak detection using the well-calibrated model.

**Keywords** Water distribution system • Calibration • Genetic algorithm

## 1 Introduction

A hydraulic model for the computation of flow rates and pressures in a water distribution system (WDS) needs a good calibration. The calibration is of great importance to the design, analysis, and operation of WDSs. A real-world WDS carries a large

---

Z. Cai (✉) · R. Dziedzic · S. Samuel Li

Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

e-mail: [caiziyanca@163.com](mailto:caiziyanca@163.com)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_46](https://doi.org/10.1007/978-3-031-34593-7_46)

volume of water on a daily basis, and the water demand differs at different times of the day. The hydraulic model of an existing WDS must respond accordingly in real time and mimic the actual status and behavior of the system; this is to ensure the usefulness of the model for investigating the effects of proposed changes [8].

The calibration of a WDS model involves adjustments of network parameters (e.g., nodal demand multiplier and pipe roughness) in order to achieve a close agreement between model predictions (e.g., flow rates and pressures) and field measurements. Ormsbee and Lingireddy [9] proposed a seven-step calibration process: (1) identify the intended use of the model; (2) determine initial estimates of network parameters; (3) collect calibration data including nodal pressures and pipe flows; (4) evaluate the model predictions; (5) perform a macro-level calibration; (6) perform a sensitivity analysis; (7) perform a micro-level calibration. In the last step, a steady-state and extended period calibration should be considered for most cases. Depending on the calculation methods, calibrations may be grouped into three categories: (1) iterative (trial-and-error) procedures; (2) explicit techniques (also called hydraulic simulation models); (3) implicit calibration (or optimization models) [11].

The iterative procedures are usually applied in a static hydraulic analysis, which are implemented by using pressure heads and/or flow rates from solving the mass and energy conservation equations, to update unknown network parameters. To reduce an excessively large number of network parameters, skeletonization is usually required. However, the iterative procedures suffer a very slow convergence rate [1] and thus are not suitable for the calibration of a large WDS.

The explicit techniques are applied to optimize one or more network parameters, based on the addition of one or more continuity or energy equations for the same network. Ormsbee and Wood [6] employed this approach for the determination of head-loss adjustment factors and head-loss calibration coefficients.

The implicit calibration refers to a nonlinear optimization problem, to be solved by using an optimization technique coupled with a hydraulic model solver. Implicit calibrations of hydraulic models have been applied in a large number of studies [2, 4, 5, 7, 12, 13, 15].

In the optimization problem, a nonlinear objective function,  $F$ , is proposed. This function is subjected to both linear and nonlinear constraints. The process of optimization uses the optimization tool (e.g., GA) to generate and update network parameters. These parameters are subjected to the explicit constraints and are put onto the computer simulation model of a WDS. Then, EPANET is called to predict variables, such as nodal pressures and pipe flow rates. Finally, the optimization tool employs the objective function to minimize the discrepancies between the measured and predicted variables.

The implicit calibration problem can be formulated as follows [7]:

$$\text{Minimize: } F(X), \quad (1)$$

$$\text{Constraints: } g(X) = 0, \quad (2)$$

$$C_L < X < C_U, \quad (3)$$

where  $X$  is a vector of network parameters;  $g(X)$  is a set of implicit network constraints (network mass balance and energy equations);  $C_L$  and  $C_U$  are the explicit constraints of upper and lower limits on network parameters.  $F$  is a function of  $X$ .

Many parameters carry uncertainties and thus affect nodal pressure heads and pipe flow rates. The pipe roughness coefficient and the demand multiplier are two of these parameters. These two parameters reportedly cause a mismatch of pressure heads and flow rates between a real-world WDS and a hydraulic model of the WDS in most cases of static calibration. The roughness coefficient varies only over a long period of time, whereas the demand multiplier can vary in a very short period of time and is therefore a sensitive parameter. To minimize overprediction/underprediction errors of network parameters [14], other conditions such as tank levels and the pump operation status should also be adjusted. The tank levels and pump operation status are usually implemented as boundary conditions to ensure that the simulated variables are under the same conditions as the measurements are collected [15].

Using a genetic algorithm (GA), this paper implements the implicit calibration at each time step during a simulation for an extensive period of time and selects water demand multipliers as the calibration parameter. The optimization of the multipliers is fed by pipe flow rates and nodal pressures [5]. This paper demonstrates that the implicit approach along with GAs is effective for model calibration, and multipliers expressed in real numbers lead to more accurate simulation results, closer to measurements, than multipliers given in a string of bits. The remaining part of this paper describes the methodologies, followed by discussion of results from a case study, before conclusions are drawn.

## 2 Hydraulic Model Calibration

### 2.1 Network Parameter

The network parameter selected to be calibrated is demand multipliers at each time step during a hydraulic model simulation. The entire WDS is divided into several demand zones based on different types of customers, and a demand multiplier is computed for each demand zone. Therefore, there are two demand multipliers at each time step  $t$ : residential demand multiplier,  $M_{r,t}$ , and commercial demand multipliers,  $M_{c,t}$ .

The explicit bound constraint used to set limits on each demand multiplier is given by

$$M_{\min} \leq M \leq M_{\max}, \quad (4)$$



where  $M_{\min}$  and  $M_{\max}$  are the lower and upper limits on decision variables, respectively. When setting the limits, a larger range of  $M$  values should be given, compared to the range of  $M$  values in the uncalibrated model, to avoid being trapped in the local optimization [16] and used to generate the simulated variables. In the uncalibrated model, the original ranges of  $M_{r,t}$ , and  $M_{c,t}$  values are  $M_{r,t} \approx 0.2\text{--}1.5$  and  $M_{c,t} \approx 0.5\text{--}1.25$ . In this paper, the bounded ranges are set to  $M_{r,t} \approx 0.1\text{--}1.6$  and  $M_{c,t} \approx 0.4\text{--}1.3$ .

## 2.2 Objective Function

The objective function applies a least-squares method to minimize the discrepancies between nodal pressures and pipe flow rates at the measurement locations, which is given by

$$\text{Minimize } F = \sum_{i=1}^n W_P (P_{oi,t} - P_{si,t})^2 + \sum_{j=1}^m W_Q (Q_{oj,t} - Q_{sj,t})^2, \quad (5)$$

where  $P_{oi,t}$  and  $P_{si,t}$  are, respectively, the measured and model-simulated pressure heads at the  $i$ -th node at the  $t$ -th timestep;  $Q_{oj,t}$  and  $Q_{sj,t}$  are, respectively, the measured and model-simulated flow rates in the  $j$ -th pipe at the  $t$ -th timestep;  $n$  and  $m$  are, respectively, the location IDs of nodal pressure head measurements and pipe flow rate measurements;  $W_P$  and  $W_Q$  are, respectively, the weighting factors for nodal pressure head and pipe flow rate.

In Eq. (5), the field measurements ( $P_o$ ,  $Q_o$ ) at the time step  $t$  were collected by the sensors, while the simulated values ( $P_s$ ,  $Q_s$ ) at the given time step were obtained from a hydraulic simulation toolkit OWA-EPANET. The goodness-of-fit of calibration using a GA is the reciprocal of  $F$  in Eq. (5). The steps of retrieving the model-simulated data are as follows: (1) assign demand multipliers randomly generated by GA to each time pattern in the L-Town WDS model; (2) EPANET, which is a WDS modeling software developed by the US Environmental Protection Agency's (EPA) Water Supply and Water Resources Division, is called to predict variables in the network, including locations where nodal pressures, and pipe flow rates were measured.

In Lingireddy and Ormsbee [5], normalization weights for nodal pressure head and pipe flow rate are given by

$$W_P = [100/\text{Max}(P_{oi,t})]^2 \quad \text{for } i = 1, \dots, n; \quad t = 1, \dots, T, \quad (6)$$

$$W_Q = [100/\text{Max}(Q_{oj,t})]^2 \quad \text{for } j = 1, \dots, m; \quad t = 1, \dots, T. \quad (7)$$

### 2.3 Constraints

The explicit bound constraints for network parameters are described in Sect. 2.1.

The implicit hydraulic constraints consist of mass conservation and energy equations for each node and pipe. EPANET employs the global gradient algorithm, which is an iterative technique, to simultaneously solve the two equations for nodal pressure heads and pipe flow rates at a particular point in time [10].

The mass conservation at a node requires that the total inflow,  $Q_{in}$ , minus the total outflow,  $Q_{out}$ , at a node must be equal to the external inflow or demand,  $Q_e$ , at the given node, which is expressed as

$$\sum Q_{in} - \sum Q_{out} = Q_e. \quad (8)$$

The conservation of energy in a pipe connecting two nodes,  $i$  and  $j$ , requires that the sum of head losses,  $h_{ij}$ , over the given pipe equals the difference in hydraulic heads between the two nodes

$$h_{ij} = h_i - h_j, \quad (9)$$

where  $h_i$  and  $h_j$  are the hydraulic heads at node  $i$  and  $j$ , respectively.

The conservation of energy in a pump connecting nodes  $i$  and  $j$  is expressed as

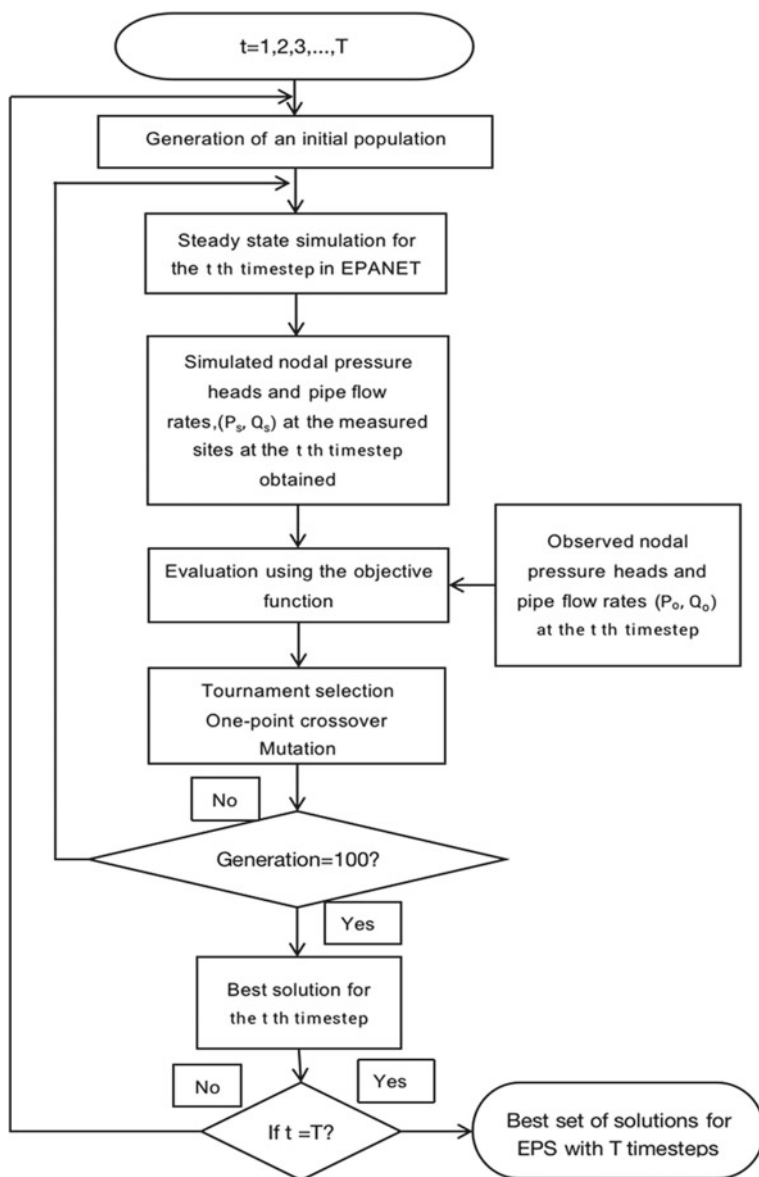
$$h_{ij} = -\omega^2 \left( h_0 - C_1 (q_{ij}/\omega)^{C_2} \right), \quad (10)$$

where  $\omega$  is the pump relative speed factor;  $h_0$  is the pump's usual head when not operating;  $C_1$  and  $C_2$  are pump curve coefficients [3].

### 2.4 Genetic Algorithm

The GA for the calibration of demand multipliers is coded in the Python language. As shown in Fig. 1, GAs start with generating an initial population with 20 potential solutions. After assigning one of two demand multipliers in the solution to each node of the model successfully, a steady-state simulation using the hydraulic model (EPANET) starts. At time step  $t$ , the simulated nodal pressure heads and pipe flow rates ( $P_s$ ,  $Q_s$ ) at the measurement locations are extracted from the simulation results. Then, the objective function (Eq. 5) is applied to determine the discrepancies between the measured and simulated values. The goodness-of-fit of each potential solution is defined by taking an inverse of the difference. The solutions with the two highest scores (or goodness-of-fit) are selected as parents, and they, together with their offspring, form a new population that inherits the features of previous population by performing selection, crossover, and mutation operator. The calibration process is repeated until the optimal vector of two demand multipliers is returned

or a maximum of 100 generations is reached. This whole GA process aims to run a steady-state simulation at each time step  $t$  in an extended period simulation (EPS). For a simulation with  $T$  time steps, the GA calibration needs to be implemented  $T$  times as real-time model calibration.



**Fig. 1** Flowchart for the calibration of demand multipliers using genetic algorithm

**Table 1** Parameters and their values in the GA

Parameter	Description
Maximum number of generations	100
Selection strategy	Tournament selection with a size of 3
The number of crossover points	1
Crossover rate	0.8
Mutation strategy	Bitwise mutation
Mutation rate	0.03

The GA used in this study performed a tournament selection with a tournament size of three due to its better convergence than roulette wheel selection [17]. A crossover operator applied one-crossover point and the crossover rate of 0.8. The mutation possibility is 0.03. Table 1 shows parameters used in the GA.

### 3 Results

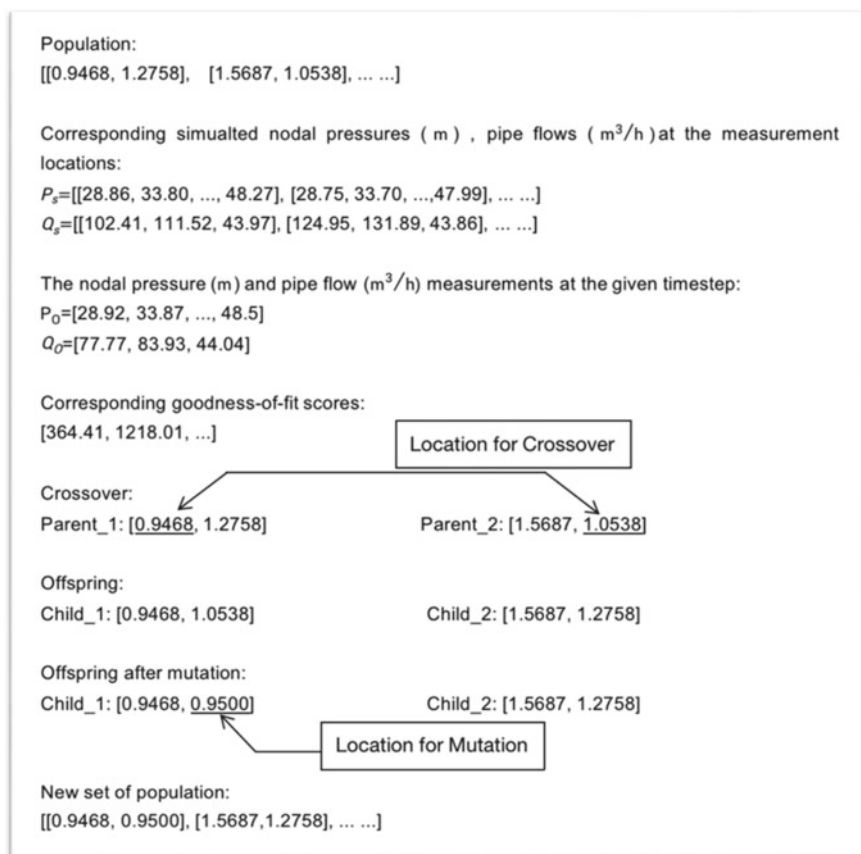
#### 3.1 Description of Water Distribution System

The L-Town water distribution system (Fig. 2) consisted of 785 junctions, including 782 nodes, 2 reservoirs, 1 tank, and 906 links (905 pipes plus 1 pump). In the network, one tank level sensor, three flow sensors, and 33 pressure sensors were installed. These sensors transmitted their measurements every 5 min to the L-Town water utility's Supervisory Control and Data Acquisition (SCADA) system.

#### 3.2 Results of Each Step of the GA

The first two candidate solutions in the first run of simulation at the first time step were taken as an example to illustrate the output of each step of the GA (Fig. 3). Each candidate solution was measured based on its the weighted sum of squared error between its corresponding sets of simulated nodal pressure heads and pipe flow rates from EPANET and its measured counterparts at the given time step. The first candidate solution [0.9468, 1.2758] resulted in simulated pressure heads [28.86, 33.80, ..., 48.27] (m) at the 33 nodes and pipe flow rates [102.41, 111.52, 43.97] ( $\text{m}^3/\text{h}$ ) in the three pipes, while the second candidate solution [1.5687, 1.0538] led to simulated pressure heads [28.75, 33.70, ..., 47.99] (m) and pipe flow rates [124.95,





**Fig. 3** Process and output of each step of the GA

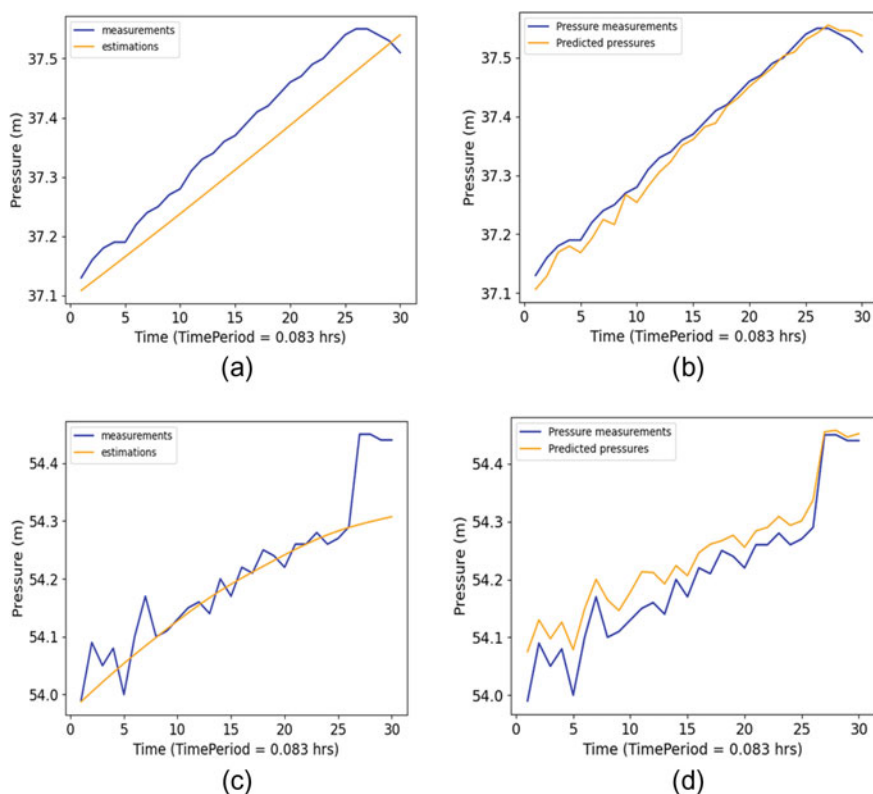
affects the quality of L-Town WDS model calibration. It took around 10 min of CPU time to obtain the optimal solutions for the first 30 time steps on a personal computer with a Ryzen-7 1.80 GHz processor.

A comparison of pressure heads between before calibration and after calibration is shown in Fig. 4. After calibration, the variation trends of simulated pressure heads at nodes 31 and 114 are consistent with those of the measured pressure heads. During the first 30 time steps, the largest and smallest differences between simulated and measured pressure heads are, respectively, 0.076 and 0.005 m at node 31 and  $0.156$  and  $2.17 \times 10^{-5}$  m at node 114, before the model calibration, in comparison to about 0.036 and 0.000 m at node 31 and to 0.093 and 0.006 m at node 114, after the model calibration. At node 31, the largest and smallest pressure head differences decrease by 0.040 and 0.005 m, respectively. At node n114, the largest discrepancy drops by 0.063 m after calibration. Although the smallest difference at node 114 is larger after the calibration than before, this increment is negligible and has no adverse impacts

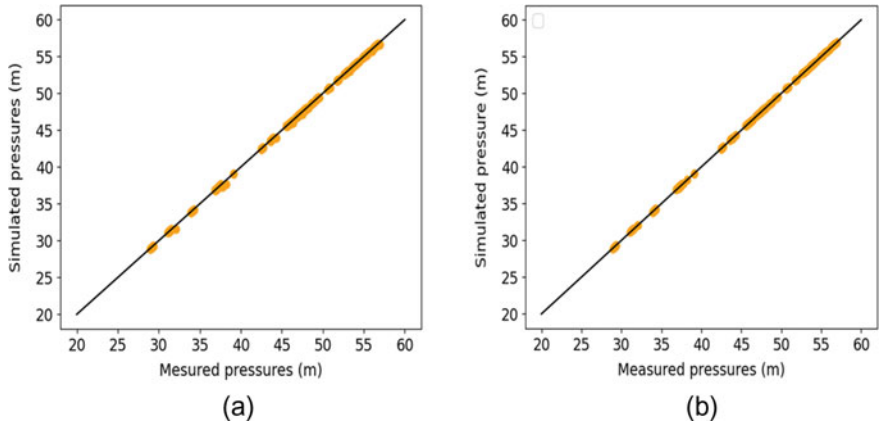
on the quality of calibration. When the demand multipliers are expressed in binary format, after calibration, the largest and smallest differences at the two nodes are, respectively, 0.055 and 0.005 m and 0.11 and 0.03 m during the first 30 time steps.

In Fig. 5, the pressure heads at the 33 nodes with sensors for the first 30 time steps are compared between before and after calibration. Before calibration, the simulated pressure heads are slightly below the measured values (Fig. 6a); after calibration, the simulated pressure heads agree well with the measurements (Fig. 6b). For instance, after calibration, the simulated pressure head at node 740 is around 43.80 m at the first time step, matching the measured value of 43.81 m, while before calibration, the simulated pressure head is approximately 43.77 m.

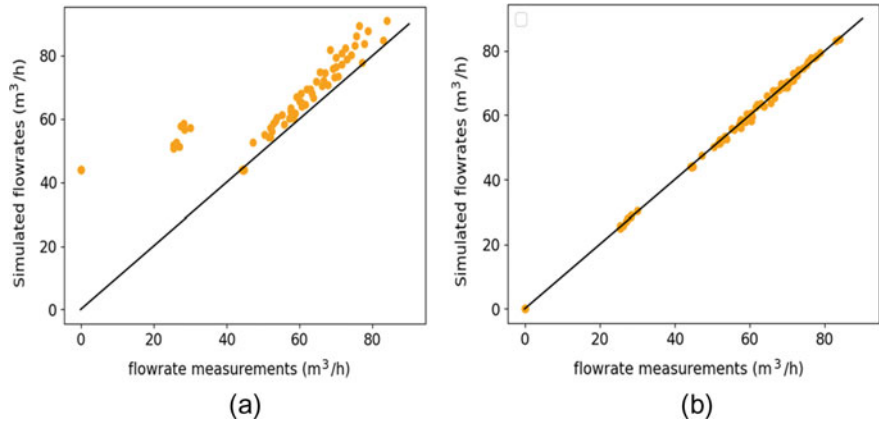
In Fig. 6, the flow rates in the three pipes with sensors before and after calibration at the first 30 time steps are plotted. Before calibration, the simulated flow rates are larger than the measured values (Fig. 6a). After calibration, a good agreement



**Fig. 4** Time series of measured and simulated pressures: **a, c** Before calibration; **b, d** after calibration. Panels **a** and **b** are for node 31, and panels **c** and **d** are for node 114



**Fig. 5** Pressure heads at the 33 nodes with sensors: **a** Before calibration; **b** after calibration



**Fig. 6** Flow rates in the three pipes with sensors: **a** Before calibration; **b** after calibration

between the simulated and measured flow rates is achieved. For example, the simulated flow rate in pipe 227 at the first time step is about 83.85 m<sup>3</sup>/h before calibration, while after calibration, the simulated flow rate is around 77.86 m<sup>3</sup>/h, compared with the measured flow rate of 77.77 m<sup>3</sup>/h.

In Table 2, the uncalibrated and calibrated demand multipliers ( $M_r$ ,  $M_c$ ) at each time step are summarized. In the process of calibration of demand multipliers, the global values of  $M_r$  and  $M_c$  are simultaneously optimized. The weighted sum of squared differences (WSSD) between simulated and measured pressure heads and pipe flow rates at the measurement locations is compared between before calibration (WSSD<sub>b</sub>) and after calibration (WSSD<sub>a</sub>). The results confirm the suitability of calibration.



**Table 2** GA calibration model results

Time step	Uncalibrated $M_r$	Uncalibrated $M_c$	WSSD <sub>b</sub>	Calibrated $M_r$	Calibrated $M_c$	WSSD <sub>a</sub>
1	0.7729	0.9174	22.60	0.9280	0.5076	1.11
2	0.7480	0.9053	70.39	0.6741	0.6911	1.55
3	0.7230	0.8929	44.61	0.6691	0.6911	0.85
4	0.6980	0.8802	54.12	0.7741	0.4635	0.78
5	0.6751	0.8673	1.34	0.6082	0.9282	0.85
6	0.6525	0.8541	28.81	0.4989	0.8497	0.98
7	0.6305	0.8406	70.90	0.7644	0.2204	4.08
8	0.6089	0.8269	12.51	0.7644	0.4656	0.58
9	0.5874	0.8128	17.32	0.7644	0.3918	1.43
10	0.5659	0.7988	11.09	0.7644	0.3918	0.48
11	0.5446	0.7848	22.60	0.4831	0.6756	0.41
12	0.5236	0.7714	29.58	0.3885	0.7318	0.33
13	0.5031	0.7585	5.65	0.4605	0.7318	0.43
14	0.4833	0.7465	20.44	0.5781	0.4284	0.43
15	0.4643	0.7351	8.42	0.4584	0.6186	0.38
16	0.4459	0.7242	21.65	0.4584	0.4973	0.56
17	0.4279	0.7136	8.25	0.4903	0.5256	0.48
18	0.4103	0.7029	28.29	0.3776	0.5256	0.30
19	0.3929	0.6919	18.45	0.2179	0.7502	0.37
20	0.3757	0.6807	4.48	0.4206	0.5153	0.42
21	0.3590	0.6693	16.20	0.2028	0.7205	0.52
22	0.3430	0.6580	9.65	0.2516	0.6505	0.47
23	0.3279	0.6473	15.52	0.2914	0.5337	0.27
24	0.3139	0.6374	2.29	0.1933	0.7528	0.38
25	0.3013	0.6287	4.07	0.1933	0.7146	0.54
26	0.2899	0.6211	16.29	0.1123	0.7040	0.50
27	0.2795	0.6146	930.93	0.2397	0.4896	0.10
28	0.2700	0.6088	932.85	0.2397	0.4750	0.03
29	0.2610	0.6034	855.41	0.2335	0.5613	0.10
30	0.2523	0.5980	880.87	0.2228	0.5177	0.02

4 Conclusions

The GA, coupled with an implicit model, is successfully applied to calibrate the demand multiplier at each time step during a model simulation. The calibration provides a good basis for a large simulation duration. The trend of simulated nodal pressures over time is consistent with that of measurements. Therefore, the model

mimics realistically the physical and operational behaviors of a WDS. It is shown that demand multipliers expressed in real numbers lead to simulation results in better agreement with measurements than expressed in a string of bits. The running time spent for each steady-state step of the first 30 time steps is about 10 min, showing the efficiency of application of the proposed method on the real-time calibration.

Different GA parameters can be adjusted to realistically simulate a real-world WDS. The more accurate the parameters, the less discrepancies between the simulation results and measurements.

Further research efforts should be made to detect leakages of the L-Town WDS, using the well-calibrated model from this study. At each time step, uncertainties caused by demand multipliers have been minimized. The remaining differences between the measured and simulated values of nodal pressures and pipe flow rates should mainly result from leakages. Therefore, the calibration of emitter coefficients at each node would be the priority of further research.

**Acknowledgements** This study received financial support from Concordia University through Faculty Research Support held by R. Dziedzic and S.S. Li.

## References

1. Bhave PR (1988) Calibrating water distribution network models. *J Environ Eng ASCE* 114(1):120–136
2. Do NC, Simpson AR, Deuerlein JW, Piller O (2016) Calibration of water demand multipliers in water distribution systems using genetic algorithms. *J Water Resources Plann Manag ASCE* 142(11):04016044
3. Hossain S, Hewa GA, Chow CWK, Cook D (2021) Modelling and incorporating the variable demand pattern to the calibration of water distribution system hydraulic model. *Water* 13(20):2890
4. Jamasb M, Tabesh M, Rahimi M (2009) Calibration of EPANET using genetic algorithm. In: *Proceedings of 10th annual water distribution systems analysis conference*, 17–20 Aug 2009, Kruger National Park, South Africa
5. Lingireddy S, Ormsbee LE (2002) Hydraulic network calibration using genetic optimization. *Civ Eng Environ Syst* 19(1):13–39
6. Ormsbee LE, Wood DJ (1986) Explicit pipe network calibration. *J Water Resources Plann Manag ASCE* 112(2):166–182
7. Ormsbee LE (1989) Implicit network calibration. *J Water Resources Plann Manag ASCE* 115(2):243–257
8. Ormsbee LE, Chase DV, Grayman W (1992) Network modeling for small water distribution systems. In: *Proceedings of the American water works association 1992 computer conference*, 12–15 Apr 1992, Nashville, TN
9. Ormsbee LE, Lingireddy S (1997) Calibrating hydraulic network models. *J Am Water Works Assoc* 89(2):42–50
10. Rossman LA (2000) EPANET2. EPA/600/R-00/057, Water Supply and Water Resources Division, National Risk Management Research Laboratory, Office of Research and Development, U.S.EPA, Cincinnati
11. Savic DA, Kapelan ZS, Jonkergouw PMR (2009) Quo Vadis water distribution model calibration? *Urban Water J* 6(1):3–22

12. Savic DA, Walters GA (1995) Genetic algorithm techniques for calibrating network models. Tech. Rep. 95/12. Center for Systems and Control Engineering, University of Exeter, UK
13. Shu S, Zhang D (2010) Calibrating water distribution system model automatically by genetic algorithms. In: Proceedings of the IEEE international conference on intelligent computing and integrated systems, 22–24 Oct 2010, Guilin, Guangxi, China, pp 16–19
14. Walski TM (2001) Understanding the adjustments for water distribution system model calibration. *J Indian Water Works Assoc* 33(2):151–157
15. Wu ZY, Walski T, Mankowski R, Herrin G, Gurrieri R (2002) Calibrating water distribution model via genetic algorithm. In: Proceedings of the AWWA information management and technology conference and exposition, 16–19 Apr 2002, Kansas City, Missouri, USA, pp 1–10
16. Yu Z, Tian Y, Zheng Y, Zhao X (2009) Calibration of pipe roughness coefficient based on manning formula and genetic algorithms. *Trans Tianjin Univ* 15(1):70–74
17. Zhong J, Hu X, Gu M, Zhang J (2005) Comparison of performance between different selection strategies on simple genetic algorithms. In: Proceedings of the IEEE international conference on computational intelligence for modeling, control and automation and international conference on intelligent agents, web technologies and internet commerce, 28–30 Nov 2005, Vienna, Austria, vol 2, pp 1115–1121

# Flow Field Characteristics of Particle-Laden Swirling Jets



F. Sharif and A. H. Azimi

**Abstract** Particle-laden turbulent jets have various engineering applications such as effluent discharge, wastewater disposal, and marine bed capping. The flow field of particle-laden swirling jets impinged into stagnant water was investigated using the combined high-speed imaging and Particle Image Velocimetry (PIV) measurements. A swirling chamber was designed with an inner diameter of 32 mm encompassed by six internal spiral water paths each having 4 mm inner diameters. Sand particles with a median diameter of 0.386 mm and an initial concentration of 60% were added through a funnel into the chamber to mix with the swirling water, and the mixture was discharged into stagnant water through an outlet nozzle with a diameter of  $d_o = 6$  mm. The axial and radial distributions of velocity components were measured for two different swirling numbers of 0.50 and 0.65. It was found that sand–water swirling jets expanded moderately under induced recirculatory motion, formed a wider spray cone, and high radial velocity of sand particles. Due to formation of Precessing Vortex Core (PVC), the horizontal velocity profiles in the radial direction,  $r$ , experienced a near-field peak at  $r/d_o = 2$ . The axial velocity decay rate of jets with swirling numbers of 0.50 was found to be dissipated rapidly by increasing the swirling intensity from 0.50 to 0.65. The swirling motion increased the radial velocity distribution in comparison to the non-swirling sand–water jets. Two different clockwise and counterclockwise rotating vortices were observed throughout the recirculation zone which linked the internal and external shear layers. The flow oscillation and inter-scale dominant structures of the swirling jets were also investigated using the Spectral Proper Orthogonal Decomposition method (SPOD) and extracting energy-ranked spectra. It was found that a significant portion of turbulence kinetic energy of the jet was concentrated in the first five low-frequency modes.

**Keywords** Flow field characteristics · Particle-laden swirling jets

---

F. Sharif · A. H. Azimi (✉)

Department of Civil Engineering, Lakehead University, Thunder Bay, ON, Canada

e-mail: [azimi@lakeheadu.ca](mailto:azimi@lakeheadu.ca)

F. Sharif

e-mail: [fsharif@lakeheadu.ca](mailto:fsharif@lakeheadu.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_47](https://doi.org/10.1007/978-3-031-34593-7_47)

## 1 Introduction

A single-phase turbulent jet can be generated by discharging fluid from a nozzle into an ambient with similar physical properties and density (e.g., air in air, water in water). A mixture of water jet with sand particles is called slurry or particle-laden jets. The presence of sand particles can form a jet effluent having a combined density greater than single-phase jet and led to more complexity of the flow system. Particle-laden turbulent jets are consisted of a central main carrier phase and a discrete phase. They have a wide range of engineering applications such as in wastewater disposals, marine bed capping, and mixing processes [1, 2, 4, 14, 15, 19–22] (Manzouri and Azimi 2019a, b). A swirling jet can be formed when an issuing flow rotates under the azimuthal velocity. The efficiency and mixing capability of the turbulent jet can be enhanced under the swilling intensity due to formation of recirculation zone. The recirculation zone in the shear layer is characterized by strong anisotropic turbulence behavior [29]. Swirling jets have several industrial and engineering applications in bluff-body combustion systems and the design of jet pumps [7, 9, 23, 24, 31]. Introducing swirling motion and formation of recirculation zone (RZ) leads to large-scale instabilities known as vortex breakdown or Precessing Vortex Core (PVC) [18, 28]. Vortex breakdown occurs when the swirl intensity is reached to a specific point and is accompanied by an abrupt change in the structure of a vortex core [11]. The common dominant instability in swirling flows is the formation and evolution of Precessing Vortex Core (PVC) at a certain distance and frequency level corresponding to swirling strength [7, 17] (Syred 2006).

Dynamics and flow mixing properties of single-phase and two-phase gas–solid swirling jets were investigated by several researchers as a function of swirling number [3, 7, 17]. The swirling number is defined as the ratio of the axial flux of angular momentum to the axial flux of axial momentum [7]. The swirling number can be used to characterize vortex breakdown and particle concentration beyond the vortex core region [5, 13]. It was indicated that increasing the swirling number enhanced particle dispersion toward the PVC region [6]. The Stokes number is defined as the ratio of particle response time to the fluid characteristic time scale. The study of gas–solid swirling jets with a moderate swirling number of 0.47 and Stokes number,  $S_t$ , greater than one indicated that particles are mostly concentrated in the recirculation zone [8]. It was indicated that introducing solid particles in single-phase swirling jets attenuated the flow circulation and vortex formation [9]. Anisotropic particle dispersion was reported in the motion of two-phase swirling jets with ultra-light particles [10]. Most recently, the dynamics of sand–water swirling jets with swirling numbers ranging from 0.45 to 0.60 was studied by Sharif and Azimi [23]. They proposed new correlations for concentration and velocity distributions in the axial and radial directions of the jet as a function of swirling number. It was found that adding the swirling motion enhanced the mixing and spreading rate of the jets in comparison to non-swirling ones.

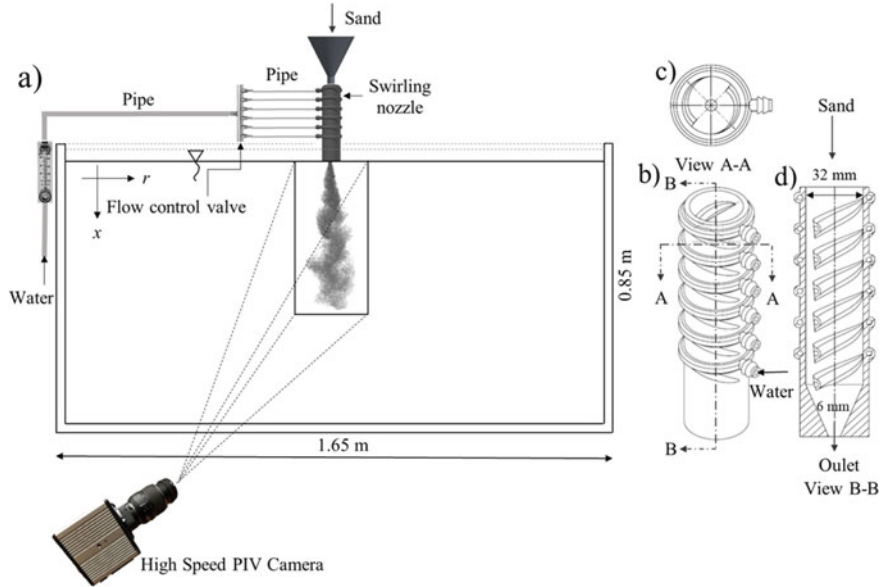
The spatial and dynamic significance of flow structures can be studied by the Spectral Proper Orthogonal Decomposition (SPOD) method. The SPOD method effectively isolates and differentiates the flow coherent structures from the stochastic motions by decomposing the flow field into space-frequency and ranking the modes according to their energy level [25]. The traditional methods such as POD or DMD are not capable of isolating flow coherent motions and energy level. It was indicated that the temporal coefficients of the SPOD modes can be characterized by the Kelvin–Helmholtz (KH) wave packets and linked to the coherent structures in the near-field region of turbulent jets [25, 33].

The previous research study and the analysis of data provided fundamental insight for sand–water swirling jets in stagnant water [23]. The main objective of the present study is to provide more detailed information about the flow field using the planar PIV measurements. The properties of particle velocity components with different swirling numbers were extracted from the PIV measurements. The secondary objective of this paper is to study the low-ranked coherent flow configuration of the swirling sand–water jet using the Spectral Proper Orthogonal Decomposition technique. The SPOD coefficients provide more insight into the frequency-based spatial modes. The results of the SPOD methods were presented to characterize the energy level contained in each mode.

## 2 Experimental Setup

Laboratory experiments were conducted in the Multiphase Flow Research Laboratory (MFRL) at Lakehead University to study the flow features of swirling sand–water jets. As shown in Fig. 1, a rectangular tank made of glass with dimensions of 1.65 m long, 0.85 m wide, and 0.85 m deep, was filled with tap water. The ambient temperature of water was kept constant with a value of  $20\text{ }^{\circ}\text{C} \pm 1\text{ }^{\circ}\text{C}$ .

The swirling jet setup which consists of a cylinder with an inner diameter of 32 mm surrounded by built-in spiral ports was designed. In this system, water flows through the ports and mixes with sand particles that were released from upper part of the cylinder through the funnel. Then, the swirling sand–water mixture impinged through a 6 mm outlet nozzle of the cylinder into stagnant water. The initial flow was set to form a jet with Reynolds numbers of  $Re = 7260$  and  $10,463$ . The chosen flow discharge provided swirling numbers in the range of 0.50 and 0.65. Sand particles with a median diameter of  $D_{50} = 0.386\text{ mm}$  and density of  $\rho_s = 2540\text{ kg/m}^3$  were implemented. The Stokes numbers of the selected tests were calculated as  $S_t = 1.1$  and  $3.7$ . The sand's initial concentration of 60% by volume with an initial mass flux of  $\dot{m}_o = 28\text{ g/s}$  was used. The central jet was connected to the tap water through an accurate flow meter (LZT M-15, UXCEL, China), and it was issued from the inner nozzle. A high-speed camera (Photron-FASTCAM, 1024PCI-100KC) with a resolution of  $1024 \times 1024$  pixels was used to capture PIV images.



**Fig. 1** Experimental setup and swirling cylinder assembly

The camera was equipped with a 15–55 mm AF-S Nikkor, 1:3.5–5.6 GII (Nikon, Japan) lens to capture images. Raw images were captured with a frame rate of 250 frames per second and a shutter speed of 0.004 s with a measurement time of 25 s to reach a steady-state swirling jet. The PIVlab software was utilized which is capable of data analysis including data preprocessing, image evaluation, and postprocessing. The image processing algorithm is based on the Discrete Fourier Transform (DFT) cross-correlation. The two-step iteration algorithm with an interrogation window of  $64 \times 64$  pixels and a subwindow of  $32 \times 32$  pixels was employed. The entrainment and mixing became sufficient to acquire high-quality PIV images for  $x/d_o \geq 20$  (i.e.,  $x = 0.12$  m). Experimental details and non-dimensional parameters are listed in Table 1.

**Table 1** Initial experimental details of sand–water swirling jets

No	Test no	$u_{wo}$ (m/s)	$S_t$	$S_w$	Re
1	A0	0	1.1	0	0
2	A1	1.0	2.6	0.50	7260
3	A2	1.7	3.7	0.65	10,463

### 3 Spectral Proper Orthogonal Decomposition (SPOD)

The SPOD method can represent complex dynamics of flow field by a small number of space-temporal and considers as a variant of POD. Therefore, SPOD provides orthogonal modes at discrete frequencies by applying Welch's decomposition method that is optimally ranked in terms of energy and that evolves coherently in both space and time [25, 30] (Schmidt et al. 2020). Welch [30] proposed an averaging technique to the converting time history dataset from the time domain to the frequency base known as the Power Spectrum Density (PSD). Therefore, temporal coefficients of the orthogonal modes can be explained as an increasing range of frequency spectrum. As a first, a single data matrix consisting of time series of snapshots,  $P$ , was adjusted by  $N_p$  snapshots and  $M_p$  pixels. As the second step, the matrix  $P$  first is segmented into different blocks. The length of each block can be 64, 128, and 256 corresponding to an overlap of 50%. However, by choosing larger block sizes and overlapping them, the variance of data is increased accordingly. Then, each block reshapes based on frequency domain by performing Welch periodogram method. By applying POD to the ensemble of frequency-based blocks,  $P_B$ , the spatial modes of each block can be extracted as a linear weighted sum of the modes as [26]:

$$P_{Bi}(x, t) = \sum_{j=1}^{r_p} \alpha_{ij} \phi_j(x), \quad (1)$$

where  $x$  and  $t$  are the independent variables denoting space and time, respectively,  $i = 1$  to  $N_b$ ,  $j = 1$  to  $r_p < N_b$ . Besides, the coefficient  $\alpha_{ij}(t)$  can be defined as:

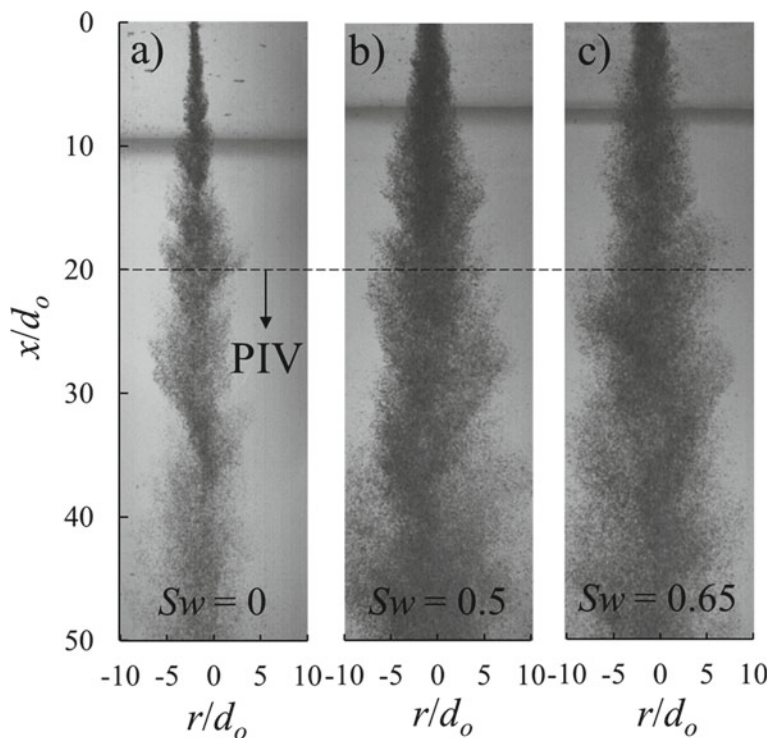
$$\alpha_{ij}(t) = \beta_{ij}(t) \lambda_j^{0.5}, \quad (2)$$

where  $\beta_{ij}$  is the temporal coefficients and  $\lambda_j$  is called the eigenvalue. The orthogonal eigenvectors or spatial modes,  $\phi_j(x)$ , represent the ensemble-averaged spatial flow features of the main time series of snapshots. The eigenvalues in decreasing order ( $\lambda_1 > \lambda_2 > \lambda_3 \dots$ ) correspond to the SPOD of each block.

### 4 Results

Instantaneous high-speed visualizations of the sand–water swirling jet were depicted in Fig. 2. The images clearly showed the changing character of the jet by adding swirling motion of 0.50 and 0.65 in Fig. 2b, c in comparison to  $S_w = 0$  (see Fig. 2a). As can be seen, particle dispersion and jet spreading were more pronounced in swirling jets since the flow was evolved very quickly in the recirculation region due to the presence of tangential velocity.



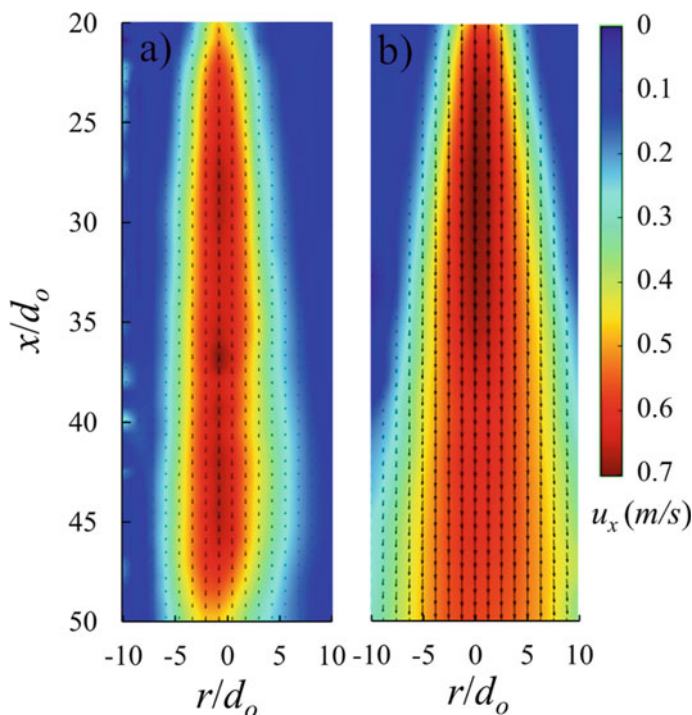


**Fig. 2** Instantaneous high-speed images of sand–water swirling jets: **a**  $S_w = 0$ ; **b**  $S_w = 0.50$ ; **c**  $S_w = 0.65$

Contour plots of the streamwise mean axial velocity in the symmetrical plane were depicted for jets with  $S_w = 0$  and  $0.65$  (i.e.,  $20 \leq x/d_o \leq 50$ ) in Fig. 3. As can be seen in Fig. 3a, the non-swirling jet propagated into the quiescent ambient water without much radial spreading. In contrast, by introducing swirling intensity, the more radial cone shape spreading and velocity can be seen in Fig. 3b. It was indicated that the swirling motion increased the radial velocity distribution in comparison to non-swirling jets. In both cases, the shear layers were symmetric about the nozzle centerline. The maximum mean velocity of the recirculation zone at  $20 \leq x/d_o \leq 35$  can be easily distinguished due to the existence of high swirling motion in Fig. 3b.

The decay of the axial mean velocity normalized with the initial velocity at the nozzle was shown in Fig. 4. The normalized axial velocity correlation of sand–water swirling jet measured by the optical probe was also added for comparison [23]. As can be seen in Fig. 4 increasing the swirling number led to a higher velocity decay rate. The maximum data divergence from the correlation was found to be  $\pm 15\%$ .

More insight into the temporal evolution of the vertical velocity components in transverse profiles is shown in Fig. 5. The normalized vertical velocity components in the radial direction of sand–water swirling jets by Sharif and Azimi [23] were also included in this figure. A more peak velocity profile can be seen at the jet with  $S_w$

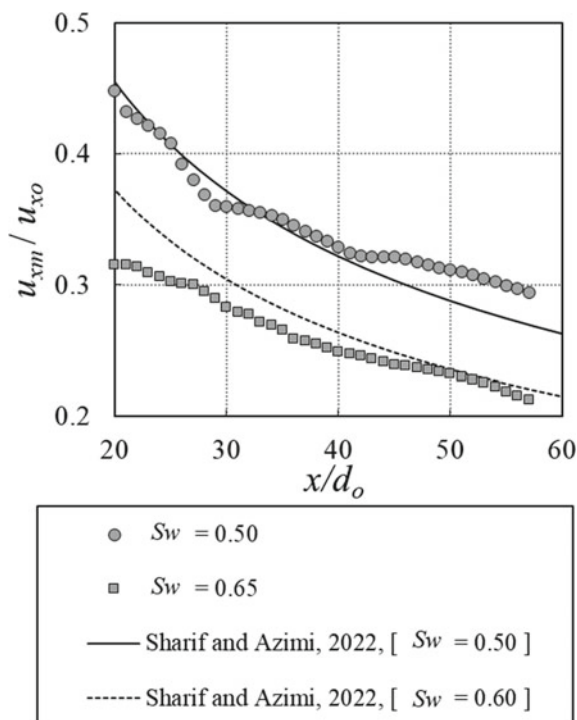


**Fig. 3** Comparison of the contour plots of mean vertical velocity component: **a**  $S_w = 0$ ; **b**  $S_w = 0.65$

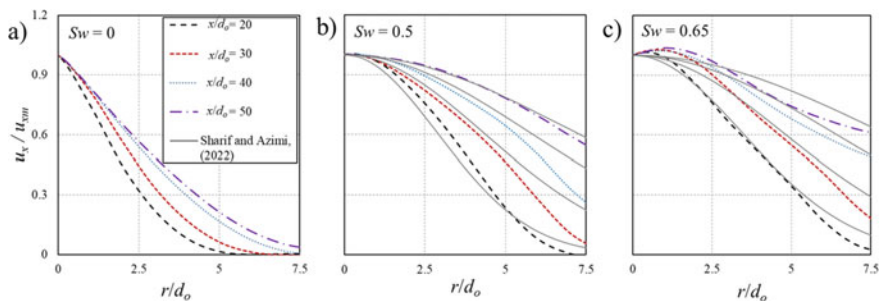
$= 0$  (see Fig. 5a) due to smaller spreading rate of the jet. The experimental results indicated that the velocity profiles widened by adding swirling intensities of 0.5 and 0.65. For exhibited swirling intensities, the velocity distribution became flattened by increasing  $x/d_o$  and can be modeled by Gaussian distribution. In addition, the PIV measurements had a good agreement with the proposed model of Sharif and Azimi [23] for jets with  $S_w = 0.5$  and  $r/d_o < 5$  (see Fig. 5b). However, increasing the swirling number led to a maximum %15 divergence compare to the included model at  $x/d_o > 30$  (see Fig. 5c).

The effect of swirling number on radial velocity component of sand–water swirling and the non-swirling jet was shown in Fig. 6. The results of PIV analysis indicated higher radial velocity in swirling jets with higher swirling intensities in comparison to non-swirling jet. A profile peak can easily be distinguished for both sand–water swirling jets of 0.50 and 0.65 at  $r/d_o = 2$  and can be attributed to the presence of PVC.

To perform a more quantitative study of PVC structure and progression in sand–water swirling jets, the vortex structure at the near field of the nozzle needs to be investigated. Therefore, the effect of swirling number on vortex structure of sand–water swirling jets was illustrated by the ensemble-averaged azimuthal vortex locator

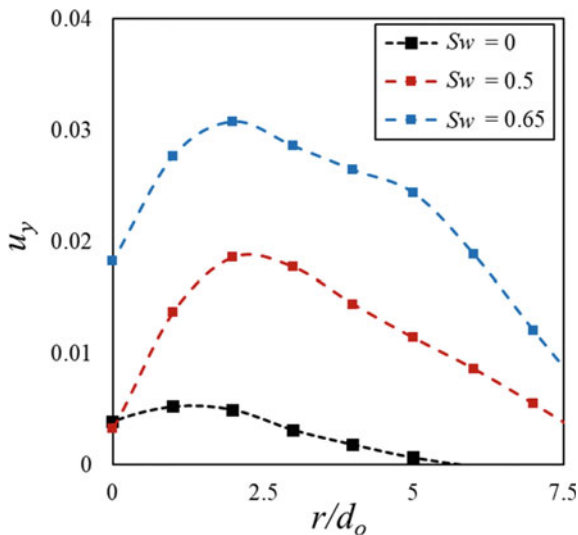


**Fig. 4** Effect of swirling number on normalized vertical velocity component in axial direction



**Fig. 5** Effects of swirling number on normalized vertical velocity component in radial direction for  $20 \leq x/d_o \leq 50$ : **a**  $Sw = 0$ ; **b**  $Sw = 0.50$ ; **c**  $Sw = 0.65$

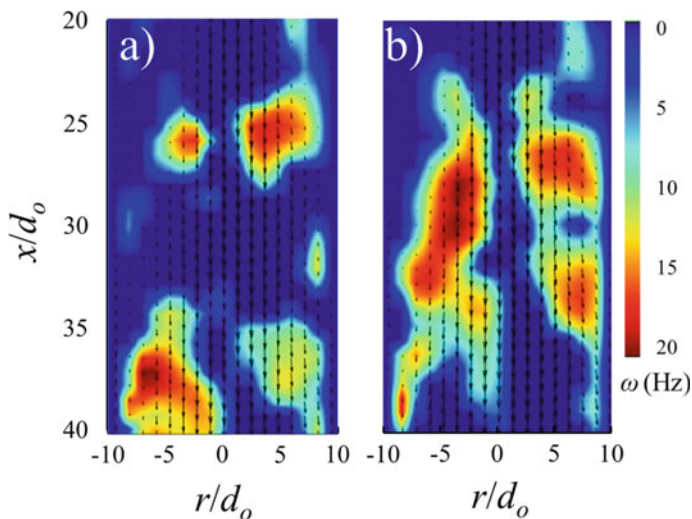
contour plots and shown in Fig. 7. As can be seen, a strong radial spreading clockwise and counterclockwise rotating vortices were indicated throughout the recirculation zone which can be linked to the internal and external shear layers due to particle migration toward the PVC region. It was found that the vortex centers were concentrated in the region of  $20 < x/d_o < 40$  and more intensity was observed in jets with



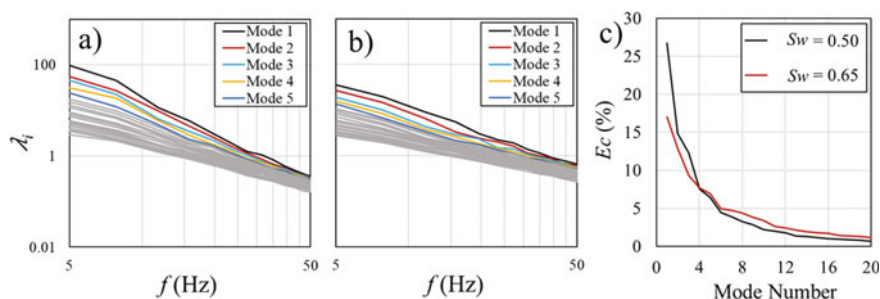
**Fig. 6** Effects of swirling number on horizontal velocity component in radial direction with  $x/d_o = 20$ : **a**  $S_w = 0$ ; **b**  $S_w = 0.50$ ; **c**  $S_w = 0.65$

the swirling number of 0.65 (see Fig. 7b). This is due to the swirling motion of water and the enhanced instability of the shear layer.

The SPOD analysis was performed for the series of 2500 high-speed snapshots to investigate the low-rank dynamic of particles, flow oscillation, and coherent structure.



**Fig. 7** Effects of swirling number on vortex structures: **a**  $S_w = 0.50$ ; **b**  $S_w = 0.65$



**Fig. 8** SPOD eigenvalue spectra of the first seven modes: **a**  $S_w = 0.5$ ; **b**  $S_w = 0.65$ ; **c** energy contribution

The block size was chosen as 256 data with a 50% overlap. The SPOD eigenvalue spectra or eigenvalues,  $\lambda$ , of the first ten modes and energy contribution of each mode were shown in Fig. 8. Large separation between the eigenvalues is a good indicator of dominant mode and energy contribution.

As can be seen in Fig. 8a, the large gap of the first mode and second mode indicated that most of the particle oscillation turbulent kinetic energy is stored in mode 1. However, increasing the swirling number to 0.65 decreased the distance of SPOD spectra of the first ten modes (see Fig. 8b). The separation between eigenspectra of the modes can be linked to dominant flow vortex shedding [16]. To quantify the effect of swirling number on the contribution of each mode at each frequency to the total energy, the accumulated energy,  $E_c$ , was shown in Fig. 8c. As can be seen, the first five modes had the majority of the kinetic energy as 67% and 52% the swirling number of 0.50 and 0.65, respectively, indicating that increasing the swirling intensity reduces particle's oscillation and vortex shedding. In general, low-frequency modes contained most of the flow turbulent kinetic energy, and the energy is transferred successively to the smaller scales and high-frequency modes [32].

## 5 Conclusion

A series of laboratory experiments were carried out to investigate the effects of swirling number on the flow feature of sand–water swirling jets in stagnant water. A combination of high-speed imaging and Particle Image Velocimetry (PIV) measurement techniques was implemented to capture flow features and velocity components of sand–water swirling jets in the axial and radial directions for two different swirling numbers of 0.50 and 0.65. A wider spray cone in velocity contours was observed due to formation of recirculation zone in sand–water swirling jets which was expanded moderately by increasing the swirling number. The axial alternation

of vertical velocity components in axial direction was found to be increased by increasing the swirling intensity. A near-field hump was observed at  $r/d_o = 2$  of the horizontal velocity profiles in the radial direction due to the presence of the PVC region.

The vertical velocity component in the radial direction was also measured. It was found that for all cases, the velocity distribution can be indicated by Gaussian distributions and became flattened as the distance from the nozzle,  $x/d_o$ , increases. The two different clockwise and counterclockwise rotating vortices of the shear layer were observed at the region of  $20 < x/d_o < 40$ . The different vortices were formed as a result of particle migration toward the PVC region. The SPOD analysis was performed for a series of high-speed snapshot images to study the low-rank dynamic of particles and flow oscillations. It was found that increasing the swirling number led to lower vortex shedding and particle fluctuations. In addition, the gap between the first five eigenspectra curves of SPOD modes indicated that a significant portion of the turbulence kinetic energy and flow vortex shedding is stored in the first five low-frequency modes of swirling sand–water jets.

## References

1. Azimi AH, Zhu DZ, Rajaratnam N (2012) Experimental study of sand jet front in water. *Int J Multiph Flow* 40:19–37
2. Azimi AH, Zhu DZ, Rajaratnam N (2012) Computational investigation of vertical slurry jets in water. *Int J Multiph Flow* 47:94–114
3. Billant P, Chomaz JM, Huerre P (1998) Experimental study of vortex breakdown in swirling jets. *J Fluid Mech* 376:183–219
4. Bühler J, Papantoniou D (2001) On the motion of suspension thermals and particle swarms. *J Hydraulic Res* 39(6):643–653
5. Eaton JK, Fessler JR (1994) Preferential concentration of particles by turbulence. *Int J Multiph Flow* 20:169–209
6. Gomes MSP, Vincent JH (2002) The effect of inertia on the dispersion of particles in the flow around a two-dimensional flat plate. *Chem Eng Sci* 57:1319–1329
7. Gupta AK, Lilley DG, Syred N (1984) Swirl flows. Abacus Press, Tunbridge Wells, p 488
8. Kazemi S, Adib M, Amani E (2018) Numerical study of advanced dispersion models in particle-laden swirling flows. *Int J Multiph Flow* 101:167–185
9. Liu Y, Zhou LX, Xu CX (2010) Numerical simulation of instantaneous flow structure of swirling and non-swirling coaxial-jet particle-laden turbulence flows. *Phys A* 389:5380–5389
10. Liu Y, Zhang L, Chen Z, Zhou L (2020) Numerical investigation on mixture particle dispersion characteristics in swirling particle-laden combustion chamber. *Int Commun Heat Mass Transf* 117:104720
11. Hall MG (1972) Vortex breakdown. *Annu Rev Fluid Mech* 4:195–218
12. MATLAB [Computer software]. MathWorks, Natick, MA
13. Meliga P, Gallaire F, Chomaz JM (2012) A weakly nonlinear mechanism for mode selection in swirling jets. *J Fluid Mech* 699:216–262
14. Moghadaripour M, Azimi AH, Elyasi S (2017) Experimental study of particle clouds in stagnant water. *ASCE J Eng Mech* 143:04017082
15. Moghadaripour M, Azimi AH, Elyasi S (2017) Experimental study of oblique particle clouds in water. *Int J Multiphase Flow* 91:101–119

16. Nidhan S, Chongsiripinyo K, Schmidt O, Sarkar S (2020) Spectral POD analysis of the turbulent wake of a disk at  $Re = 50,000$ . *Phys Rev Fluids* 5:124606
17. Oberleithner CO, Paschereit RS, Wygnanski I (2012) Formation of turbulent vortex breakdown: Intermittency, criticality, and global instability. *AIAA J* 50:1437–1452
18. O'Doherty T, Negro O (2001) Vortex breakdown: a review. In: *Progress in energy and combustion science*. [https://doi.org/10.1016/S0360-1285\(00\)00022-8](https://doi.org/10.1016/S0360-1285(00)00022-8)
19. Pakzad L, Azimi AH (2017) Investigations on the dynamics of particle clouds in stagnant water using response surface methodology. *Can J Civ Eng CSCE* 44(2):117–128
20. Sharif F, Azimi AH (2020) Particle cloud dynamics in stagnant water. *Int J Multiph Flow* 125:101–119
21. Sharif F, Azimi AH (2021) Effects of velocity ratio on dynamics of sand-water coaxial jets water. *Int J Multiph Flow* 140:103643
22. Sharif F, Azimi AH (2021) Experimental study of sand-water coaxial jets with low-velocity ratio. In: *Hydrotechnical specialty conference, CSCE*
23. Sharif F, Azimi AH (2022) Experimental study of sand-water swirling jets in stagnant water. *Exp Fluids* 63:26, 23p
24. Schetz JA (1980) *Injection and mixing in turbulent flow*. AIAA, New York, USA
25. Schmidt O, Towne A, Colonius T, Cavalieri A, Jordan P, Brès G (2017) Wavepackets and trapped acoustic modes in a turbulent jet: coherent structure education and global stability. *J Fluid Mech* 825:1153–1181
26. Sirovich L (1987) Turbulence and the dynamics of coherent structures I. Coherent structures. *Q Appl Math* 45:561–571
27. Thielicke W, Stamhuis EJ (2014) PIVlab—towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB. *J Open Res Softw* 2:30
28. Vanierschot M, Van Den Bulck E (2008) Influence of swirl on the initial merging zone of a turbulent annular jet. *Phys Fluids* 20:105104
29. Vanierschot M, Van Dyck K, Sas PV, Den Bulck E (2014) Symmetry breaking and vortex precession in low-swirling annular jets. *Phys Fluids* 26:105110
30. Welch PD (1967) The use of fast Fourier transform for the estimation of power spectra; a method based on time averaging over short, modified periodograms. *IEEE Trans Audio Electroacoust* 15:70–73
31. Wicker RB, Eaton JK (1994) Near field of a coaxial jet with and without axial excitation. *AIAA J* 32:542
32. Xiaobai L, Guang Ch, Liang XF, Liu D, Xiong XH (2021) Research on spectral estimation parameters for application of spectral proper orthogonal decomposition in train wake flows. *Phys Fluids* 125103
33. Zhang Y, Vanierschot M (2021) Determination of single and double helical structures in a swirling jet by spectral proper orthogonal decomposition. *Phys Fluids* 33:015115

# A Detailed 2D Hydraulic Model for the Lower Fraser River



Junying Qu

**Abstract** The Lower Fraser River (LFR) refers to the Fraser River reach from the Trans-Canada Highway 1 Bridge at Hope, British Columbia to the river mouth at the Salish Sea. Its floodplains and the coastal areas adjacent to the river mouth are inhabited by approximately 3.1 million people. The LFR also has a long history of flooding. A two-dimensional (2D) hydrodynamic model was developed for the LFR using the software MIKE 21 model with flexible mesh. The model extent includes the LFR and its key tributaries—the Harrison and Pitt Rivers. It was calibrated and validated using the 1972, 2012, and 2017–2020 flood events with the flow range from 5000 m<sup>3</sup>/s to 13,000 m<sup>3</sup>/s. Two roughness scenarios were developed to adapt the flow range: the roughness scenario for low river levels and the roughness scenario for high river levels. The model results for the 2018 flood event were in excellent agreement with the observed data since the digital elevation model, which was used for the 2D model, was built based on the river bathymetry data surveyed in less than one year before the 2018 freshet. The model results for the other flood events were also in good (2012 and 2017–2020) or reasonable (1972) agreement with the observed data. The model will be used to calculate the LFR design flood profiles. It can also be used for many other LFR hydraulic modeling applications in a regional scale with a broad flow range.

**Keywords** 2D hydraulic model • Lower Fraser River

## 1 Introduction

The Lower Fraser River (LFR) refers to the Fraser River reach from the Trans-Canada Hwy #1 Bridge at Hope, British Columbia (BC) to the river mouth at the Salish Sea. The Fraser River is the largest river in the Province of BC. It travels more than 1300 km from the Rocky Mountains to the sea and drains a catchment of

---

J. Qu (✉)

Flood Safety Section, Water Management Branch, Ministry of Forests, British Columbia, Canada  
e-mail: [junying.qu@gov.bc.ca](mailto:junying.qu@gov.bc.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_48](https://doi.org/10.1007/978-3-031-34593-7_48)

775



approximately 234,000 km<sup>2</sup>. From Quesnel, BC (approximately 620 km measuring on the river from the sea) to Hope, the river follows a steep course through the mountains and plateau where it picks up gravel and finer sediment. From Hope to Sumas Mountain (approximately 95 km measuring on the river from the sea or 10 km upstream of Mission, BC), the river's gravel transport capacity is greatly reduced due to the declining river slope, so that islands and gravel bars are formed in this approximately 72 km long river reach. From Sumas Mountain to the sea, the river slope is so mild that finer gravel and sand are deposited on the riverbed and in the delta area in the approximately 95 km long reach. The Fraser River reach from Hope to Mission is traditionally known as the "Gravel Reach", and the reach from Mission to the river mouth known as the "Sand Reach". That is to say, the LFR consists of these two reaches. The LFR plays a significant role on BC's economic prosperity. Its floodplains and the coastal areas adjacent to the river mouth are inhabited by approximately 3.1 million people (including the districts of Metro Vancouver and the Fraser River Valley), which is 60% of BC's total population (2020 Census).

The LFR has a long history of flooding. The largest flood event on record occurred in May and June 1894. After that flood, many dikes were constructed or upgraded continuously along the LFR in more than a century. The second largest flood event occurred in 1948, which caused multiple dike breaches and massive flooding in Chilliwack and the other areas. Several notable flood events (1972, 1997, 2012, etc.) occurred with less magnitude after 1948, but still caused economic losses despite a growing system of upgraded dikes. Because of continuous development and population growth in the LFR floodplain and the delta area, vulnerabilities of the Fraser River flood to the Lower Mainland of BC are more severe. According to the key findings in Phase 1 (2014–2016) of the Lower Mainland Flood Management Strategy (LMFMS) completed by Fraser Basin Council (FBC) (<https://floodwise.ca/flood-strategy/phase-1-highlights/>): *"If a major Fraser River or coastal flood were to occur between now and 2100, it would trigger losses estimated at \$20–30 billion, which would be the largest disaster in Canadian history"*.

Numerical models have been widely used as important tools by the Province of BC and many other organizations for flood risk assessment, flood mitigation studies, and flood level forecasting for the LFR. In 2000/2001, a one-dimensional (1D) hydraulic model was developed by the UMA consultants for the gravel reach from Laidlaw to Mission [12, 13], and a design flood profile for this reach was created. From 2006 to 2008, NHC was retained by FBC and developed a 1D hydraulic model for the sand reach from Mission to the sea in 2006 and later merged it with the UMA's model in 2008, so that an integrated 1D model was created using the software MIKE 11 [7]. As a result, the design flood profile for the sand reach was created in 2006 [5]. In 2014, the Province extended the integrated model from Laidlaw to Hope, updated the model geometry based on the LiDAR data and bathymetric data surveyed in 2008, 2011, and 2012, and updated the design flood profile for the gravel reach from Mission to Hope [1]. The 2014 model has been used for the LFR freshet flood level forecasting since then. In 2019, NHC was contracted by FBC for Phase 2 of the LMFMS initiative and developed a hydraulic model for the LFR floodplain using the software HEC-RAS 2D. The model domain included the LFR main channel,

the major tributaries, and the floodplain from Hope to the sea [10]. It was used for flood risk identification and flood hazard delineation, dike breach scenario studies and flood mitigation modeling for the LFR floodplain. The DEM developed for the HEC-RAS 2D model was shared with the Province and used in this project.

The objective of this study was to develop a detailed 2D hydrodynamic model for the LFR main stem so that it can be used to update the LFR design flood profile at the current design conditions and create the design flood profiles at the design conditions of 2050 and 2100 with consideration of climate change and sea level rise. The profiles will be used for dike design and other flood protection infrastructure design by the Province, districts and local municipalities, First Nations and many other organizations to ensure the communities alongside the LFR are more resilient to the future.

## 2 Model Development

### 2.1 Model Domain

The extent of the 2D model was defined based on the computational domain of the existing 1D model owned by the Province [1]. It covers: (1) the LFR main stem from downstream of the Trans-Canada Hwy #1 Bridge at Hope to the river mouth at the Salish Sea; (2) the key tributaries including the Harrison River and Lake, the Pitt River and Lake, and the Vedder River up to the railway bridge near Yarrow Village, BC; (3) the LFR and Harrison River floodplains that are not protected by the river dikes.

The model extent was defined under the assumption that the flows from the Fraser, Harrison, and Pitt Rivers would not overtop the river and sea dikes. This assumption was consistent with the previous design flood profile calculations using the 1D model from 2006 to 2014 ([5, 7], the Province 2014). Secondly, this model will be used to calculate the LFR design flood profiles to ensure all the dikes can protect the communities from extreme flood events; so, the modeled flows should not overtop the dikes under design conditions. Therefore, the computational domain was bordered by the river dikes or high lands on both sides of the rivers. For example, the Lulu Island, Sea Island, Westham Island, Barnston Island, and Nicoman Island were excluded from the model since they were all protected by the river or sea dikes; the urban areas and agricultural lands in Chilliwack and Abbotsford protected by the diking system were also excluded from the model. Most of the LFR floodplains were not covered by the model, which made it different from FBC's HEC-RAS 2D model.

## 2.2 *Digital Elevation Model*

A digital elevation model (DEM) for 2D hydraulic models must integrate seamlessly the elevations of river bottom surface and the topography of riverbanks and floodplains to define the physical geometry for the model domain. Its accuracy has a significant impact on the performance of the 2D model. The DEM for the LFR main conduit was developed by NHC in [9] and then extended to the LFR floodplain in 2019 [10]. To support the DEM development, the LFR gravel reach and Harrison River channels were surveyed at a section spacing of 200 m in the summer 2017. The river bathymetry in the sand reach was mostly surveyed by Public Works and Government Services Canada (PWGSC) from 2014 to 2017, which was a combination of high-density bathymetric points collected via multi-beam sonar and survey sections at a spacing of 50 m. There were only a few locations not surveyed by the PWGSC in the sand reach, and they were surveyed at a spacing of 200 m in the summer 2017. The LiDAR data collected by GeoBC in the summer 2016 was used to build the DEM for the riverbanks and floodplains. More information about the other data sources used for the DEM development can be found in NHC's 2019 report. Then, the Province updated the DEM using the new bathymetry survey data collected from 2018 to 2021, which included: (1) the river bathymetry data surveyed at a spacing of 50 m during the 2018 and 2021 freshets for the reaches where the river geometry features were of high interest but missed in the 2017 survey at a spacing of 200 m, (2) the LiDAR data collected for dry riverbeds in the gravel reach in March 2019, and (3) the river bathymetry data newly collected by PWGSC from 2017 to 2020. The final DEM was generated at the resolution of 1 m by 1 m in the horizontal directions.

The DEM has limitations for its accuracy. First, the gravel reach was mainly surveyed in sections at a spacing of 200 m, and the river bottom elevations between any two survey sections were estimated. Secondly, there are floating logs and many types of structures (boat houses, wharfs, etc.) on the LFR, the Harrison River, and the Pitt River; the river bathymetry adjacent or under the floating logs and structures was not able to be surveyed and had to be estimated in the DEM. Lastly, the river bathymetry of local drains and water ponds in the floodplains were not available and had to be estimated in the DEM. Therefore, the DEM is only suitable for the LFR hydraulic modeling at a regional scale, and cautions should be exercised when the DEM is used to model specific local river reaches where high accurate river geometry is required.

## 2.3 Mesh Design

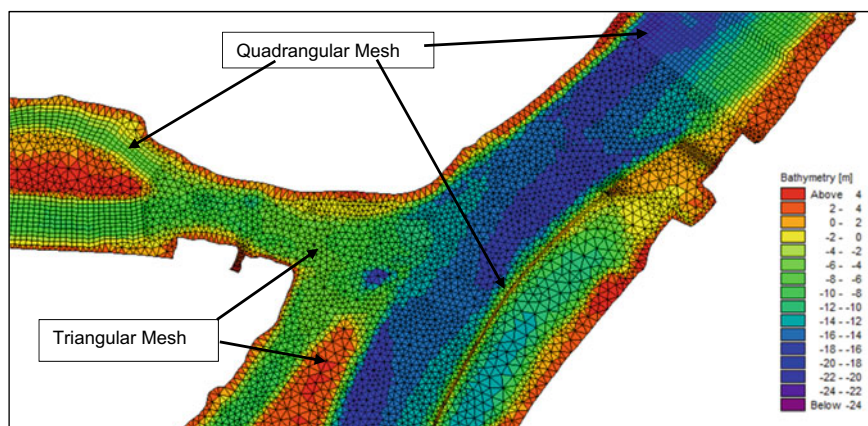
### 2.3.1 Mesh Design Methodology

Mesh design is a state of art for 2D hydraulic models. An ideal mesh design should capture the model geometry with high accuracy, avoid potential model instabilities, and ensure model runs efficiently. In this study, the mesh was generated using MIKE Zero Mesh Generator, which is part of the MIKE 21 software package. A combination of triangular mesh and quadrangular mesh was used for mesh design. Triangular mesh is flexible and can be easily used to model any kind of complex geometries. However, sometimes triangular mesh has instability issues, especially when the mesh angles are less than  $30^\circ$ . Quadrangular mesh is stable and with good water volume conservation for numerical models, especially for solving the 2D shallow equations using the finite volume method. Using quadrangular mesh can reduce the number of mesh elements compared to using triangular mesh. However, the shape of quadrangular mesh limits its flexibility of modeling irregular geometries. The combination of these two mesh styles can make use of their pros and avoids their cons. Therefore, quadrangular mesh was used in the areas with simple channel shape, such as the channelized river reaches, railways, highways, wave deflectors while triangular mesh was used in the areas where quadrangular mesh is difficult to be applied, such as in irregular river channels and floodplains.

To capture the geometry features such as the riverbanks, islands and sand bars, floodplains, lakes, highways, railways, and wave deflectors the linework of these features were imported to the Mesh Editor as break lines. The various sizes of mesh elements were carefully designed according to the river features and model interests. Generally, the mesh sizing of 20 m in the length scale was used to model river channels. There are 20 or more mesh elements in the lateral direction for the Fraser River main channel and 6 or more mesh elements in the lateral direction for side channels or tributary channels. Figure 1 shows the application of mesh design methods to various geometry features.

During the mesh development, it was found that the MIKE Zero Mesh Generator became inefficient when the mesh file size was greater than 4 megabytes. So, a mesh file was created for each of the two LFR reaches. Then, the two mesh files were merged into one and the mesh nodal elevations were interpreted from the DEM using in-house programs.

The mesh elements were designed generally with a high resolution to capture the geometry of river channels and banks with a good accuracy. However, the design also had limitations. First, the mesh elements did not capture the geometries of the structures on the river, such as floating houses, wharfs, and bridge because their geometries were either not available or too small. Most bridge piers were included in the model as hydraulic structures to count flow energy losses caused by the piers [4]. The impact of those structures on the modeled flood levels were considered minor because most of them are located downstream of the LFR trifurcation, where the water level is highly affected by the tidal levels. Secondly, the mesh elements could



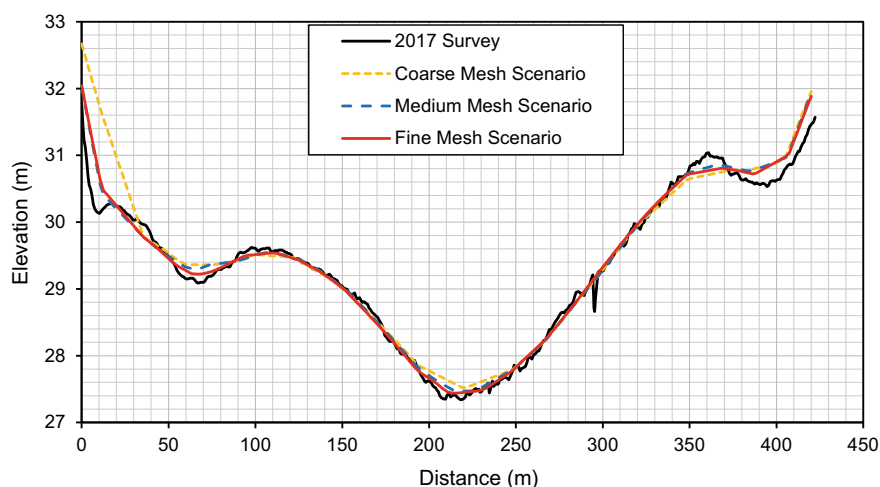
**Fig. 1** Application of triangular and quadrangular mesh elements at the Fraser River trifurcation

not model accurately the river dunes vastly distributed in the sand reach. The dunes were captured by the multi-beam sonar survey conducted by PWGSC from 2014 to 2020 and included in the DEM. The dune shapes, alignments, wavelengths, and heights were quite different from location to location on the riverbed. As such, it was not possible to model them with a high accuracy even using a fine mesh resolution.

### 2.3.2 Mesh Sensitivity

To evaluate the sensitivity of geometry representation and model results to mesh resolution, three mesh files were created with coarse, medium, and fine mesh sizes. In average, the triangular mesh sizing was reduced approximately 50% from the coarse mesh design scenario to the medium mesh design scenario and another 30% to the fine mesh design scenario; the quadrangular mesh sizing was reduced approximately 10% from the coarse mesh design scenario to the medium mesh design scenario and another 14% to the fine mesh design scenario. In the end, the total number of mesh elements was increased 22% from the coarse mesh scenario to the medium mesh scenario and another 20% to the fine mesh scenario. The fine mesh scenario has 602,147 mesh nodes and 1,097,654 mesh elements, which is the most detailed 2D hydraulic model developed for the LFR so far.

The comparison of profiles at a selected LFR cross section in Fig. 2 showed that the three mesh scenarios were overall interpreted the DEM with a high accuracy in the main channel. The model accuracy was low at both ends of the river section for all three mesh scenarios because: (1) the river bathymetry had abrupt changes in elevations; (2) the relatively coarse mesh elements were used near the shorelines for those scenarios. A large amount of mesh elements would be required to improve the model accuracy at the riverbanks using much detailed mesh elements, which could be not valuable at a point view of computational efficiency; however, this can



**Fig. 2** Comparison of profiles for three mesh resolution scenarios at a Fraser River cross section (approximately 700 m downstream of Trans-Canada Hwy #1 Bridge at Hope, BC)

be studied further in the future model development. Overall, the fine mesh scenario showed the best fit to the survey data and the coarse mesh scenario was the worst. The sensitivity of model results to mesh resolutions was evaluated using the 2018 flood event in the model calibration process, which will be discussed in Sect. 3.5.3.

## 2.4 Roughness Coefficients

For 2D hydrodynamic models, the bed resistance is the most important friction to be considered. Generally, the roughness coefficient (Manning's  $n$  value) is the only parameter that can be adjusted and determined through the process of model calibrations and validations once the river geometry has been determined by the DEM. The initial roughness coefficients for the model were first assigned based on the land use maps available on DataBC, the existing 1D hydraulic model, knowledge on the study area and general guidance related to river roughness settings. The final values were determined through the model calibration process using the 2018 flood event and validated using several other flood events, which will be discussed in next section.

### 3 Model Calibration and Validations

#### 3.1 Selected Flood Events

Once the 2D hydraulic model is developed, its performance (or accuracy) must be evaluated through the calibration and validation processes, so that model parameters can be adjusted and validated to ensure the model performance satisfies pre-agreed criteria. Due to the high interest of the LFR, many gauge stations along the river were established and owned by Water Survey Canada (WSC), the Province, or the local municipalities. Their locations can be found from NHC's 2019 report. The flow and water level records for the Fraser River at Hope can be dated back to more than 100 years ago (starting from 1912). Since 2007, the Province has been running the flood level forecasting model during each freshet and collected observed water level and flow records from various data sources, which is a big treasure for the model calibration and validations.

The 2018 flood event was selected for the model calibration because: (1) it is the first flood event after the river bathymetry of the gravel reach was surveyed and the largest flood event after 2012; (2) there was a lot of observed flow and water level data collected during the freshet. The freshet events in 2012, 2017, and 2020 were selected for model validations. Additionally, the freshet events in 2019 and 1972 were selected for further model validations to test the model performance in a broad flow range. Especially, the 1972 flood event was the highest without recorded dike breaches in the LFR floodplain. Although the water level and flow records in 1972 were very limited; the LFR river geometry was supposed to change a lot in 50 years, and it is still valuable to verify the model's capability in modeling such a large flood magnitude. Table 1 lists the ranks of these 6 flood events based on the flow records available since 1894.

**Table 1** Historical lower Fraser River flood event ranks (up to 2020)

Rank no	Year	Maximum daily flow (m <sup>3</sup> /s)	Return period and comment
1	1894	17,000	
2	1948	15,200	
3	1972	12,900	Near 1 in 50-year flood event
5	2012	11,700	1 in 20-year flood event
11	2018	10,875	Slightly below 1 in 10-year flood event
16	2020	10,646	Slightly below 1 in 10-year with a long last freshet
34	2017	9702	Below 1 in 5-year flood event
93	2019	7060	Below 1 in 2-year flood event

## **3.2 *Boundary Conditions***

### **3.2.1 Inflows**

The recorded flows from the following WSC gauge stations were used for model calibration and validations:

- Fraser River at Hope (WSC Gauge 08MF005) from 1912 to present
- Lillooet River at Tenas Narrows (WSC Gauge 08MG027) from 2012 to present (approximately 5-h flow traveling time to Harrison Lake)
- Chilliwack River at Vedder Crossing (WSC Gauge 08MH001) from 1911 to present.

Besides the Lillooet and Chilliwack Rivers, inflows from the key tributaries in the LFR watershed were considered including Silverhope Creek, Wahleach Creek, Ruby Creek, Norrish Creek, Chehalis River (joining to Harrison River), Sumas River (joining to Vedder River), Stave River, Pitt River, Aloutte River (joining to Pitt River), and Coquitlam River. The Stave, Aloutte, and Coquitlam Rivers are regulated by the dams. There are no gauged flows available for the other tributaries. Since the flows from these rivers are generally low during the freshets and have no significant impact to the Fraser River levels during the freshet, their mean monthly flows provided in the NHC's 2009 freshet forecasting report [8] were used for model calibration and validations. Additionally, the local flows from the Harrison Lake watershed from the Lillooet River at Tenas Narrows to the Harrison Lake outlet were also considered as they always have a great contribution to the Harrison River flows. The RFC has provided estimated Harrison Lake local flows in each freshet since 2017. For the 1972 and 2012 flood events, the Harrison Lake local flows were estimated based on the flow records available at the WSC Gauge for Harrison Lake at Hot Springs (08MG013) from 1951 to 2018, which was located right downstream of the Harrison Lake outlet.

### **3.2.2 Sea Levels**

The downstream boundary of the 2D model was extended approximately 10 km into the Salish Sea. The observed tide levels at the stations Sandheads Light and Point Atkinson were downloaded from the website of Canadian Hydrographic Service (CHS) for model calibration and validations. The former station is closer to the model boundary than the latter one, but the latter has more tidal level records than the former.



### **3.3 Initial Conditions**

To secure model stability, initial water levels and flow conditions were carefully assigned before performing a run at the desired boundary conditions. The preliminary initial water levels including the LFR, the Harrison Lake and River, and the Vedder River were estimated based on the model results of the previous 1D model runs or the preliminary 2D model tests. The initial inflows for the Fraser River at Hope always started from a low magnitude (such as  $500 \text{ m}^3/\text{s}$ ) and slowly increased to normal to obtain a stable flow condition at the upstream boundary. It was found adding a large flow (such as  $6000 \text{ m}^3/\text{s}$ ) at the beginning of a simulation could cause instability issues. No instability issues were found for the tributary inflows. A warm-up run was conducted with the initial water levels and inflows to set appropriate start water level and flow conditions in the river channels for a target run with desired boundary conditions.

### **3.4 Model Performance Evaluation Criteria**

Based on the accuracy analysis for recorded water levels, recorded/measured flows, river geometry modeling and uncertainties of tributary inflows, the evaluation criteria were defined to measure the model performance. The following four levels were used to evaluate errors of modeled water levels:

- Excellent: errors within  $\pm 0.15 \text{ m}$
- Good: errors within  $\pm 0.2 \text{ m}$
- Reasonable: errors within  $\pm 0.3 \text{ m}$
- Not good: errors exceeding the range of  $\pm 0.3 \text{ m}$ .

The following three levels were used to evaluate errors of modeled flows:

- Excellent: errors within 5%
- Good: errors within 10%
- Not good: errors exceeding the range of 10%.

Please note that the criteria above were established based on the common sense in hydraulic modeling in river engineering since there are no standard criteria available to measure the performance of 2D hydraulic models.

### 3.5 Model Calibration

#### 3.5.1 Observed Data

The 2018 flood event was selected for the model calibration. The Fraser River at Hope (WSC Gauge Station 08MF005) reached an instantaneous peak flow of  $11,000 \text{ m}^3/\text{s}$  on May 20, 2018; the Fraser River at Mission (WSC Gauge Station 08MH024) reached the instantaneous peak water level of 6.03 m on the same day. There was a lot of water level and flow data recorded or surveyed during the freshet including flows/water levels at continuous gauge stations, water level readings from staff gauges, water surface profile surveys, and flow measurements [10].

#### 3.5.2 Roughness Coefficient Scenario Tests

The model roughness coefficients were optimized based on trial and error through the process of model calibration. In the end, two roughness scenarios including the Roughness Scenario for Low River Levels (RSLRL) and the Roughness Scenario for Higher River Levels (RSHRL) were created to model a broad flow range with a high accuracy. Table 2 listed the Manning's  $n$  values adopted for different river reaches. The Manning's  $n$  values of 0.026/0.027 were assigned to the LFR reach from Surrey (Barnston Island) to Abbotsford (Matsqui Island) in RSLRL, and 0.023/0.024 in the RSHRL for the same reach. The RSLRL can be applied at the river levels at Mission below 5.5 m, and the RSHRL for river levels at Mission above 5.5 m. However, the model calibration and validation results showed that the reference level of 5.5 m is an approximate value, and the selection of the roughness scenarios must be based on model result analysis and comparisons to the observed water levels.

As shown in Table 2, in the existing MIKE 11 model, the changing Manning's  $n$  values from 0.028 to 0.027 were also assigned with river flow changes for the same LFR reach (from Barnston Island to Matsqui Island). Generally, higher Manning's  $n$  values for 1D models are expected for the reasons: (1) to reduce flow conveyance of the river channels for a compensation of overestimation of river geometry, (2) to consider energy losses caused by 2D or 3D flow features such as contraction/expansion, or secondary flows. Generally, the Manning's  $n$  values used in 2D models are closer to the actual  $n$  values and lower than the values used in 1D models. This was discussed in the reference document FHWA in 2019.

Since there were no specific studies to investigate the reasons for changing roughness values in the LFR reach from Barnston Island to Matsqui Island, the author can only think the following two reasons:

- It is common for river channels with a variation of roughness values depending on flow depth. For example, this was discussed in the reference document for 2D Hydraulic Modeling for Highways in the River Environment prepared by FHWA in [3], which recommended a minimum of two Manning's  $n$  values be used for 2D hydraulic models when applying a depth varied Manning's  $n$ .

**Table 2** Reclassification of water body and related roughness coefficients

Land cover attribute	Manning's n in MIKE 21—low river levels	Manning's n in MIKE 21—higher river levels	Manning's n in existing MIKE 11 model
Hope Reach	0.03	Same as low river levels	0.033
Hunter Reach	0.025		0.033
Laidlaw Reach	0.026		0.03
Seabirds Island Reach	0.023		0.03
Agassiz Reach	0.022		0.03
Kent Channel Right	0.022		0.031
Kent Reach	0.021		0.031
Lower Kent Reach	0.02		0.031
MintoUS Reach	0.023		0.032
Harrison River Confluence	0.021		0.032
Minto Reach	0.02		0.031
Chilliwack Reach	0.02		0.028
Cannor Reach	0.02		0.03
Mission Reach	0.023		0.029/0.03
Harrison Lake	0.02		0.02
Harrison River Up	0.025		0.035
Harrison River Morris Up	0.025		0.035
Harrison River Morris Down	0.024		0.035
Harrison River HWY7	0.023		0.032
Harrison River Down	0.023		0.032
Harrison Bay	0.02		0.02
Vedder River	0.028		0.032
Vedder Channel	0.024		0.032
Vedder Mouth	0.023		0.032
Whonnock Reach	0.026	0.023	0.027/0.028
Maple Ridge Reach	0.027	0.024	0.027/0.028
Barnston Reach	0.026	0.023	0.027/0.028
Douglas Reach	0.022	Same as low river levels	0.028
Port Mann Reach	0.022		0.028
New Westminster Reach	0.022		0.031
South Arm	0.022		0.033
Annacis Reach	0.027		0.033

(continued)

**Table 2** (continued)

Land cover attribute	Manning's $n$ in MIKE 21—low river levels	Manning's $n$ in MIKE 21—higher river levels	Manning's $n$ in existing MIKE 11 model
Ladner Reach	0.022		0.031
Canoe Pass	0.022		0.033
North Arm	0.027		0.032/0.035
Middle Arm	0.027		0.032
Pitt Lake	0.02		0.02
Pitt River	0.021		0.03
Ocean	0.015		0.015

- The river dunes play an important role on the roughness features. As mentioned in Sect. 2, many dunes are distributed in the sand reach (of course including the reach from Barnston Island to Matsqui Island). The characteristics of flow over river dunes was well described by Best in 2. The flow over dune generates thick boundary layers and results in much more energy losses than the bed material itself. The thicker the boundary layer, the more energy losses are generated. The factors that influence the boundary layer thickness include dune height, wavelength, flow velocity, and water depth. Generally, the deeper the water, the energy losses are less affected by river dunes, which explains why smaller Manning's  $n$  values work better at higher river levels.

The river reaches listed in Table 2 could be divided into more sub-reaches with more refined Manning's  $n$  values. However, it was found more optimizations would not help improve the model performance significantly due to the limitations of the DEM accuracy and the observed data.

### 3.5.3 Mesh Sensitivity Tests

As described in Sect. 2.3.2, three mesh scenarios (coarse, medium, and fine) were created to evaluate the sensitivity of model results to mesh resolution. The same boundary conditions and roughness settings were applied to the model tests. Table 3 compared the modeled peak levels to the observed peak levels at the continuous gauge stations for the three mesh scenarios. The model results with the fine mesh showed the best performance since all the peak level differences were within the range of  $\pm 0.15$  m; the differences exceeded the range of  $\pm 0.15$  m at four gauges in the run with the medium mesh; the peak level differences exceeded the range of  $\pm 0.15$  m at many gauges in the run with the coarse mesh. Figure 3 compared the modeled water level hydrographs to the observed hydrographs at the WSC gauge station for the Fraser River at Whonnock (08MH044), which showed that the model results with the fine mesh were in the best agreement with the observed data, and the

coarse mesh was the worst. That is to say, the higher mesh resolution had the higher model accuracy, which could be explained that the fine mesh scenario simulated the river bottom bathymetry at the best accuracy. Therefore, the fine mesh scenario was finally chosen for model calibration and validations.

Please note that the model results for the runs with the coarse and medium mesh scenarios could be improved if the roughness coefficients were carefully assigned to them based on trail and errors. However, according to the model tests for the medium mesh scenario, it was found that its model results were always less accurate than the fine mesh scenario at the gauges: Fraser River at Bayes and Pitt River at Fenton PS listed in Table 3. In other word, the fine mesh scenario always worked better than the medium mesh scenario in that reach.

The trend of the model result comparisons also indicated that the mesh sizing could be refined further. However, this was not continued due to the accuracy limitation of the DEM. As mentioned in Sect. 2, the DEM was developed based on the river bathymetry survey data at a section spacing of 200 m for the gravel reach and the mixture of high resolution survey data and the survey data at a section spacing of 50 m for the sand reach. In the fine mesh scenario, the average mesh size was approximately 20 m in length, which already had higher resolution than the bathymetry survey data. It was anticipated that more detailed mesh would not significantly improve the model accuracy but would cost more computer resources and use more computational time. The higher mesh resolution for the 2D model can be further studied once the DEM accuracy is improved in future.

### 3.5.4 Final Model Result Analysis

#### 1. Water Level Comparison

With the fine mesh scenario and the application of the two roughness scenarios, the model results for the 2018 flood event had excellent agreement with the observed water levels. Table 3 shows the comparison of the modeled peak levels with the observed peak water levels at 37 continuous gauge stations. The peak water level differences are all within the error range of  $\pm 0.15$  m.

Figures 4, 5, 6 and 7 show the comparisons of modeled water level hydrographs with observed data at the selected WSC gauge stations, which were used to evaluate the model performance on modeling water level changes in response to flow changes in the temporal order. The water level hydrographs calculated using the existing MIKE 11 model with the same boundary conditions were also included in these figures. Figure 4 shows the good agreement between the simulation data and the observed data at the gauge station Fraser River at Steveston, which indicates that the 2D model correctly simulated the tidal level changes given at the downstream boundary conditions. Figure 5 shows that the modeled levels are in excellent agreement with the observed data at the gauge station for Fraser River at Mission at the flood levels below 5.0 m on the rising limb; this confirms the performance of the roughness scenario RSLRL at low river level conditions. Once the flood level rose

**Table 3** Peak water levels for mesh sensitivity comparison at continuous gauge stations

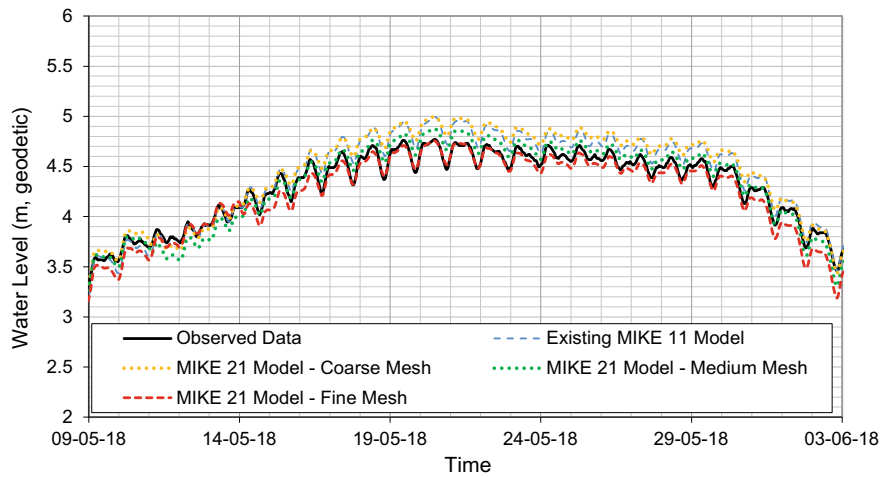
Continuous gauge station	Water Level (m)						
	Observed	Coarse mesh		Medium mesh		Fine mesh	
		Modeled	Difference	Modeled	Difference	Modeled	Difference
Fraser R. at Hope Bridge—08MF005	37.33	37.40	0.07	37.29	-0.04	37.22	-0.11
	30.83	31.00	0.17	30.88	0.05	30.87	0.04
North of Hunter Creek—Cont	26.85	26.87	0.02	26.85	0.00	26.82	-0.03
Fraser R. at Laidlaw—08MF072	21.80	21.87	0.07	21.84	0.04	21.81	0.01
North of Herring Island—Cont	16.97	16.85	-0.12	16.87	-0.10	16.82	-0.15
Fraser R. near Agassiz—08MF035	13.11	13.10	-0.01	13.03	-0.08	12.98	-0.13
Carey Point—Cont	12.35	12.55	0.20	12.46	0.11	12.39	0.04
Lower Kent	12.32	12.34	0.02	12.26	-0.06	12.21	-0.11
Harrison L. near Harrison Hotsprings—08MG012	11.98	12.09	0.11	12.01	0.03	11.96	-0.02
Harrison R. below Morris Creek—08MG022	11.80	11.93	0.13	11.85	0.05	11.79	-0.01
Harrison R. at Harrison Mills—08MG014	10.96	11.08	0.12	10.99	0.03	10.94	-0.02
Fraser R. at Bell Slough	10.88	10.85	-0.03	10.81	-0.07	10.73	-0.15
Fraser R. near Harrison Mills—08MF073	9.07	9.22	0.15	9.14	0.07	9.07	0.00
Hope Slough at Young Road	8.50	8.67	0.17	8.58	0.08	8.48	-0.02
Chilliwack Creek PS (Wolfe Road)	7.64	7.92	0.28	7.81	0.17	7.66	0.02
Fraser R. at Cannon—08MF038	7.36	7.53	0.17	7.40	0.04	7.23	-0.13
McGillivray Slough PS	6.03	6.30	0.27	6.19	0.16	6.03	0.00
Fraser R. at Mission—08MH024	6.00	6.27	0.27	6.15	0.15	5.98	-0.02
Matsqui Slough Discharge	4.77	5.00	0.23	4.87	0.10	4.75	-0.02
Fraser R. at Whonnock—08MH044	4.22	4.43	0.21	4.31	0.09	4.19	-0.03
Salmon River Confluence	3.92	4.25	0.33	4.14	0.22	4.02	0.10
Albion PS							

(continued)

Table 3 (continued)

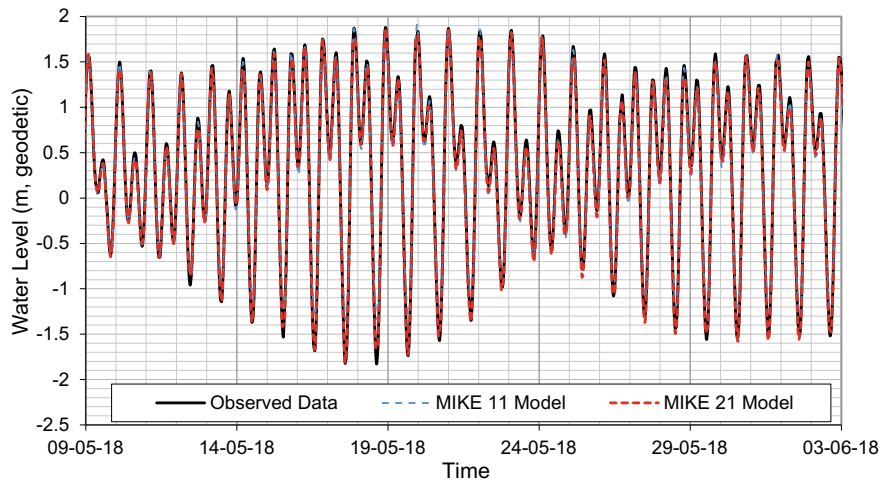
Continuous gauge station	Water Level (m)						
	Observed	Coarse mesh		Difference	Medium mesh		Fine mesh
		Modeled	Difference		Modeled	Difference	
192nd Street	3.33	3.52	0.19	3.45	0.12	3.34	0.01
Baynes PS	3.04	3.27	0.23	3.22	0.18	3.11	0.07
Pitt R. near Port Coquitlam—08MH035	2.62	2.82	0.20	2.75	0.13	2.66	0.04
Pitt River—Fenton PS	2.61	2.82	0.21	2.75	0.14	2.66	0.05
Pitt River at Argue St	2.62	2.73	0.11	2.68	0.06	2.62	0.00
Fraser R. at Port Mann PS—08MH126	2.57	2.63	0.06	2.57	0.00	2.53	− 0.04
New Westminster	2.39	2.45	0.06	2.40	0.01	2.37	− 0.02
Manson	2.41	2.45	0.04	2.40	− 0.01	2.36	− 0.05
9600 River Road	2.27	2.24	− 0.03	2.21	− 0.06	2.17	− 0.10
Nelson Road	1.96	2.10	0.14	2.08	0.12	2.06	0.10
No. 6 Road	1.95	2.01	0.06	1.99	0.04	1.98	0.03
62B & River Road	2.09	2.02	− 0.07	2.00	− 0.09	1.99	− 0.10
Fraser R. at Deas Island Tunnel—08MH053	1.83	1.98	0.15	1.97	0.14	1.96	0.13
Elliot & River Road	1.96	1.95	− 0.01	1.95	− 0.01	1.94	− 0.02
Fraser R. at Steveston—08MH028	1.88	1.87	− 0.01	1.87	− 0.01	1.86	− 0.02
Canoe Channel—3395 River Road	1.84	1.90	0.06	1.90	0.06	1.90	0.06
North Arm—Queensborough	2.09	2.20	0.11	2.17	0.08	2.11	0.02
North Arm—Fraser R. at Byrne Creek	1.98	2.10	0.12	2.09	0.11	2.04	0.06

Note The cells shaded in green indicate the water level difference within ± 0.15 m



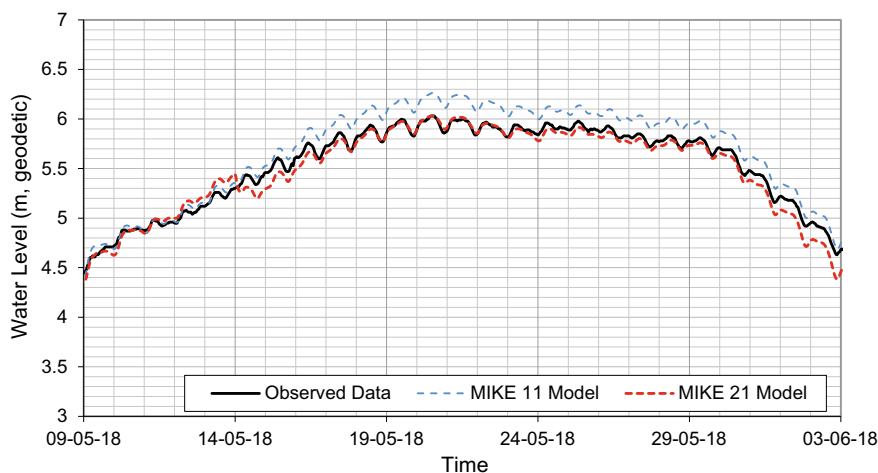
**Fig. 3** Comparison of water level hydrographs for Fraser River at Whonnock for mesh sensitivity analysis

above 5 m at Mission, the modeled water levels started showing over-estimated if still using RSLRL; then, the roughness scenario RSLRL was applied after May 14, 2018 and the modeled flood levels were in excellent agreement with the observed data at high flood levels. Figures 6 and 7 confirm the good model performance for the Fraser River at Hope and the Harrison Lake near Harrison Hot Springs. Figures 4, 5, 6 and 7 also show that the 2D model have provided the modeled water levels with a higher accuracy than the existing MIKE 11 model.

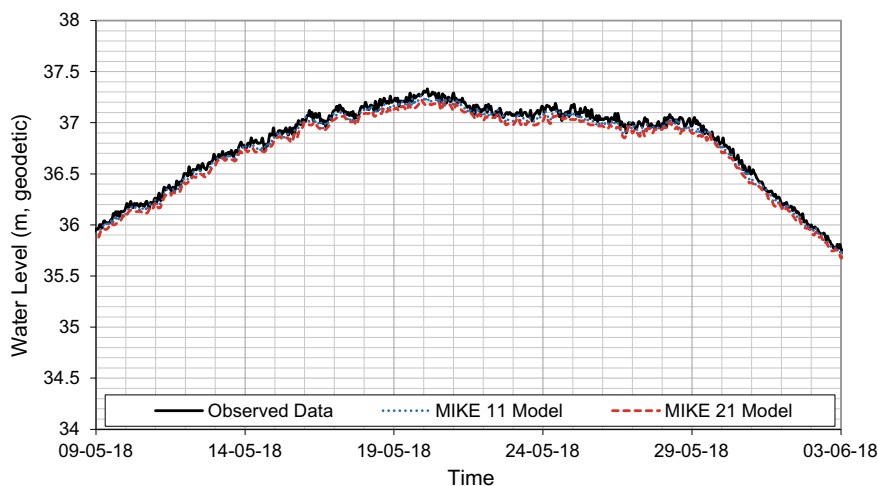


**Fig. 4** Comparison of water level hydrographs for the Fraser River at Steveston (08MH028)



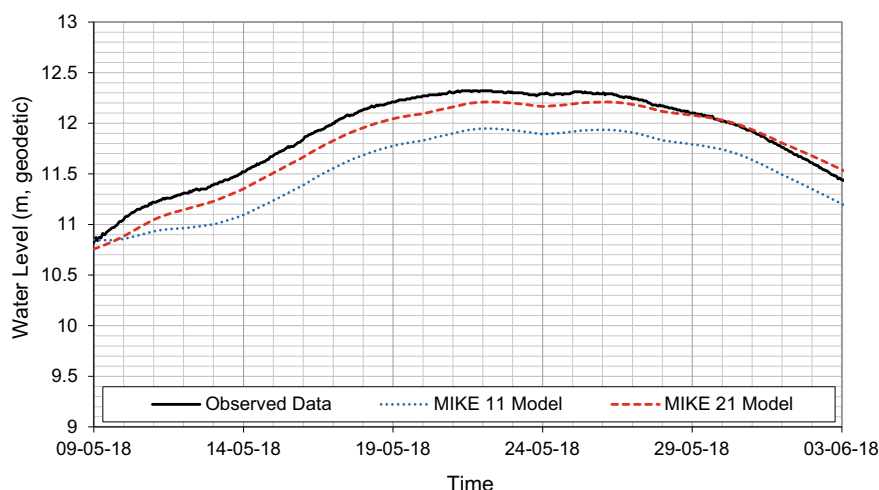


**Fig. 5** Comparison of water level hydrographs for the Fraser River at Mission (08MH024)



**Fig. 6** Comparison of water level hydrographs for the Fraser River at Hope (08MF005)

The modeled water level was compared with the observed data at 19 staff gauges. It was found that water level differences between the simulation data and the observed data were within  $\pm 0.15$  m at the staff gauges located adjacent to the Fraser, Harrison or Vedder Rivers. Large water differences were found at the staff gauges located on the side channels or sloughs far away from these three rivers. This was believed because of the poor DEM accuracy for those side channels or sloughs. The modeled water levels were also compared with the surveyed water surface profiles in the



**Fig. 7** Comparison of water level hydrographs for the Harrison Lake near Harrison Hot Springs (08MG012)

longitudinal direction for the Fraser and Harrison Rivers, and they were generally in good agreement.

## 2. Flow Comparison

The modeled flows are compared with two sets of flow measurements that were conducted on May 22–23 and May 30–31, 2018, respectively. Table 4 compares the modeled flows with the measured flows on May 22–23, 2018. Among the 34 flow split measurement sections,

- The differences between the modeled and measured are within  $\pm 5\%$  at 16 sections: L1, L4, L7, L8, L10, L15, L16, L17, L21, L22, L26, L28, L29, L31, L32, and L34, which indicates that the modeled flows are in excellent agreement with the measured.
- The flow differences are within  $\pm 10\%$  but exceed the range of  $\pm 5\%$  at 8 sections: including sections L3, L6, L9, L13, L23, L25, L27, and L33, which indicates that the modeled flows are in good agreement with the measured.
- The flow differences are within  $\pm 20\%$  but exceed the range of  $\pm 10\%$  at 4 sections: L11, L14, L20, and L24, which indicates that the modeled flows are not in good agreement with the measured.
- The flow differences exceed the range of  $\pm 20\%$  at the 6 sections: L2, L5, L12, L18, L19, and L30, which indicates that the modeled flows are not in agreement with the measured flows at those locations.

The sections with large flow differences are all located on the side channels of the Fraser River due to the poor DEM accuracy. The comparison of the modeled flows with the flows measured on May 30–31, 2018 showed the similar trend as Table 4.

**Table 4** Comparison of modeled flows to measured on May 22–23, 2018

Section	Date and time	Discharge (m <sup>3</sup> /s)		Discharge Difference (%)
		Measured	Simulated	
L 1	2018–05–22 11:20	9595	9746	1.6
L 2	2018–05–22 12:08	1344	922	– 31.4
L 3	2018–05–22 11:54	8499	8971	5.6
L 4	2018–05–22 12:29	6737	6902	2.4
L 5	2018–05–22 12:53	393	224	– 43.0
L 6	2018–05–22 13:18	9807	10,354	5.6
L 7	2018–05–22 13:43	5943	6012	1.2
L 8	2018–05–22 15:41	5518	5631	2.1
L 9	2018–05–23 9:28	7247	7700	6.2
L 10	2018–05–22 13:58	8905	9281	4.2
L 11	2018–05–22 14:36	6716	7519	12.0
L 12	2018–05–22 15:17	710	426	– 40.0
L 13	2018–05–22 16:01	4477	4120	– 8.0
L 14	2018–05–22 16:31	1741	1482	– 14.9
L 15	2018–05–22 16:14	8109	8448	4.2
L 16	2018–05–22 16:42	6811	7124	4.6
L 17	2018–05–22 16:58	6102	6391	4.7
L 18	2018–05–22 17:15	523	254	– 51.4
L 19	2018–05–23 8:42	382	227	– 40.6
L 20	2018–05–22 17:31	2167	1878	– 13.4
L 21	2018–05–23 9:55	10,041	10,088	0.5
L 22	2018–05–23 10:21	6639	6675	0.6
L 23	2018–05–23 10:47	7282	7760	6.6
L 24	2018–05–23 11:32	730	621	– 15.0
L 25	2018–05–23 11:41	2252	2103	– 6.7
L 26	2018–05–23 12:03	2745	2742	– 0.1
L 27	2018–05–23 12:48	6118	5505	– 10.0
L 28	2018–05–23 13:04	4859	4882	0.5
L 29	2018–05–23 13:22	11,030	11,415	3.5
L 30	2018–05–23 13:55	1821	1434	– 21.2
L 31	2018–05–23 14:14	3854	3742	– 2.9
L 32	2018–05–23 12:16	8671	8726	0.6
L 33	2018–05–23 9:05	4032	3738	– 7.3
L 34	2018–05–23 11:12	3482	3314	– 4.8

*Total flow of two sections*

(continued)

**Table 4** (continued)

Section	Date and time	Discharge (m <sup>3</sup> /s)		Discharge Difference (%)
		Measured	Simulated	
L2 + L3		9843	9892	0.5
L10 + L12		9615	9707	1.0
L8 + L13		9995	9751	– 2.4
L14 + L15		9851	9930	0.8
L17 + L33		10,135	10,129	– 0.1

### 3.6 Model Validations

As mentioned earlier, the 2D model was validated with an extensive set of observed flow and water level data based on the flood events of 1972, 2012, 2017, 2019, and 2020 with the Fraser River flows at Hope from 5000 m<sup>3</sup>/s (2019 flood) to 13,000 m<sup>3</sup>/s (1972 flood). The modeled peak flood levels were with a high accuracy for the 2012 and 2017–2020 flood events compared to the observed data, and the water level changes were successfully modeled for both the rising and recession limbs of these flood events. For example, Table 5 compared the modeled peak levels to the observed at the continuous gauge stations for the 2012 flood event; Fig. 8 shows the water level hydrographs comparison for the Fraser River at Mission (08MH024) for modeling the 2020 flood event. More model validation results will be published in the Province’s report “Lower Fraser River–2D Hydraulic Model Development and Design Flood Profile Update (in progressing)”.

## 4 Conclusion and Recommendations

### 4.1 Conclusions

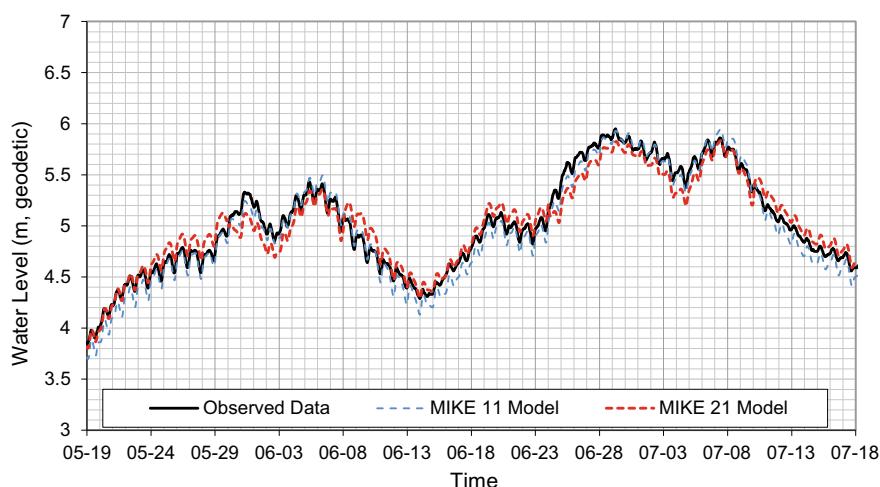
A 2D hydrodynamic model was developed for the LFR using the software MIKE 21 with flexible mesh. The mode includes the Fraser River main stem from the Trans-Canada Hwy #1 Bridge at Hope, BC to the Fraser River mouth at the Salish Sea and the key tributaries: the Harrison River and Lake, the Pitt River and Lake, and the Vedder River. Two roughness scenarios (RSLRL and RSHRL) were created for the model and have been proved effective at modeling the LFR in a broad flow range. The RSLRL can be applied at the river levels at Mission below 5.5 m, and the RSHRL can be applied at river levels at Mission above 5.5 m. It should be in caution that the reference level of 5.5 m is an approximate value, and the selection of the roughness scenarios must be based on model result analysis and comparisons to the observed water levels.

**Table 5** Peak water levels comparison at continuous gauge stations for model validation using the 2012 freshet flood event

Continuous Gauge Station	Peak Water Level—2012 Flood (m)		
	Observed	Modeled	Difference
Fraser R. at Hope Bridge—08MF005	37.65	37.57	− 0.08
North of Hunter Creek—Cont	31.25	31.20	− 0.05
Fraser R. at Laidlaw—08MF072	27.20	27.12	− 0.08
North of Herrling Island—Cont	21.87	22.06	0.19
Fraser R. near Agassiz—08MF035	17.21	17.07	− 0.14
Lower Kent	12.48	12.59	0.11
Harrison L. near Harrison Hotsprings—08MG012	12.44	12.37	− 0.07
Harrison R. below Morris Creek—08MG022	12.10	12.12	0.03
Fraser R. at Bell Slough	11.24	11.16	− 0.08
Fraser R. near Harrison Mills—08MF073	11.00	10.87	− 0.13
Chilliwack Creek PS (Wolfe Road)	8.86	8.74	− 0.12
Fraser R. at Cannor—08MF038	8.12	7.90	− 0.22
Fraser R. at Mission—08MH024	6.42	6.32	− 0.10
Fraser R. at Whonnock—08MH044	5.12	5.01	− 0.11
Albion PS	4.23	4.23	0.00
Baynes PS	3.34	3.27	− 0.07
Fraser R. at Port Mann PS—08MH126	2.71	2.57	− 0.14
New Westminster	2.47	2.35	− 0.12
Manson	2.41	2.33	− 0.08
Nelson Road	1.97	1.97	0.00
Fraser R. at Steveston—08MH028	1.97	1.84	− 0.12

*Note* The cells shaded in green indicate the water level difference within  $\pm 0.15$  m

The model was calibrated and validated with an extensive set of observed flow and water level data including the flood events of 1972, 2012, 2017 to 2020. The Fraser River flows at Hope were tested from 5000 m<sup>3</sup>/s (2019 flood) to 13,000 m<sup>3</sup>/s (1972 flood). The model has an excellent performance in simulating the peak flood levels and has a good performance in simulating the water level changes at both the rising and recession limbs for those flood events. It overall provides higher accuracy on modeling flood levels than the existing 1D model, especially on modeling the Harrison River. It can be used to calculate the LFR design flood profiles with extreme flood events. With the applications of the two roughness scenarios, it can also be used to model the LFR in a broad flow range in a regional scale.



**Fig. 8** Comparison of water level hydrographs for the Fraser River at Mission (08MH024) for model validation using the 2020 freshet flood event

## 4.2 Recommendations

The model should be only used in a regional scale but not a local scale. This is because of the lack of the river bathymetry data in the shallow water areas and side channels of the LFR, Harrison River and local drains. Due to the LFR channel dynamics, the DEM and the 2D model should be updated with new river bathymetry survey data at least every 10 years or after a significant flood event. In the next DEM update, more detailed bathymetry data should be collected for the LFR gravel reach and the Harrison River, especially for the shallow water areas and side channels. The river bathymetry data for local drains or sloughs should be also collected and incorporated into the DEM.

Additionally, higher mesh resolution can be tested to capture the river geometry and model flow features with higher accuracy if the DEM accuracy is improved. The reasons causing the change of roughness coefficients (Manning's  $n$  values) with river levels in the Fraser River reach from Barnston Island to Matsqui Island should be further investigated.

**Acknowledgements** Fraser Basin Council shared the DEM and the related reports for its completion of the Lower Mainland Flood Management Strategy initiative–Phase 2 in 2019. Public Works and Government Services Canada (PWGSC) shared the Fraser River bathymetry data collected from 2014 to 2020. GeoBC shared the 2016 LiDAR dataset for the DEM development.

## References

1. BC FLNRORD (2014) Fraser river design flood level update—hope to mission, final report, flood safety section, March 2014
2. Best J (2005) The fluid dynamics of river dunes: a review and some future research directions. *J Geophys Res* 110:F04S02. <https://doi.org/10.1029/2004JF000218>
3. Federal Highway Administration Resource Center (2019) Two-dimensional hydraulic modeling for highways in the river environment—reference document. Publication No. FHWA-HIF-19-061, October 2019, USA
4. MIKE by DHI (2019) MIKE flow model FM hydrodynamic module user guide
5. Northwest Hydraulic Consultant (2006) Lower Fraser River hydraulic model, final report prepared for Fraser Basin Council, December 2006
6. Northwest Hydraulic Consultant (2007) Fraser River 2007—freshet flood level forecasting, final report prepared for BC ministry of environment
7. Northwest Hydraulic Consultant (2008) Fraser river hydraulic model update, final report prepared for BC ministry of environment
8. Northwest Hydraulic Consultant (2010) Lower Fraser river 2009 Freshet flood level forecasting, final report prepared for BC ministry of environment
9. Northwest Hydraulic Consultant (2018) Lower Fraser River bathymetric survey and digital elevation model, final report for Fraser Basin Council, April 2018
10. Northwest Hydraulic Consultant (2019) Hydraulic modeling and mapping in BC's lower mainland—a lower mainland flood management strategy project draft report prepared for Fraser Basin Council.
11. The U.S. Department of Transportation—Federal Highway Administration (2019) Two-dimensional hydraulic modeling for highways in the river environment—reference document. Publication No. FHWA-HIF-19-061, October 2019
12. UMA Engineering Ltd (2000) Fraser River gravel reach hydraulic modelling study. Report prepared for City of Chilliwack
13. UMA Engineering Ltd (2001) Fraser and Harrison Rivers—hydrologic and hydraulic investigations. Report prepared for City of Chilliwack

# **Construction Management: History of Civil Engineering**



# A Brief History of the Pattullo Bridge



Jivan Johal and F. Michael Bartlett

**Abstract** The Pattullo Bridge over the Fraser River is the only major steel through-arch highway bridge remaining in British Columbia. Opened in 1937, the bridge replaced the narrow traffic lane above the railway tracks on the New Westminster Rail Bridge, improving vehicular traffic volumes. It connects the former BC capital of New Westminster with the region of Surrey and, along the Pacific Highway, the USA. Col. W. G. Swan (1885–1970) led the bridge design team, and the structural steel was fabricated and erected by the Dominion Bridge Company, both iconic names in the history of bridge engineering in BC. The Pattullo Bridge was the first major crossing into the Southern Greater Vancouver Area during a time of rapid societal expansion and has served a pivotal role, for almost 90 years, connecting suburban areas with metropolitan Vancouver. The impending replacement of the Pattullo Bridge heightens its historic importance: steel through-arch designs in BC, and so a part of British Columbia’s structural engineering history will become extinct.

**Keyword** Pattullo Bridge

## 1 Introduction

The Pattullo Bridge over the Fraser River, Fig. 1, connects Bridgeview, Surrey, and New Westminster in the Lower Mainland of British Columbia. The bridge is named to honor Premier Thomas Dufferin “Duff” Pattullo, who ceremoniously opened the bridge on November 15, 1937.

---

J. Johal (✉)

Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

e-mail: [jivan.s.johal@gmail.com](mailto:jivan.s.johal@gmail.com)

F. M. Bartlett

Department of Civil and Environmental Engineering, Western University, London, ON, Canada

© Canadian Society for Civil Engineering 2023

801

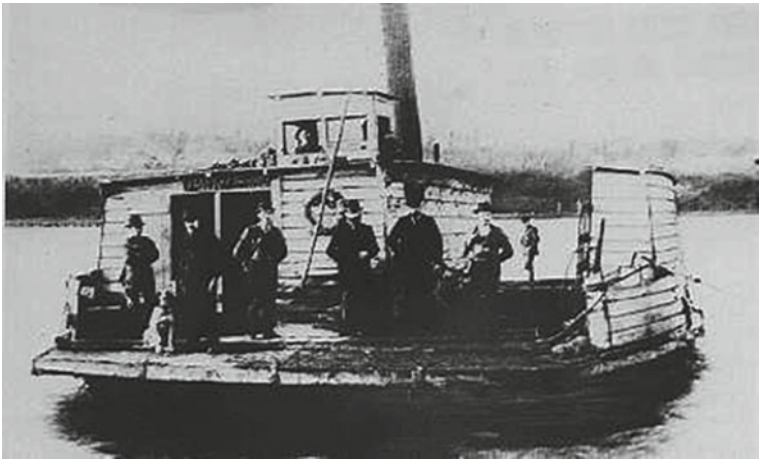
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363, [https://doi.org/10.1007/978-3-031-34593-7\\_49](https://doi.org/10.1007/978-3-031-34593-7_49)



**Fig. 1** Pattullo Bridge in 2021 with New Westminster Rail Bridge in lower foreground. *Source J. Johal*

## 2 Historic Background

New Westminster was first settled in 1859 as BC’s capital. It served as a vital hub for river traffic headed upstream to the Cariboo Region, which was experiencing a gold rush. The K. D. K. ferry, Fig. 2, was established as the first river crossing in 1898. It connected New Westminster and Surrey, to address an influx of economic growth and farming activities in the region as settlement grew along the southern banks of the Fraser River and into Surrey [1].



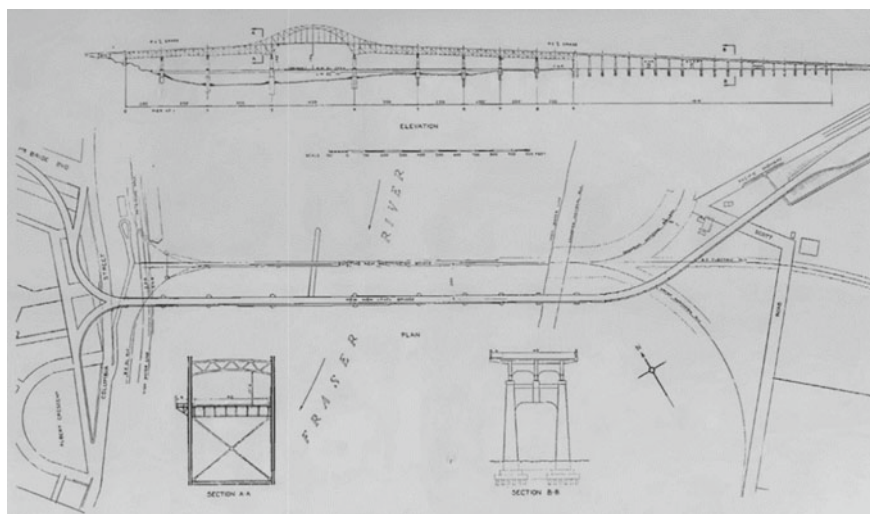
**Fig. 2** K. D. K Ferry, 1900s. *Source Surrey historical society*

With the establishment of Vancouver as a major West Coast port, the city and its surrounding area became a hub for transportation and commerce. The ferry connected the agrarian areas of Brownsville, on the south shore, to the City of Vancouver via Kingsway. By 1900, the ferry was incapable of handling the increasing traffic volumes. The New Westminster Rail Bridge, shown in the foreground of Fig. 1, opened in 1904 to accommodate the newly constructed Great Northern and New Westminster Southern Railways. To relieve the ferry, it featured a 15 ft. (4.6 m) wide roadway [2], supported on its upper chords. The toll for the upper roadway was 25¢, and farmers were very annoyed to be charged 25¢ per head for cattle herded across the bridge when they would only pay 25¢ per load if the cattle was trucked across [3].

Growing vehicular traffic volumes quickly exceeded the capacity of the New Westminster Rail Bridge, particularly given the frequent interruptions caused by opening the swing arm to accommodate river traffic. Moreover, the construction of the Pacific Coast Highway to Blaine WA, a two-lane gravel road, later paved in 1923, and the Trans-Canada highway, paved east to Chilliwack by 1935 [4], markedly increased traffic volumes. And [1] notes “the Pattullo Bridge and its road networks were essential to the growth of tourism. The booming 1920s made it possible for more and more families to afford an automobile and holiday trips were no longer dependent on public transportation”.

Thus, the need for a new high-level crossing was established. In 1933, an enquiry in accordance with the Navigable Waters Protection Act recommended that the new bridge can be located 200 ft (61 m) downstream of the New Westminster Rail Bridge, with the deep-water piers aligned closely with the existing railway bridge piers, Fig. 3. In November 1935 (sic), the Provincial Minister of Public Works commissioned W. G. Swan to “make surveys, obtain borings and test pile data, and proceed with the preparation of plans and specifications and estimate the cost of the new project” [2]. On April 30, 1935, a call for tenders was issued “on a unit price basis for all foundation work, combined with a lump sum price for the steel superstructure” [2]. On the closing date, two tenders were received and the lower, \$3,250,000 from the Dominion Bridge Company Limited (DB), was accepted. The Northern Construction Company and J. W. Stewart Ltd. were listed as subcontractors for the substructure and approach spans, respectively. The contract was signed on August 21, 1935 [2].

As shown in Fig. 3, the north end of the bridge consists of a series of Warren deck trusses of 200 and 250 ft (61 and 76 m) spans. The main span is a 450 ft (137 m) tied through-arch, with adjacent spans of 350 ft (107 m) “similar to the Honore Mercer Bridge at Montreal” [2]. The approaches from the south consist of 1400 ft (427 m) of reinforced concrete girder spans “reinforced by light welded truss construction” [2] and the approaches from the north are two conventional 70 ft (21 m) reinforced concrete spans. The clearance under the main span is “150 ft (46 m) above normal freshet level”. The deck is a 40 ft (12 m) wide reinforced concrete slab, that accommodates four lanes of traffic, and there is a six-foot (1.8 m) wide sidewalk, on the downstream side only. [2] notes that “the bridge deck is designed for uniform live



**Fig. 3** General plan of bridge [2]

loading of 80–100 lb. per sq. ft. (2.8–4.8 kPa) and concentrated loading of 20-ton (178 kN) trucks, with three-fourth of the load on the rear axle. As far as applicable, Canadian Engineering Standards Association specifications govern”.

Swan’s selection of an arch bridge created an architecturally appealing design for the location and the through-arch readily provided for the necessary large clearance below the deck. The relatively simple Warren trusses yielded cost savings. Swan also opted for large piers for Piers 1 through 4 because the silty nature of the riverbed and the adjacent soils would require deep foundations with excessively long piles.

### 3 COL. W. G. Swan: Designer

William George Swan (1885–1970), born in Kincardine, ON, received a B.A.Sc. in 1908 and C.E. in 1911 from the School of Practical Science at the University of Toronto [5]. He moved west in 1910 to become a divisional engineer for CNR, responsible for the first 100 miles (160 km) of the main line from New Westminster to Yale. His service in World War I began in 1915 and was exemplary: He was promoted to Major, was twice mentioned in dispatches, and was awarded the Distinguished Service Order and the French Croix de Guerre [6]. He returned to supervise extensive construction work as Chief Engineer of the Vancouver Harbour Commission (1920–1925), served as the first elected President of the Association of Professional Engineers of British Columbia (1921) and, starting in 1925, founded his own consulting firm. In 1930, Hiram Wooster and H. A. Rhodes joined his practice, and in 1945, they formed the firm Swan, Rhodes, and Wooster. After Rhode’s

death in 1954, the firm continued as Swan Wooster. He was called up for service in the Second World War, with the rank of Colonel in the Corps of Royal Canadian Engineers, and served as Chief Engineer of Pacific Command, earning the Order of the British Empire in 1945. He was awarded the University of Toronto Engineering Award for Achievement in 1952, the Engineering Institute of Canada (Fig. 4).

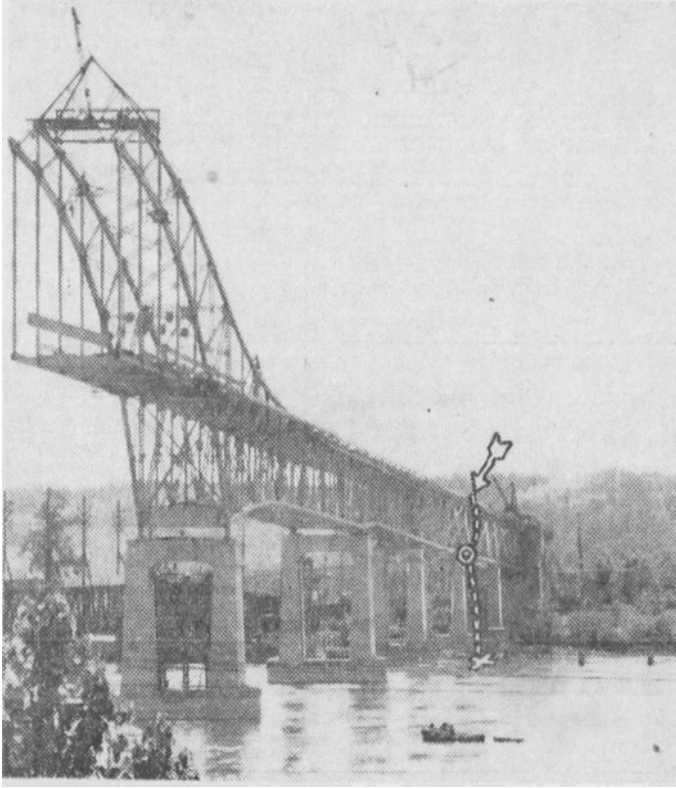
Julian C. Smith Medal for Achievement in the Development of Canada in 1955, and an Honorary Doctorate of Science by the University of British Columbia in 1956. He was very engaged in community service and retired in 1967.

Swan noted that although the onset of the depression had weakened the influx of projects and staff, firm's contributions to the Pattullo and Lions Gate Bridge had "put us on the map again... ..since then we never looked back" [7]. However, Swan's successful experience at the Pattullo Bridge became severely tainted on July 28, 1937. His only child, William MacKenzie Swan, was a fourth year Civil Engineering student at the University of British Columbia. Employed that summer as an inspector of the Pattullo Bridge construction, Bill tragically fell 30 m that afternoon and subsequently

**Fig. 4** Col. W. G. Swan, 1956. *Source* City of Vancouver archives







**Fig. 5** Trajectory of Bill Swan's fatal fall. *Source* Vancouver Sun 1937

died of the internal injuries sustained. Figure 5, from the front page of the July 29 *Vancouver Sun*, shows the location where he fell from (arrow) his fall (dotted line) and the boggy ground where he landed (X). Jamieson [8] notes “That Swan likely obtained the job for his only son must have been a burden that he carried with him for the rest of his life”.

#### **4 Dominion Bridge Co. Ltd.: Fabricator and Erector**

By the mid-1930s, the Dominion Bridge Company Ltd. was the most iconic steel fabricator and erector in Canada. The firm was spawned from the Wrought Iron Bridge Company (WIBC) of Canton Ohio. Sir John A. Macdonald's National Policy of 1878 created high tariffs that effectively prevented American fabricators like WIBC from selling bridges in Canada. With a new Canadian railway to the Pacific Ocean proposed, WIBC essentially created a branch plant in Canada, the Toronto Bridge Company, to be able to tender such projects. The key corporate officers were

all from WIBC, including Job Abbott, president, John Thom, secretary-treasurer, and Abbott's staff including his brother Ira Abbott, Phelps Johnson, and General Luke Lyman [9]. "This company was a small concern and only fairly successful, since its very first contract, a one hundred and eighty foot span of 42 tons, for London, Ontario, (at the price of \$4000) was dropped into the river during erection, and the company was held liable for its repair and replacement." [9]. In 1882, Abbott, as president of the company, secured funding from three Scottish investors for a new venture titled the "Dominion Bridge Company Ltd." to be sited at Lachine, in the vicinity of Montreal. In 1890, Abbot resigned from the Dominion Bridge Company to participate in field operations for the construction of the Bridgeport Bridge across the Ohio River at Wheeling WV, which was fabricated by WIBC. Phelps Johnson remained in Canada, rising up the ranks to become Chief Engineer in 1898 and subsequently served as General Manager (1892–1904), Director (1903–1926), General Manager and Chief Engineer (1904–1919), Managing Director (1910–1913), and President (1913–1919) [10].

Dominion Bridge expanded from its base at Lachine to open new branches in Toronto (1901), Winnipeg (1904), Ottawa (1909), Vancouver (1927), Calgary (1929), and Amherst, N.S. [9]. The Vancouver facility was initially a warehouse on False Creek, and the 1929 Annual Report of the company [11] notes:

The plant at False Creek has been worked to capacity throughout the year. A new site of about thirty-five (35) acres, in the district of Burnaby about four miles from the False Creek plant, has been purchased, and construction is now proceeding on a plant which will give additional capacity of about two thousand tons per month. The property is well situated and is so laid out that as demand increases the plant may be extended to an eventual capacity of three to four times the initial development.

The "Pacific Division", headed by Allan S. Gentles, had by 1939 completed "some very interesting steel structures... outstanding among which are the Empress Hotel Extension at Victoria; the Marine Building, Royal Bank Building, Canadian National Hotel, at Vancouver; the Burrard Street Bridge, reconstruction of the Second Narrows Vertical Lift Bridge; and the New Westminster Tied Arch Bridge over the Fraser River" [9]. Steel for the Pattullo Bridge was fabricated at the Burnaby plant, located near the neighborhoods of Willingdon and Renfrew Heights along Boundary Road.

Walter Pruden (1919–2005) joined Dominion Bridge in 1936 as an apprentice, assigned to assist journeymen and fitters with laying out the structure at the Burnaby plant [12]. He recalled that the project was "wonderful for increasing employment", observing that the number of DB plant employees increased from approximately 350 men to 600 over the course of the project, with many more employed onsite. He also recalled that the Provincial Government missed the first progress payment, forcing the company to carry on using their own resources. Perhaps fortunately for the project, when the second progress payment was due, the government was able to pay both the first and second progress payments [12]. Pruden spent 40 years with the company, eventually retiring in 1976, having been promoted to Chief Welding Inspector for the last decade of his career.

It is perhaps puzzling that very little information seems available describing the work of Dominion Bridge in fabricating and erecting the superstructure. There are no published papers in the *Engineering Journal* about it—the only mention [13] is of a presentation made by A.L. Carruthers, Bridge Engineer for the Province of British Columbia, to a meeting of the Victoria Section of the Engineering Institute of Canada on December 16, 1937. There is a 16-page supplement about the bridge in the November 13, 1937, edition of the *Vancouver Sun*—much of [2] *Engineering Journal* article [2] is reproduced verbatim, and the roles of Northern Construction Co. and J. W. Stewart Ltd. are described in detail, but there is virtually no mention of the superstructure. The last page of the supplement is clearly a full-page advertisement paid for by Dominion Bridge—but it is mostly photographs with very little text.

## 5 Construction

Swan [2] describes the substructure construction in detail. Steel caissons were constructed at a work area one mile (1.6 km) upstream of the bridge site where an existing wharf was repaired and launching ways constructed. The depth of water at the dock and in the channel to the bridge site was 30 ft (9.1 m), so the caissons were “completed to a stage where they drew 29 ft (8.8 m) before towing to place”. Sensibly, the contractors opted to build the simplest caissons first: The sequence was therefore Caissons 1, 4, 3, and finally 2 (as shown in Fig. 3, Pier 1 is the first in-water pier at the north end of the bridge, and the tied-arch spans between Piers 3 and 4). The construction milestones for each caisson were: “Start Cutting Edge”—presumably, fabrication of the steel cutting edge; “Launched”; “Towed to Site”; “Grounded”—ballasted to sink to the river bottom; and “Sunk to Finish Elevation”—after the excavation of material from inside the caisson was complete.

The processes for Caissons 1, 2, and 3 were relatively uneventful, but “Caisson No. 4 had a rather hectic experience: it grounded in towing, suffered an extreme freshet, and finished 19 inches (480 mm) from location as maximum error”. As clearly evident in Fig. 6, the “second heaviest freshet in the history of the Fraser River” caused scouring to depths of “20–40 ft (6–12 m) below the original bed level. The cutting edge of Pier 4 caisson was exposed, and the caisson was barely prevented from overturning by placing hundreds of tons of rock under the lower side, which was eight feet (2.4 m) out of level. One end of the cutting edge was three feet (1 m) out of position.” It is a credit to the contractors that the remedial measures they adopted, including “removing debris”, “jetting along the outside of the north portion of the cutting edge”, and resuming sinking operations, reduced the maximum error to 19 inches (480 mm).

Piers 5 through 8 are supported on 35 ft (10.7 m) piles. However, Pier 5, supporting the south end of the southern 350 ft (106 m) span, was subjected to even more distress than Pier 4. The cofferdam wall was punctured and the cofferdam flooded: “for a time it was feared that the work on Pier 5 would be destroyed” [2]. The bearing value of the foundation piling was eventually restored by: backfilling around the exposed





**Fig. 6** Caisson 4 in distress. *Source* [2]

portions of the piles with pit run sand and gravel, driving new sheet piles to add semi-circular ends to the rectangular Pier 5 cofferdam and installing additional piles in these regions; placing a 5 ft (3.2 m) tremie concrete seal below the surface of the original river bottom; and pressure grouting the void beneath the tremie seal. The newly rounded ends of the pier reduced the turbulence of the water flowing past it, mitigating the risk of scour, and rip rap was placed to further mitigate this risk. The remedial measures for Pier 5's cost \$35,000 (almost \$700,000 today) [5].

The rest of the viaduct for the approach span on the south (Surrey) side was completed with relative ease as the piers were not under water.

Construction of the approach spans commenced after the piers were completed.

The southern approach spans were reinforced concrete girders and cross-beams reinforced by light welded trusses known as the "Kane System" [2], Fig. 7. According to [14]: "the principles involved in this construction were conceived and developed by Mr. C. (Charles) S. Kane, the Montreal representative of the Dominion Bridge Company Ltd". The system was originally developed for building construction, where the use of concrete floors is desirable to provide inexpensive fire resistance. Kane's innovations included: making the concrete slab and beam stem into a load-carrying element; and, prefabricating the reinforcing cage as a light truss that could be erected as a unit and support the weight of the fluid concrete and necessary formwork without additional falsework.

Jacobsen [14] lists ten buildings constructed or under construction in the Montreal area between 1930 and 1932 that feature Kane System reinforcement. Appendix I of the thesis notes "the elimination of shoring, together with its other features, renders highly advantageous its adaption to bridge work". This certainly would have been attractive for the construction of the Pattullo southern approach spans.

**Fig. 7** Kane system reinforcement in south approach spans. *Source* Wikipedia



The Western Bridge Company Ltd., with facilities at the south shore of False Creek on First Avenue between Main and Cambie Streets, fabricated “a considerable share” of the Pattullo Bridge steelwork, including “a number of the largest trusses that to date have been fabricated in the province” [15]. The members are mostly built-up cross-sections, assembled by riveting and welding in the shop and subsequent riveting in the field. The main chords of the arch are built up from steel plates and angles to form box sections, with oval holes in the plates to reduce weight and allow access for riveting. The web members connecting the two chords of each arch have a Pratt truss configuration. The Warren truss members are typically laced, with end batten plates, to save weight. Nonetheless, some 5300 tons (4800 tons) of structural steel was required, as well as 1815 tons (1650 tons) of reinforcing steel and 62,000 cubic yards (47,400 m<sup>3</sup>) of concrete [16].

The structural steel was erected using Stiffleg Derricks, Fig. 8, and a smaller traveler running on the top chord of the tied arch, Fig. 5. Photographs of the time (e.g., [12]) show falsework under the exterior steel spans 1–2 and 8–9, Fig. 3, only, indicating that the other spans were initially cantilevered as erection progressed to eliminate the need for temporary supports in the river. This is confirmed by [12] who lists the lack of falsework as one of the unique features of the bridge. This is further confirmed by the presence of temporary steel attached to the top chord of the Warren trusses in Fig. 8, which increases the negative moment over the leftmost pier to facilitate cantilevering the adjacent span under construction.



**Fig. 8** Construction of Span 2–3, ca. 1937. *Source* City of Vancouver archives

The bridge was illuminated with sodium vapor lamps installed by Mott Electric of New Westminster, the first installation of such lamps in Canada [17]. By the end of construction, the Pattullo Bridge cast a large and impressive shadow over its predecessor, the New Westminster Rail Bridge.

## 6 Opening Ceremony: November 15, 1937

The opening ceremony was grand; both marine and land parades were held on the opening day, with long lines of tugboats, fishing boats, yachts, and other craft traveling below the central arch. On the main deck, floats, bands, and a line of mounted officers crossed in a ceremonious affair. The businesses of New Westminster displayed decorated shopfronts, and school children were given a half day off to witness the event.

Shortly after noon, Premier Duff Pattullo ceremoniously opened the bridge, using an oxy-acetylene torch to cut a steel bar linking two lengths of chain draped across the roadway, Fig. 9 [18]. Walter Pruden, present at the celebration, recalled that Mr. Harry Daly, the Chief Welding Inspector for the Dominion Bridge Company's BC plant, held the premier's hand and guided him with the cutting torch [12]. Pattullo said, to commemorate the occasion, "I trust that this massive structure may be the symbol of a breadth of vision and practical measures for fulfilment thereunder in the future expansion and upbuilding of our glorious province" [18]. There were, however, two minor mishaps in the opening ceremony, which took place in "rain chilled to

**Fig. 9** Premier Pattullo opens the bridge, November 15, 1937. *Source* Vancouver *Sun*



the marrow by a wind and temperature which is hardly in keeping with a bridge opening". The speech by Allan S. Gentles of Dominion Bridge was interrupted by the tugs and fishing boats letting off their horns. A canvas tarpaulin over the speakers' platform tore suddenly, causing several "high top-hatted government officials" to be drenched.

## 7 Operation and Economic Impact

The economic impact of the Pattullo Bridge has been immense. It was estimated that 65% of the capital cost supported construction workers, providing 1000 jobs for two years at the peak of the Great Depression [19]. It was estimated that almost 10 million people (through walking, personal commuting, public transportation, and other uses) would cross the bridge within its first year of its opening and that almost one million tons of commercial goods and other freight would be taken across the bridge [5]. Users initially paid a toll of 25¢ per crossing until 1952 to recover the capital expenses of the project.

The Pattullo Bridge facilitated significant development in Surrey and Delta and remains today an essential link between these cities and New Westminster, the Trans-Canada Highway, and the large, densely populated, economic centers of Burnaby and

Vancouver. As growth and development expand deeper into the Lower Mainland, additional crossings of the Fraser River at Port Mann and Delta, the Port Mann and Alex Fraser Bridges, respectively, were constructed to strengthen connections initially created by the Pattullo Bridge.

In 1998, the newly created South Coast British Columbia Transportation Authority (colloquially known as Translink) became the owner of the bridge, with responsibilities for maintenance and operation. The narrow lanes have been hazardous: road safety pylons were installed to divide oncoming lanes of traffic, previously only separated by paint. These two center lanes are typically closed overnight in response to a severe history of head-on accidents and related fatalities. Chain link fencing was installed to protect pedestrians using the walkway. Expensive retrofits to address seismic and other deficiencies have been carried out. Safety nets have been installed above the Columbia Street underpasses in New Westminster to protect from falling deteriorated concrete [20].

In January 2009, the entire creosote-treated timber approach span on the Surrey side was destroyed by fire. Structural girders and components from the recently completed Canada Line were integrated into the approach. On January 26, just 8 days after the fire, the bridge was re-opened to the public [21].

Currently, a replacement cable-stayed bridge is under construction. The existing Pattullo Bridge is expected to continue operating until the mid 2020s and will demolished once the new bridge is operational.

## 8 Historical Significance

The opening of the Burrard Street Bridge in 1932 and the Pattullo Bridge in 1937 perhaps marks the start of a new generation of large-scale high-level bridges with efficient approach spans throughout the Lower Mainland of British Columbia. Earlier bridges were primarily wooden trestles (1889 Granville Street Bridge), low swing-arm bridges (1909 Granville Street Bridge, 1904 New Westminster Rail Bridge), or low-level structures (1910 Westham Island Bridge, 1929 Capilano Bridge). The trend toward large-scale high-level structures continued with construction of the 1954 Granville St. Bridge, the 1960s Narrows Bridge (renamed the Ironworker's Memorial Bridge in 1994), the 1964 Port Mann Bridge. With the replacement of the Port Mann Bridge (1964–2012), the Pattullo Bridge is now the only remaining major through-arch bridge in the region. Its impending replacement will cause the extinction of this architectural style and will diminish the legacies of prominent names within the historic BC engineering community, including Col. W. G. Swan and the Dominion Bridge Company, Ltd., Fig. 10.

The scale of the Pattullo Bridge was very significant for the Lower Mainland, and at the time of its construction, it was the longest tied-arch span in Canada. There were many longer tied-arch bridges in America, as shown in Table 1—the Pattullo Bridge would have been tied for the 17th longest of the bridges on this list.



**Fig. 10** Official 1937 plaque, photographed in 2021. *Source* J. Johal

**Table 1** Longest through-arch bridges in Canada and USA

Rank	Bridge	Location	Date	Main span		Reference
				(ft)	(m)	
1	Bayonne	New York NY	1931	1675	510.5	[22]
2	Hell gate	New York NY	1916	978	298.1	[23]
4	West end	Pittsburgh PA	1932	780	237.7	[24]
5	McKees rocks	McKees Rocks PA	1931	750	228.6	[25]
6	South side point	Pittsburgh PA	1927	670	204.2	[26]
7	Detroit-superior	Cleveland OH	1918	620	189.0	[27]
8	Bourne	Bourne MA	1935	616	187.8	[28]
9	South grand island	Grand Island NY	1935	600	182.9	[29]
9	Yaquina bay	Newport OR	1936	600	182.9	[30]
11	Old trails	Tiprock AZ	1916	592	180.4	[31]
12	Tacony-palmyra	Philadelphia PA	1929	558	170.1	[32]
13	Tyngsborough	Tyngsborough MA	1930	547	166.7	[33]
14	Bellows falls arch	Bellows Falls VT	1905	540	164.6	[34]
15	Brady street	Pittsburgh PA	1896	520	158.5	[35]
16	Jerome street	McKeesport PA	1936	455	138.7	[36]
17	Broadway	Little Rock AR	1923	450	137.2	[37]
17	Pattullo	New Westminster BC	1937	450	137.2	
–	Honoré mercier	Montreal QC	1934	302	92.1	[38]



Nathan Holth describing the Pattullo Bridge on his [Historicbridges.org](https://www.historicbridges.org) website [39] considers the Pattullo Bridge to be “one of the largest and oldest bridges remaining in Greater Vancouver. It also has a high level of historic integrity with no major alterations to the arch or deck truss spans”. He laments the plan to replace the bridge, noting “Greater Vancouver has very few bridges of any heritage value” and concluding “the Pattullo Bridge is a beautiful and pristine heritage bridge that should be preserved not destroyed”.

## 9 Summary

The Pattullo Bridge over the Fraser River is the only major steel through-arch highway bridge remaining in British Columbia. Opened in 1937, the bridge replaced the narrow traffic lane above the railway tracks on the New Westminster Rail Bridge, improving vehicular traffic volumes. It connects the former BC capital of New Westminster with the region of Surrey and, along the Pacific Highway, the USA. Col. W. G. Swan (1885–1970) led the bridge design team, and the structural steel was fabricated and erected by the Dominion Bridge Company, both iconic names in the history of bridge engineering in BC. The Pattullo Bridge was the first major crossing into the Southern Greater Vancouver Area during a time of rapid societal expansion and has served a pivotal role, for almost 90 years, connecting suburban areas with metropolitan Vancouver. The impending replacement of the Pattullo Bridge heightens its historic importance: steel through-arch designs in BC, and so a part of British Columbia’s structural engineering history, will become extinct.

**Acknowledgements** The paper has been improved through thoughtful insight and comments provided by Eric Jamieson and Kevin Baskin—though the authors assume all responsibilities for any errors it may contain. Financial support from the Natural Science and Engineering Research Council to the second author is gratefully acknowledged.

## References

1. Cook D (2018) Pattullo bridge replacement project: historical heritage study. Denise Cook Design, North Vancouver, BC
2. Swan WG (1937) The substructure of the new highway bridge over the Fraser River at New Westminster, BC. *Eng J* 20:768–777
3. Wikipedia (2022a) New Westminster bridge. [https://en.wikipedia.org/wiki/New\\_Westminster\\_Bridge](https://en.wikipedia.org/wiki/New_Westminster_Bridge). Accessed 14 Jan 2022
4. Hayes D (2005) Historical atlas of Vancouver and the lower Fraser valley. Douglas & McIntyre Ltd., Vancouver, BC
5. Anon (1937a) Pattullo bridge: engineering triumph. *Vancouver Sun*, November 13. New Bridge Supplement on the Official Opening Pattullo Bridge, pp 2–16

6. UBC Archives (1956) The degree of doctor of science, (*Honoris causa*) conferred at congregation, May 15, 1956: William George Swan. <https://www.library.ubc.ca/archives/hdcites/hdcites4.html>. Accessed 14 Jan 2022
7. Anon (1968) Engineer retires after six decades serving British Columbia. *BC Profess Eng* **147**:15–16
8. Jamieson E (2008) Tragedy at second narrows. Harbour Publishing Co., Ltd., Madeira Park, BC
9. Shearwood FP (n.d.) (probably ca. 1939). A dominion that spans the dominion. No publisher.
10. Millard JR (2005) Johnson, Phelps. Dictionary of Canadian biography, vol 15. University of Toronto/Université Laval. [http://www.biographi.ca/en/bio/johnson\\_phelps\\_15E.html](http://www.biographi.ca/en/bio/johnson_phelps_15E.html). Accessed 01 Jan 2022
11. Dominion Bridge Co. Ltd. (1929) Report and statement DOMINION BRIDGE COMPANY LTD for the year ended 31st October 1929. Dominion Bridge Co. Ltd., Lachine PQ. [http://digital.library.mcgill.ca/hrcorpreports/pdfs/D/Dominion\\_Bridge\\_Co\\_Ltd\\_1929.pdf](http://digital.library.mcgill.ca/hrcorpreports/pdfs/D/Dominion_Bridge_Co_Ltd_1929.pdf). Accessed 14 Jan 2022
12. Pruden W (1998) Walter pruden: Pattullo Bridge (oral history recorded March 4). New Westminster Archives, New Westminster, BC. <http://archives.newwestcity.ca/permalink/41082/>. Accessed 15 Jan 2022
13. Reid K (1938) Victoria branch. *Eng J* **29**:167
14. Jacobsen ER (1932) The kane system of composite construction. Masters of Engineering Thesis, McGill University
15. Anon (1937b) Largest trusses B.C. Made. *Vancouver Sun*. New Bridge Supplement on the Official Opening Pattullo Bridge, pp 3–16
16. Anon (1937c) 106,000 Barrels of cement; 7100 Tons of Steel Used in New Structure. *Vancouver Sun*, November 13. New Bridge Supplement on the Official Opening Pattullo Bridge, pp 12–16
17. Anon (1937d) Sodium vapor lamps on span light themselves. *Vancouver Sun*, November 13. New Bridge Supplement on the Official Opening Pattullo Bridge, pp 4–16
18. Anon (1937e) Premier pattullo cuts chain with torch; British Columbia's great \$4,000,000 bridge across fraser river is ready for traffic. *Vancouver Sun*, November 15, p 1
19. Anon (1937f) 2 Millions went to workers. *Vancouver Sun*, November 13. New Bridge Supplement on the Official Opening Pattullo Bridge, pp 3–16
20. Ministry of Transportation and Infrastructure (MoTI) (2018) Pattullo bridge replacement project: business case. Business-Case.pdf (pattullobridgereplacement.ca). Accessed 16 Jan 2022
21. Woolford D, Matson D (2009) The Pattullo bridge emergency repair. In: Transportation association of Canada annual conference, Vancouver BC, p 16
22. Wikipedia (2022b) Bayonne bridge. [https://en.wikipedia.org/wiki/Bayonne\\_Bridge](https://en.wikipedia.org/wiki/Bayonne_Bridge). Accessed 07 Jan 2022
23. Wikipedia (2022c) Hell gate bridge. [https://en.wikipedia.org/wiki/Hell\\_Gate\\_Bridge](https://en.wikipedia.org/wiki/Hell_Gate_Bridge). Accessed 07 Jan 2022
24. Wikipedia (2022d) West end bridge (Pittsburgh). [https://en.wikipedia.org/wiki/West\\_End\\_Bridge\\_\(Pittsburgh\)](https://en.wikipedia.org/wiki/West_End_Bridge_(Pittsburgh)). Accessed 07 Jan 2022
25. Wikipedia (2022e) McKees rocks bridge. [https://en.wikipedia.org/wiki/McKees\\_Rocks\\_Bridge](https://en.wikipedia.org/wiki/McKees_Rocks_Bridge). Accessed 07 Jan 2022
26. Wikipedia (2022f) Point bridge (Pittsburgh). [https://en.wikipedia.org/wiki/Point\\_Bridge\\_\(Pittsburgh\)](https://en.wikipedia.org/wiki/Point_Bridge_(Pittsburgh)). Accessed 07 Jan 2022
27. Wikipedia (2022g) Detroit-superior bridge. [https://en.wikipedia.org/wiki/Detroit%E2%80%93Superior\\_Bridge](https://en.wikipedia.org/wiki/Detroit%E2%80%93Superior_Bridge). Accessed 07 Jan 2022
28. Wikipedia (2022h) Bourne bridge. [https://en.wikipedia.org/wiki/Bourne\\_Bridge](https://en.wikipedia.org/wiki/Bourne_Bridge). Accessed 07 Jan 2022
29. Wikipedia (2022i) South Grand Island bridge. [https://en.wikipedia.org/wiki/South\\_Grand\\_Island\\_Bridge](https://en.wikipedia.org/wiki/South_Grand_Island_Bridge). Accessed 07 Jan 2022
30. Wikipedia (2022j) Yaquina Bay bridge. [https://en.wikipedia.org/wiki/Yaquina\\_Bay\\_Bridge](https://en.wikipedia.org/wiki/Yaquina_Bay_Bridge). Accessed 07 Jan 2022



31. Wikipedia (2022k) Old trails bridge. [https://en.wikipedia.org/wiki/Old\\_Trails\\_Bridge](https://en.wikipedia.org/wiki/Old_Trails_Bridge). Accessed 07 Jan 2022
32. Wikipedia (2022l) Tacony–Palmyra bridge. [https://en.wikipedia.org/wiki/Tacony%E2%80%9393Palmyra\\_Bridge](https://en.wikipedia.org/wiki/Tacony%E2%80%9393Palmyra_Bridge). Accessed 07 Jan 2022
33. Wikipedia (2022m) Tyngsborough bridge. [https://en.wikipedia.org/wiki/Tyngsborough\\_Bridge](https://en.wikipedia.org/wiki/Tyngsborough_Bridge). Accessed 07 Jan 2022
34. Wikipedia (2022n) Arch bridge (bellows falls). [https://en.wikipedia.org/wiki/Arch\\_Bridge\\_\(Bellows\\_Falls\)](https://en.wikipedia.org/wiki/Arch_Bridge_(Bellows_Falls)). Accessed 07 Jan 2022
35. Wikipedia (2022o) Brady street bridge. [https://en.wikipedia.org/wiki/Brady\\_Street\\_Bridge](https://en.wikipedia.org/wiki/Brady_Street_Bridge). Accessed 07 Jan 2022
36. Wikipedia (2022p) Jerome street bridge. [https://en.wikipedia.org/wiki/Jerome\\_Street\\_Bridge](https://en.wikipedia.org/wiki/Jerome_Street_Bridge). Accessed 07 Jan 2022
37. Wikipedia (2022q) Broadway bridge (little rock). [https://en.wikipedia.org/wiki/Broadway\\_Bridge\\_\(Little\\_Rock\)](https://en.wikipedia.org/wiki/Broadway_Bridge_(Little_Rock)). Accessed 07 Jan 2022
38. Wikipedia (2022r) Honoré Mercier bridge. [https://en.wikipedia.org/wiki/Honor%C3%A9\\_Mercier\\_Bridge](https://en.wikipedia.org/wiki/Honor%C3%A9_Mercier_Bridge). Accessed 07 Jan 2022
39. Holth N (2014) Patullo bridge. <https://historicbridges.org/bridges/browser/?bridgebrowser=britishcolumbia/pattullo/>. Accessed 17 Jan 2022

# Britannia Mine: A Canadian Innovation with a Lasting Environmental Impact



Ali A. Mahmood and F. Michael Bartlett

**Abstract** The Britannia Mine, situated on the east shore of Howe Sound, 45 km (28 miles) north of Vancouver, produced more copper than any other mine in the British Empire between 1925 and 1930. Dr. A. A. Forbes originally discovered minerals there in 1888. When it ceased operations in 1974, it had produced over 517,000 metric tonnes (mt) of copper, 125,000 mt of zinc, and significant quantities of lead, cadmium, silver, gold and pyrite. The mine applied and improved a froth-flotation system that was particularly efficient in separating and concentrating the ore: the Britannia deep-cell flotation system helped triple the yield at one of their mills. The steep local mountain slopes were used to generate hydroelectric power that provided compressed air for the mine ventilation system, and to transport ore through the concentrator by gravity. Old rails were recycled to make grinding balls for the mills, many years before recycling technology and the circular economy were recognized to be desirable practices. They used IBM punch cards for time keeping in 1929. The mine also leaves an environmental legacy, however, as one of the largest sources of metal pollution in North America. Remediation efforts to protect Howe Sound and the Squamish River from acid rock drainage will be necessary for the foreseeable future.

**Keywords** Acid rock drainage · Froth flotation · Hydro-electric power

## 1 Introduction

Figure 1 shows the location of the historic Britannia Mine on the east shore of Howe Sound, about 45 km (28 mi) north of Vancouver, British Columbia, Canada. Dr. A. A. Forbes, a physician with an interest in prospecting, first discovered copper ore there in

---

A. A. Mahmood (✉)

Advanced Studies Research Centre, Drummondville, QC, Canada

e-mail: [ali@asrcentre.com](mailto:ali@asrcentre.com)

F. M. Bartlett

Department of Civil and Environmental Engineering, Western University, London, ON, Canada

© Canadian Society for Civil Engineering 2023

819

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_50](https://doi.org/10.1007/978-3-031-34593-7_50)

**Fig. 1** Location of Britannia Mine [2]



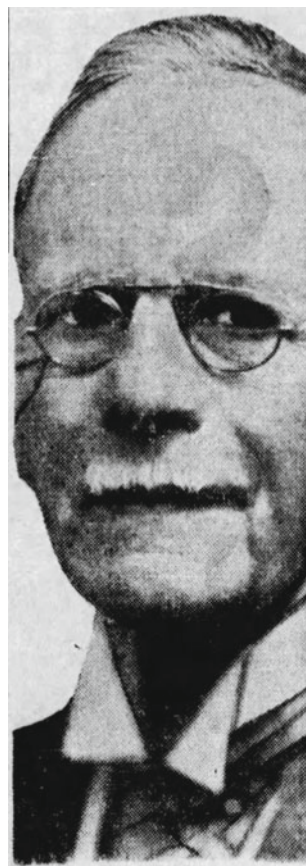
1888. Large-scale mining occurred between 1903 and 1974: it is estimated that 60,000 people from 50 countries came to live and work in the area as mine employees [1]. The Britannia Mine has had a profound social, economic and environmental impact on the area. The present study will briefly shed light on the history, significance, and environmental legacy of the Britannia Mine.

## 2 Early History: Dr. Forbes’ “Famous Buck”

Dr. Alexander Allen Forbes (c. 1850–1935), Fig. 2, was a man of many talents. Originally from Aberdeen, Scotland, he joined the British navy at age 13, cultivating a taste for chemistry while serving for 10 years before moving to Connecticut where he married and studied to become a school teacher [3]. He subsequently attended medical school at New York University, and, accompanied by his wife Annie, arrived in Vancouver in 1886. They lived at a small cottage he built at what is now Hopkins Landing, where he medically tended the Howe Sound First Nations people, often travelling by canoe [3].

Forbes’ interest in chemistry evolved to an interest in prospecting in his spare time. In 1888 a local “dog” fisherman named Granger offered to show him, for a \$400 fee, the location of rock containing interesting deposits in the Britannia area. Snow on the mountain hampered their investigations. Days later, however, Forbes shot a buck on the mountain and the animal kicked considerably before dying, resulting in the uncovering of good copper prospects [3]. Forbes bought in the area and paid Granger his \$400 [4].

**Fig. 2** Dr. A. A. Forbes [3]



Forbes was not successful in establishing any financial backers for his find in Britannia, so he later sold his claim to a “Mr. Turner” [4]. Forbes also had interests on property on Texada Island, which he subsequently “sold out at a good figure” [3]. In about 1900, he and his wife moved to Minnesota, but they returned in 1912 when he became a company doctor in Powell River.

In 1898, a trapper named Oliver Furry staked claims in the area as agent for friends and family of Thomas T. Turner, a Vancouver furrier, and Joseph Boscowitz and his sons Leopold and David, based in Victoria. In 1899 the Boscowitz group began development work, and while early results were encouraging, the group were not experienced in mining and sought an experienced partner. In late 1899, Howard Walters of Libby, Montana, purchased a 70% interest in the venture and enlisted Vancouver partners to form the Britannia Copper Syndicate Ltd [4].

The mine was brought into production in 1904 by George Robinson, a mining engineer from Butte Montana, who acquired an interest in the syndicate, buying out Walters and Boscowitz. New York financier Henry Stern incorporated the Howe

Sound Company under the laws of the State of Maine, which in 1904 acquired all the shares of the Britannia Copper Syndicate [4]. Ore shipments to a company-owned smelter at Crofton began in 1905. The Britannia Power Co. Ltd was formed to operate a hydroelectric power plant and dam on the site. In 1908 the assets of the Britannia Copper Syndicate, the Britannia Smelting Company and the Britannia Power Co. Ltd. were assumed by the Britannia Mining and Smelting Company, with the Howe Sound Company as its holding company [4].

### 3 Production Prior to World War I

The Britannia Mine consists of various ore bodies located at various elevations in Britannia Mountain, as shown in Fig. 3, (later Mount Sheer) and requires transportation systems to bring the ore to concentrators located at Britannia Beach. A two-stage aerial tramway, Fig. 4, that was 5150 m long and dropped 980 m vertically, carried ore from the Jane Mine to the concentrator at Britannia Beach. Both the tramway and concentrator were constructed in 1904 [5]. The miners lived at “Jane Camp”, next to the mine. By 1909, however, the outlook was bleak, due to difficulties to produce a satisfactory concentrate from the Jane ore, and low copper prices. But development continued—in particular, a horizontal tunnel was constructed, labeled as the “2200 or Main Tunnel” in Fig. 3, to bring ore to Tunnel Camp at the top of the tramway. By 1911 a tunnel to the Fairview ore zone provided a supply of higher grade ore that was more readily concentrated [4].

In 1912, a second concentrator, Mill No. 2, was commissioned at Britannia Beach. The price of copper dropped significantly in 1913, however, and mining operations were cut by half [4].

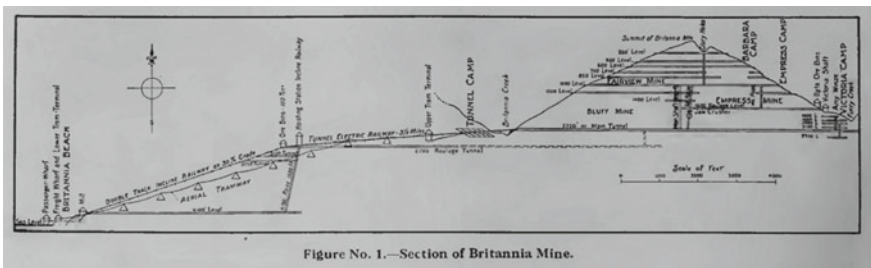


Fig. 3 Section of Britannia Mine [6]



**Fig. 4** The 5 km long aerial tramway in 1911 [1]

## 4 Production During World War I

The outbreak of World War I caused the price of copper to rebound and activity at the mine resumed. At 12:03 am on March 22, 1915, a severe debris avalanche hit Jane Camp: 57 lives were lost, and the top leg of the tramway and many buildings were destroyed. “As far as can be learned a whole peak crashed down, loosening thousands of tons of huge boulders and bringing earth, trees and snow upon the illfated village. It was swift and sudden and the unfortunates were snuffed out in their sleep” [7]. Jane Camp was subsequently abandoned.

Fortunately, bunkhouses and mine facilities constructed at the Tunnel Townsite, shown as “Tunnel Camp” in Fig. 3, allowed production to continue after loss of Jane Camp. A 5.6 km narrow-gauge electric surface railway was constructed in 1915 to carry ore from the tunnel portal, shown as “Upper Train Terminal” in Fig. 3, to the top of the incline, “Hoisting Station Inclined Railway” in Fig. 3. From there, cable-drawn rail cars carried the ore down a 27% grade to ore bins at Mill No. 2, a vertical drop of 500 m.

In 1916, Britannia Mine became the largest copper mine in the British Empire [4]. New horizontal tunnels were constructed, “2700 Haulage Tunnel” and “4100’ Level” in Fig. 3, to carry the ore to the concentrator—eventually they replaced the electric railway and incline, which were relegated to carrying passengers and supplies. In 1917, the price of copper hit 29.2 ¢/lb—and production increased from 200,000 metric tonnes (mt) in 1915 to 663,000 mt in 1918 [4]. The 1918 Spanish Flu Pandemic devastated the Britannia community, killing approximately 40: the Dance Hall at Tunnel Camp was used as a temporary hospital. In 1919, Mill No. 1 was dismantled.

The price of copper collapsed at the end of World War I. By November 1920, the concentrator, tramways, electric railway and incline were mothballed, and the mines payroll was reduced by 40% [4].

## 5 The Roaring Twenties

Hard times continued into 1921. The price of copper dropped to 12.6 ¢/lb. In March, Mill No. 2 was destroyed by fire. On October 28, a log and debris obstruction on the upper reach of Britannia Creek broke, causing a flash flood that destroyed part of the Britannia Beach community, killing 37 and hospitalizing another 15. The scale of the devastation is clear from Fig. 5. Phone lines were destroyed, and two men tasked with informing Vancouver of the disaster had to row ~ 29 km to find a working telephone [4].

Production was suspended for all of 1922 to facilitate recovery and rebuilding. A new concentrator, Mill No. 3 was constructed. A newly constructed vertical shaft, marked “Raise 1500 ft.” in Fig. 3, allowed ore to be transferred by gravity from the 2700 portal to the 4100 tunnel, where trams took it out to the mill storage bins.

Production resumed. The price of copper ranged between 12 and 14 ¢/lb. for most of the decade but rose to 18 ¢/lb. in 1929. The Britannia Mine was again the largest producer of copper in the British Empire. In 1930, ore production exceeded two million metric tons [4].

Mine Manager C. P. Browning, in an address to the Vancouver Branch of the Engineering Institute of Canada on May 11, 1927, described the different mining techniques adopted [9]:

The Britannia mines, so-called, are comprised of five distinct mines running from west to east. The type of deposits vary in these five sections, which means that different systems of mining must be resorted to in order to make a successful extraction of the mineralized portion.



**Fig. 5** Destruction of Britannia Beach community by 1921 flood [8]



The four systems of mining, all in use at Britannia, are as follows:

1. **Glory Hole:** Outside method, only used when working from the top. This is the cheapest method, with good light and ventilation, and immediate extraction.
2. **Shrinkage:** This consists of service raises, chutes and connecting through chutes, the blasting being done in the solid material by the men standing on the broken material. This method ties up broken ore which, if water runs through it, will oxidize it. It also requires good walls. The cost of this method is between 75 cents and \$1.00 per ton delivered to the main haulage level.
3. **Square Set:** This method is used where the walls are poor and require, after mining, to be filled. The standard set in use is 10- by 10-inch (25 by 25 cm) timbers with 6-foot (1.8 m) centres.
4. **Rill Stopes:** The inclined stope is characteristic of this type of mining. In some cases, where the walls are exceedingly bad, timbering is resorted to. The cost per ton in types (3) and (4) is about the same, averaging \$3.00 per ton.

The level intervals vary, depending on the type of mining employed. In good standing ground, 150–200 ft (45–60 m), and in poor standing ground 100–150 ft (30–45 m).

## 6 Later Years

The activity at the Britannia Mine fluctuated with the market price of copper. In 1930, it fell to 13 ¢/lb. and in 1932 to 5.8 ¢/lb, so by 1933 the mill was operating at 20% capacity, with the workforce reduced from 1000 in 1930 to 400. Layoffs were minimized by allocating reduced employment in staggered shifts [4]. Eventually production returned to roughly two million metric tonnes per year between 1937 and 1940. World War II created a labour shortage as workers enlisted or sought better-paying jobs in other industries contributing to the war effort. Production dropped to 514,000 mt in 1945 [4].

In 1946, the Britannia local of the International Union of Mine, Mill and Smelter Workers (Canada) went on strike, closing the mine for over three months. Production that year dropped to less than 400,000 mt, and the New York office of Howe Sound Co. decided to close the mine if new production quotas could not be met. Fortunately metal prices recovered, and the mine remained active and viable until 1957, when prices collapsed again. Direct subsidies from the federal and provincial governments were required to keep the mine operational on a reduced scale, but operations were temporarily suspended in March 1958. Operations resumed in 1959 and continued at a moderately active level through 1960 and 1961 [4].

In 1962, the Howe Sound Company decided to get out of the mining business, and sold the Britannia operation to the Anaconda Company (Canada) Ltd., a wholly owned subsidiary of Anaconda American Brass Limited, in January 1963 for \$5,000,000 [4]. The new owners got off to a rocky start when the workers struck again in 1964, postponing production for five months to May 1965. Gradually the reserves depleted, and the last production shift went underground in November 1974 [4].



Smitheringale [4] notes:

During nearly 70 years of almost continuous operation, Britannia Mines produced 517,000 mt of copper, 125,290 mt of zinc and lesser, but economically important, quantities of gold, silver, lead and cadmium. The network of tunnels, shafts, raises and development drifts connecting the workings extended more than 1.8 km vertically and 7 km horizontally, and is estimated to have exceeded 200 km total length.

## 7 Technological Advances

The Britannia Mine featured several technological advances that will be briefly described herein.

### 7.1 *Froth Flotation*

The ore at Britannia Mine required fine grinding to liberate the chalcopyrite, and the available recovery procedure in 1905, gravity separation, was not efficient for finely ground material. In 1909 or 1910, experiments were conducted at Mill No. 1 using Elmore vacuum flotation cells, and froth flotation was investigated there in 1911. In 1912 a new flotation circuit was added to Mill No. 1—the first successful application of froth flotation in British Columbia and possibly in Canada [4]. The adoption of this technology was essential to the financial success of the mill, given the nature of the ore it produced.

The froth flotation procedure involves mixing the finely ground ore in water and adding a “frothing agent” (in this case, pine oil) and “collectors” (potassium xanthate). When air is blown into the mixture, the ore particles stick to the air bubbles and rise, while the waste particles sink. The ore-air bubbles are then skimmed off the surface of the mixture, recovering the ore. The mill workers even developed their own “Britannia deep cell” throughout the ‘20 and ‘30 s, which yielded a recovery rate of 90%, quite high for that time [1].

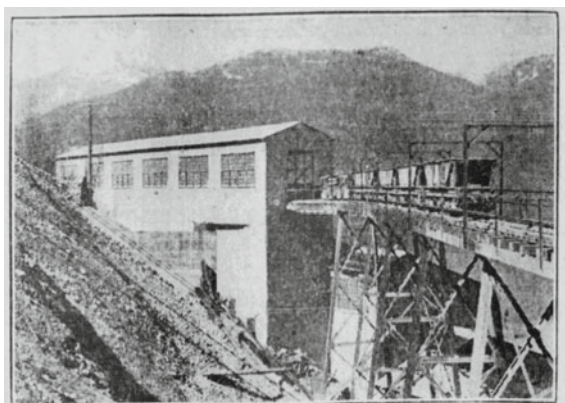
### 7.2 *Mill No. 3*

The concentrator at Mill No. 3 features several innovations. One is its application of froth flotation to separate the ore as described previously. It is also remarkable because its hillside location allowed gravity to transport the ore as it traveled vertically through the concentrating process. Figure 6 shows an ore train arriving at the upper storey of the mill. With the train in position, gates at the bottoms of the cars open and the ore drops into the storage areas, Fig. 7.

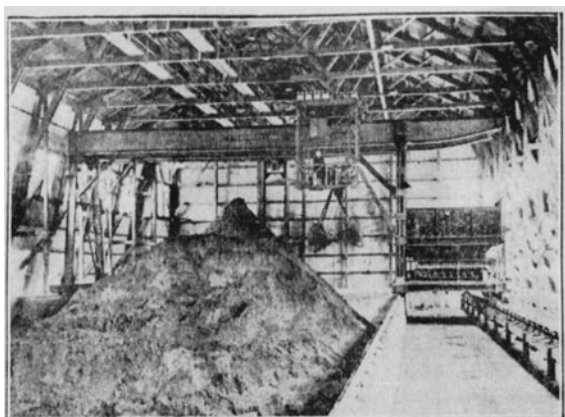
Figure 8 is a schematic diagram of the concentrator mill. The ore is dumped into the 500 ton (450 mt) coarse ore bins at the top of the mill and is driven by gravity

through the rest of the process. Browning [9] noted that “milling charges were 50 cents per ton, including all direct and indirect costs.” The concentrates were stored in concrete bins of 9000 mt capacity. They were removed from the bins using a crane with a clam-shell, and taken by conveyor belt to the vessel at the Britannia Beach wharf. Smitheringale [4], citing [11] notes: “In reference to Britannia’s gravity-fed Mill No. 3, 46 years after it was built, the Director of Mining Research for Anaconda stated ‘Although this design would not be used today because of high construction costs and inefficient use of labour and supervision on the many operating levels, the metallurgical efficiency of the plant is good, even by modern standards’”.

**Fig. 6** Ore train arriving at mill [10]



**Fig. 7** Ore dumped into storage area [10]



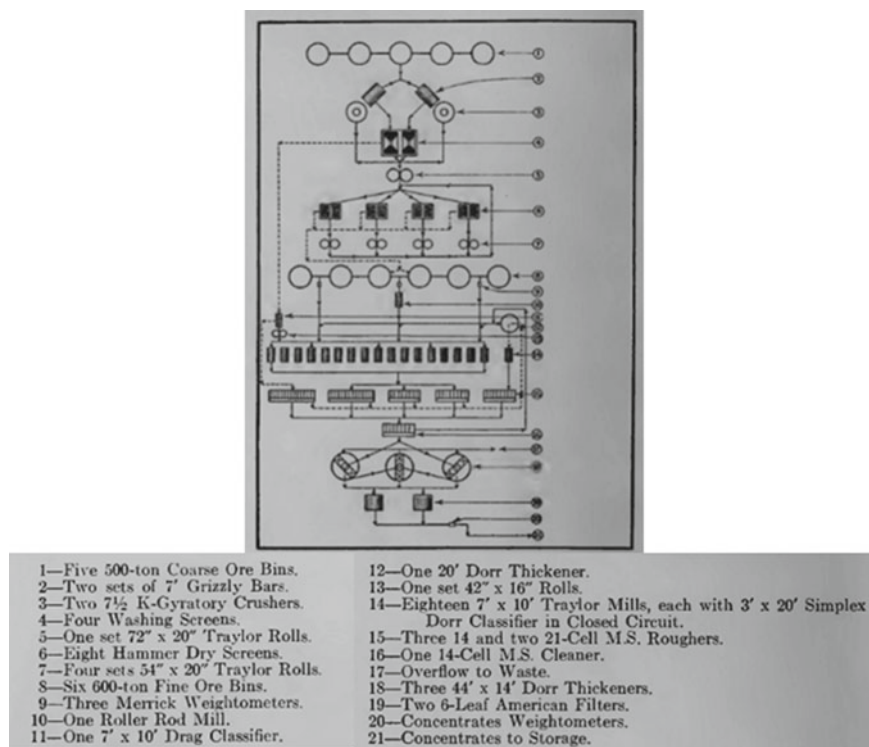


Fig. 8 Schematic diagram of concentrator mill [6]

### 7.3 Hydroelectric Power

At the turn of the last century, most mining operations used coal to generate electricity but at the Britannia Mine, hydroelectric power was used. Browning [9] noted that “the needs for the whole property are 4500 hp (3360 kW)”. By 1922, powerhouses at the Townsite and the Beach generated hydroelectricity using water from six storage dams on Britannia and Furry Creeks [4]. The capacity was sufficient during the wet season but the water supply was insufficient for dry seasons, so in 1923 a 32,000-V transmission line from the BC Electric Railway Company was constructed from North Vancouver along the coast, Fig. 9. The highway between these communities was completed in 1958 [1].

**Fig. 9** Transmission lines from North Vancouver [12]



## 7.4 Copper Precipitation Plants

Rain and snow falling at the open-pit “glory holes” at the top of Britannia Mountain seeped through fractured rock and became copper-bearing. Mining using the shrinkage system exacerbated the problem. Oxidation of sulphides produced metal-bearing acid-mine drainage (see Sect. 9) in zones up to 60 m deep [4]. The water then drained from the mine, principally along the 2200 and 4100 horizontal shafts.

A “launders” copper-extraction system was installed in 1925 to extract copper from the mine’s waters. The mine water was directed to flow over wooden “launders”—box-like troughs containing iron and tin can scraps on gratings. A galvanic reaction occurred, where the iron dissolved in the water and the copper precipitated out, accumulating as a sludge at the bottom of the launders. The sludge was subsequently removed, dried, packaged and sent to the smelter [4].

The total length of the launders in the first plant was 273 m, and a plant constructed in 1956 at the Beach had five launders, each 91.4 m long. “The precipitation plants required only a few men per shift and little equipment to operate them, so the cost/lb of copper recovered was considerably less than for copper recovered from ore” [4]. The copper precipitation plants generated revenue and reduced the hazardous effects of Acid Rock Drainage (see Sect. 9) by removing the dissolved copper from the water. The Beach launders continued to operate until 1979 [1], well after the mine closed.

## 7.5 *Other Innovations*

Other significant technological innovations adopted to control costs and optimize construction are as follows [1]:

- Old rails were recycled to make grinding balls for the mill. After proving successful, this procedure was soon adopted in other mines.
- IBM punch cards were used in their time-keeping system as early as 1929.

## 8 **Life at the Mine**

The Britannia communities at the Beach and Townsite were true company towns, constructed on company-owned land. Either one worked for the company or for a service, such as a bank, that operated with the company's permission. The company provided schools, free medical and dental care, well-stocked stores and recreational facilities at both the Beach and Townsite communities. The Beach store offered a full line of dry goods that included "fashionable clothing, furniture, and silver and china table settings" [4]. By 1916, the Tunnel Camp Hospital contained "an operating room, a three-bed ward, a dispensary and nurses quarters" [4]. The Britannia Mining and Smelting Co. Ltd. wanted to provide good amenities for its workers to try to reduce the high turnover and associated training costs—though perhaps the workers didn't always see it that way [13].

For the first 45 years of operation, the two communities were isolated from the outside world and from each other. For example, in May 1912, the Terminal Steam Navigation Co. Ltd. announced that the S.S. Britannia would commence daily service for passengers, baggage and small freight between Vancouver, Britannia, Newport and Squamish [14]. Although the power transmission line from North Vancouver was completed in 1923, the road from Squamish was completed in 1949, the railway from Vancouver in 1956, and the road from Vancouver in 1957.

The trip on electric railway, the incline, and the 347-step climb from the Beach to the bottom of the incline, Fig. 3, took 45-min one-way. It was the only means of travel between the Beach and Townsite until a connecting road was constructed in 1952. The high school was at the Townsite, so students from the Beach made the trip daily. "The Incline and the Skip (electric railway) were integral components of life at Britannia, and although the ride was a nuisance to some, it was an experience that was remembered" [4].

## 9 Acid Rock Drainage: The Mine's Environmental Legacy

The ore extraction and processing work at Britannia Mine has come, however, with a significant environmental cost. The site has been characterised by the [15] as “one of the largest metal pollution sources in North America—having significant impact on local waterways like Howe Sound and the Squamish River”. The cause of this pollution is a hazardous side effect of mining activities known as Acid Rock Drainage (ARD) or Acid Mine Drainage (AMD).

Acid Rock Drainage (ARD) occurs naturally when metal sulfides are exposed to the combined action of water and oxygen from rain and air. Producers of ARD in Britannia include pyrite (iron sulfide), chalcopyrite (copper ore), galena (lead) and sphalerite (zinc). A chemical reaction ensues, producing sulfuric acid that dissolves metals. Both the acid and the dissolved metals percolate through the ground and surface water systems, polluting water resources and the local environment. The tunnelling and processing activities at Britannia Mine have exacerbated the ARD problem, to the extent that the Province of British Columbia was required to intervene. Design of remedial measures commenced in 2003, and the Britannia Mine Remediation Project will continue to 2026 with a total budget of \$99.3 M [16]. Since 2006, a water treatment plant built and operated by EPCOR has used lime slurry to increase the pH of the ARD from 3.8 to 9.3 before releasing it to Britannia Creek [4].

## 10 Historical Significance

On November 20, 1987, Parks Canada officially recognized the Britannia Mines Concentrator (Mill No. 3) as a National Historic Site of Canada [17]. The innovative architecture and process design of the gravity-fed concentrator, and its geographic location, among other factors, were key factors in the recognition. This structure incorporated new, at the time, milling and processing techniques, particularly the bulk froth flotation process, advancing the business to process 2500 tons of ore per day. This had made Britannia the largest copper ore concentrate producer in the British Empire between 1925 and 1930.

## 11 Conclusions

Over its 70 year operating history, Britannia Mine has been a unique Canadian feat not only because it was the biggest copper ore producing mine in the British Empire in the years 1925–1930, but also for its many technical innovations. This includes the froth flotation technique to separate copper from the ore, which was essential financially given the relatively low-grade ore the mine produced. Other innovations include its

early application of hydroelectric power generation, its launders to precipitate copper from the mines water, and the use of IBM punch cards for time keeping in 1929. All of these made Britannia an authentic part of Canadian civil and mining engineering history.

Several tragedies occurred, however, that were often marked by significant loss of life, due as much to precarious location of the communities that supported the mine and concentrator operations as to the risk inherent in carrying out those operations. The members of these communities proved their perseverance and mettle. The mine is also recognized as a very serious source of metal pollution, and expensive steps to mitigate this will have to be in place for quite some time.

In 1987, Parks Canada recognized the history and significance of Britannia Mine by officially recognizing Mill No. 3 as a National Historic Site of Canada, thereby giving duly deserved credit to the ingenuity, innovations, perseverance and rich history of the mine, concentrator, and the people who operated them.

**Acknowledgements** Financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) to the second author is gratefully acknowledged.

## References

1. Britanniaminemuseum.ca (2022) The A–Z of Britannia mines. <https://cdn.shopify.com/s/files/1/0084/3851/3782/files/a-to-z-britannia-low-res.pdf?8223101623668257082>. Accessed 17 Feb 2022
2. Wikipedia.com (2015) Britannia beach. [https://en.wikipedia.org/wiki/Britannia\\_Beach\\_](https://en.wikipedia.org/wiki/Britannia_Beach_). Accessed 17 Feb 2022
3. Vancouver Province (1935) Dr. A. A. Forbes, 85, pioneer physician, passes. 04 December, p 15
4. Smitheringale WG (2011) Great mining camps of Canada 5. Britannia Mines, British Columbia. <https://journals.lib.unb.ca/index.php/gc/article/view/18783/20600>. Accessed 07 March 2022
5. Montgomery WB (1970) A brief history of the Britannia mine and its several communities. *West Miner* 5:33–37
6. Wheatley EA (1926) Inspection of Britannia mine. *Trans Eng Instit Canada* 9(11):490
7. Vancouver Province (1915) Village is buried under rocks and snow. 23 March, p 10.
8. Vancouver Sun (1935) Britannia disaster fourteen years ago. 23 November, p 27
9. Browning CP (1927) Britannia mine. *Trans Eng Instit Canada* 10(5):368
10. Vancouver Sun (1924) Scenes at Britannia mine. 19 June, p 8
11. Stewart RM (1968) Britannia operations: anaconda mining research department. Unpublished report, p 24
12. Buchan PH (1923) Vancouver branch trip to Britannia mines. *Trans Eng Instit Canada* 5(8):372
13. Rollwagon K (2006) “That touch of paternalism”: cultivating community in the company town of Britannia beach 1920–1958. *BC Stud* 151:39–67
14. Vancouver Sun (1912) Terminal Steam Navigation Co. Ltd. May 15, p 12
15. Province of British Columbia (2021) Britannia mine. <https://www2.gov.bc.ca/gov/content/environment/air-land-water/site-remediation/remediation-project-profiles/britannia-mine>. Accessed 07 March 2022

16. Azevedo B, O'Hara G (2007) A review of the management of the Britannia mine remediation project: counting the pennies. British Columbia Mine Reclamation Symposium, University of British Columbia. <https://open.library.ubc.ca/soa/cIRcle/collections/59367/items/1.0042516>. Accessed 07 March 2022
17. Historicplaces.ca (2022) Britannia mines concentrator national historic site of Canada. <https://www.historicplaces.ca/en/rep-reg/place-lieu.aspx?id=7561&pid=0>. Accessed 18 Feb 2022



# A Brief Historical Review of the Defunct La Colle Falls Hydropower Project Near Prince Albert, Saskatchewan



Jim Kells and Van Pul Paul

**Abstract** In the era pre-1950, electrical energy in Saskatchewan was produced locally by municipalities that could afford to install electrical energy generation systems. In 1906, the City of Prince Albert, Saskatchewan gave consideration to developing a hydropower project on the North Saskatchewan River at a site located some 45 km downstream from the City. The site is known as La Colle Falls. The hope at the time was to encourage industry to develop in the City. The La Colle Falls project was conceived as a run-of-river hydropower project to be comprised of a low-height Ambursen style dam, a hydropower plant located a short distance downstream on a power canal, and a navigation lock to permit the passage of sternwheeler vessels on the river. Initial design work on the project started in 1909, with site investigations beginning in 1911. Construction commenced in 1912, but shortly thereafter the City ran into issues with financing for the project and work was stopped in August 1913. Although there were some attempts to get it going again the following year, the breakout of World War I essentially prevented any further progress. By this time, the City was almost bankrupt, but ultimately it was able to pay off the accumulated debt over the following 50 years. Today, the partially-completed works stand as a reminder of the often tenuous nature of engineering projects. In this paper, an overview is provided of the La Colle Falls project and of some of the key players involved with getting the project underway. Included are comments on the technical aspects of the project, such as the project location and style of dam, brief discussion of the several critical decisions that impacted the project's success, and a brief summary of the project as it remains today.

**Keyword** Defunct La Colle Falls hydropower project

---

J. Kells (✉)

Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, SK, Canada

e-mail: [Jim.Kells@usask.ca](mailto:Jim.Kells@usask.ca)

V. P. Paul

BSc Surveying, Saskatoon, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_51](https://doi.org/10.1007/978-3-031-34593-7_51)

## 1 Introduction

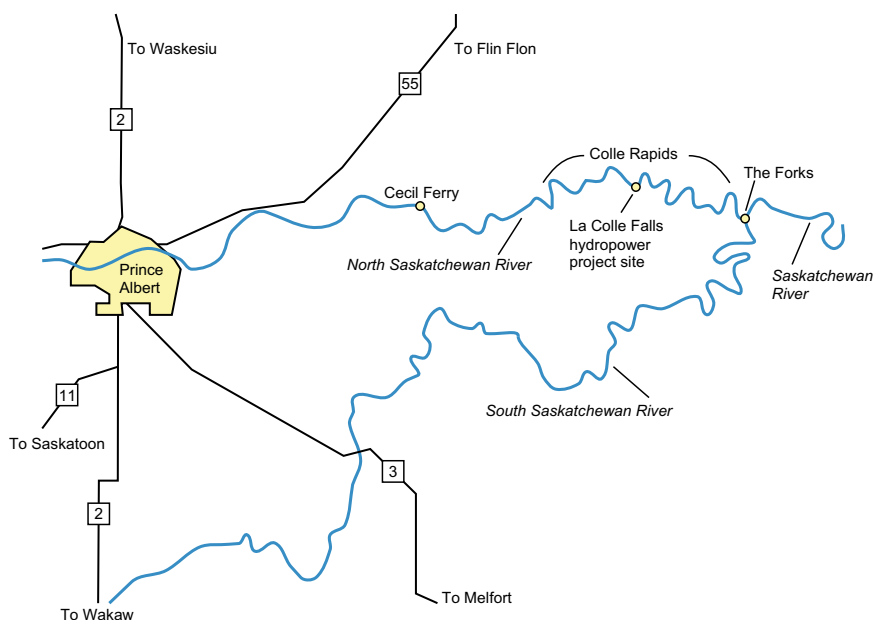
### *1.1 Historical Context*

Prince Albert, Saskatchewan is a small Prairie city that was founded on the banks of the North Saskatchewan River in 1866. Prior to 1905, the City was a major voyageur route sandwiched between what was rapidly being settled as the Grain Belt to the south and the Boreal Forest with its timber riches to the north. It was the provisional capital of the District of Saskatchewan. In the early 1900s, its political leaders and business elite vied for economic supremacy in the Prairies, rivaling Edmonton and Calgary. This vision was turned upside down in 1905, when the Dominion Government made a decision that changed the territorial landscape of Western Canada. Regina became the capital of the “redesigned” Saskatchewan province, Saskatoon got a university and Prince Albert was left with just a federal prison! The railway companies delivered the ultimate blow by routing their transcontinental railroads through Regina and Saskatoon, thus leaving Prince Albert as only a lonely railway terminus. However, the City’s power brokers continued to believe in the future of Prince Albert. If the railway was only going to be a secondary dead end in the system, the river could nonetheless provide a future for the transport of grain and timber to faraway lands. In order to develop the industrial potential of the City, the boosters intended to adopt the new power source conquering the world at the time: electricity. With the river on its doorstep, they turned their thoughts to the production of cheap hydroelectricity, thus obviating the need for importing expensive coal. And thus was born the La Colle Falls hydropower project and the coining of the expression White Coal City in reference to the type of power to be provided to Prince Albert.

### *1.2 Geographical Setting*

The La Colle Falls hydropower project is located on the North Saskatchewan River at a site some 45 km downstream from Prince Albert and just over 15 km upstream of the confluence of the North and South Saskatchewan rivers, a site known as The Forks (Fig. 1). It is founded within an approximately 20 km long reach of the river comprising some 20 sets of rapids, with the first set of rapids being located about 38 km downstream from Prince Albert. The rapids were non-navigable by the sternwheeler vessels that plied the river at that time, and thus river travel between Edmonton, Alberta and Grand Rapids, Manitoba at the lower end of the Saskatchewan River (entry to Lake Winnipeg) was not possible.

La Colle Falls was apparently named after fur trader John Cole, although there are various versions as to how the name was obtained. Cole reportedly built a post near the rapids in 1776, although the post is thought to have never been occupied. Cole had a history of mistreating the local Indigenous people and was killed in a skirmish in 1779.



**Fig. 1** Location plan showing Prince Albert and the reach of the North Saskatchewan River comprising Colle Rapids and the site of the La Colle Falls hydropower project

### 1.3 Project Initiation

In the summer of 1906, a young, civil engineer from Toronto named Charles Mitchell was retained to provide an assessment of the power potential of the rivers in the Prince Albert district. Mitchell had been involved with a number of hydropower projects, including on the Bow River system in Alberta and the large Niagara power project in Ontario. Earlier that year, he had also investigated a hydropower project for the City of Saskatoon, but that concept was rejected because of being too expensive for the City's financial resources. As part of his work, Mitchell identified La Colle Falls as being a potentially viable site for a run-of-the-river hydropower project. The City leaders were taken with Mitchell's enthusiasm, and in 1909 they asked him to elaborate on his plans for the project.

Mitchell's concept was for a two-stage project to best align with the City's growing electrical power needs. The basic concept involved a solid low-height dam, which was amenable to being raised as the demand for power grew, that would create additional head and divert some of the river flow into a power canal leading to a powerhouse located a short distance downstream. After long deliberations with much consideration of the financial implications, the City Council agreed to Mitchell's proposal. Charles and his partner brother, Percival, an electrical engineer, then proceeded to develop the detailed design. (Hereafter, unless otherwise indicated, all reference to Mitchell is to Charles).

At the time of first reaching out to the Dominion government for assistance with the project, the City leaders encountered a potential major technical obstacle. On the drafting tables of Public Works engineers lay the embryo of a navigable Saskatchewan River from Edmonton to Winnipeg, and perhaps even down to the Nelson River to a soon-to-be-built port on the shores of Hudson Bay. This system would connect the heart of the wheat-growing Canadian Prairies to the lucrative markets of Europe. In some respects, it was a Canadian version of what was happening along the Mississippi River in the USA at the time. The (expensive) compromise for the City of Prince Albert was to have a navigation lock built to bypass the hydropower dam.

By spring 1911, the Mitchells had their blueprints ready. The major change from their original proposal was that the navigation lock was moved from adjacent to the dam to adjacent to the powerhouse along the power canal. With that change, shipping could occur via the power canal, which would circumvent the withdrawal section of the river at the dam and thereby concentrate all major technical equipment in one location. (Ultimately, the navigation lock was returned to its original location prior to the commencement of project construction). City Council moved quickly to advance the project, and by July tenders were in at City Hall. However, the results were not encouraging. Local contractors did not believe that Prince Albert would be able to finance and complete the project on budget. In response, Council engaged the Toronto engineering firm of Smith, Kerry & Chace (SKC) to review the Mitchell plans.

The SKC consultants took the case to heart. Chace, and shortly afterwards Smith, visited the site and later recommended a number of changes on top of a detailed study of the local market for power and its expected evolution. Even more importantly, Smith recommended the adoption of an Ambursen type of dam constructed immediately to its final height instead of the staged approach suggested by Mitchell. While this approach would speed up construction, it also increased the total cost. The SKC dam would have had a height of 8.23 m versus Mitchell's proposed dam height of 6.25 m at its second and final stage. In turn, the higher dam would allow the use of vertical shaft turbines rather than horizontal ones at the powerhouse. But it all came at an increased cost: \$775,000 versus \$429,000 for Mitchell's first-stage original estimate and \$654,000 for his two-stage approach. Another reason for the cost increase was SKC's revelation of a large flood, with a discharge of  $5154 \text{ m}^3/\text{s}$ , having been recorded at Edmonton in August 1899. In turn, SKC recommended that the dam be designed to withstand a flood of  $7000 \text{ m}^3/\text{s}$ , accounting for downstream tributary inflows, which was more than double the design flood magnitude of  $3100 \text{ m}^3/\text{s}$  used by Mitchell.

### ***1.4 River Hydrology and Hydraulics: A Hydropower Perspective***

One of the significant matters for any hydropower project is good data on the river hydrology. In the case of the North Saskatchewan River in the early part of the last century, however, there were scarcely any data available with which to undertake an analysis. A key issue was the low flow during the winter period, which was also an important time for the use of electricity. In his initial analysis, Mitchell based his work on a discharge measurement of 317 m<sup>3</sup>/s that he made at La Colle Falls in September 1909. From this, he somehow deduced a likely low winter discharge of 158 m<sup>3</sup>/s, which he determined could be even as low as 130 m<sup>3</sup>/s in some years. It was only later that he learned of a lower winter discharge of 42.5 m<sup>3</sup>/s, but which was reported to be suspect because of the difficulty in making a good measurement under ice cover conditions.

It is assumed that part of the issue with Mitchell's thinking was based on a typical runoff analysis between locations along a river being directly related to the drainage area ratio between the stations, in this case between an upstream station (Edmonton?) and that at La Colle Falls. From this, Mitchell inferred that the discharge would be much higher at La Colle Falls than that obtained at the upstream measurement station. However, it would seem that Mitchell did not appreciate the nature of the drainage catchment. In the case of the North Saskatchewan River, the majority of the river flow is the result of runoff from the mountain and foothill region southwest of Edmonton, with a relatively small amount of the river flow being the result of runoff from the vast area of drainage basin downstream from Edmonton. Needless to say, given that power is directly related to discharge, Mitchell's power estimate during the low flow winter period was about three to four times more than was likely to be experienced.

Another matter that may have been overlooked is that of the power potential over the range of river discharges and associated heads that would be experienced at La Colle Falls. Little on this can be found in the archives. The only indication of such a need for this type of information is that reportedly indicated by H.C. Beatty, then the secretary of the Prince Albert Board of Trade. The power equation includes the product of discharge and head, and because of the nature of head decreasing as discharge increases the power produced will vary accordingly. Indeed, there is an optimal point for power production, which is somewhere between a low discharge and high head condition and a high discharge and low head condition. Determining the average annual power production capability thus requires an estimate of the average annual hydrograph and related determination of the corresponding variation in head. Based on the lack of evidence to this effect, it would appear that this may not have been taken into account for the La Colle Falls project.

## 2 Project Development

### 2.1 Overview of the Site Layout

At the project location, the valley wall on the south side is relatively steep and high, while on the north side of the river is the floodplain of the meandering channel. The eventual project plan was for a dam just upstream of Rapids #9, with a power canal conveying water to a powerhouse located near the lower end of this set of rapids. After a lot of design revisions, the final design concept for the site comprised five main elements: overflow dam (Ambursen), navigation lock, diversion headgate, power canal, and powerhouse. The general arrangements of the elements are shown schematically in Fig. 2. As indicated, only the navigation lock and about 40% of the dam were substantially completed and they remain the key artifacts from the project to this day. There is some evidence remaining of the diversion headgate foundations and powerhouse excavations as well as the excavated channel for the power canal including the tailrace.

With the dam raising the upstream water level by 6.70 m, a total head of about 8.50 m was to be available for power production at the powerhouse. The diversion headgate would be used to control the flow from the river into the power canal. Of course, as described above, the actual head at any given time would vary with the discharge in the river, with the available head decreasing as the discharge increased. Moreover, during the winter periods, aside from the relatively low discharge in the river, the tailwater elevation would be increased slightly due to the hydraulic resistance offered by the river ice cover, thus serving to slightly reduce the head available for power production at that time.

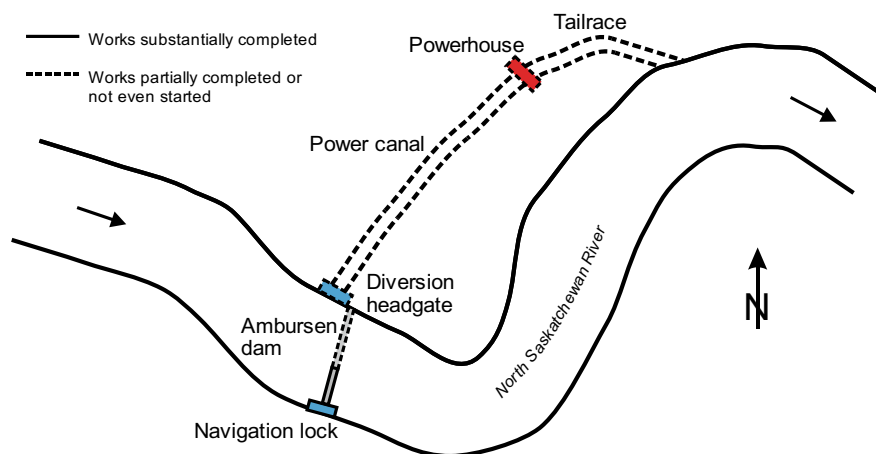
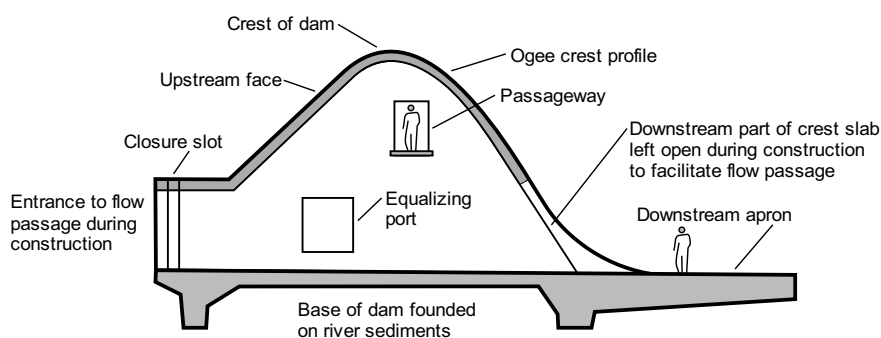


Fig. 2 Site layout for the La Colle Falls hydropower development

## 2.2 The Ambursen Dam

The Ambursen style of dam is classified as a hollow dam insofar as the interior space is either empty or partially filled with gravel ballast to improve sliding stability. It was developed by Nils Andersen, a Norwegian-American civil engineer and inventor, who founded the Ambursen Hydraulic Construction Company in 1903. The dam was used in many installations in the early part of the twentieth century. This style of dam was made possible after the development of steel reinforced concrete. The dam required a significantly lesser amount of material than a traditional gravity dam, thus making it more cost effective. For the La Colle Falls project, the dam was constructed on a concrete base slab on which was positioned a series of base panels to which cast-in-place concrete buttresses spaced at 4.5 m on centerline were founded. In turn, the buttresses supported a relatively thin reinforced concrete slab having a 45° upstream face, an ogee-shaped crest and a concave up downstream face. A schematic illustration of the dam in cross section is shown in Fig. 3. Ports were left open on the upstream side to allow the river flow to pass through the dam while it was under construction, and similarly the concrete slab was left unfinished on the downstream side so as to provide for the throughflow. A photo showing the upstream ports is given in Fig. 4. When completed, the intention was to close off the upstream ports and complete the slab on the downstream face and thus force all of the river flow to pass over the crest of the dam. Of course, as history played out, the dam was never finished and the ports were never closed.

The dam construction was complicated when the entire site was flooded in April 1913. Construction was halted for two weeks. Later, on June 21, the cofferdam protecting the in-river work disintegrated completely. Fortunately, by this time, 20 buttress base panels had been completed, which allowed work to continue by using a temporary wooden floor located just above the flood level. In the large flood of April 1915, which was almost two years after the dam construction had been halted



**Fig. 3** Schematic illustration of the Ambursen type dam as designed for the La Colle Falls hydropower project



**Fig. 4** Oblique view of the upstream side of the dam at La Colle Falls (taken November 2021). Note the outer end of the incomplete dam on the left side of the photograph and the throughflow port openings along the front face of the dam. The port openings had slots on either side so as to facilitate the insertion of stop logs once the dam was to be made operational

in August 2013, the outer two buttresses of the dam and concrete capping were undermined and collapsed into the river.

### **2.3 Navigation Lock**

With the construction of the dam by the City, several sets of the upstream rapids in the river would disappear and consequently navigation would be facilitated in that reach, while in the downstream reach the removal of boulders and some river training would address the navigation issues there. The inclusion of a lock in the dam, of course, would facilitate vessel passage around the dam. This enhancement to navigation in the river was never acknowledged by the Dominion government. However, the inclusion of a lock resulted in considerable impact on the project cost and schedule, and ultimately became a millstone around the City's neck. Due to vacillations of the Dominion government, the size of the lock was changed three times over the course of the project design, including during construction, with the size of the lock chamber increasing with each change in order to accommodate what was judged to be a more efficient vessel size on the river. As initially proposed, the lock dimensions were 36.5 m long by 7.6 m wide, which was later increased. The as-built lock dimensions are 50.3 m long by 12.2 m wide. However, the Dominion government would never provide a written guarantee to pay for the cost of this complex concrete structure, the funding of which was essential for the project's financial success. The lock as seen today is shown in Fig. 5.

Instead of waiting for a decision from the Dominion government, the City decided to press ahead with the lock construction. Indeed, it was the first major structure





**Fig. 5** Photograph of the lock at La Colle Falls viewing downstream (taken November 2021). The upper miter gate doors would close against the V-shaped concrete step in the foreground. The lock wall on the right was never brought up to full height. The guard rail located along this wall of the lock has simply been added as a safety precaution at the now-abandoned site

constructed for the project as the overall construction was to proceed from the south side of the river to the north. Fortunately, the flood of April 1913 did not affect the lock construction as it was located on what was still dry land. However, the massive spring flood of 1915 rushed through the lock chamber and took with it two meters of soil from the upper ship entrance.

When the work was halted in August 1913, the concrete walls of the lock were only partially completed. Some of the mechanical equipment, such as the metal lock doors, had already been ordered. Based on a recent inspection of the project, it appears obvious that shortcuts had been taken during construction. As an example, the various keys between concrete sections had not been properly secured with tar-soaked rope for water tightness and the consistency in the mixing and curing of the concrete was evidently an issue. Such important details were presumably offered on the altar of speed to finish the project. This leads to two questions: were the dam engineers aware of the intricacies of lock construction, and how long would the structure have withstood the rigors of repeated filling and emptying of the lock chamber?

As a final comment related to the lock, it may be noted in Fig. 2 that the dam across the river was oriented with a slight skew to the river flow (i.e., north end is slightly upstream of the south end). The angled orientation was intended to provide for better approach conditions to the lock chamber for vessels moving upstream on the river.

### 3 Project Review

A month after construction had been halted, Mitchell presented a detailed report outlining his reasons for what had become massive cost overruns. The result turned into a municipal inquiry. Among other, the inquiry found blame for sloppy oversight of the delivery chain, especially coal, between Prince Albert and the construction site. It was determined that the amount of coal leaving Prince Albert was not the same amount that arrived on site. Suspected losses included spillage and possibly theft, although there was no conclusive evidence on either account. Meanwhile, a general sentiment grew in the City to scapegoat Mitchell for the failure of the development. In retrospect, what seems odd is that no one focused on the key matter of the City not proceeding with Mitchell's 1909 proposal for a staged project development, which would have avoided overburdening the City with debt.

After the project had been halted in August 1913, the J.G. White Engineering Corporation of New York was brought in to assess the situation. In their report, issued on January 30, 1914, several significant issues were raised, including that of the underestimated low discharge that would likely occur during the winter months, the higher costs of project financing, the over-estimated likelihood that the added electrical energy production capabilities would attract industrial development as anticipated, and so on. Their recommendation was to continue with steam-generated electrical power until such time as the additional demand was realized so as to justify the cost of the La Colle Falls project.

Over the years following, there were a couple of attempts to revive the development through private investors. Initially, the First World War intervened, driving British investors away, after which the aftermath of the war conflict hardened the conditions for investment in the project. Eventually, due to advancing technology, the blueprints had become stale. It would be more attractive to build a new, larger hydropower plant at a more suitable location on the river. In due course, the project that did develop is now known as the E.B. Campbell power station, which has been in operation since 1963. It is located some 234 km downstream from Prince Albert.

### 4 Project Financing: The City's Long-Term Debt and Related Impacts

Throughout the duration of the project, the matter of project costs and the raising of funds to cover those costs was a major issue. What started out as a project estimated to cost \$429,000 eventually resulted in a debt to the City of more than \$3 million (equivalent to about \$350 million today). The City ultimately managed to avoid bankruptcy by renegotiating the debt burden to an acceptable level with the bond holders. Nonetheless, as things played out, it took the City more than 50 years to pay off the accumulated debt, with the last debentures finally burned on December 31,

1965. The impact of the debt burden was significant for the City. For much of the 50-year repayment period, the City had to forego a large number of typical municipal improvements, including street paving, sidewalk installation, and the erection of street signs.

Much of the project funding was raised through debentures that were sold to investors through London, UK investment houses. In many cases, the debentures were sold at a discount and for relatively high rates of annual interest in order to attract sufficient funds to keep the project in operation. On one of the fundraising trips to London, the \$100 debentures were priced at about \$80 with an associated annual interest rate of 4.5%. Clearly, the investor community was concerned with the risk of the project, which eventually was born out in an adverse way. Soon after, in order to avert future financial calamities by Saskatchewan municipalities, the provincial government set up a municipal oversight board. Realizing that public investment in power production was too large of an undertaking for most municipalities on their own, the Saskatchewan Power Commission was established, which later became SaskPower. Unlike Alberta, in Saskatchewan there had always been the general perception that (electrical) power production should be a public undertaking, not a private enterprise.

## **5 Key Players: A Bit of Personal History**

There were a number of key players in the La Colle Falls project, some already mentioned. Here, a small amount of additional commentary is provided for a few of them.

As mentioned earlier, Brothers Charles and Percival Mitchell were the designers of the project. Percival never visited the site as he was only responsible for the design of the electrical equipment. Charles, on the other hand, was the main architect of the project. He was a civil engineer from Toronto (U. of T. graduate), had served as city engineer for Niagara Falls, and had spent four years working on the Niagara power project for the Ontario Power Company as well as on hydropower projects on the Bow River in Alberta. These attributes, along with some of his other related experiences, made him attractive to City council. Mitchell was first retained by the City in 1906 to provide a feasibility assessment of hydropower development on rivers in the Prince Albert region, which later included La Colle Falls. However, it was not until April 1909 that La Colle Falls got any traction. After first completing a feasibility assessment, Mitchell was thereafter retained to develop detailed engineering plans for the project. Following the failure of the La Colle Falls project in late 1913, Charles enlisted in the World War I effort, where he served as Chief of Intelligence, with intelligence responsibilities related to topographical work describing the conditions in front of the British line at Ypres. After the war, he obtained a position as the Dean of Engineering at the University of Toronto, where he served for a lengthy period of time from 1919 to 1941. He passed away just a few months after stepping down as Dean.

Frank Creighton was one of three Canadian engineers in the United States-Canadian commission planning the construction of the St. Lawrence Seaway. He was hired by the City as city engineer in about 1906, and later was made City Commissioner. Creighton's hiring was regarded by some City officials as not a good choice. However, in due course, the City appointed him to be the supervising general manager for the La Colle Falls project. Creighton was reported to be on quite friendly terms with Mitchell, which allegedly prejudiced his views of any criticisms leveled at Mitchell at various times throughout the project development. He left Prince Albert after the project was shut down, moved his family to Winnipeg, and then proceeded to enlist with the Canadian Expeditionary Force in 1914. Creighton was sent overseas where he saw battle in World War I. He was posted to the Western Front and was seriously injured in the Battle of Mount Sorell in France. Creighton died of his wounds on June 16, 1916. He is buried in Belgium.

Smith, Kerry, and Chace (SKC) of Toronto was the engineering firm brought in by the City in July 1911 to review Mitchell's plans for La Colle Falls. Cecil Smith was the one that recommended that the style of dam be changed from a solid concrete structure as proposed by Mitchell to the Ambursen style of dam. He also recommended that the dam be built at once to its full height rather than adopting the staged approach suggested by Mitchell. SKC also made other recommendations that resulted in additional test drilling for the foundation conditions at various locations near the selected site, which added to the project cost. Unfortunately, Smith passed away in early July 1912, and William Chace took over for SKC. Soon after, when the work had already started at the site, Chace brought the project to a halt when he raised a flag about the design flood discharge for the project, which he deemed to be much too low.

After the design changes proposed by SKC, another eminent consulting engineer was later called in by the City, namely Isham Randolph from Chicago. Randolph was well-respected in the engineering community, in part because of his extensive involvement in the design of the Chicago Ship Canal connecting Lake Michigan to the Mississippi River. While construction was already in full swing, he too advised a number of important changes, although he was of the view that the project was viable.

The other person that deserves some mention here is H.C. Beatty, even though his connection to the project was relatively contentious. Beatty was secretary of the Prince Albert Board of Trade and editor of the newly-founded Prince Albert Herald newspaper. He has been described as being shrewd, energetic and thorough, and throughout the early stages of the project he was the lone vocal opponent of the entire undertaking, particularly with the work of Mitchell. Interestingly, however, it was Beatty that introduced Mitchell to the La Colle Falls project at Prince Albert, having met him in relation to a similar project concept being considered for Saskatoon. After doing some of his own research into hydropower developments, Beatty took strong exception to some of Mitchell's early work in developing a plan with associated costs. In turn, he accused the City council, and the mayor in particular, of not being prudent in hiring an engineer with so little relevant experience. After he raised some question pertaining to the suppression of certain information, Beatty leveled charges

of graft and perjury against the mayor and some of the aldermen, which resulted in the start of an inquiry. In short order, the inquiry was deemed a farce. Beatty left the City soon after and was never heard from again.

## 6 The La Colle Falls Project Today

After 110 years since construction was first initiated, the La Colle Falls project lies abandoned in the North Saskatchewan River. The dam remains largely intact, with only the last two sections of the Ambursen dam having fallen into the river due to undercutting by erosion as the river flow was concentrated in order to bypass the obstruction. The navigation lock remains as it was left in 1913, which is about 70% complete. While deterioration of the concrete is evident, it is in remarkably good condition given its age and the working conditions in place at the time of the project construction. Graffiti artists have left their mark in many places, most notably on the walls of the navigation lock. A photograph of the dam as it is today is shown in Fig. 6.

In 2007, Jason Hurd published his M. Arch. thesis on a potential use of the project as a suitable location for the development of a tourist attraction complete with spa. His intention was to turn what currently exists as a negative outcome for the residents of Prince Albert to one that shifts to a positive perceptive experience in the context of the project and its beautiful setting on the North Saskatchewan River. As proposed, the spa would have a multitude of amenities, including a series of pools, sauna, and steamroom. Massage rooms, meditation space, and multipurpose activity rooms, as well as a café and small gallery, rounded out the spa plan. One of the architectural



**Fig. 6** Photograph of the dam at La Colle Falls in an oblique upstream view (taken November 2021). The navigation lock is located to the left, the incomplete dam in the center, and the open river to the right

amenities was a small roof garden overlooking the lock and river. To date, nothing has arisen from Hurd's work.

The City of Prince Albert has also considered plans for how the site might be developed into a historic site, even though the memory of the failed project remains somewhat raw for the City's older citizens. In 2009, a Federal grant was applied for through the Prince Albert tourism office. The project, entitled "Saskatchewan Rivers Historical Tours", included the La Colle Falls site as one of the (educational) destinations of an eco-tourism corridor. The grant did not materialize and the project was later abandoned.

Also about 2009, a proposal was submitted to the Saskatchewan Heritage Foundation for funding of an archaeological survey of the site. As the wording in the grant application implied the excavation of historical remnants, and the proponent was not an archeologist, the application was denied. Sadly, the intention had been only for a topographical survey to identify possible sites that might be suitable for archeological excavation. The only result from this work was meetings with interested parties, a few articles in periodicals and a radio interview. It did serve, however, to keep the possibility of a future option for a historical project in the forefront.

## 7 Concluding Remarks

Today, the partially-completed La Colle Falls hydropower project stands as a reminder of the often tenuous nature of engineering projects. Of course, the lack of data available for a full analytical assessment of the proposed project played a key role in the failure that ultimately occurred. However, incomplete agreements for funding, such as for the navigation lock, inadequate oversight of the materials handling (e.g., coal shipments), and even a lack of an adequate design team with suitable experience (i.e., the Mitchell brothers) all played a role. As things played out, the project occurred on what turned out to be the cusp of major changes to the provincial transportation network, with the expansion of the railways more-or-less making inland waters navigation obsolete in this part of the world. However, that the City fathers dared to dream big for their City is commendable. And even Mitchell's efforts to bring the project to a successful conclusion are also admirable.

## Selected Bibliography

1. Abrams GWD (1965) A history of Prince Albert, Saskatchewan to 1914. Master of Arts (History), University of Saskatchewan, April, 363p
2. Abrams GWD (1976) Prince Albert: the first century 1866–1966. 2nd printing. Modern Press, Saskatoon, 389p
3. Chalupiak T (2015) Environmental protection plan: La Colle Falls hydroelectric dam and lock reclamation, NW Quarter, Section 30, T49 R22 W2M, Saskatchewan. Report prepared by North Forks Group: An Environmental Company, report prepared for Teresa Bomersbach, Instructor

and Kelly Ljunggren, Instructor, Saskatchewan Polytechnic, Palliser Campus, Moose Jaw, SK, 45p. Plus appendices

4. Hurd JJ (2007) *The conscious landscape: reinterpreting the La Colle Falls hydro dam*. Master of Architecture (Architecture) thesis, University of Waterloo, 179p
5. J.G. White Engineering Corporation (1914) *Report on the Prince Albert hydro-electric development*, New York, 30p
6. Mitchell CH, Mitchell PH (1909) *Report to the City of Prince Albert on the hydro-electric power development at La Colle Falls on the North Saskatchewan River*. Times Print, 31p
7. Van Pul P (2020) *La Colle Falls: A dam too far, hydroelectric expectations, errors and lessons*. Unpublished manuscript on the history of the La Colle Falls project, 11 chapters
8. Various newspaper articles: 1909–1915. Prince Albert Herald, Prince Albert Times, and Saskatoon Phoenix

# Biennial Update of the Activities of the CSCE National History Committee



F. Michael Bartlett

**Abstract** This short submission highlights the activities of the CSCE National History Committee (NHC) since the previous update in 2020. New National and International Historic Sites have been created, specifically the Kinsol Trestle and David Thompson's Surveying and Mapping of the Northwest of North America, respectively. Existing Historic Sites are now being regularly monitored. A major initiative has been rewriting the online descriptions of the Historic Sites using a standard template: The new descriptions together comprise roughly 55,000 words, 450 images, and 240 links to online information. Based on these new descriptions, weekly "Today in Canadian Civil Engineering History", blurbs have appeared in the CSCE eBulletin since October 2020. The NHC is also participating in the Engineering Institute of Canada's "Oral History Interviews to Preserve Canadian Engineering Achievements" initiative: in the summer of 2021, ten male and six female engineers were interviewed. The NHC has organized: tripartite webinars with the ASCE History and Heritage Committee and the ICE Panel on Historic Engineering Works; a special session commemorating the 80th anniversary of the construction of the Alaska Highway for the 2022 Whistler Conference; and a session on historic bridges for the 2022 Short and Medium Span Bridge Conference. Individuals interested in contributing to or participating with the CSCE National History Committee are warmly encouraged to contact the author.

**Keywords** CSCE National History Committee · Biennial update of the activities

---

F. M. Bartlett (✉)

Department of Civil and Environmental Engineering, Western University, London, ON, Canada  
e-mail: [f.m.bartlett@uwo.ca](mailto:f.m.bartlett@uwo.ca)

CSCE National History Committee, London, ON, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_52](https://doi.org/10.1007/978-3-031-34593-7_52)

851



## 1 Overview

The National History Committee (NHC) was created in 1983 “to increase public and professional awareness of Civil Engineers and Civil Engineering as an integral part of Canadian history, heritage and society.” In 2002, the CSCE was awarded the Pierre Berton Award from Canada’s National History Society for “helping popularize Canadian History and bringing it to a wider audience.” The committee currently consists of: 24 active members from Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia; two Section representatives; and one student member.

The Committee’s website is <https://csce.ca/en/committees/history/>.

## 2 Civil Engineering Historic Sites

Creating and monitoring CSCE Civil Engineering Historic Sites have been a “mainstay” of the NHC’s activities. To date, 78 National, International, and Regional Historic Sites have been recognized. Since 1983, it has been customary to commemorate one or more sites at each CSCE Annual Conference.

### ***2.1 New and Recent Civil Engineering Historic Sites and Landmarks***

Two new Civil Engineering Historic Site plaques should be unveiled at the CSCE 2022 Annual Conference in Whistler. Plaques for two previously identified Civil Engineering Historic Sites were unveiled as part of the proceedings of the 2021 Annual Conference, held virtually. Details of the new sites are briefly summarized in this section.

#### **2.1.1 David Thompson’s Surveying and Mapping of the Northwest of North America, 2022**

David Thompson (1770–1857)—surveyor, map-maker, explorer, and fur trader for both the Hudson’s Bay and North West Companies—was considered by J.B. Tyrrell to be “the greatest land geographer that the world has produced”, despite his serious visual impairment. Often accompanied by his Métis wife, Charlotte Small, he surveyed and mapped a vast region stretching from 45°N to 60°N latitude and from the western shores of Hudson Bay to the Pacific Ocean between 1792 and 1812. His 1814 Great Map, compiled from his surveys and those of Alexander Mackenzie,

**Fig. 1** Thompson Statue in Sandpoint, Idaho (Holly Ellis, Construction Project Manager for City of Sandpoint)



Simon Fraser, George Vancouver, and Thompson's teacher Philip Turnor, laid the groundwork for development of the Northwest of North America (Fig. 1).

David Thompson's achievements will be jointly recognized as an American Society of Civil Engineers International Civil Engineering Historic Landmark and as a CSCE International Civil Engineering Historic Site. The Minnesota and Montana ASCE Sections are co-sponsors with CSCE of this recognition. After the plaque is unveiled, it will be displayed at the Fort William Historic Park in Thunder Bay, ON, where a replica of Thompson's Great Map is on display in the Great Hall. Funding will be sought to install identical plaques at Invermere, BC, and Fort Carlton Provincial Park, SK. Gilbert and Bartlett [1] provide further details of the nomination.

### **2.1.2 Kinsol Trestle, 2022**

The massive Kinsol Trestle near Shawnigan Lake, BC, completed by the Canadian National Railway in 1920, is also recognized as a CSCE National Historic Civil Engineering Site. It is noteworthy for: (1) the scale and complexity of its original design and construction; (2) the operational and engineering challenges during its long railway service life; and (3) the innovative rehabilitation design and construction

**Fig. 2** Kinsol Trestle near the time of completion (<https://www.historicplaces.ca/en/rep-reg/image-image.aspx?id=18478#1>)



to repurpose the trestle and extend its service and heritage value on the Cowichan Valley Trail which is part of the Trans Canada Trail. Built with a seven-degree curve, it is 44 m high, 188 m long and so remains today as one of the largest and highest wooden trestle bridges in the world. The Kinsol Trestle represents an enormous feat of engineering and construction (Fig. 2).

After the plaque is unveiled, it will be displayed at the trestle. Baskin and Bartlett [2] provide further details of the nomination.

### 2.1.3 Middle Road Bridge, 2009/2021

The CSCE Board of Directors approved the designation of the Middle Road Bridge as a Civil Engineering Historic Site in 2009. Constructed in 1909, primarily to carry farming traffic, it was the first reinforced concrete arch-truss bridge in North America. The Toronto-based firm of Barber and Young designed the structure, following the principle that “mathematics and aesthetics go hand-in-hand”. James Franklin Barber (1875–1935) was a very prominent bridge designer of over 200 bridges in Ontario between 1908 and 1920. Clarence Richard Young (1869–1964) joined the Department of Civil Engineering at the University of Toronto in 1907. The builder, Octavius Laing Hicks (1852–1930), was a widely known bridge contractor around Toronto (Fig. 3).



**Fig. 3** Middle road bridge, ca. 1909 (Wikimedia, Public Domain)

The plaque is scheduled to be erected onsite in the spring of 2022. Mahmood [3] provides further details of the nomination.

#### **2.1.4 Niagara Power Generating Stations, 2005/2011**

The CSCE Board of Directors approved the designation of the Niagara Power Generating Stations as a National Civil Engineering Historic Site in 2004. The stations represent formidable historic Canadian and American milestones in the generation of hydroelectric power. In particular, the turn-of-the-last-century stations were among the first to generate alternating current (AC) power, which is more readily transmitted than direct current (DC) and has since become the worldwide standard. Unique design features deter ice from entering the conduits and penstocks and so prevent damage to the turbines. The careful consideration of the aesthetics of these installations has preserved the immense popularity of Niagara Falls as an international tourist destination [4] (Fig. 4).

We are optimistic that the plaque will be erected at the Canadian Niagara Power Company generating station, which has been recently converted into a museum, in the spring or summer of 2022. Mahmood and Bartlett [4] provide further details of the nomination.



**Fig. 4** 1913 map showing early Canadian and American installations (arcgis.com 2017)

## ***2.2 Monitoring Historic Sites and Landmarks***

Monitors have been assigned to visit and report on the condition of the existing historic plaques, and after a first cycle of such visits in 2018/2019, a second cycle has been initiated. The regional coordinators of this effort are H. Helmer-Smith, C. Sexsmith, N. Van Engelen, C. Katsanis, and B. Higgins. Two plaques have never existed (Lethbridge Viaduct, Bridges of Niagara), two have been forcibly removed (Ottawa River Canals, Albert Street Memorial Bridge), and two are, at least temporarily, lost (Grand Rapids Tramway, Hamilton Pumping Station). The Red River Floodway plaque in Winnipeg has been moved to an accessible site at Duff Roblin Provincial Park. An initiative to relocate the Victoria Bridge Plaque to a more appropriate site at Parc Jean-Drapeau in Montreal is underway.

## ***2.3 Revision of Online Descriptions of Historic Sites and Landmarks***

Descriptions of the various National Historic Sites have been revised using a standard template to achieve better consistency. The new descriptions together comprise roughly 55,000 words, 450 images, and 240 links to online information. The template fields are as follows:

1. Site name (nearest City/Town, Province/Territory).
2. Site photograph—captioned, with thin border.
3. Site location (latitude and longitude, succinct driving directions).
4. Plaque location (include photo, captioned, with thin border).
5. Plaque close up (to show all text—in case plaque needs to be replaced).



6. Description (physical description, date of construction, engineering personnel involved—100 words maximum).
7. Historic significance (succinct—100 words maximum).
8. Plaque wording (useful for web users with visual limitations).
9. Photographs of plaque unveiling ceremony (if available).
10. Link to online documentation (nomination document as PDF, other references).

## 2.4 Regular Short Article in CSCE eBulletin

The revised online descriptions of the National Historic Sites have been used to develop short “This Week in Canadian Civil Engineering History” articles for the CSCE eBulletin since October 2020. Each article consists of a title that includes the date of the featured event, an image with caption, a brief description (the target is 100 words), and, for those desiring more information, a link to the online description of the National Historic Site or other relevant webpage. Figure 5 shows an example of such an article.

## 3 New Subcommittees and Task Groups

A number of new subcommittees and task groups have been created and are currently active. Almost all National History Committee members are also affiliated with a subcommittee or task group.

**December 31, 1928:**

### **CNR President Sir Henry Thornton Announces Bessborough Hotel to be Built in Saskatoon**

When the Canadian Pacific Railway opened the Hotel Saskatchewan in downtown Regina in 1927, civic leaders in Saskatoon pressed the Canadian National Railway to build a “railway hotel” in their city. Excavation of the site, using a steam thawer and a gasoline-powered excavator, began in February 1930. Construction was completed in 1932 but the Great Depression delayed the opening until December 10, 1935. The hotel is named after His Excellency the 9<sup>th</sup> Earl of Bessborough, the 14<sup>th</sup> Governor General of Canada. The City of Saskatoon exempted the railway from property tax on the hotel for a quarter century.

[Link to Megan Hubert's History of the Bessborough Hotel 2016 M.A. Thesis](#)

**HELP!** We need more dates between December and April. Share your historic Canadian Civil Engineering dates with CSCE National History Committee Chair Mike Bartlett, [f.m.bartlett@uwo.ca](mailto:f.m.bartlett@uwo.ca)



*Bessborough Hotel under construction, 01 January 1931 (Wikimedia commons, public domain)*

**Fig. 5** Sample “this week in Canadian civil engineering history” short blurb

### ***3.1 Historic Bridge Task Group***

The Historic Bridge Task Group was formed in October 2020 and consists of ten members. Its Terms of Reference are as follows:

1. Develop criteria for listing a bridge on the CSCE National History Committee's Inventory of Historic Civil Engineering Sites.
2. Review and modify as necessary the fields used to classify listed bridges.
3. Review the bridges currently listed, in the context of these criteria, to determine whether each listing should be maintained, and indicate the current status of each listing.
4. Search federal, provincial, municipal, and other databases for additional bridges to add to the list.
5. Identify bridges that particularly warrant designation as CSCE International or National Historic Sites.
6. Disseminate findings to the National History Committee, the CSCE, and the public at large.
7. Be an information resource for the designation of appropriate historic bridges as heritage bridges.
8. Be an information resource for the reuse of historic bridges by collecting and disseminating relevant information.

Item 7 reflects the Task Group's perspective that its role is to inform stakeholders of the historic aspects of bridges but not advocate for the preservation of such structures.

There are currently 63 bridges listed on the CSCE NHC's Inventory of Historic Civil Engineering Sites (<https://cscehistory.ca/inventory-of-historic-civil-engineering-sites/>). Previously, there were no formal criteria for adding a bridge to this list, so the Task Group developed the following criteria:

1. Significant contribution to bridge engineering technology (e.g., Mosquito Creek Bridge: first prestressed concrete bridge in Canada).
2. Rare example of previously common style/structural system.
3. Significant historical events associated with a particular structure.
4. Original reason for the bridge being constructed, such as significant railway crossing (Victoria Bridge in Montreal, first railway crossing of St. Lawrence River), property development (Lions Gate Bridge), etc.
5. Other unique historical aspects (involvement by prominent architects, engineers, politicians).

The Task Group agreed that the age of the bridge should not be a criterion, but in keeping with the criteria should not include the age of the bridge but, in keeping with the criteria for nominating CSCE Civil Engineering Historic Sites, a structure should normally have been completed 50 years prior to being added to the list.

Based on these criteria, the Task Group has developed a template for classifying historic bridges using the following fields: Name; Location; Owner; Designer; Builder; Date of Construction; Structural Engineering Information; Feature Bridged;

Current Historic/Heritage Designation; Original/Current Use; Potential for Re-use; Historic Significance; and Online Links. Table 1 lists the 20 bridges that have been classified using these fields. In addition, 31 concrete bowstring trusses in Ontario have been classified, including the Freeport Bridge in Kitchener and the Main Street Bridge in Cambridge, Fig. 6a, b, respectively.

**Table 1** Historic bridges classified using template

Bridge name	Opening date	Location	Author	Note
Alexandra suspension	1926	Near Spuzzum, BC	K. Baskin	a
Hagwilget suspension	1931	Near Hazelton, BC	M. Bartlett	a
Brilliant suspension	1916	Near Castlegar, BC	M. Bartlett	a
Churn creek suspension	1913	West of 100 Mile House, BC	K. Baskin	a
Burrard street	1932	Vancouver, BC	S. Lefebvre	a
Pattullo bridge	1937	New Westminster/Surrey, BC	J. Johal	a
Okanagan lake floating	1958	Kelowna, BC	K. Backtash	a
George massey tunnel	1959	Richmond/Delta, BC	K. Backtash	a
Peace river	1960	Taylor, BC	K. Baskin	a
Kinsol (Koksilah R.) Trestle	1920	Shawnigan Lake, BC	S. Lefebvre	a
New westminster rail	1904	New Westminster/Surrey, BC	Baskin/Bartlett	a
Cisco steel arch	c. 1915	Near Lytton, BC	Baskin/Bartlett	a
Niagara canyon rail	1911 (1884)	Near Victoria, BC	K. Baskin	b
Walhachin	1911	Near Walhachin, BC	K. Baskin	
Rudy Johnson	1968 (1948)	Near Williams Lake, BC	K. Baskin	b
Sasagui Rapids	1963	District of Mystery Lake, MB	D. Ennis	
Old finch avenue bailey	1962	Rouge River, Toronto, ON	L. Newton	
L'Île-d'Orléans	1935	Québec, QC	M. Doyer	
Jolicoeur	1930	Montréal, QC	M. Doyer	
Papineau-Leblanc	1969	Montréal, QC	M. Doyer	

<sup>a</sup>Shortlisted candidate for 2022 National Historic Site Nomination

<sup>b</sup>Bridge dismantled and moved; date of original construction in parentheses





**Fig. 6** **a** Freeport bridge, Kitchener ON, 1926 (Morgan Herrell); **b** Main street bridge, Cambridge ON, 1931 (Morgan Herrell)

### 3.2 Publications Subcommittee

At the November 2021 National History Committee meeting, it was agreed in principle to create a publications subcommittee to enhance the previous informal collaboration of Mahmood and Bartlett. The draft Terms of Reference for the Publications Subcommittee are as follows:

“The CSCE NHC Publications Subcommittee shall support members of the National History Committee, and others, who are drafting publications concerning Canadian civil engineering history. These include, but are not limited to:

- ‘This Week in Canadian Civil Engineering History’ blurbs for the eBulletin, which consist of a date and title, an image with caption, and a 100-word description.
  - Papers presented at History Sessions at CSCE Annual Conferences—typically 8–10 pages long.
  - Papers submitted to the *Canadian Journal of Civil Engineering*—typically 9000 word equivalents long.
  - Articles submitted to the *Canadian Civil Engineer* magazine—typically 500–600 words with one photo or graphic or 400 words with two photos or graphics.
1. Members of the Publications Subcommittee shall be Members in good standing of the National History Committee.
  2. Anyone submitting a manuscript for possible publication in a CSCE journal, conference proceeding, magazine, or eBulletin is invited to submit a near-final draft of that manuscript to the CSCE NHC Publications Subcommittee for review and comment.
  3. The Publications Subcommittee shall review and respond to such submissions in a timely manner.”

### 3.3 Sesquicentennial Poster Task Group

As part of the celebration of Canada’s sesquicentennial in 2017, 35 posters were created with images and text depicting the history of civil engineering with an

emphasis on significant developments in Canada, an initiative conceived and led by Alan Perks. Originally displayed at Ottawa City Hall, the posters have subsequently been exhibited at the 2018 Annual Conference in Fredericton and the 2019 Conference in Montreal. The exhibit is currently being sent to be displayed at the University of Northern British Columbia, to help commemorate the creation of the new CSCE Northern British Columbia Section, and will then go to the 2022 Annual Conference at Whistler.

The Sesquicentennial Poster Task Group (A. Perks, A. MacKenzie, P. Wright) has an oversight role to ensure that the posters are maintained in good condition, that suitable venues can be found for their display, and that the themes and content of new posters are appropriate.

### ***3.4 Gordon Plewes Award Selection Committee***

The Gordon Plewes Award is given to individuals who have made particularly noteworthy contributions to the study and understanding of the history of the civil engineering in Canada, or of civil engineering achievements by Canadians elsewhere. The NHC has the responsibility to forward recommendations for recipients of this award to the CSCE Honours and Fellowships Committee. Alistair MacKenzie is the recipient of the 2021 Gordon Plewes Award.

### ***3.5 Support for the 2021 and 2022 Annual Conferences***

In addition to the historic site plaque unveilings, the following four papers were presented virtually at the 2021 Annual Conference, in a session moderated by S. Arbuckle and M. Bartlett:

1. Kells and Sexsmith [5]: “A Brief Historical Review of the Gardiner Dam and the South Saskatchewan River Project”.
2. Bouzari et al. [6]: “Civil Engineering History Overshadowed by Politics”.
3. Mahmood and Bartlett [4]: “The Niagara Power Generating Stations: a Major Milestone in the Use of Hydro-electrical Energy”.
4. Bartlett [3]: “A Brief History of the Middle Road Bridge”.

The details of the history-related activities at the 2022 Annual Conference are still being finalized, coordinated by K. Baskin. At this time, it is envisaged that they will include the unveiling of two plaques to commemorate the Historic Sites described in Sect. 2 above. There are two history sessions scheduled, an open paper session and a session organized jointly with members of the ASCE History and Heritage Committee to commemorate the 80th Anniversary of the Construction of the Alaska Highway. We are also working with the Yukon Transportation Museum in Whitehorse, who will host an online portal with information concerning the construction

and operation of the Alaska Highway. The CSCE, through the Western Region, the 2022 Conference Local Organizing Committee, and the History Program Fund, is donating \$1000 to the Yukon Transportation Museum to support creating the online portal.

## 4 Oral History Interviews to Preserve Civil Engineering Achievements

In partnership with six other Engineering Societies and the Engineering Institute of Canada (EIC), an initiative to research, record, and transcribe oral histories of eminent Canadian engineers has commenced. Each member society has agreed to pay \$750 per year for two years, and both the EIC and MITACS have matched these funds to create a total project budget of \$45,000. A Master of Arts in Public History Intern was hired in the summer of 2021, who interviewed nine male and five female engineers. One additional male has since been interviewed. The civil engineers interviewed included Susan Tighe (see Fig. 7) and Nick Isyumov (recommended by CSCE), Sarah Devereaux (recommended by the Canadian Society of Senior Engineers), Kwan Yee Lo (recommended by the Canadian Geotechnical Society), and Peter Lighthall and Denise Leahy (recommended by the Canadian Dam Association). Sufficient funds are available to hire two interns in the summer of 2022, when it is hoped to conduct 30 interviews.

It is intended that audio recordings and written transcripts of the interviews will be made available online through the Engineering Institute of Canada website. Video recordings, audio recordings, and written transcripts will also be archived in the Engineering Institute of Canada fonds at the Ontario Tech University Library. The interviews range from roughly 50–75 min in length, so it is likely that shorter one to three minute “snippets” will also be made available on the EIC website.



**Fig. 7** CSCE past President Dr. Susan Tighe, left, and interviewer Robin Marshall, right

The oral history interviews capture a diverse wealth of Canadian engineering history. The interviewees or “Narrators” describe major projects, such as Toronto’s CN Tower, the SkyDome (now the Rogers Centre), and the Red River Floodway. There are also excellent descriptions of now-superseded technologies and practices: for example, consider the following excerpt from Robin Marshall’s (“R.M.”) interview with Geotechnical Consulting Engineer Dr. Denise Leahy (“D.L.”) on August 24, 2021:

R.M.: “Throughout your career, how have you noticed that your field has changed? You’ve mentioned a few things already but I was wondering if you had any other things that you’ve noticed that have evolved?”

D.L.: “Well, of course, it really changed a lot, just due to computers. In the beginning, drafting methods were completely different. There was a room with drafting tables and they would make blueprints and the lettering was important and they have all this material to do the lettering, and they had electrical erasers so that they could do modification to a drawing. They had different colored pencils for different kinds of materials (smiles). The draftsmen were not very happy if you had too many changes to a drawing—you better make sure that you knew what you wanted, before you asked for a new drawing.”

D.L.: “It was the beginning of photocopying and that cost a lot. So, we were very spare. We didn’t spend too much paper and it was written in the proposal how many copies of the report the client would get.”

D.L.: “And then for the draftsmen AutoCAD came, and that was such a challenge for these guys, and many draftsmen retired at that point because it was too big of a step. But those that did learn that AutoCAD, the younger ones, became really some of the best draftsmen I knew. They had this better connection to the project because, when it was on paper on the drafting tables, the project engineer had to really explain to the draftsmen, how to do things. While now they can become sometimes more executive than that. Okay, you just do that and if it’s not right, oh, we’ll try something else and because it’s not seen as much time consuming and paper consuming and energy consuming. So, that was a change.”

## **5 Engagement with ASCE and ICE History Organizations**

On May 11, 2021, a tripartite online webinar was held, organized by representatives of the ASCE History and Heritage Committee, the Institution of Civil Engineers Panel on Historic Engineering Works, and the CSCE National History Committee. Various presenters briefly summarized current initiatives taken by these three organizations. There were some interesting differences but also several common themes: all three groups are, for example, trying to improve their websites and monitor the plaques commemorating their Historic Sites. The CSCE NHC has created a small Task Group (V. Ayan, N. Shrive) to facilitate future tripartite webinars.

Representatives of the ASCE History and Heritage Committee now attend monthly meeting of the CSCE National History Committee, and vice versa.

## 6 Past and Future Objectives

Approximately two years ago, [7] reiterated the following objectives for the National History Committee:

1. Development of a system that will safeguard and maintain the plaques located on the Society's Historic Sites by designating a "custodian" for each site.
2. Further development of communications through:
  - a. Enhancement of the NHC website—upgrading the details, descriptions, and images of the Historic Sites.
  - b. Creating a social network presence through LinkedIn, Facebook, Twitter, etc.
  - c. Increasing the visibility of the committee by having members interact with local media outlets.
3. Further development of relationship with University of Ontario Institute of Technology (now corporately branded "Ontario Tech University") for storage and digitization of archival material.
4. Further interactions with like programs run by fellow Engineering organizations, initially ASCE and ICE.
5. Further development of the Oral History Program.
6. Engaging CSCE Regions and particularly sections, in the activities of the NHC
7. Engaging representatives of relevant governmental and corporate organizations in NHC activities as consulting members of the NHC.

At this time, Objectives 1, 3, 4, and 5 have been met, although ongoing work is required to maintain these initiatives. Objective 2, particularly 2a, remains a work-in-progress, and although the descriptions of the Historic Sites have been markedly improved, their migration to the current CSCE web server has not been satisfactory. While the National History Committee membership continues to grow and the engagement of members in the NHC's activities continues to improve, it would be desirable to engage further sections and relevant external representatives in the committee's activities.

It seems prudent to continue to strive toward achieving these objectives.

It is also a pleasure to report the appointment of two new Vice Chairs of the National History Committee in 2021. Ali Mahmood has agreed to serve as Vice Chair (Annual Conference Activities) and Wes Wilson has agreed to serve as Vice Chair (Projects and Communications). We are also grateful to Bruce Higgins for his continuing service as the NHC Secretary.

## 7 Conclusions

The National History Committee has been committed to implementing the Society's National History Program continuously since 1982. However, we remain mindful of a further comment of founding chair W. Gordon Plewes following the May 1982

meeting that “the ultimate success of the Society’s History Program depends on the interest and action of individual members throughout the country and its organization must extend to all Regions and Sections”.

**Acknowledgements** Funding of \$1500 from the History Program Fund, matched with funding from other Canadian Engineering Societies, \$11,250 from the Engineering Institute of Canada, and \$22,500 from MITACS to support the “Oral History Interviews to Preserve Canadian Engineering Achievements” initiative is gratefully acknowledged. Total funding of \$1000 from the CSCE Western Region, the 2022 Annual Conference Organizing Committee and the History Program fund to support the creation of an online portal commemorating the 80th Anniversary of the Construction of the Alaska Highway at the Yukon Transportation Museum in Whitehorse is also gratefully acknowledged. Finally, the continued active engagement of the members of the National History Committee is warmly and sincerely appreciated.

## References

1. Gilbert DR, Bartlett FM (2022) David Thompson’s surveying and mapping of the Northwest of North America. In: Proceedings, CSCE 2022 annual conference. Springer, New York
2. Baskin KR, Bartlett FM (2022) A brief history of the Kinsol Trestle. In: Proceedings, CSCE 2022 annual conference. Springer, Whistler
3. Bartlett FM (2021) A brief history of the middle road bridge. In: Proceedings, CSCE 2021 annual conference. Springer, New York
4. Mahmood AA, Bartlett FM (2021) The Niagara power generating stations: a major milestone in the use of hydro-electrical energy. In: Proceedings, CSCE 2021 annual conference. Springer, New York
5. Kells J, Sexsmith C (2021) A brief historical review of the Gardiner dam and the South Saskatchewan river project. In: Proceedings, CSCE 2021 annual conference. Springer, New York
6. Bouzari N, Ngabire S-M, Van Engelen N (2021) Civil engineering history overshadowed by politics. In: Proceedings, CSCE 2021 annual conference. Springer, New York
7. Bartlett FM (2020) Biannual update of the activities of the CSCE National History Committee. *Can Civil Eng* 36(7):28–29

# Cariboo Wagon Road—A User's Perspective



W. C. Sexsmith

**Abstract** The Cariboo Wagon Road was built by the Royal Engineers in the 1860s. The road linked Yale, the head of navigation on the Fraser River, with Barkerville, the center of the Cariboo gold rush. The Cariboo Road was the primary link between the Cariboo gold fields and the outside world for several decades. The Author's Great Grandfather, the Reverend William V Sexsmith was the Wesleyan Methodist Minister in Barkerville from 1877 to 1881. In 1879 he traveled the length of the Cariboo Road from Barkerville to Yale and then by steamboat to Victoria to attend church meetings. He then traveled to Ontario by steam ship to San Francisco, CA, and railway to Napanee, ON. The purpose of the trip was to visit family (he was born in what is now Ontario) and more importantly to get married. He and his new bride then returned to Barkerville via the same route. Great Grandfather Sexsmith kept journals from time to time while he lived in British Columbia. One of those journals covers his 1879 trip from Barkerville to Victoria. During that trip, which took many days, he traveled by cutter, horseback, stage coach, steam boat, and on foot. Another journal covers the railway trip with his new bride from Napanee to San Francisco.

**Keyword** Cariboo Wagon Road

## 1 Introduction

Gold was first discovered in the Fraser Canyon in the mid-1850s. Initially, the Hudson Bay Company, which controlled the region, discouraged prospecting as they did not want mining to disrupt their lucrative fur trading business. They wanted to keep industry and agriculture out of British Columbia. However, they eventually began to accept locally mined gold in trade at their posts and in February 1858 sent 800 oz (22.7 kg) of raw gold to San Francisco to be minted.

---

W. C. Sexsmith (✉)

Saskatoon Water, Transportation and Utilities Department, City of Saskatoon, Saskatoon, SK, Canada

e-mail: [calvinsexsmith@sasktel.net](mailto:calvinsexsmith@sasktel.net)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_53](https://doi.org/10.1007/978-3-031-34593-7_53)

Word of the gold shipment quickly spreads, and by the end of summer, over 30,000 gold seekers had headed for the Fraser River. They quickly discovered the difficulties in traveling through the Fraser Canyon, the cost of supplies, and not as much gold as they expected. Most of the gold seekers left as quickly as they arrived and only about 10% of them remained.

Those who remained pushed further inland, moving up the tributaries of the Fraser and into the Cariboo region. Reports of gold being found in the Cariboo attracted more prospectors in 1860 and 1861, but although gold was found, it was not in great quantities. Finally, in August 1862, a major strike was made on Williams Creek near where the town of Barkerville would soon spring up, triggering the Cariboo Gold Rush.

## 2 Development of the Cariboo Wagon Road

Travel to the Cariboo was very difficult. For the most part, the waterways were not navigable and there were no roads over which wheeled vehicles could travel. This meant that travel was by foot and supplies were moved by pack animal. In 1858 a trail had been developed to Lillooet that bypassed the Fraser Canyon, and by 1860 it had been developed into a series of Wagon Roads linking navigable lakes where steam boats would take over. In 1861 engineer and contractor Gustavus Blin Wright built a 47 mile (76 km) Wagon Road from Lillooet to what would become Clinton (Fig. 1).

The route via Lillooet proved to be unsatisfactory. Between Port Douglas on the lower Fraser and Lillooet, freight had to be trans-loaded eight times between wagons and steam boats and the grades on Wright's road to Clinton were very steep. At this point in 1861, Governor Douglas of the colony of British Columbia proposed an 18 foot (5.5 m) wide Wagon Road from Yale, the head of navigation on the Fraser River, approximately 400 miles (645 km) to the Cariboo gold fields.

The Royal Engineers began to survey the route in October 1861 and the same month the first contract was let to build the Boston Bar to Lytton section. The Royal Engineers built the most challenging portions of the route through the Fraser Canyon starting at Yale. The rest of the construction was contracted out. Gustavus Wright who had built the Lillooet to Clinton road (sometimes called the old Cariboo Road) was awarded the contract from Clinton to Soda Creek, a distance of 177 miles (285 km).

By September 1863 the road was completed from Yale to Soda Creek. From Soda Creek, steam boats were in service to Quesnel, a short distance from the gold fields which were accessed by a variety of pack trails. In 1864 Gustavus Wright began construction north of Soda Creek and the road reached Barkerville in 1865. Table 1 summarizes the lengths of the various segments of the Cariboo Road.

There were many engineering and construction challenges building the Cariboo Road. Most of these were in the Fraser Canyon section constructed by the Royal Engineers. The road had to be blasted out of the rock walls of the canyon and sometimes suspended over the river on stilts and cribbing. An engineering landmark





**Table 1** Distances on the Cariboo Road

From	To	Miles	Kilometers
Yale	Spuzzum	12	19
Spuzzum	Boston Bar	12	19
Boston Bar	Lytton	29	47
Lytton	Spences Bridge	29	47
Spences Bridge	Clinton	57	92
Clinton	70 Mile House	23	37
70 Mile House	100 Mile House	30	48
100 Mile House	Alexandria	107	172
Alexandria	Quesnel	35	56
Quesnel	Barkerville	43	69
Total		377	606

### 3 Biography of the Reverend William V Sexsmith

William V Sexsmith was born near Napanee in what is now Ontario in 1840 and was one of nine children of Irish protestant immigrants Mathew and Eliza (nee Grant) Sexsmith. After completing his education, he taught school for a few years before being ordained as a Wesleyan Methodist Minister. Following his ordination, the church sent him to Vancouver Island in 1872 where he served at Maple Bay on Vancouver Island and on Salt Spring Island. In 1877 he was assigned to Barkerville in the Cariboo region. While he was living in Barkerville, he made a trip back to Ontario to marry Hannah Robinson in 1879. The couple returned to Barkerville and later relocated to Wellington on Vancouver Island in 1881. In 1884 they returned to Ontario where Reverend Sexsmith served a number of churches near the Bay of Quinte. For health reasons, they moved to Saskatoon, Saskatchewan, in 1914. William passed away in 1919 and Hannah in 1938. They are buried in the Riverside Cemetery in Napanee, Ontario.

### 4 The Voyage from Barkerville to Victoria

In March and April of 1879 Reverend Sexsmith traveled from Barkerville to Victoria to attend the British Columbia District meeting of the Wesleyan Methodist Church. After attending the district meeting, he traveled to Ontario to visit friends and family and more importantly to get married. He returned with his bride to Barkerville in August of 1879. Table 2 summarizes the various stages of his journey.

While he lived in British Columbia, Reverend Sexsmith kept journals from time to time. One of these journals covers most of his voyage from Barkerville to Victoria with daily entries over a three-week period. The journal entries cover his trip over

**Table 2** Reverend Sexsmith's trip itinerary

Date	From	To	Mode
March 25	Barkerville	Beaver Pass	Horse and Cutter
March 26	Beaver Pass	Mr. Boyd's (Cold Spring House)	Horse and Cutter
March 27	Mr. Boyd's	Quesnel	Horse and Cutter
March 28 to March 30	Stayed in Quesnel		
March 31	Quesnel	Mr. Mc Innes'	Horseback
April 1	Mr. Mc Innes'	Mr. Pinchhick's	Horseback
April 2	Mr. Pinchhick's	Blue Tent Ranch	Horseback
April 3	Blue Tent Ranch	100 Mile House	Horseback
April 4	100 Mile House	70 Mile House	Borrowed Horse
April 5	70 Mile House	Clinton	On Foot
April 6 to April 8	Stayed in Clinton		
April 9	Clinton	Lytton	Stage Coach
April 10	Lytton	Yale	Stage Coach
April 11	Yale	Sumas	Steam Boat Reliance

the entire length of the Cariboo Wagon Road from Barkerville to Yale and beyond to Sumas in the Fraser Valley approximately 90 km east of the mouth of the Fraser River. As happens from time to time in his journals, the entries abruptly end in Sumas on April 14 and do not resume until August 9 two days before his marriage to Hannah Robinson in Napanee, Ontario.

The trip from Barkerville to Yale took him 17 days, including stops of three days in each of Quesnel and Clinton making for a total of 11 days on the road for the 377 mile (606 km) trip. Both of the stop overs included a Sunday, being a minister it seems likely that he was loath to travel on the Sabbath. Today, the highway distance between Barkerville and Yale is 577 km and the trip can be made by car in about 8½ h.

Reverend Sexsmith originally planned to travel to Yale on horseback. However, circumstances caused him to use a variety of transportation modes. He was accompanied on the trip by a dog named Bronti who belonged to a Mr. Fraser. The first three days from Barkerville to Quesnel were by horse and cutter (a cutter is a small sleigh) as there was still sufficient snow. Although the snow was still deep when he departed Quesnel, he left the cutter behind and proceeded on horseback for the next four days to 100 Mile House.

At 100 Mile House, he discovered that his horse had developed saddle sores and was unable to continue. He was able to borrow two horses belonging to the stage coach company at Hundred Mile House and rode one while leading the other on the next leg of his trip to 70 Mile House the following day. Other than mentioning that he turned his injured horse loose in the valley at 100 Mile House, there is no indication if he left it in the care of the proprietor of 100 Mile House or left it to fend for itself.

Apparently his arrangement with the stage coach horses went only as far as 70 Mile House and the next day he walked the 23 miles to Clinton. He reported being very sore and tired upon arrival in Clinton. The final two days of his trip to Yale were by stage coach. He reports leaving Clinton in a new stage with six horses and changing stages at the breakfast stop in Cache Creek. It appears that he started his final day on foot again walking 10 miles to the stables where he re-joined the stage. He reports that the stage was full with eight passengers inside and two more plus the driver outside.

On April 11, the day after he arrived in Yale, he boarded the steam boat *Reliance* for the trip to Sumas where he stayed until at least April 14. Although the journal entries end on April 14, we know he continued on to Victoria, likely by steam boat, to attend the church meeting and later continued on to San Francisco, California, by steam ship and Napanee, Ontario by train. He does record that the fare to Frisco was \$20.00.

## 5 Road and Weather Conditions

Reverend Sexsmith makes almost daily comments on the weather and the state of the road. As the trip begins in early spring, the snow is beginning to melt on the northern portions of the road and is mostly gone further south. The melting snow frequently made the road very soft and difficult to travel on. He often starts early in the morning so as to travel while the road is still frozen. Further south he frequently encountered rain and muddy road conditions.

On the first leg of the trip from Barkerville to Quesnell, he generally describes the road as in fair to good condition in the morning but deteriorating as the snow softens in the afternoon. At one point, he describes the road as nearly impassible. On the second day, he stops traveling for the day at 1:00 p.m. due to road conditions. The worst weather of this leg was the first afternoon when he encountered snow and sleet; he describes that part of the trip as cold, wet, and disagreeable.

He describes the snow in Quesnell as deep with the only place clear of snow being the bridge over the Fraser River. While he was there, the mail contractor left for Barkerville at 2:00 a.m. but found the road too soft even before dawn and returned to Quesnell. The contractor had an Indian (sic) take the mail through on foot.

Road conditions improve on the next leg to Clinton. He reports that the snow is very deep for the first 20 miles (32 km) south of Quesnell but does not have difficulty. By the time he reaches Williams Lake, the road is clear of snow and relatively hard. Further south in the Lac La Hache Valley, he reports rain and a muddy road. Between 100 Mile House and 70 Mile House, he reports the road as being very bad. The last day of this leg, he is on foot and reports that the walking is good until the sun strengthens enough to melt the frost leaving the road in a horrible mess.

The final leg of the trip from Clinton to Yale was by stage coach and he does not comment on road conditions. The weather starts out cold and rainy, but for most of this leg the weather is fine showing the mountains to good advantage. He frequently comments on the scenery and for the most part is impressed.

## 6 Road Houses

As trips on the Cariboo Road took several days, many road houses were established along the route providing meals and lodging for travelers. Typically they were spaced 10–15 miles (16–24 km) apart, although some were as close as 3 miles (5 km) to each other.

The road houses were built by road contractors and by others who believed that operating a road house was easier and/or more lucrative than prospecting. Most road houses were located where there was land for pastures and gardens as they produced locally most of the food that they served. Only non-perishable staples such as flour, sugar, salt, coffee, tea, and liquor would be imported. The quality of the meals and quantity of food were usually quite good as those serving inferior meals would soon go out of business.

Reverend Sexsmith mentions several road houses including Beaver Pass, Cottonwood House, 150 Mile House, Blue Tent Ranch, 100 Mile House, 70 Mile House, and California House in Yale. The mileage by which some houses are identified is measured from Lillooet on the original route of the road, not Yale. Sometimes, he identifies his stopping places by family name. In most cases, it is not clear if these people are friends who gave him a bed for the night in their homes or the proprietors of road houses. He does clearly indicate that Beaver Pass is run by Mrs. Hyde, 100 Mile House by Mrs. Pratt, and 70 Mile House by Mr. Saul. The second night he stays at Mr. Boyd's place 5 miles (8 km) from Cottonwood House. The Boyd family owned both Cottonwood House and nearby Cold Spring House and did not relocate to Cottonwood House until 1886 so it is possible that he was at Cold Spring House. At Sumas he stays at the home of Brother Thompson who is a church member. Generally if the house was at a village, he mentions the name of the village but not the house in locations such as Quesnel, Clinton, and Lytton.

He also lists his expenses at some of the places he stayed or ate at as listed below:

- Beaver Pass \$2.00—food and lodging
- Boyd's \$1.50—food and lodging
- Quesnel \$9.00—food and lodging
- Clinton \$4.00—Lodging
- Soda Creek \$1.00—Food
- Cook's Ferry \$1.00—Food
- Lytton \$12.00—Food and lodging

- Boston Bar \$1.00—Food.

The cost at Lytton appears to be out of line with the other costs as he was only there one night, while he was three nights at each of Quesnel and Clinton. As the reference source here is a transcription of the original, it may be due to a transcribing error and might be only \$2.00 which would be in line with other costs.

For the most part, he does not comment on the quality of the accommodations with the exceptions of Lytton which he describes as wretched and the California House in Yale which he describes as first class. He further goes on to say that he has heard that the California House is the best in British Columbia.

## 7 People and Places

Reverend Sexsmith mentions many people by name in his journal. Some of these are road-house operators, while others are friends and/or members of the Methodist Church. At one point when he visits an elderly man named Mr. Price living alone on a ranch south of Quesnel, he introduces himself to Mr. Price as “The Minister of the Cariboo” implying that he is the only Methodist minister in the Cariboo. He does hold a church service while he is in Quesnel and later baptizes the daughter of a Mr. Cummings who lived south of Soda Creek. There was a Methodist Minister stationed in Clinton, and Reverend Sexsmith assists him in his duties and also visits the local school.

For the most part, he does not use people’s first names. Generally, he does this only when differentiating between people with the same surname. In addition to his church duties, he also delivers letters and messages to people living along the road.

He makes some mention of the towns and villages describing Quesnel as quiet and business is dull and Litton (sic) as wretched. He is much more impressed with the scenery of the British Columbia interior and frequently comments on its beauty. At one point, he reflects on his two years in the Cariboo stating that he likes it very much, both the scenery and the situations.

## 8 Return to Barkerville

Reverend Sexsmith starts writing his journal again on August 9. The first few entries concern his goodbyes to his family and his marriage to Hannah Robinson. The journal entries continue through their train trip from Napanee to San Francisco but then again abruptly end the day of their arrival in San Francisco and do not resume until after they have arrived in Barkerville.

Family oral history is that they traveled by steamship to Victoria, steamboat to Yale, and stage coach to Barkerville. Their son, and the author's grandfather, William P. Sexsmith did relate a story he heard from them that at one point on the return trip the stage coach went off the road forcing the passengers to walk some distance while waiting for another stage coach to pick them up.

In another journal, written in 1880 and 1881, he does talk about short trips on the road between Barkerville and Quesnel. These entries mostly describe the people he met and the duties he performed with no comment on what travel was like.

## 9 Subsequent Developments

The Cariboo Wagon Road was the main means of transportation to and from the Cariboo for many years. During the construction of the Canadian Pacific Railway, many parts of the road through the Fraser Canyon were damaged or destroyed. Following the completion of the CPR in 1886, the southern terminus of the road was moved to Ashcroft.

The railway was late to arrive in the Cariboo with the Pacific Great Eastern Railway reaching Clinton from Squamish and Lillooet in 1915 and Quesnel in 1921. Between 1924 and 1926 the province of British Columbia re-established a road through the Fraser Canyon creating a gravel highway from Hope to Prince George called the Cariboo Highway closely following the route of the Cariboo Road from Yale to Quesnel. Today, the route through the Fraser Canyon to Cache Creek is part of the Trans-Canada Highway (#1) and from Cache Creek to Quesnel the Cariboo Highway (#97). Highway #26 runs from Quesnel to Barkerville also closely following most of the route of the Cariboo Road.

# Ballantyne Pier History



Willy Yung

**Abstract** Ballantyne Pier was originally constructed between 1921 and 1923, as a major expansion to the Port of Vancouver to alleviate shortages of dock facilities, which had become sharply apparent after the First World War. The new pier also provided port access for the new Canadian National Railway, as the Vancouver shoreline had been dominated by the Canadian Pacific Railway up to this point. In addition to cargo-handling facilities, the new pier included the first grain elevator built for export of grain from Canada's West Coast. The Ballantyne Pier shed's façade was built in a three-bay brick and stone Classical Revival style, a popular style of the time, which was found in banks, schools, government offices and churches in that era. When completed in 1923, it was considered by some as one of the most technically advanced port facilities in the world. In 1935, Ballantyne Pier was the site of a clash between police and striking workers, an event known as the "Battle of Ballantyne Pier", which ultimately led to the creation of the International Longshore and Warehouse Union (ILWU). Up until the early 1990s, Ballantyne Pier served as a bulk export facility. In 1992, the pier was redeveloped into a two-berth cruise ship terminal. Years later, the water lot between the west berth and adjacent Centennial Pier (Centerm) was filled in to allow expansion of Centerm's container handling capacity. To meet continued growth in the Port of Vancouver's container sector, the Vancouver Fraser Port Authority is once again redeveloping the combined Ballantyne Pier and Centerm sites. The historic Ballantyne Pier shed façade, including a portion of the original shed, is being retained and incorporated into Centerm's new state-of-the-art container operations facility building.

**Keyword** Ballantyne Pier history

---

W. Yung (✉)  
Vancouver Fraser Port Authority, Vancouver, BC, Canada  
e-mail: [willy.yung@portvancouver.com](mailto:willy.yung@portvancouver.com)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_54](https://doi.org/10.1007/978-3-031-34593-7_54)

877



# 1 History of Ballantyne Pier

Almost 100 years ago, in the wake of the First World War, the Port of Vancouver faced a similar challenge as today: a shortage of available dock space. The solution then was to build a new pier that would enable the port to keep pace with increased shipping traffic. The Ballantyne Pier, as shown in Fig. 1, was thus born. Located on the south shore of Burrard Inlet in Vancouver, B.C., as shown in Fig. 2, Ballantyne Pier was constructed between 1921 and 1923 for freight handling and storage. It also supported the growing resource trade as the local economy recovered following the war. The facility contained four storage sheds, rail tracks, grain conveyor galleries and dockside cargo cranes.

Ballantyne Pier was named after Canadian politician, Charles Colquhoun Ballantyne (August 9, 1867–October 19, 1950). Ballantyne was a millionaire, who at one time owned Sherwin Williams Paints in Montreal. He was president of the Canadian Manufacturer’s Association and a member of the Montreal Harbour Board. He also raised and commanded the first Battalion Grenadier Guards. In October 1917, Ballantyne was appointed minister of public works, minister of marine and fisheries and minister of the naval service by Sir Robert Borden’s First World War Union Government. He became a cabinet minister prior to being elected to the Canadian House of Commons in the federal election of December 1917.

The 1920s was a decade of intense growth for the Port of Vancouver, with the development of piers and docks. With grain shipment from the Prairies steadily increasing, Vancouver became Canada’s second largest port (behind the Port of

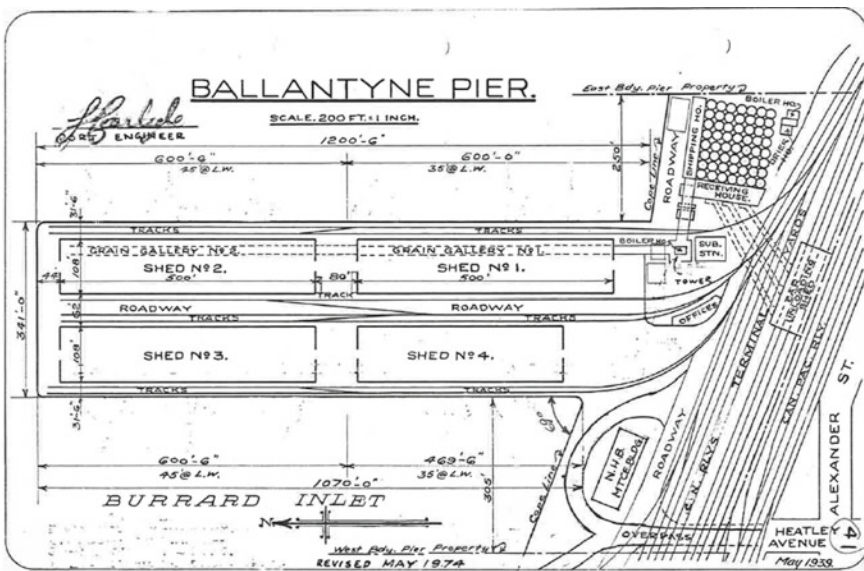


Fig. 1 Ballantyne Pier—general arrangement drawing





**Fig. 3** Police guarding the Heatley Street entrance to Ballantyne Pier, 1930s, Wikimedia



**Fig. 4** Police chasing protestors through Vancouver's East End, 1935, Wikimedia

For the next 50 years, the Ballantyne Pier played a vital role in the Port of Vancouver's economic development—becoming Canada's largest port and the third largest in North America, handling the transportation of goods between Canada and more than 170 economies. Figure 5 shows an aerial view of the pier from the 1940s.



**Fig. 5** Aerial view of Ballantyne Pier from northeast, circa 1940s, CVA Air P29.4

## **2 Original 1920s' Construction of Ballantyne Pier**

Ballantyne Pier originally consisted of a 1200-foot long by 340-foot wide (366 m long by 104 m wide) concrete pier structure with four two-story concrete storage sheds. The pier was built by the Northern Construction Company. Port of Vancouver records indicate that the cost of the project, including equipment and land, was just over \$7.1 million.

Construction involved dredging of 167,000 cubic yards (127,800 cubic meters) of the seabed and bedrock, with some of the dredged material reused as the core fill for the pier, and placement of 615,000 cubic yards (470,600 cubic meters) of fills. The outer portion of the pier was supported on seven-foot diameter by four-foot to seventeen and one half-foot long (2.1 m by 1.2–5.3 m) precast reinforced concrete hollow sections, while the inner portion was supported on fir piles, both founded on the bedrock as shown in Fig. 6. The hollow concrete sections were interconnected to form a continuous column and infilled with concrete. The superstructure of the pier comprised precast concrete braces which supported cast-in-place concrete girders, beams and deck slabs.

Construction of the Ballantyne Pier, as shown in Figs. 7, 8 and 9, was supervised and managed by the consulting engineer to the Vancouver Harbour Board, Andrew Don Swan of Montreal, Que. Andrew Swan was also responsible for the design of the pier structure and the four reinforced concrete freight sheds; he was supported locally by William George Swan, the Chief Engineer for the Vancouver Harbour Board from 1920 to 1925. William Swan eventually founded the consulting firm



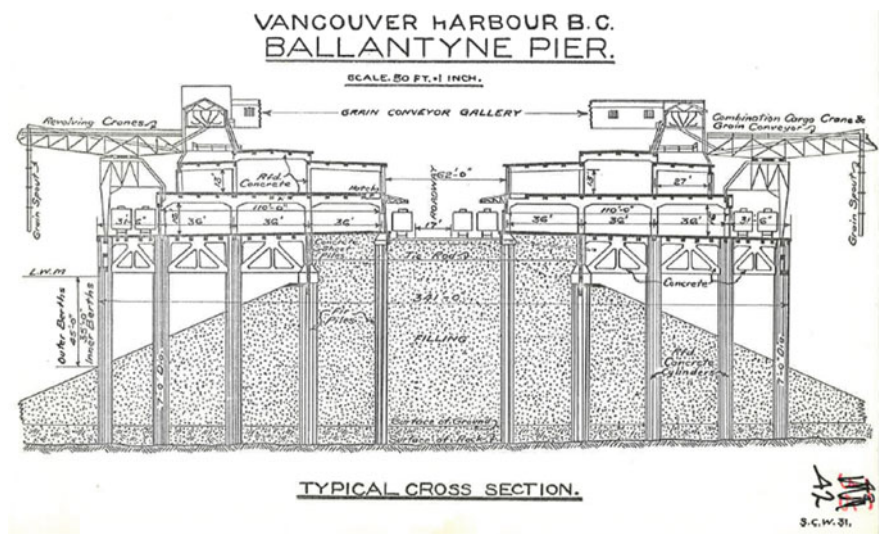


Fig. 6 Typical cross-section

Swan-Wooster and was also elected as President of the Association of Professional Engineers of British Columbia in 1921.



Fig. 7 Precast concrete sections and braces to be lowered into place, circa 1922, CVA 1376-314



**Fig. 8** Deck slab soffit falsework and formwork under construction, circa 1922, CVA 1376-315



**Fig. 9** Stiff-leg derricks to lift and rail tracks to transport materials on site, circa 1922, CVA 1376-316

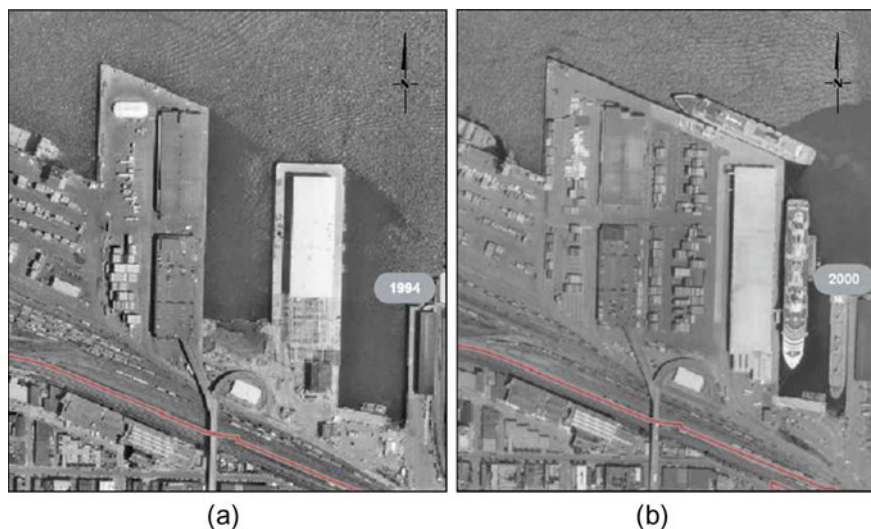
### **3 1990s' Redevelopment of Ballantyne Pier**

By the end of the century, Ballantyne Pier had fallen into disrepair and its use progressively lessened over time. In the early 1990s, the pier was completely gutted and renovated into a dual-use facility for break-bulk export and the growth of the cruise ship sector. Three of the original sheds and part of the fourth were demolished to make room for a warehouse. The warehouse floor was used for wood pulp storage, while the upper east and west wings became the cruise ship terminal's passenger concourses.



Figure 12a was taken during the \$23 million 1994 renovation of Ballantyne Pier into a dual-use facility. Structural steel erection is in progress at the south portion of the warehouse, while roof cladding has been completed at the north portion. The warehouse columns are supported on new precast concrete pile and cast-in-place concrete pilecap foundations; the warehouse floor is of roller-compacted concrete. Renovations to the pier included a two-foot to three-foot (0.6–0.9 m) thick reinforced concrete overlay on the east berth apron, rehabilitation of the concrete braces, girders, beams and deck slab on the west apron, a mechanically stabilized earth (MSE) wall to retain the core fill, post-tensioning of the reinforced concrete precast pier columns, installation of vertical gravel drains and placement of slope protection. These improvements increased the pier's live load capacity and seismic performance. New bollards, fenders and camels were also added to accommodate the cruise vessels calling at the facility. At completion of the two-year project, over one hundred and fifty 914 mm diameter precast concrete piles were driven, 20,000 cubic meters of reinforced cast-in-place concrete were placed and 14,500 tons of warehouse structural steel were erected.

Figure 12b from the year 2000 shows the infill between Ballantyne Pier and Centerm completed and in use. The land reclamation comprised a pile and deck extension to the north end of Ballantyne Pier, concrete caissons along the north edge of the infill to create a new berth face, fill placement behind the caissons and asphalt overlay to create an expanded container storage area for Centerm. At completion of the four-year project, 1050 m<sup>2</sup> of steel pile and concrete deck structure were built, three 41.7 m long × 15.5 m wide × 20.25 m high precast concrete caissons were placed and 32,350 m<sup>2</sup> of area was infilled and capped with asphalt pavement.



**Fig. 12** Ballantyne Pier and Centerm aerial photos: **a** 1994; **b** 2000



#### 4 Incorporating Ballantyne Pier into the Centerm Expansion Project

To meet anticipated near-term demand for containers to be shipped through Vancouver, the nearly one-hundred-year-old Ballantyne Pier is being rehabilitated and consolidated with Centerm through the Centerm Expansion Project and South Shore Access Project. Through an increase of ~15% to the terminal footprint and a reconfiguration of the terminal layout, terminal capacity will increase from a peak capacity of 900,000 20-foot (6.1 m) equivalent unit containers (TEUs) to 1.5 million TEUs annually. The reconfiguration of the terminal will allow achievement of a 67% increase in throughput capacity when complete. The Project, shown in Fig. 13, will also improve port road and rail infrastructure. The improvements to port roads include extending Waterfront Road to connect it to Centennial Road and building the Centennial Road overpass. The new overpass eliminates road and rail conflicts and improves container truck movements and traffic flow, thereby reducing delays caused by trucks idling at train crossings. Changes to Waterfront Road will provide a continuous port road between the Vancouver Convention Center and Highway 1.



Fig. 13 Centerm expansion project and south shore access project components

The Project's marine works include:

- Expanding the land area at the west end of Centerm and east end of Ballantyne Pier.
- Dredging, constructing new rock dikes and infilling open water areas within the dikes.
- Dredging to enhance a navigational turning basin for the Canada Place cruise ship berth in the area between Centerm and the SeaBus south terminal.

The Project's on-terminal works include:

- Removing existing warehouse structures and rehabilitating the Ballantyne Pier (while retaining the Ballantyne Pier building's heritage façade and conversion into a new container operations facility).
- Reconfiguring the terminal intermodal yard to extend the existing tracks.
- Removing the Heatley Avenue overpass.
- Reconfiguring the terminal container yard and entrance area.
- Establishing new storage facilities and increasing terminal parking.
- Upgrading terminal control systems and yard equipment.
- Two new quay cranes, five new electric rail-mounted gantry cranes and diesel-powered internal transfer vehicles.

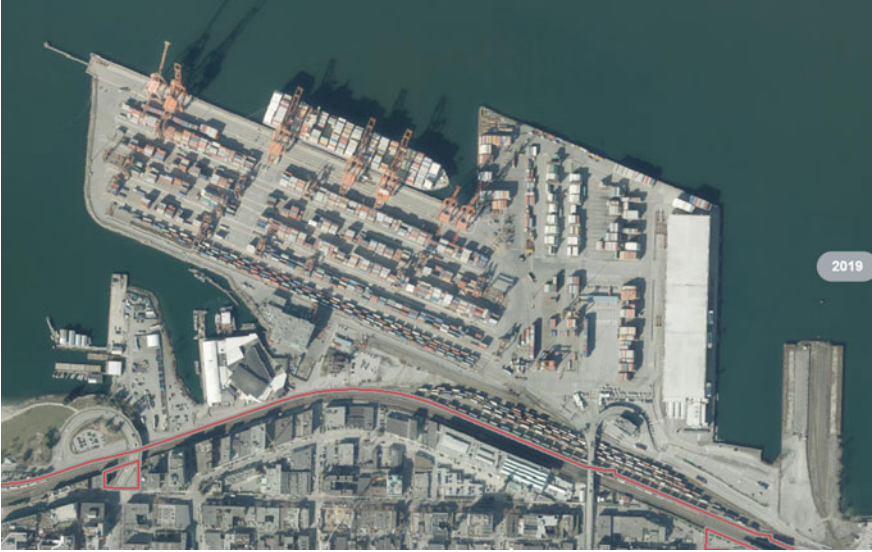
The Project's off-terminal works include:

- Constructing an overpass to the entrance of the terminal (Centennial Road overpass).
- Removal of the Southern Railway rail crossing at Centennial Road.
- Rail yard modifications to the rail yard south of the terminal.
- Westward extension to Waterfront Road.

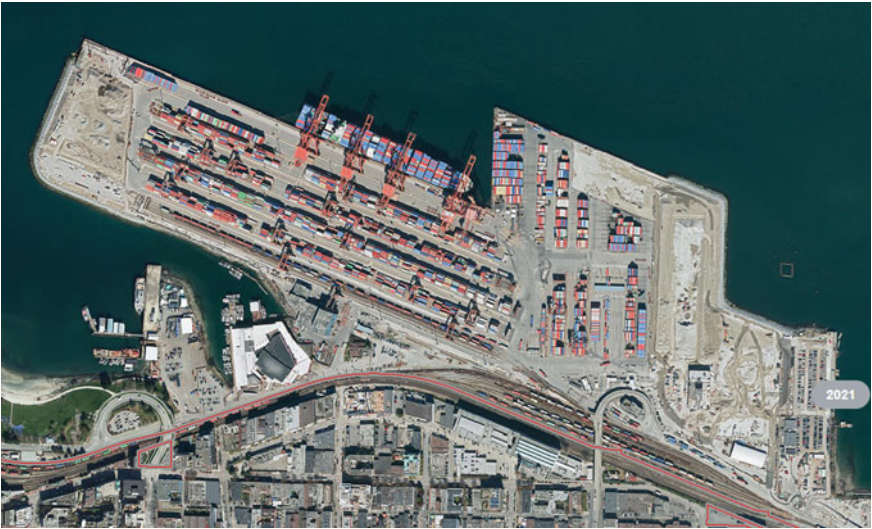
The layouts of Centerm and Ballantyne Pier prior to the start of the Centerm Expansion Project and South Shore Access Project are shown in Fig. 14.

The changes to Centerm and Ballantyne Pier made through the Centerm Expansion Project and South Shore Access Project are depicted in Fig. 15.

Through the Centerm Expansion Project and South Shore Access Project, the iconic Ballantyne Pier building, a City of Vancouver registered historic building, is being renovated once again. It will begin its new life as the container operations' facility of DP World Vancouver, the terminal operator for Centerm. Figure 16 shows the building's main entrance and Fig. 17 shows two views of the partially renovated interior. The Ballantyne Pier facility has been around for almost 100 years. While the only constant may be change, it is good to have places that stay the same and bridge the past to the future.



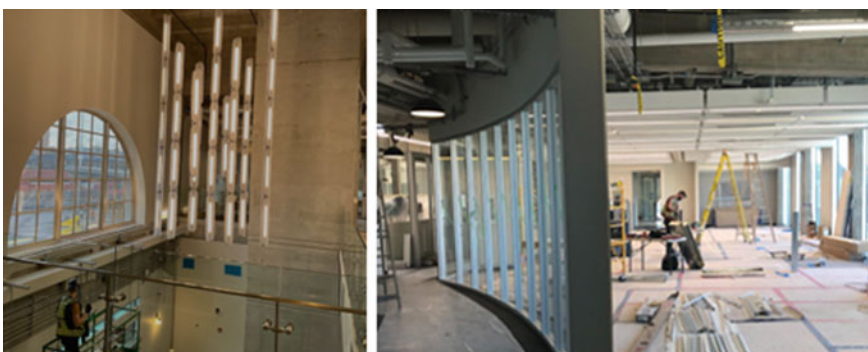
**Fig. 14** Ballantyne Pier and Centerm aerial photo—year 2019 (pre-construction)



**Fig. 15** Ballantyne Pier and Centerm aerial photo—year 2021 (under construction)



**Fig. 16** DP world Vancouver container operations' facility—main entrance



**Fig. 17** DP World Vancouver container operations' facility—interior

**Acknowledgements** The Vancouver Fraser Port Authority's Container Terminal Construction and Engineering departments and the Centerm Expansion Project and South Shore Access Project teams are gratefully acknowledged for their contributions to this paper.

## Reference

1. Delgado JP (2005) Waterfront: the illustrated maritime story of greater Vancouver. Stanton Atkins and Dosil/Vancouver Maritime Museum, Vancouver BC

# **Construction Management: Miscellaneous Topics**

# Post-fire Damage Assessment of Buildings at the Wildland Urban Interface



Ahmad Abo-El-Ezz, Faten AlShaikh, Azarm Farzam, Marc-Olivier Côté, and Marie-José Nollet

**Abstract** Wildfires are considered one of the costliest natural hazards in Canada. Significant fire events that occurred had threatened and destroyed buildings at the Wildland Urban Interface (WUI). Standard methods for wildfire risk assessment include hazards' analysis, inventory of exposed buildings and vulnerability analysis that correlates expected losses to fire intensity measure and distance from forest boundary. On the other hand, there is limited research on buildings' vulnerability assessment to wildfire impacts and scarcity of models that correlate the likely response and expected loss of different types of buildings to varying levels of fire intensity. This article presents a methodology for geospatial data collection of post-fire buildings damage at Canadian WUI communities with the objective of developing community-scale empirical building fire vulnerability models that can be integrated in community-scale wildfire risk assessment tools. In this study, the empirical fire vulnerability model is developed in terms of the loss rate defined by the proportion of buildings burned as a percentage of the total exposed buildings as a function of the distance from forest edge and the corresponding fire intensity. The methodology consists of consecutive steps including geospatial digitization of burned and survived buildings from post-fire open-source satellite imagery; characterization of building types and occupancy based on open-source municipal databases; estimation of distances to burned forest boundary based on burn scar satellite imagery and the measurement of distance increments to buildings. The buildings data are then combined to develop an empirical fire vulnerability model. The methodology is demonstrated by a case study WUI community in Canada that was exposed to a damaging wildfire event.

**Keywords** Post-fire damage assessment · Buildings · Wildland urban interface

---

A. Abo-El-Ezz (✉) · F. AlShaikh · A. Farzam · M.-O. Côté · M.-J. Nollet  
Department of Construction Engineering, École de Technologie Supérieure, Montréal, QC,  
Canada  
e-mail: [ahmad.abo-el-ezz@etsmtl.ca](mailto:ahmad.abo-el-ezz@etsmtl.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_55](https://doi.org/10.1007/978-3-031-34593-7_55)

893

## 1 Introduction

Wildfires are considered one of the costliest natural hazards in Canada. Significant fire events that occurred had threatened and destroyed buildings at the Wildland Urban Interface (WUI) such as the 2016 Fort MacMurray fire that destroyed 1600 buildings [1] causing \$3.6 billion of insured damage [2], the 2017 British Columbia fire events that destroyed 300 buildings [3] causing \$127 million in insured damage [4] and the most recent Lytton fire in 2021 that destroyed 90% of the buildings in the community and causing \$102 million of insured damage. Standard methods for wildfire risk assessment include hazards' analysis, inventory of exposed buildings and vulnerability analysis that correlates expected losses to fire intensity measure [5]. Research efforts on wildfire risks have been focusing on ignition management and intensity reduction [6], however, little is known about vulnerability of structures to wildfires in Canada. One indicator is the limited understanding of fire propagation and structural response at WUI. Research on WUI disasters [7, 8] has described the WUI problem as a structural ignition problem as losses are primarily determined by structure ignition conditions. If structures are safeguarded against ignition sources, property loss and costs incurred (not to mention potential loss of life) can be avoided. There is ongoing work on the likelihood and exposure components of the wildfire risk framework in Canada [6]. Existing tools for fire vulnerability such as [9] and [10] have been focused on building-by-building assessment using a scoring methodology to assign different levels of vulnerability using local site-specific information on roof, siding, window materials, vegetation conditions and combustible surrounding each building and indicators of municipal fire suppression capacity. On the other hand, there is a need for tools that integrate fire vulnerability functions that can provide community-level estimates of potential damage and loss with specific consideration to their distance from forest edge where ember spotting into urban areas phenomena is frequently observed [11]. These community-scale vulnerability functions are particularly relevant for first-order assessment of fire risk at a community scale for preparing mitigation and emergency management planning scenarios. This article presents a methodology for geospatial data collection of post-fire buildings damage at Canadian WUI communities with the objective of developing empirical building fire vulnerability functions that can be integrated in community-scale wildfire risk assessment tools. In this study, the empirical fire vulnerability model is developed in terms of the loss rate defined by the proportion of buildings burned as a percentage of the total exposed buildings as a function of the distance from forest edge.

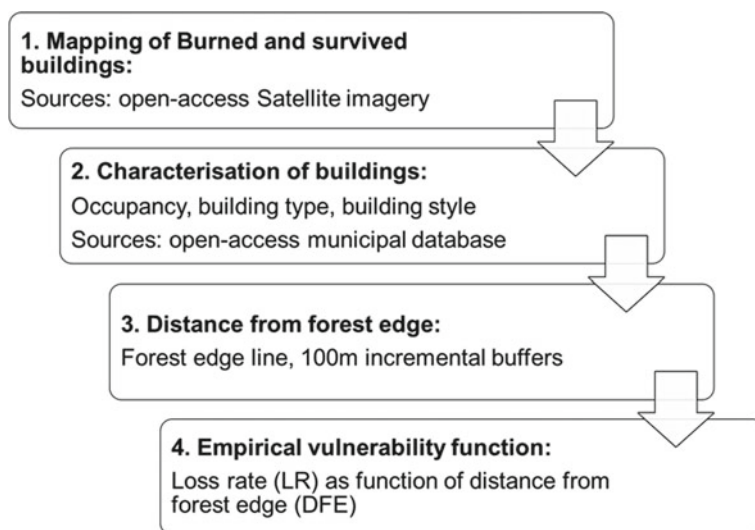


## 2 Methodology

The methodology for the development of community-scale empirical vulnerability functions consists of consecutive steps (Fig. 1) including geospatial digitization of burned and survived buildings from post-fire open-source satellite imagery; characterization of building types and occupancy based on open-source databases; estimation of distances to forest edge based on burn scar satellite imagery and the measurement of distance increments to buildings. The buildings' data are then combined to develop an empirical fire vulnerability model. The methodology is demonstrated by a case study WUI community of Slave Lake in Alberta (Fig. 2), Canada, that was exposed to a damaging wildfire event in May 2011 [12]. The fire which originated 15 km outside of the community was quickly pushed by 80 km per hour winds. The fire forced the complete evacuation of Slave Lake's 7000 residents. Such population displacement was considered the largest in the province's history at the time. The fire destroyed significant number of buildings including the Town Hall and the library buildings. Insurable damage was estimated at C\$750 million, making it the second costliest insured disaster in the country's history at the time.

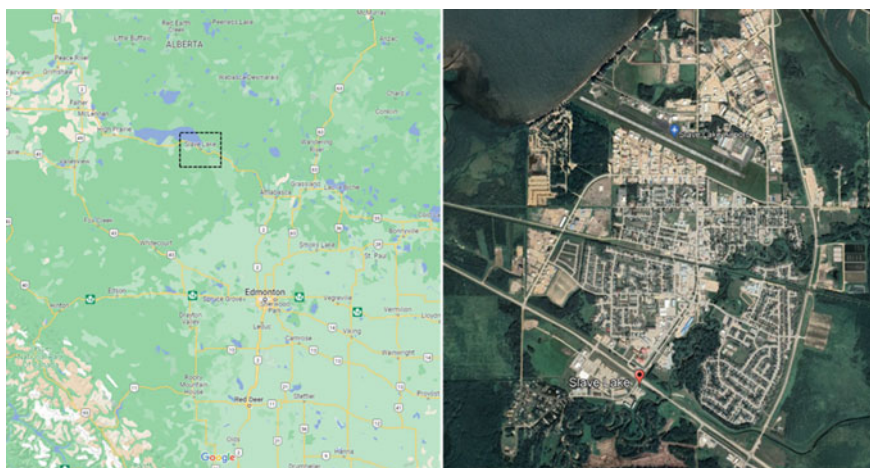
### 2.1 Mapping of Burned and Survived Buildings

In this step, the exposed buildings including those that have been burned and survived the fire are identified and mapped. This allows for the assessment of the spatial



**Fig. 1** Methodology used for the development of community-scale empirical vulnerability functions





**Fig. 2** Location map of the case study community of Slave Lake, Alberta. *Source* Google maps© and Google earth pro©

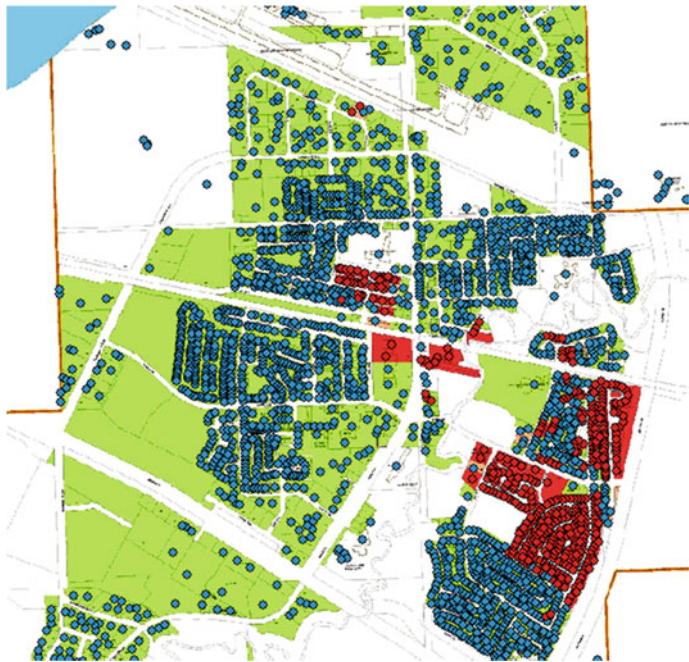
distribution of damage and the estimation of the extent of exposure to wildfire ember attack into the urban areas. Google earth pro© was used to find satellite imagery that was taken after the fire using the function “show historic imagery”. The images were then downloaded and georeferenced using ArcGIS© software. The centroid of each building in the community was digitized as a point feature and was then identified as either destroyed or survived base on visual interpretation of the images (Fig. 3). In total, 2247 buildings were digitized out of which there were 450 buildings burned. It was observed that the fire burned buildings that were within 1.6 km from the forest edge. This is less than the 2 km threshold that has been considered in the literature as an upper bound for maximum ember spotting distance potential based on different CWFIS fuel types [13]. This highlighted the significant of the impact of ember spotting for the estimates of potential exposure to ember ignition in WUI communities. The damage mapping results were validated with the official fire damage map of the town of Slave Lake (Fig. 4) showing damaged land parcels which was released on May 18, 2011, by the municipal emergency operation center [14].

## 2.2 Characterization of Buildings

The characterization of buildings is typically conducted to provide a classification system that includes occupancy (e.g., residential, commercial, government, industrial); building type (e.g., main building, outbuilding such as garage or storage sheds);



**Fig. 3** Satellite image (Google earth pro©.) of part of the community of Slave Lake taken after the fire shown on the left image and the georeferenced destroyed (in red) and survived (in blue) buildings shown on the right image in ArcGIS © software



**Fig. 4** Validation of burned building locations digitized in ArcGIS© with fire damage map released by the town of Slave Lake. *Source* [14]

residential buildings style (e.g., detached single family, attached townhouse, apartments' building); In this study, the occupancy classes were considered as either residential or other (non-residential). Building-specific parameters such as construction materials of roofs, walls, windows; fuel load in buildings; ventilation aspects; age of the buildings and existing damage are not currently implemented in the methodology due to the lack of availability of detailed information about exposed buildings but should be included in future developments for improved risk assessment capabilities. The building classification process was conducted using the town of Slave Lake 2010 zoning map which was available online [15]. The survey showed that 88% of total number of buildings were residential occupancy with 86% of these buildings classified as single-family houses.

### ***2.3 Distance from Forest Edge***

In this step, the forest edge line is identified by considering the boundary of the burned forest, and distance increments from that line are noted for references. The objective of this step is to identify the relative location of buildings that burned and survived with respect to distance from forest edge given the significant impact of ember spotting into urban areas. Previous studies on ember spotting have investigated the maximum distance that ember can travel in forested areas. It has been observed that the ember density is typically reduced with distance from the fire front. On the other hand, it did not consider the potential of these embers to ignite structures in urban areas. The few studies that have evaluated the ignition potential of structures from ember attack in urban areas which were mainly conducted in Australia. In this case study, the burn scar Landsat images, available from a technical report on the fire incident, were used to identify forest edge line [12]. Incremental polygons with distance range of 100 m were then drawn in ArcGIS until it intersected the last burned building (Fig. 5). It was observed that the last burned building was located 1600 m away for the forest edge. As compared to previous studies in Australia, the furthest impact of ember spotting into urban areas was observed up to 700 m [16]. Figure 6 shows the percentiles of all burned buildings as a function of distance from forest edge. It is observed that 50% of all burned buildings were located within 300 m from the forest edge and 90% were within 800 m. Such observation is of particular interest in risk-based fire planning where specific level of fire protection is to be mandated for buildings based on the distance from the forest edge to reduce the loss potential from ember spotting.

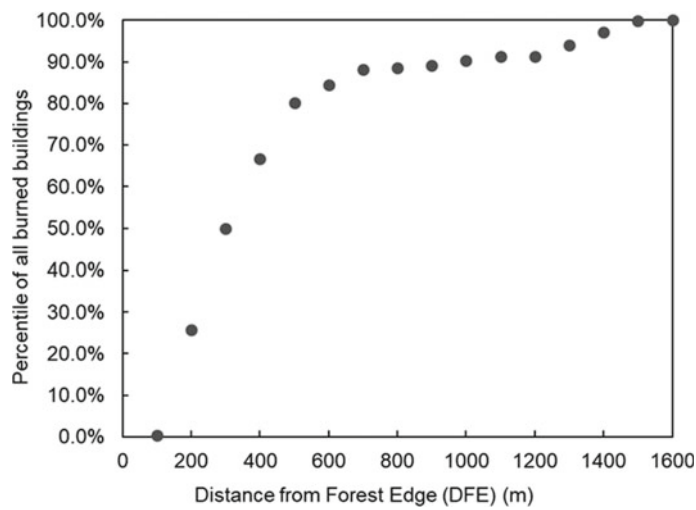
### ***2.4 Empirical Vulnerability Function***

In this step, an empirical relationship between the loss rates (LRs) defined as the proportion of buildings that were damaged to the total number of buildings at each



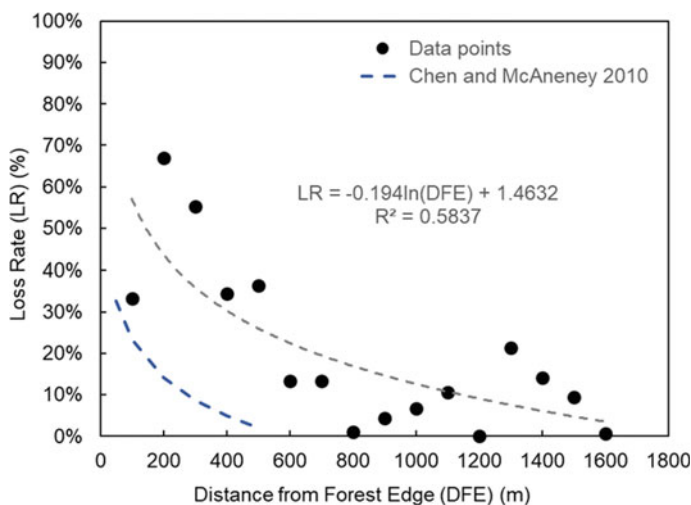


**Fig. 5** Extent of damage buildings into the community (left) and forest edge line and considered 100 m wide polygons showing relative distance increments of buildings (right)



**Fig. 6** Cumulative distribution of the percentile of all burned buildings as a function of distance from forest edge

distance increment from the forest edge (DFE) is developed. Figure 7 shows the empirical vulnerability function corresponding to the obtained data points. The results show an incremental reduction of the LR with increasing distance from forest edge. This is compatible with previous observations from Australia in terms of LR trend as well as the expected reduction of ember density further away from the



**Fig. 7** Empirical vulnerability function showing the correlation between LR of buildings with DFE (m)

burning fire front. An additional factor was the proximity of buildings that increased the potential for building-to-building fire spread. For example, some clustering of damage was observed beyond 1 km from the forest edge which could be explained by an ignition of few buildings from ember spotting that spread to other adjacent buildings. The empirical vulnerability function can also be interpreted as a measure of probability of damage of buildings given a fire event occurrence in relationship to its distance from forest edge. For example, a building located 500 m from the forest edge would have a 30% probability of being damage if a wildfire event occurred. In comparison with the empirical model developed by [17] based on fire damage data from Australia, it is observed that their model would provide an underestimation of potential damage to buildings in Canada based on the results from the case study in this paper. This underscores the significance of developing region-specific vulnerability functions that implicitly consider the forest vegetation types and their fire behavior (e.g., ember spotting potential) as well as typical building types and spatial distribution within the community.

### 3 Conclusions and Outlook

This paper presented a methodology for the development of an empirical fire vulnerability model correlating the expected loss rates of buildings as a proportion of total number of exposed buildings and the distance from fire front edge. The methodology consisted of consecutive steps including geospatial digitization of burned

and survived buildings from post-fire open-source satellite imagery; characterization of building types and occupancy based on open-source databases; estimation of distances to forest edge based on burn scar satellite imagery and the measurement of distance increments to buildings. The buildings data are then combined to develop an empirical fire vulnerability model. The methodology was demonstrated by a case study WUI community of Slave Lake in Alberta, Canada, that was exposed to a damaging wildfire event in May 2011. The results showed that there is an incremental reduction of buildings loss rates with increasing distance from forest edge. Such observation is compatible with previous studies in Australia showing similar trend. On the other hand, the maximum distance into community where burned buildings were reported was around 1600 m from the interface forest edge compared to 700 m in previous studies in Australia but still within the empirical limit of 2000 m of maximum travel distance of ember spotting. Moreover, it was observed that 90% of all burned buildings were located within 800 m from the forest edge. This observation highlights the importance of considering fire mitigation measures not only for buildings adjacent to the forest interface but should consider risk-based distance criteria for improved fire protection and retrofit measures from the interface line given the unpredictable nature of ember spotting phenomena and the current limitations of their modeling. Future research in the framework of this project will include the assessment of impact of building-to-building distance on fire spread and the quantification of building-specific vulnerability parameters such as construction materials of roofs, walls, windows and vegetation conditions or combustibles near the building. Other parameters such as fuel load in buildings; ventilation aspects; age of the buildings and existing damage in buildings before the fire should be considered for improved risk assessment capabilities.

**Acknowledgements** This study was supported by funding from the Canadian Forest Service, Natural Resources Canada.

## References

1. Alberta Government (2016) Home again: recovery after the Wood Buffalo wildfire. <https://open.alberta.ca/dataset/3c8f8b73-d7a5-42b0-85b2-12367c7d82bf/resource/147e872d-10a1-491f-826a-10e803c40bfe/download/2016-home-again-recovery-after-wood-buffalo-wildfire.pdf>
2. Canadian Underwriter (2016) <https://www.canadianunderwriter.ca/catastrophes/nearly-3-6-billion-insured-losses-fort-mcmurray-wildfire-catiq-1004096018/>
3. Global News (2017) B.C. wildfires destroy over 300 buildings. <https://globalnews.ca/news/3641081/b-c-wildfires-destroy-over-300-buildings/>
4. IBC (2017) <http://www.abc.ca/bc/resources/media-centre/media-releases/british-columbia-wildfires-cause-more-than-127-million-in-insured-damage>
5. Scott JH, Thompson MP, Calkin DE (2013) A wildfire risk assessment framework for land and resource management. In: Generation of the technical report RMRS-GTR-315. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, p 83. <https://doi.org/10.2737/rmrs-gtr-315>

6. Johnston LM, Wang X, Erni S, Taylor SW, McFayden CB, Oliver JA, Stockdale C, Christianson A, Boulanger Y, Gauthier S, Arseneault D (2020) Wildland fire risk research in Canada. *Environ Rev* 28(2):164–186
7. Calkin DE, Cohen JD, Finney MA, Thompson MP (2014) How risk management can prevent future wildfire disasters in the wildland-urban interface. *Proc Natl Acad Sci* 111(2):746–751
8. Cohen JD (2000) Preventing disaster: home ignitability in the wildland-urban interface. *J For* 98(3):15–21. <https://doi.org/10.1093/jof/98.3.15>
9. Alberta Infrastructure (2013) Guideline for wildfire protection of institutional building in forested regions in Alberta. <https://www.alberta.ca/assets/documents/tr/tr-wildfireprotection.pdf>
10. FireSmart Canada (2022) [https://firesmartcanada.ca/wp-content/uploads/2022/01/FSC\\_NRP\\_HIZ-ScoreCard\\_Form-final-03-15-TM-1.pdf](https://firesmartcanada.ca/wp-content/uploads/2022/01/FSC_NRP_HIZ-ScoreCard_Form-final-03-15-TM-1.pdf)
11. Kramer HA, Mockrin MH, Alexandre PM, Radeloff VC (2019) High wildfire damage in interface communities in California. *Int J Wildland Fire* 28(9):641–650. <https://doi.org/10.1071/WF18108>
12. Alberta Government (2011) Flat top complex, Wildfire science documentation report. <https://open.alberta.ca/dataset/dee39e20-8832-4175-8f3a-a313dade523f/resource/37b00384-79d7-4ea3-aa82-12b6ab64fef6/download/flattopcomplex-sciencedocreport-2012.pdf>
13. FLNROR (2017) Provincial strategic threat analysis: 2017 update. BC Wildfire Service; Forests, Lands, Natural Resource Operations and Rural Development, BC
14. National Post (2011) <https://nationalpost.com/news/canada/slave-lake-fire-destroyed-374-properties-damaged-52/>
15. TSL (2010) Town of Slave Lake zoning map. <https://www.slavelake.ca/DocumentCenter/View/197/Zoning-Map-PDF>
16. Chen K, McAneney J (2004) Quantifying bushfire penetration into urban areas in Australia. *Geophys Res Lett* 31(12):147
17. Chen K, McAneney J (2010) Bushfire penetration into urban areas in Australia: a spatial analysis. Report for the Bushfire CRC. <https://www.bushfirecrc.com/sites/default/files/managed/resource/bushfire-penetration-urban-areas.pdf>

# Fire Building Codes in Developed and Developing Countries: A Case Study of Canada and Costa Rica



Sara Guevara Arce, Hannah Carton, and John Gales

**Abstract** In Canada, the development of fire codes and standards is based upon governmental resources aiming to advance requirements deemed insufficient or potentially non-reflective of current design trends. This has led to recent advancements of fire-safe engineered timber and urban interface protection from wildfires. It has at this time minimized attention to the marginalized populations exposed to fire hazards, such as the numerous informal settlement fires observed across Canada in the last few years. Developing countries, generally, do not have the financial or administrative resources to advance their own fire codes and standards in such a manner. In general, developing countries adopt codes and standards from more established sources to fit their necessities. However, these codes and standards may not be wholly representative of their current design trends or fire problems. For example, in Costa Rica, adopting the National Fire Protection Association (NFPA) Standard 101 has caused confusion around the required fire protection needed for acceptable design as it was not created with Costa Rica in mind. The adopted NFPA code, more importantly, does not consider how to protect marginalized populations living in informal settlements. Informal settlements are a growing issue in both developing and developed countries. Informal settlements are at a greater risk of fire and other hazards due to the physical characteristics of the structures and lack of regulation thereof, socio-economic vulnerability of the residents, and the political and institutional marginalization of the settlements. Considering Costa Rica and Canada, the present study serves three purposes: to compare Developed (Canada) and Developing (Costa Rica) code advancement approaches; to determine code applicability to informal settlements; and to compare each country's approach to informal settlements.

**Keyword** Fire building codes

---

S. Guevara Arce · H. Carton · J. Gales (✉)  
York University, Toronto, Canada  
e-mail: [jgales@yorku.ca](mailto:jgales@yorku.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_56](https://doi.org/10.1007/978-3-031-34593-7_56)

903



## 1 Introduction

Building design codes and regulations are meant to protect human lives and structural integrity of the building in the event of a hazard through providing standards that buildings must adhere to [5]. Codes have historically been prescriptive but have recently begun transitioning to performance-based codes [33]. In Canada, codes are developed based upon national governmental resources aiming to advance requirements deemed insufficient or potentially non-reflective of current infrastructure design trends. However, developing countries, generally, do not have the financial or administrative resources to advance their own fire codes and standards in such a manner and instead adopt codes and standards from more established sources to fit their necessities. However, building and fire codes are typically developed with specific scopes, are only applicable to conventional infrastructure, and do not address unconventional buildings and non-traditional housing such as informal settlements.

Informal settlements are commonly defined as residential areas where residents have no guaranteed permanency, a lack of basic services and amenities, and housing does not comply with building codes and regulations. They are a growing issue in both developed and developing countries due to a lack of affordable housing, increased urban population growth and economic vulnerability [39]. Informal settlements are at a greater risk of fire and other hazards due to the socio-economic vulnerability of the residents, the physical characteristics of the structures, and the political and institutional marginalization of the settlements [41]. The UN has targeted informal settlements through the Sustainable Development Goals (SDGs), notably SDG #11: make cities, inclusive, safe, resilient, and sustainable, but also through SDG #1: end poverty in all forms and SDG #10: reduce inequality within and among countries [40].

The objective of this paper addresses important interrelated themes to the above: (1) compare Developed (Canada) and Developing (Costa Rica) code advancement approaches, (2) determine code applicability to informal settlements, and (3) compare each country's approach to informal settlements.

## 2 Code Advancement Processes

These countries were chosen for this study due to the differences between how they are addressing the housing and informal settlement issues they have been experiencing. Canada is a developed country with a known affordable housing crisis and informal settlements which can be compared with a country who is also dealing with similar issues, though in different forms. Canada has accessibility to greater economic resources to address these issues, while for Costa Rica to obtain solutions must take different approaches, due to the relative lack of monetary resources to work in housing issues.

Costa Rica was chosen for comparison due to the housing solution efforts that the government has been doing through the years, aiming to provide decent housing to all Costa Ricans. There are records since 1904 about government actions focused on solving housing problems affecting the country [26]. In 1949, it was established in the Costa Rican Political Constitution that the State has the obligation of providing acceptable housing to the population with limited economic resources [26]. These efforts have placed Costa Rica on top of the region, as Costa Rica is the country with the lowest percentage of urban population living in informal settlements in Latin America [37].

## **2.1 *Canada***

In Canada, building and fire safety regulation are under the jurisdiction of the provinces and territories who can choose to adopt or enforce the national model codes or develop their own codes [6]. Canada has five national building model codes: National Building Code of Canada (NBC), National Fire Code of Canada (NFC), National Plumbing Code of Canada (NPC), National Energy Code of Canada for Buildings (NECB), and the National Farm Building Code of Canada (NFBC). The NBC and NFC are the two main model codes that concern the fire safety of newly constructed and existing buildings. The NBC and NFC are designed to complement each other, where the NBC covers fire safety requirements at the time of building construction and reconstruction, while the NFC generally covers the ongoing operation and maintenance of fire safety systems of buildings in-use [7].

As model codes, they were created to provide the different building code jurisdictions with a model to follow, to encourage consistency and compatibility across the jurisdictions [6]. The codes are developed by the Canadian Commission on Building and Fire Codes (CCBFC) and updated every five years through a broad-based consensus process [35]. Proposed code changes are submitted to the CCBFC where they are examined by the appropriate expert standing committee within CCBFC for merit. Committee members are appointed to the committees through nominations under regulatory, industry, or general interest categories. If the proposed change is accepted, it undergoes a series of reviews by relevant standing committees within CCBFC, provincial and territorial authorities, and the public. Comments received from the reviews are taken into consideration before the proposed change is submitted to CCBFC for final approval. If approved, the change is included in the next edition of the national code [35].

## **2.2 *Costa Rica***

The current regulations that govern Costa Rica regarding fires and fire safety were created in 2002 from Law 8228 and its Decree No. 37615-MP [1]. It allows the

Costa Rican Fire Corps to issue technical standardization, which will be mandatory for individuals or legal entities, either public or private, in matters of security, fire protection, and human security [3]. After the creation of the law and decree, the Fire Corps was obliged to create a guide that helps professionals apply the regulations, mainly concerning the processing of plans, which resulted in the creation of the “Manual of technical provisions for human safety and fire protection” in 2005, where the entire regulatory package of the National Fire Protection Association (NFPA) was adopted [4]. In the next years, four more editions were published, mainly updating the information and making it more comprehensible for the users [42], as the adoption of NFPA standards has caused confusion in the required fire protection needed for acceptable design. The last version was published in 2020; in this edition, the document converts from a manual into a regulation [1], but it must be highlighted that none of these documents addressed how to protect marginalized populations living in informal settlements.

### 3 Informal Settlements

Nearly 1 billion people or 32% of the world’s urban population live in informal dwellings and around 88% (881,080,000) informal settlement dwellers are estimated to be living in developing countries [40]. This means that issues with informal settlements are not equally distributed between developed and developing countries, and it is magnified in the latter. Lack of affordable housing is a growing problem, related to population growth which is estimated to increase by 2 billion in the next 30 years [43]. The growth is expected to lead to increased use and density of existing informal settlements and/or the creation of new ones.

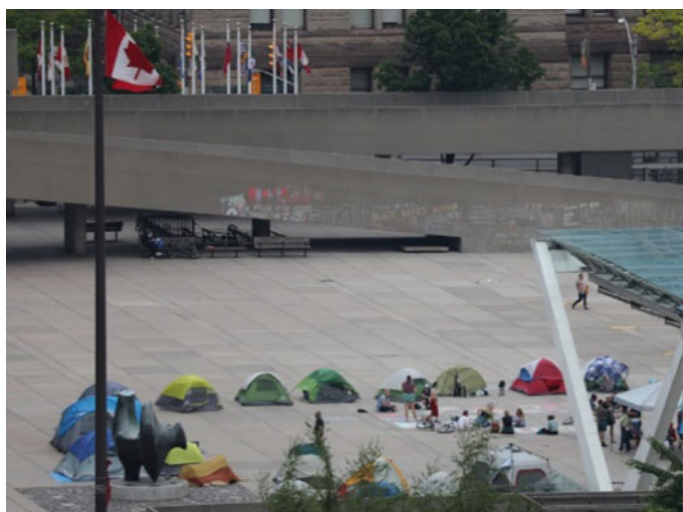
Informal settlements have a higher fire risk due to poor electrical connections and the use of open flames for cooking, warmth, and lighting which create ignition sources that generate a latent risk. Additionally, the high density of dwellings and short distances between them, the use of combustible construction materials, the topography of the land, among other factors, are features that can worsen a fire emergency, turning a fire rapidly into a conflagration. Koker et al. found that a dwelling inside an informal settlement can reach flashover within a minute or less after ignition and that downwind neighbor structures can be ignited in less than a minute [31].

#### 3.1 *Canada*

Homelessness in Canada has been a growing concern since the 1980s due to less spending and support for social services and affordable housing, and rapid declines in job availability and permanency, resulting in approximately 235,000 individuals

experiencing homelessness per year and 35,000 on any given night [23]. It is estimated that 0.010% of the urban population in Canada live in informal dwellings [38], which corresponds to approximately 3000 people. This is without considering the effects of the COVID-19 pandemic where the number of encampments soared across Canada as people decided to abandon overcrowded shelters or felt they were unwelcome to rest in friend's homes [13]. The City of Toronto estimates that in 2021, there were 421 tents and other temporary structures located within the city at 100 different locations, mainly parks and right-of-way passages [10]. The 2021 Streets Need Assessment estimated that approximately 7347 individuals were experiencing homelessness in Toronto, with 163 individuals residing in encampments, however, both figures are likely below the actual figure as only public spaces were included and hidden homelessness was excluded [11] (Fig. 1).

In 2020, Toronto Fire Services responded to 253 fire incidents in encampments. This represented a 247% increase over incidents in the same period in 2019 [10], however that number includes incidents of suspected fires, not only actual fires which occurred [19]. Fire risks in homeless encampments include open flames, unsafe wiring, and power generators as well as unsafe fuel and other flammable material storage [10]. Many of these risks are associated with Canada's colder climate and the need to stay warm along with a lack of fire prevention training or suppression equipment such as fire extinguishers [24]. There is little to no active or passive fire protections within homeless encampments.



**Fig. 1** Temporary informal settlement during Summer 2020 located in Central Toronto (authors' photo)

### 3.2 *Costa Rica*

In 2018, around 3.9% of the urban population in Costa Rica were living in informal settlements [37], which meant approximately 195,000 dwellers. In the same year, the Engineering Unit of the Costa Rican Fire Corps reported a total of 50 fires in informal settlements [18], but this data might not reflect the real number of fires as several times inhabitants try to extinguish the fire by themselves without calling the emergency services [25]. The houses inside informal settlements are in constant risk of fire, as they are surrounded by potential causes of fires, including short circuits due to electrical connections in disrepair, cooking and lighting with open flame, poor maintenance and manipulation of gas cylinders, attempted controlled burns of wires or dry grass, candles or wood burning without surveillance, as well as children playing with matches, candles, or left alone in their houses [27]. In addition to the constant risk of fire, there is a lack of active and passive fire protection, which may worsen the severity of the fire and increases the likelihood of having greater losses.

In 2019, two large fires affected two different informal settlements in Costa Rica. The first one occurred on April 13th in La Carpio, located in La Uruca district of San Jose. Seven people died in a house where 15 people were living. Some of the survivors escaped through the ceiling as the fire started in the only available entrance. The cause of the fire was arson [26].

The second fire occurred on September 16, 2019, in El Pochote, located in Hospital district of San Jose. In this event, 40 houses located in 2400 m<sup>2</sup> were burned, leaving approximately 216 people homeless. The cause of the fire was a short-circuit [26], and Fig. 2 shows the fire footprint.



**Fig. 2** El Pochote fire footprint (by permission from Fernández [21])



**Fig. 3** Costa Rican multi-story informal dwellings versus smaller Canadian tent encampments (author's photos)

### ***3.3 Differences Between Informal Settlements***

There is a marked difference between the informal settlements of Costa Rica and the homeless encampments in Canada. Comparing the level of informality, Canada can be considered more informal than Costa Rican informal dwellings, as in the latter, there are physical houses instead of mainly tents. These houses can be multi-storied (see Fig. 3). Tents are the primary form of shelter in Canadian informal settlements as they are the most convenient and available option to residents as municipalities prevent larger shelters, such as tiny houses, from being constructed using legislative restrictions. Additionally, the size of informal settlements differs between the two countries as an informal settlement in Costa Rica can reach up to 25,000 people (around 5000 families), [34] almost three and a half times the estimated amount people experiencing homelessness throughout Toronto. As Canada's municipalities leverage zoning by-laws and building code provisions to prevent large temporary shelters, homeless encampments tend to be scattered in different areas of cities. In Toronto, tent encampments were found at 59 unique park locations and 41 right-of-way passages [10].

Although living in illegal tenure, the people in Costa Rican informal dwellings have a fixed space in which can remain for years or even decades. For example, La Carpio, the biggest informal settlement in Costa Rica was established in 1993 [34], whereas in Canada, the primary method of dealing with encampments is to forcefully evict them from the area using building and fire code provisions and municipal zoning by-laws, preventing permanent and/or longer-term encampments [20]. Additionally in Costa Rica, the residents manage to get basic services as water and electricity, though mostly in an illegal manner. In some cases, national companies develop projects to minimize the informality condition and enhance their living situation. For instance, the National Company of Force and Light (CNFL) started a project of grouped metering and network shielding, providing informal dwellers with a





**Fig. 4** Grouped metering (author's photos)

minimum of passive fire protection and legal electricity at an affordable price. In Fig. 4, the left picture shows the cabinets installed for electricity distribution, and the right picture depicts the connections that the community did in order to legalize their services.

Another difference is the ability to trace and census them, although in both cases it is difficult due to their nomad behavior. It is more probable for people to stay longer in informal buildings than in tents, as the latter poses more ease for location changes. In Canada, the issue of homelessness, affordable housing, and social services falls across multiple jurisdictions which means that there is no centralized system for surveying the number of people experiencing homelessness. Employment and Social Development Canada (ESDC) conducts voluntary community Point-in-Time counts every few years to count a community's sheltered and unsheltered homeless [17]. However, these counts are voluntary and may not specify whether "unsheltered" means that they live in an encampment depending on the municipality [15]. There are also no surveys about the different forms of shelter within homeless encampments.

## 4 Building Codes and Informal Settlements

Informal settlements exist in an unclear legal area as they are often situated on land the residents do not own. Additionally, due to the nature of informal settlements, the buildings and structures used are often whatever is available and constructed to provide shelter, without any regards to building or fire regulations. While there are existing frameworks on informal settlements that propose to create policies in order to update land use and ensure informal settlements meet minimum safety requirements [2, 20], there have not been any proposals to incorporate informal settlements in building and fire safety regulations themselves.

In Canada, homeless encampments are not considered permanent buildings and do not directly fall under the authority of building codes. Tents are the most common form of shelter in homeless encampments as one of the few available forms of shelter. Larger shelters, such as tiny homes, are considered permanent shelters and as such fall under building and fire code regulations which municipalities have utilized in order to have them dismantled as they lack active and passive fire protections. Additionally,

as a vulnerable population, many of whom lack financial means, any building or structural regulations are unlikely to be followed due to a lack of ability. There are currently no official processes in place to regulate the safety of tent encampments. Municipalities have commonly employed zoning by-laws to criminalize temporary shelters in city spaces and allow police to forcefully evict residents of encampments using those powers. However, these by-laws have come under legal scrutiny and been found to violate Section 7 of the *Canadian Charter of Rights and Freedoms* that stipulates all Canadians have a right to “life, liberty and security of the person” which the Supreme Court of Canada has interpreted it to include the right to housing [20]. While there are programs in place for municipalities to provide access to sleeping bags, propane stoves as well as burn bins, in order to reduce the use of open flames for warmth [24, 36], there are few initiatives in place to provide housing or supply residents with safe(r) building materials in encampments, outside of directing them to existing services such as city shelters and social housing programs.

Some cities have started initiatives using tiny homes as an alternative to homelessness and a form of affordable housing [32, 44]. However, this move has not been without resistance. Tiny homes have largely been promoted as a sustainable alternative to traditional forms of housing, however, they occupy a gray area within building codes and pose a number of safety risks, including fire [22]. While some types of tiny homes can fit under existing building codes and regulations, there are prescriptive requirements that tiny houses are non-compliant with by nature of the smaller space [9]. Tiny homes lack of fire protection outside of fire alarms and the smaller space impacts egress as well as fire spread within the home [32]. There is also very little research on the material behavior in fire of tiny homes and on fire spread within and between tiny homes. It has been observed that fires spread quickly within the homes with very little time for effective fire suppression. In March 2022, there was a fire within a tiny home village for the homeless in Lake Merritt, California, where it was reported that three tiny homes were destroyed within ten minutes as fire spread between the tiny homes. One resident stated “the walls were just melting” around her during the fire [30].

The City of Toronto settled a lawsuit against an individual who was constructing tiny wooden shelters for homeless individuals. These shelters consisted of a single insulated room with one window and a door, designed to protect against freezing temperatures [29]. The City’s position at the time was that the tiny shelters were not legal dwellings and violated municipal by-laws against structures on City-owned land [12]. While not included as part of the legal proceedings, the City also had concerns regarding the fire safety of the tiny shelters as they were made of combustible materials, and had not been inspected by Toronto Fire Services to confirm fire safety features [12, 29].

In Costa Rica, there are no legal and administrative provisions to facilitate the regulation of informal settlements, and this generates problems regarding access to basic services, health, accessibility, and security. Additionally, it makes it impossible for the Costa Rican government to apply the corresponding controls and to collect payments for the services provided [8]. However, it was found that dwellers of Costa Rican informal settlements self-regulate their communities, reinforced with



jurisdictional support, attaining a certain degree of fire safety. For instance, in Bajo Zamora, an informal settlement visited by the author, minimum distances in alleys must be respected to be able to construct there. Furthermore, community members organized an emergency brigade which is also in charge of fire prevention and mitigation. While tent encampments in Canada have been observed to have some degree of self-governance, if and how they manage fire risks and fire safety has not been explored [45].

In Costa Rica, there have been several efforts done by different institutions to study, improve, and try to formalize informal settlements. In 2007, there was an intent to tailor the national norms and regulations to make them more flexible for people living in informal settlements. This project gathered around 12 public and private stakeholders. Three workshops were carried out through the Commission of Experimental Norms to reform the Urban Development Plan (PDU) of the San Jose Municipality. The main objective was to analyze the experiences these different stakeholders had when applying proposals in informal settlements to create special norms or guidelines that can help to intervene with the urban planning of these settlements. These workshops allowed for several conclusions in different topics to be obtained, such as formalities, general and specific rules for design criteria, minimum access dimensions, design criteria in health, citizen security, and emergencies [26]. This project was abandoned as people with legal tenure began to complain, as they stated that it was unfair for people with no legal tenure of the land to receive special treatments and permissions for their constructions.

Later, in 2017, the National Company of Force and Light (CNFL)—one of the stakeholders involved in the 2007 formalization project—began grouped metering and network shielding projects in informal settlements. The main objective was to formalize the users to avoid electrical and money losses for the company, but it drove several institutions to look for solutions to improve the settlements they were intervening in. For instance, in Tejarcillos (another settlement visited by the author), after CNFL installed the electrical posts and made the connections, the San Jose Municipality intervened and increased the width of the main road. Also, Ebi, a company in charge of solid waste treatment, donated a hook lift dumpster for the community to throw their waste in one place and make it easier to collect at the site. All these improvements not only enhanced the physical features of the informal settlement but also the feeling of belonging of the dwellers, which in turn, resulted in communities committed to taking responsibilities to keep the informal settlement safe.

In 2021, a law was proposed to the Legislative Assembly aiming to transform, rehabilitate, regenerate, regularize, and enhance the condition of the informal settlements in order to provide a better quality of life for the inhabitants. The law would be carried out through the creation of intervention zones, in which urban renewal programs are carried out, to allow for the relaxation of urban regulations and the provision of public services and infrastructure, as well as the granting of property titles [8]. This law is under revision and waiting for the approval of the plenary.

## 5 Lessons Learned for Canada from Costa Rica

Code development processes between Canada and Costa Rica differ where Canada uses a broad-based consensus process to update codes, and Costa Rica adopted the NFPA Standard 101 and has been slowly updating the information to tailor it to Costa Rican buildings and infrastructure before adopting it as an official regulation. However, while informal settlements are an issue in both countries and expected to grow, neither country has indicated any move toward incorporating or recognizing the risk of informal settlement fires into building and fire codes. However, as noted previously, in Costa Rica, several actors are trying to formalize the informal settlements, although so far it has not been possible. Canada has not made any indication that they will attempt to formalize informal settlements; however, decriminalizing homeless encampments and ensuring homeless encampments meet minimum safety requirements and have access to basic services which are identified as key principles in *A National Protocol for Homeless Encampments in Canada* [20]. Canada aims to reduce homelessness by 50% in 10 years through its National Housing Strategy [14], however, it has not included regulation or improvement of informal settlements as part of its directives, instead focusing funding on creating new affordable housing options and improving its shelter system [16].

With Canadian municipalities' history of criminalizing homeless encampments, following Costa Rica's lead to conduct workshops and studies tailored specifically to informal settlements as a preliminary step is a critical one, not only to fill existing knowledge gaps from a lack of information, but also to generate trust within a marginalized community that has been found to distrust government social services and supports [28]. Additionally, creating a specific strategy on informal settlements would help address the *National Housing Strategy* goal of reducing homelessness by 50% as informal settlements, homelessness and affordable housing are all interconnected. It would also help Canada fulfill its human rights obligations under both international and national laws.

## 6 Limitations

Statistics on homelessness and homeless encampments in Canada are largely unavailable or patchworked together from different jurisdictions. Homeless encampments are often not reported as their own category and grouped with "unsheltered", and as such, any demographic data may not be entirely applicable. There is also a lack of fire incident data in homeless encampments available which makes it difficult to gauge the true risk and dangers present. In addition to lacking fire incident data, there is a research gap in investigating fire risks and fire spread within North American informal settlements as there are not many studies available. There is also a lack of

studies related to the community within informal settlements and if and how they navigate safety risks, such as fire, within encampments.

There is also a lack of information regarding Costa Rican informal settlements' physical characteristics, fire record data, features of the communities such as demographic, social, economic data, and so forth. In addition, the available information is outdated or incomplete; thus, making an accurate characterization of the diverse informal settlements of the country is extremely difficult. Similar to Canada, there is also lack of studies on the community and community organization within informal settlements which was observed during a previous site visit by the authors.

## 7 Conclusions

Neither Canada's nor Costa Rica's code development processes lend themselves easily to incorporating informal settlement regulations into their codes. As informal settlements are a globally, growing issue, both countries need to develop approaches in order to address informal settlements and informal settlement safety. Costa Rica is currently in an early stage of preliminary investigations. The stakeholders are trying to understand the issue of informal settlements and how to begin addressing it, aiming to find out what are the future steps to propose solutions for different problems related to the informal settlements. While there have been some Canadian initiatives to begin addressing homeless encampments, from a government standpoint, priority has been given to homelessness prevention, improved shelter access and quality, and increased availability of affordable housing.

It is crucial to conduct surveys and research studies to create and update information on informal settlements and their conditions to fill knowledge gaps. This would give better and more accurate recommendations for each settlement, as all of them have different characteristics that made them unique. Additionally, more research to identify and address fire risks within informal settlements is required in order to better provide recommendations to improve the fire safety of its residents. Fire spread within informal settlements and tent encampments is understudied and would contribute to understanding fire risks in informal settlements. However, while both countries are working to address the issue of informal settlements, with the anticipated increase in informal settlements due to population growth and worsening of poverty and the affordable housing crises, these measures may not be sufficient.

**Acknowledgements** Authors thank Chloe Jeanneret for assistance in copyediting the manuscript. NSERC Canada is acknowledged through its CGS scholarship.

## References

1. Araya M (2022) Email: Historia de las regulaciones de Costa Rica
2. Arup (2018) A framework for fire safety in informal settlements. Arup
3. Asamblea Legislativa (2002) Ley del Benemérito Cuerpo de Bomberos de Costa Rica. Sistema Costarricense de Información Jurídica, Costa Rica
4. Asamblea Legislativa (2005) Propuesta de Manual de disposiciones técnicas Generales al reglamento sobre Seguridad Humana y Protección Contra Incendios. Sistema Costarricense de Información Jurídica, Costa Rica
5. Calder K, Weckman E (2020) Identifying implicit risk in the national building code of Canada. In: International performance-based codes & fire safety design methods conference. The Society of Fire Protection Engineers, Auckland, New Zealand
6. Canadian Commission on Building and Fire Codes (2015a) National building code of Canada: 2015. National Research Council of Canada, Ottawa, Canada
7. Canadian Commission on Building and Fire Codes (2015b) National fire code of Canada: 2015. National Research Council of Canada, Ottawa, Canada
8. Chacon Monge LF (n.d.) Law project: ley de transformación y titulación de asentamientos humanos informales e irregulares. Asamblea Legislativa, Costa Rica
9. Chown A (2016) Tiny houses in Canada's regulatory context: issues and recommendations. Edmonton, Canada
10. City Manager (2021) COVID-19 response update: protecting people experiencing homelessness and ensuring the safety of the shelter system. Toronto, Canada
11. City of Toronto (2021a) 2021 street needs assessment. Toronto, Canada
12. City of Toronto (2021b) City of Toronto reaches settlement on tiny shelters. City of Toronto. <https://www.toronto.ca/news/city-of-toronto-reaches-settlement-on-tiny-shelters/>
13. Crawford B (2021) Encampments 'an expression of the human right to housing,' poverty advocate says. Ottawa Citizen, Ottawa, Canada
14. Employment and Social Development Canada (2017) Canada's national housing strategy. Ottawa, Canada
15. Employment and Social Development Canada (2019) Everyone counts 2018: highlights—preliminary results from the second nationally coordinated point-in-time count of homelessness in Canadian communities. Ottawa, Canada
16. Employment and Social Development Canada (2020) Reaching home: Canada's homelessness strategy directives. Employment and Social Development Canada. <https://www.canada.ca/en/employment-social-development/programs/homelessness/directives.html>. 7 Mar 2022
17. Employment and Social Development Canada (2022) Everyone counts—standards for participation in the coordinated count. <https://www.canada.ca/en/employment-social-development/programs/homelessness/resources/point-in-time.html>. 5 Mar 2022
18. Engineering Unit of the Costa Rican Fire Corps (2019) Análisis e investigación de incendios Enero—Diciembre 2018. San José, Costa Rica
19. FactCheckToronto (2021) Claim: in 2020, Toronto fire services responded to 253 fires in encampments. FactCheckToronto. <https://factchecktoronto.ca/2021/06/07/encampment-fires/>. 4 Mar 2022
20. Farha L, Schwan K (2020) A national protocol for homeless encampments in Canada. Ottawa, Canada
21. Fernández J (2019) Post-fire media. Costa Rica
22. Ford J, Gomez-Lanier L (2017) Are tiny homes here to stay? A review of literature on the tiny house movement. *Fam Consum Sci Res J* 45(4):394–405
23. Gaetz S, DeJ E, Richter T, Redman M (2016) The state of homelessness in Canada 2016. Toronto
24. Gibson V (2021) 'This kind of living arrangement outside is not safe,' acting fire chief says, as Toronto records first encampment fire fatality of the winter after early morning Corktown blaze. The Toronto Star, Toronto, Canada

25. Guevara Arce S (2019) Interviews with Barrio México and Pavas fire stations. San José, Costa Rica
26. Guevara Arce S (2020) Analysis of the existing information of La Carpio informal settlement, Roble Norte sector, to create the basis for future research in fire safety. Tecnológico de Costa Rica
27. Guevara Arce S, Jeanneret C, Gales J, Antonellis D, Vaiciulyte S (2021) Human behaviour in informal settlement fires in Costa Rica. *Saf Sci* 142
28. Herring C, Lutz M (2015) The roots and implications of the USA's homeless tent cities. *City* 19(5):689–701
29. Jones AM (2020) City of Toronto threatens to remove tiny shelters built to help the homeless, citing safety concerns. CTV News, Toronto, Canada
30. Kendall M, Lin S (2022) Oakland: fire in tiny homes for homeless residents raises safety concerns. The Mercury News, San Jose, USA
31. Koker N, Walls RS, Cicione A, Sander ZR, Löffel S, Claasen JJ, Fourie SJ, Croukamp L, Rush D (2020) 20 dwelling large-scale experiment of fire spread in informal settlements. *Fire Technol*
32. Koutalianous A, Radvak C, Sawatzky J, Jones S (2021) Tiny homes—an alternative to conventional housing. Vancouver, Canada
33. Meacham BJ (2010) Accommodating innovation in building regulation: lessons and challenges. *Build Res Inf* 38(6):686–698
34. Ministerio de Vivienda y Asentamientos Humanos (MIVAH) (n.d.) La Carpio—Reseña histórica. [https://www.mivah.go.cr/Documentos/precarios/Precario\\_Tugurio\\_GAM\\_Febrero\\_2005/SAN%20JOSE/SAN%20JOSE/LA%20CARPIO.pdf](https://www.mivah.go.cr/Documentos/precarios/Precario_Tugurio_GAM_Febrero_2005/SAN%20JOSE/SAN%20JOSE/LA%20CARPIO.pdf). 7 Mar 2022
35. National Research Council of Canada (2020) Canada's national model codes development system. <https://nrc.canada.ca/en/certifications-evaluations-standards/codes-canada/codes-development-process/canadas-national-model-codes-development-system>. 2 Mar 2022
36. Samson S (2022) Winnipeg homeless camps getting burn barrels. CBC News, Winnipeg, Canada
37. The World Bank (2018) Population living in slums (% of urban population). <https://data.worldbank.org/indicator/EN.POP.SLUM.UR.ZS?end=2018&start=1990&view=chart>. 2 Mar 2022
38. The World Bank (n.d.) Population living in slums (% of urban population)
39. UN-Habitat (2015) Habitat III issue papers—22: informal settlements. United Nations Conference on Housing and Sustainable Urban Development, UN-Habitat, Quito, Ecuador
40. UN-Habitat (2017) Human rights in cities handbook series. Volume I. The human rights based approach to housing and slum upgrading. Nairobi
41. UN-Habitat (2018) Pro-poor climate action in informal settlements. Nairobi, Kenya
42. Unidad de Ingeniería (2013) Manual de disposiciones técnicas generales sobre Seguridad Humana y Protección contra Incendios. Benemérito Cuerpo de Bomberos de Costa Rica, San José, Costa Rica
43. United Nations (n.d.) Global issues—population. <https://www.un.org/en/global-issues/population>. 2 Mar 2022
44. Wong A, Chen J, Dicipulo R, Weisse D, Sleet DA, Francescutti LH (2020) Combatting homelessness in Canada: applying lessons learned from six tiny villages to the Edmonton Bridge healing program. *Int J Environ Res Public Health* 17(17):6279
45. Young MG, Abbott N, Goebel E (2017) Telling their story of homelessness: voices of Victoria's Tent City. *J Soc Distress Homeless* 26

# Coastal Hotels and Resorts: Infrastructure Asset Management System Model



**Athnasious Ghaly, Mahmoud Amin, Tesfu Tedla, Ossama Hosny,  
and Hatem Elbehairy**

**Abstract** For the past few decades, many attempts have been made in order to develop infrastructure asset management models which simulate and optimize asset elements' condition and maintenance plan. Organizations owning/operating hotels and resorts are of great need of asset management models to develop plans that efficiently manage their valuable asset to keep it in good conditions and achieve customer satisfaction. In addition, these plans should be within a certain budget and can accurately predict future deterioration/conditions of the different elements in the hotel/resort to keep on providing the required services. Several studies show that coastal structures have the highest deterioration rates due to harsh environmental exposure. In this research, an attempt has been made to develop an infrastructure asset management model for the Coastal Hotels and Resorts that can improve resource management and the overall condition of the asset. This model was built considering different factors that might affect the user cost or the required condition of the asset such as: star rating, occupancy. Elements' qualities and types are used as input factors that predict the deterioration behavior of elements based on data acquired from literature and data collected from the field in form of surveys and inspection histories. The deterioration behaviors for the elements were determined by using different approaches; some were estimated using deterministic prediction models, while others were predicted using Markov chain or linear method. The model was applied on a case study and run for two scenarios: the first scenario was to minimize the total cost for the following 24 months while achieving a minimum overall condition; and the second was to maximize the overall condition with constrained budget. Finally, a life cycle cost analysis was conducted for the asset over a 10-year period to investigate the impact of the different material types and quality used in construction on the total life cycle cost.

**Keywords** Coastal Hotels and Resorts · Infrastructure asset management

---

A. Ghaly (✉) · M. Amin · T. Tedla · O. Hosny · H. Elbehairy  
Department of Construction Engineering, The American University in Cairo, Cairo, Egypt  
e-mail: [athnasious93@aucegypt.edu](mailto:athnasious93@aucegypt.edu)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering  
Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_57](https://doi.org/10.1007/978-3-031-34593-7_57)

917

## 1 Background and Literature Review

Coastal structures are exposed to harsh environment; therefore, they require high quality of execution. This extreme weather condition includes high variability of temperature, air humidity accompanied with high air salinity saturation. This environmental condition causes severe damages to buildings' structure and envelopes. According to Edirisinghe et al. [6], the building closer to the sea has a higher deterioration rates than buildings located away from the coastline. Coastal hospitality building, like hotels, are dynamic, complicated, and very costly to operate. Coastal hotels entail high operational and maintenance measures to ensure Guest's satisfaction [4]. For instance, having less functionality or failure of HVAC is unacceptable as it will affect the client's comfort and their overall assessment of the hotel. Thus, the hotel management should have an effective asset management system to closely monitor the deterioration of the building and maintain the high performance of the hotel. Ghazi [7] has studied the maintenance management practices and efficiency in hotels through conducting a survey. A questionnaire was distributed to 34 hotel's maintenance managers around Egypt to get data about main factors affecting maintenance process in coastal/hotel projects. Analysis of the questionnaire indicates that having a standardized maintenance management plan, computer-based information system to organize the maintenance work, and allocating sufficient fund are the main factors that influence the maintenance efficiency in hotels. Therefore, the hotel management needs a tool to assist them in keeping track of the repair/maintenance plan and determining the appropriate repair decision based on the condition performance of each element. However, there is rare research done in this area, and most of the past contributions attempted to analyze the maintenance procedures in hotels and the factors affecting renovation frequency [10].

## 2 Objective and Scope

The objective of this research was to develop an infrastructure asset management model. Coastal Hotels and Resorts that would support decision-makers in efficiently manage their resources and maximizing its overall condition of the asset. In order to reach this objective, an optimization model was built considering star rating, occupancy, and other factors that might affect the condition and deterioration of the asset, its repair cost, and both the condition and repair user cost. Two influence diagrams were developed to describe the relation between the model components and to illustrate how the model works. As shown in Fig. 1, the repair option controls many factors which should be considered before decision-making.





Facilities	Elements													
Reception	Columns	Slabs&Beams	L.B.Walls	Walls	Doors	Windows	Paint	Flooring	GRC	Lighting	Furniture	-	-	-
Restaurant	Columns	Slabs&Beams	L.B.Walls	Walls	Doors	Windows	Paint	Flooring	GRC	Lighting	Furniture	-	-	-
20 Guest Room	Columns	Slabs&Beams	L.B.Walls	Walls	Doors	Windows	Paint	Flooring	GRC	Lighting	Furniture	-	-	-
Administrative	Columns	Slabs&Beams	L.B.Walls	Walls	Doors	Windows	Paint	Flooring	GRC	Lighting	Furniture	-	-	-
Beach Seating	-	-	-	-	-	-	-	-	-	-	-	Beach Seating	Umbrellas	-
MEP	-	-	-	-	-	-	-	-	-	-	-	-	-	HVAC

**Fig. 3** Elements in each facility

## 3.2 Condition Assessment and Predictions

### 3.2.1 Condition Assessment Systems

The condition assessment uses a numerical rating system which ranges from nine to three. Nine was considered as the best condition, whereas three was taken as the worst condition. The star rating in this model, which varies from three to five stars, affects both the overall acceptable condition of the asset and the user cost.

### 3.2.2 Deterioration Predictions

The deterioration behavior for the elements was determined by different approaches, some were done using deterministic approach, and others were done using Markov or linear method. According to Marteinsson [14], the factor method, a deterministic approach, takes into account the variability of element quality, outdoor environmental factors, and other factors on the service life of individual or entire building elements. Consequently, the deterioration method of each element adopted is shown in Fig. 4.

#### Structural Elements

According to the acquired data from field and experts' surveys, different deterioration curves for the structural elements have been developed. The deterioration behavior of the concrete structural elements including the columns, slabs and beams, and loadbearing walls in this model was determined based on the quality factors of the concrete such as: cement content, water–cement ratio, usage of mineral admixtures, and cement block quality. Deterioration behavior curves were developed from the data acquired; then using regression analysis, a function was driven with polynomial function as shown in Fig. 5 for slabs and beams.

ELEMENTS	DETERIORATION CURVES
COLUMN	DETERMINISTIC
SLAB & BEAMS	DETERMINISTIC
LOAD BEARING WALLS	DETERMINISTIC
Windows	DETERMINISTIC
Doors	DETERMINISTIC
Paint	DETERMINISTIC
Walls	MARKOV CHAIN
GRC	LINEAR
Flooring	MARKOV CHAIN
Furniture	MARKOV CHAIN
Lighting	LINEAR
Beach Seating	LINEAR
Beach Umbrella	LINEAR
HVAC	MARKOV CHAIN

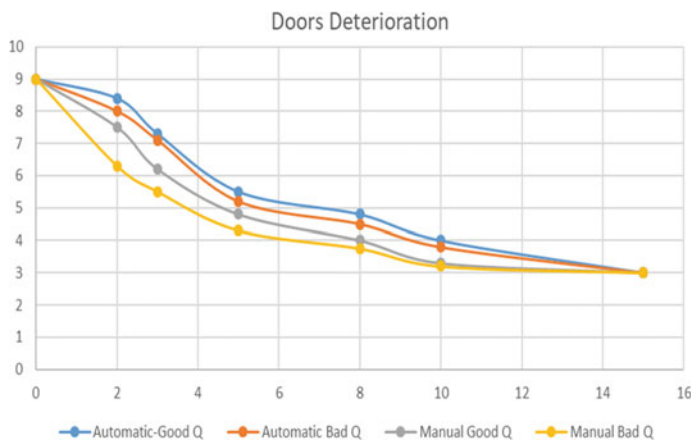
Fig. 4 Elements’ deterioration prediction methods

Life Time	Cases	Deterioration Function
10	(Slabs&Beams) - CC(200-300) & WC(0.4-0.5)	$y = 0.0072x^3 - 0.14x^2 + 0.08x + 9$
15	(Slabs&Beams) - CC(200-300) & WC(0.35-0.4)	
	(Slabs&Beams) - CC(200-300) & WC(0.4-0.5)-(M.A)	$y = 0.0001x^4 - 0.0036x^3 + 0.0103x^2 - 0.24x + 9$
20	(Slabs&Beams) - CC(300-400) & WC(0.4-0.5)	
	(Slabs&Beams) - CC(200-300) & WC(0.35-0.4)-(M.A)	$y = 4E-05x^4 - 0.001x^3 - 0.0048x^2 - 0.1258x + 9.0079$
25	(Slabs&Beams) - CC(300-350) & WC(0.3-0.35)	
	(Slabs&Beams) - CC(300-400) & WC(0.4-0.5)-(M.A)	$y = 5E-05x^4 - 0.0024x^3 + 0.0312x^2 - 0.2575x + 9.0357$
30	(Slabs&Beams) - CC(400-500) & WC(0.4-0.5)	
	(Slabs&Beams) - CC(300-400) & WC(0.35-0.4)	$y = 1E-05x^4 - 0.0007x^3 + 0.0066x^2 - 0.1286x + 9.0112$
	(Slabs&Beams) - CC(200-300) & WC(0.3-0.35)-(M.A)	
	(Slabs&Beams) - CC(300-400) & WC(0.3-0.35)	
40	(Slabs&Beams) - CC(400-500) & WC(0.4-0.5)-(M.A)	
	(Slabs&Beams) - CC(300-400) & WC(0.35-0.4) (2)	$y = 9E-07x^4 + 6E-05x^3 - 0.0075x^2 - 0.0049x + 8.9641$
50	(Slabs&Beams) - CC(400-500) & WC(0.35-0.4)	
	(Slabs&Beams) - CC(300-400) & WC(0.3-0.35)-(M.A)	$y = -1E-06x^4 + 0.0003x^3 - 0.013x^2 + 0.0631x + 8.9115$
60	(Slabs&Beams) - CC(400-500) & WC(0.3-0.35)	$y = 2E-07x^4 + 2E-05x^3 - 0.0031x^2 - 0.0171x + 8.9827$
65	(Slabs&Beams) - CC(400-500) & WC(0.35-0.4)-(M.A)	$y = 4E-05x^3 - 0.004x^2 + 0.0085x + 8.904$
75	(Slabs&Beams) - CC(400-500) & WC(0.3-0.35)-(M.A)	$y = -2E-08x^5 + 3E-06x^4 - 7E-05x^3 - 0.0029x^2 + 0.0077x + 8.9375$

Fig. 5 Slabs and beams’ sample of deterioration functions

Architectural Elements

The walls’ deterioration curves were determined from Markov series. The walls’ deterioration is depending up on different factors such as: the cement content of the mortar, quality of bricks, and availability of polymers. This is also applied to flooring and Furniture, in which deterioration behavior was predicted using Markov transition matrix from literature. On the other hand, for Paint and Doors, the deterioration curves are drawn from the field data and the curves have different ages according to the type of the door or paint. For example, the doors are assumed to be two types: automatic and manual operated, along with two different wood and coating quality. Accordingly, four curves are formed for deterioration according door type and quality, as shown



**Fig. 6** Deterioration curves of doors

in Fig. 6. Moreover, the deterioration of windows was determined from the field data. Various windows types and qualities were considered. Aluminum and wood types were also included. Aluminum quality varies according to the frame types and its ability to resist corrosion. Consequently, curves were developed to represent the conditions and service life collected.

Lighting—Beach Seating—Beach Umbrella

The deterioration rate for each of lighting, beach seating, and umbrellas are assumed to be linear. The total lifetime for lighting including fixtures and switches is assumed to be 20 years. While for the beach seating, the life span was expected to vary based on its type: 3 years for plastic and 4 years for wood. Finally, beach umbrellas have lifetime for wood and metal types, 4 and 5 years, respectively.

HVAC

The Heating Ventilation and Air Conditioner (HVAC) is one of the most critical elements in the hotel as its failure or even less functionality is not acceptable as it has a great effect on customers’ satisfaction. Therefore, it should be maintained at a higher condition and to be checked frequently. The HVAC includes several sub-components (Fig. 7). For instance, it comprises of cooling generating system, a heating generating system, distribution systems, terminal and package units, control, and others HVAC equipment. Also, it includes main components such as chillers, boiler, and other auxiliary parts. Each of these sub systems has a different deterioration rates, so it

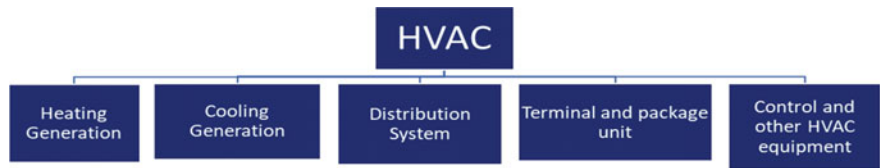


Fig. 7 HVAC sub-competent

should be calculated with a different transition matrix. However, the HVAC could be managed at high level, so it was defined as one component and only one deterioration rate was developed for the whole HVAC element. The transition matrix from Grussing [9] was adopted.

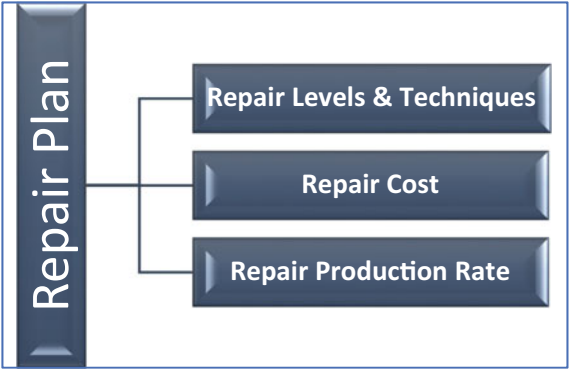
3.3 Repair Plan Model

In this model, each element has two levels of repair that can be applied depending on the condition of the element, and it is assumed that the repair restores the element to its initial condition which is 9. Each level of repair has different technique, cost, and productivity rate (Fig. 8).

3.3.1 Repair Levels and Techniques

There are different levels or types of repair that could be applied to each element in the asset; in this model, two levels of repair were used: light and heavy. For example, for columns, the light repair includes removing the concrete cover above the corroded reinforcement, then removing the corrosion layer with angle grinder or sandblasting machine, along with coating the old reinforcement with protective coating, and finally

Fig. 8 Repair plan



ELEMENTS	LIGHT REPAIR	HEAVY REPAIR
COLUMN	<ul style="list-style-type: none"> <li>* Removing Concrete Cover</li> <li>* Sandplasting the Corroded Layer of the Reinforcement</li> <li>* Applying a Corrosion Protective layer to the Reinforcement</li> </ul>	<ul style="list-style-type: none"> <li>* Removing Concrete Cover</li> <li>* Sandplasting the Corroded Layer of the Reinforcement</li> <li>* Applying a Corrosion Protective layer to the Reinforcement</li> </ul>
SLAB & BEAMS	<ul style="list-style-type: none"> <li>* Re-applying the concrete cover by Shotcreting</li> </ul>	<ul style="list-style-type: none"> <li>* Fixing a full new Reinforcement to the concrete element</li> <li>* Re-applying the concrete cover by Shotcreting</li> </ul>
LOAD BEARING WALLS	<ul style="list-style-type: none"> <li>* Cleaning and removing paint and plaster around the crack.</li> <li>* Opening the Crack with Angle Grinder.</li> <li>* Close the surface of the opened crack except at points of injection.</li> <li>* inject the crack with either epoxy or grout.</li> </ul>	<ul style="list-style-type: none"> <li>* Cleaning and removing paint and plaster around the crack.</li> <li>* Opening the Crack with Angle Grinder.</li> <li>* Close the surface of the opened crack except at points of injection.</li> <li>* inject the crack with either epoxy or grout.</li> <li>* Fixing a steel reinforced mesh to the wall.</li> <li>* Applying a layer of shotcrete.</li> </ul>
Windows	<ul style="list-style-type: none"> <li>* Repainting + changing ligh accessories</li> </ul>	<ul style="list-style-type: none"> <li>* Sanding and repairing wood/Aluminum defects.</li> <li>* Repainting + changing major accessories</li> </ul>
Doors	<ul style="list-style-type: none"> <li>* Repainting + changing ligh accessories</li> </ul>	<ul style="list-style-type: none"> <li>* Sanding and repairing wood/Aluminum defects.</li> <li>* Repainting + changing major accessories</li> </ul>
Paint	<ul style="list-style-type: none"> <li>* Cleaning &amp; Light Scratches Repainting</li> </ul>	<ul style="list-style-type: none"> <li>* Major Repainting</li> </ul>
Walls	<ul style="list-style-type: none"> <li>* Minor haircrack injection with bonding material</li> </ul>	<ul style="list-style-type: none"> <li>* major crack injection with bonding material</li> </ul>
GRC	<ul style="list-style-type: none"> <li>* Cleaning or ligh Repair for minor cracks</li> </ul>	<ul style="list-style-type: none"> <li>* Repair for major cracks</li> </ul>
Flooring	<ul style="list-style-type: none"> <li>* Raplacing minor (few) Defected or broken Tiles</li> </ul>	<ul style="list-style-type: none"> <li>* Raplacing Major (Several) Defected or broken Tiles</li> </ul>
Furniture	<ul style="list-style-type: none"> <li>* Minor Scratches repainting</li> </ul>	<ul style="list-style-type: none"> <li>* Major Scratches or breaking fixing &amp; repainting</li> </ul>
Lighting	<ul style="list-style-type: none"> <li>* Minor defect repair or bulb changing</li> </ul>	<ul style="list-style-type: none"> <li>* Major defect in lighting fixtures or connections repair</li> </ul>
Beach Seating	<ul style="list-style-type: none"> <li>* Repainting Wooden Seating or Fixing minor Plastic Crack</li> </ul>	<ul style="list-style-type: none"> <li>* Repairing Broken Wooden Seating or Replacing major broken Plastic</li> </ul>
Beach Umberalla	<ul style="list-style-type: none"> <li>* Repaint either wooden or Steel</li> </ul>	<ul style="list-style-type: none"> <li>* Repairing Broken Wooden umbrella or repairin a severely corroded Steel umbrella</li> </ul>
HVAC	<ul style="list-style-type: none"> <li>* Changing minor parts like fans, small parts of the controle unit inside facilities</li> </ul>	<ul style="list-style-type: none"> <li>* Changing major parts like motor , controle unit inside facilities</li> </ul>

**Fig. 9** Repair type and corresponding repair plan for each element

replying the concrete cover using shotcreting or batching. However, for column heavy repair, RC Jacketing is applied, with same procedure of light repair along with the addition of new reinforcement with anchored shear connector. Figure 9 shows the description of both light and heavy repairs applied to each of the 14-element type.

### 3.3.2 Repair Costs and Productivity Rates

Based on the type of repair (light/heavy) and type of element, each has a different repair costs and productivity rates, as shown in Figs. 10 and 11. Productivity rates were determined from Gordian Group [8] “Facilities maintenance and repair cost data” manual.

ELEMENTS	UNIT	Element Type	LIGHT REPAIR	HEAVY REPAIR
COLUMN	EGP/m2	-	1000	2700
SLAB & BEAMS	EGP/m2	-	650	1500
LOAD BEARING WALLS	EGP/m2	-	750	4500
Windows	EGP/UNIT	Wood	200	350
		Aluminum	150	400
Doors	EGP/UNIT	Manual	300	450
		Automatic	500	1000
Paint	EGP/m2	Oil Base	45	100
		Water Base	30	65
Walls	EGP/m2	-	30	50
GRC	EGP/m2	-	200	800
Flooring	EGP/m2	Marble	200	800
		Ceramic	50	200
Furniture	%EGP/UNIT	-	5	30
Lighting	%EGP/UNIT	-	2	4
Beach Seating	EGP/UNIT	Wood	600	4000
		Plastic	300	2000
Beach UMBERALLA	EGP/UNIT	Wood	400	1500
		Steel	300	2500
HVAC	EGP/UNIT	-	150000	300000

Fig. 10 Repair cost of each element from field

ELEMENTS	UNIT	LIGHT REPAIR	HEAVY REPAIR
COLUMN	m2/d	5	2.5
SLAB & BEAMS	m2/d	12.5	7.2
LOAD BEARING WALLS	m2/d	10	4
Windows	UNIT	1	1
Doors	UNIT	1	1
Paint	m2/d	140	70
Walls	m2/d	90	45
GRC	m2/d	3	1.5
Flooring	m2/d	50	17
Furniture	Unit	1	0.5
Lighting	Day/unit	0	1
Beach Seating	Day/unit	0	0
Beach UMBERALLA	Day/unit	0	0
HVAC	Unit	1	0.5

Fig. 11 Repair productivity rates

Condition	User Cost	Occupancy	Condition	User Cost
Reception And Restaurant	Low	3-6	8% x number of rooms x Avg room rate x (days/month)	
		6-8	3% x number of rooms x Avg room rate x (days/month)	
	Medium	3-6	13% x number of rooms x Avg room rate x (days/month)	
		6-8	6% x number of rooms x Avg room rate x (days/month)	
	High	3-6	20% x number of rooms x Avg room rate x (days/month)	
		6-8	10% x number of rooms x Avg room rate x (days/month)	
Guest Room	Low	3-6	20% x Avg room rate x (days/month)	
		6-8	10% x Avg room rate x (days/month)	
	Medium	3-6	35% x Avg room rate x (days/month)	
		6-8	15% x Avg room rate x (days/month)	
	High	3-6	50% x Avg room rate x (days/month)	
		6-8	20% x Avg room rate x (days/month)	
Adminstrative				
HVAC	Low	3-6	7.5% x number of rooms x Avg room rate x (days/month)	
		6-8	3% x number of rooms x Avg room rate x (days/month)	
	Medium	3-6	12.5% x number of rooms x Avg room rate x (days/month)	
		6-8	5% x number of rooms x Avg room rate x (days/month)	
	High	3-6	20% x number of rooms x Avg room rate x (days/month)	
		6-8	8% x number of rooms x Avg room rate x (days/month)	
Beach	Low	3-6	8% x number of rooms x Avg room rate x (days/month)	
		6-8	3% x number of rooms x Avg room rate x (days/month)	
	Medium	3-6	13% x number of rooms x Avg room rate x (days/month)	
		6-8	6% x number of rooms x Avg room rate x (days/month)	
	High	3-6	20% x number of rooms x Avg room rate x (days/month)	
		6-8	8% x number of rooms x Avg room rate x (days/month)	

Fig. 12 Condition user cost calculation

3.4 User Cost

3.4.1 Condition User Cost

The condition user cost occurs due to gradual loss of functionality of a certain element beyond its acceptable condition. The condition user cost is affected by two factors: occupancy rates and the condition state of each element. The occupancy is divided into three categories: low, medium, and high with rates of less than 30%; from 30 to 70%; and more than 70% of the hotel capacity, respectively. The condition state is grouped into two levels from 3 to 6 and from 6 to 8. The condition user cost is calculated based on the average rooms prices as shown in Fig. 12.

3.5 Repair User Cost

The repair user cost is incurred due to applying the repair activity of certain element which will execrate a loss to the hotel owners. The cost is calculated based on the maximum duration of the repair activities which take place in each month for the certain facility as shown in chart in Fig. 13.

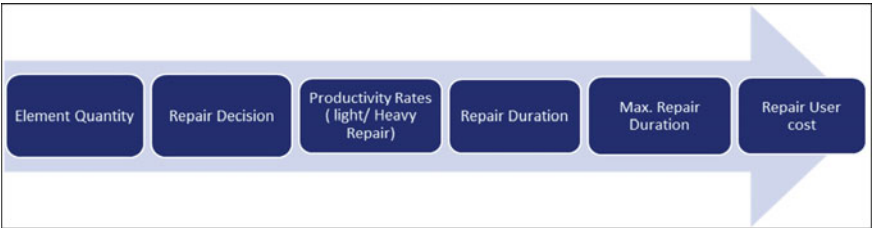


Fig. 13 Repair user cost chart

4 Optimization

In order to reach an optimum solution, two optimization approaches were applied. The first approach was to minimize the total cost, including the repair cost, repair user cost, and condition user cost, with a constraint on the overall condition state, while the second approach was to maximize the average overall condition of the hotel throughout the 24 months with a budget constraint.

5 Model Validation

5.1 User Inputs

For validating the model, the quantities, duration since previous repair, and defining distinct quality factors for each element were inserted. The case study inputs are presented in Figs. 14, 15, and 16.

Facility/Element	Quantity /Element																			
	Reception	Restaurant	GR 1	GR 2	GR 3	GR 4	GR 5	GR 6	GR 7	GR 8	GR 9	GR 10	GR 11	GR 12	GR 13	GR 14	GR 15	GR 16	GR 17	GR 18
Columns	150	200	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
Slabs&Beams	300	400	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50
L.B.Walls	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Walls	360	400	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
Doors	5	5	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Windows	6	6	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Paint	660	800	140	140	140	140	140	140	140	140	140	140	140	140	140	140	140	140	140	140
Flooring	300	400	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50
Gypsum&GHC	25	10	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Lighting	1	30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Furniture	-	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Beach Seating	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Umbrellas	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HVAC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Fig. 14 Element quantities



Facility/Element	Columns			Slabs&Baams			L.B.Walls		
	W/C	C.C (kg/m3)	M <sub>a</sub> Admix	W/C	C.C (kg/m3)	M <sub>a</sub> Admix	Block Q.	Mortor Q.	Mortar C.
Reception	0.35-0.4	350-420	No	0.35-0.4	300-350	No	Low	Low	High
Resturant	0.35-0.5	350-421	No	0.35-0.5	300-351	No	Low	Low	High
Administrative	0.35-0.6	350-422	No	0.35-0.6	300-352	No	Low	Low	High
20 Guest Rooms	0.35-0.6	350-422	No	0.35-0.6	300-352	No	Low	Low	High
Beach	-	-	-	-	-	-	-	-	-
MEP	-	-	-	-	-	-	-	-	-

Fig. 15 Quality factors

Facility/Element										
	Reception	Restaurant	GR 1	GR 2	GR 3	GR 4	GR 5	GR 6	GR 7	C
Columns	170	165	145	145	145	145	155	155	155	
Slabs&Beams	170	165	145	145	145	145	155	155	155	
L.B.Walls	170	165	145	145	145	145	155	155	155	
Walls	78	74	46	40	63	55	45	54	37	
Doors	20	20	10	17	15	20	24	19	11	
Windows	12	12	3	3	3	3	3	7	7	
Paint	11	11	6	7	9	4	3	5	2	
Flooring	65	60	24	35	19	12	36	23	9	
Gypsum&GRC	63	59	31	25	48	40	30	39	22	
Lighting	27	25	17	18	20	17	18	20	16	
Furniture	-	-	34	45	29	22	46	33	19	
Beach Seating	-	-	-	-	-	-	-	-	-	
Umbrellas	-	-	-	-	-	-	-	-	-	
HVAC	-	-	-	-	-	-	-	-	-	

Fig. 16 Duration since last repair

## 5.2 Methodology

According to the input data, the current condition state of each element is calculated based on its age and deterioration category. The age after repair is calculated linked with the repair decision, variable, which automatically determines the condition after repair, as in Fig. 17, and the associate repair cost. Consequently, the overall condition state of each facility is calculated in each month based on defined weights as shown in Fig. 18.

The condition user cost is determined based on the pervious mentioned equations as shown in Fig. 19.

The repair duration is calculated by dividing the repair quantity by the repair productivity rates. Then, the repair user cost is determined by calculating the

Element	Condition After Repair																									
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
Columns	4.0671	3.9741	3.92	3.87	3.82	3.77	3.72	3.67	3.62	3.57	3.52	3.47	3.42	3.37	3.32	3.27	3.22	3.17	3.12	3.07	3.02	2.97	2.92	2.87	2.82	2.77
Shaft/Ends	3.9067	3.9733	3.96	3.9487	3.9333	3.92	3.9067	3.8933	3.88	3.8667	3.8533	3.84	3.8267	3.8133	3.8	3.7867	3.7733	3.76	3.7467	3.7333	3.72	3.7067	3.6933	3.68	3.67	3.66
Lb/Vals	3	3.9395	3.95	3.9055	3.96	3.975	3.97	3.965	3.96	3.955	3.95	3.945	3.94	3.935	3.93	3.925	3.92	3.915	3.91	3.905	3.9	3.895	3.89	3.885	3.88	3.875
Lb/Vals	3	3.9451	3.895	3.932	3.7623	3.625	3.674	3.695	3.665	3.665	3.4507	3.397	3.4048	3.31	3.2813	3.2955	3.2239	3.182	3.2622	3.255	3.2037	3.073	3.0432	3.054	3.7637	
Lb/Vals	3	3.8375	3.775	3.6523	3.72	3.4372	3.325	3.2522	3.1	3.7673	3.728	3.728	3.745	3.7	3.6275	3.7	3.6625	3.5	3.4733	3.325	3.1	3.7523	3.75	3.75	3.75	3.75
Windows	4	3.857	3.84	3.78	7.2	7.956	7.011	9.977	6.2222	6.7228	6.3333	5.9384	3.6644	3	4.8	7.8	7.2	7.956	7.011	9.977	6.2222	6.7228	6.3333	5.9384	3.6644	3
Panels	4.3857	3	7.2	6.6	6	5.75	5.3343	5.3343	4.3429	4.7571	4.5708	4.3857	3	7.2	6.6	6	5.75	5.3343	5.3343	4.3429	4.7571	4.5708	4.3857	3	7.2	6.6
Support	3.2704	3	3.8086	3.8163	4.2428	3.2832	3.0477	3.175	3.6171	3.7447	3.2683	3.7059	3.8564	3.76	3.8035	3.51	3.4408	3.301	3.2678	3.41	3.0048	3.5032	3.807	3.5032	3.5032	
Grout/Concrete	3	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	3.975	
Lighting	3	3.975	3.95	3.925	3.9	3.875	3.85	3.825	3.8	3.775	3.75	3.725	3.7	3.675	3.65	3.625	3.6	3.575	3.55	3.525	3.5	3.475	3.45	3.425	3.4	3.4

**Fig. 17** Condition after repair of reception elements

[illegible]

**Fig. 18** Overall monthly condition state of the reception and restaurant facilities

[illegible]

**Fig. 19** Condition user cost

maximum repair duration per month and multiple it by the estimated loss due to close part of the reception or a room to execute the repair activity as shown in Fig. 20.

The total cost is calculated by adding the repair cost, condition user cost, and repair user cost.

### 5.3 Results

The model was run one time for each approach, and the parameters for each approach are represented in Fig. 22, including objective and constraints. Two approaches were utilized to determine and summarize the short-term repair plan for the next 24 months. The first approach was subjected to maintain the overall condition of each facility

Curren	Repair User Cost																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0	50000	50000	0	40000	10000	0	0	0	90000	30000	90000	0	0	40000	90000	40000	10000	40000	0	30000	40000	40000	10000	0
0	60000	30000	150000	60000	40000	0	0	0	0	40000	120000	30000	50000	0	150000	0	0	0	0	50000	40000	40000	0	30000

Fig. 20 Repair user cost

throughout the entire duration, higher than or equal 6 while minimizing the total cost. While the second approach the average overall condition of the hotel for the whole duration was minimized under a constrain on the allocated repair budget of 5.5 million, and same constraint as first approach of minimum overall condition of 6 for each facility throughout the entire duration (Fig. 21).

Optimization	
Approach 1	Approach 2
Minimizing Total Cost	Maximizing Overall Condition
Constrain: Minmum Condition of each facility /month => 6	Constrain: Minimum condition of each facility /month => 6
	Total repair Cost <= 5.5 million

Fig. 21 Optimization approaches

Optimization Output	Total 24 Months User Cost		Total 24 Months Repair Cost		Total Cost		Overall Condition	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
	EGP 6,831,000.00	EGP 6,961,000.00	EGP 3,720,220.00	EGP 5,444,600.00	EGP 10,551,220.00	EGP 12,405,600.00	7.84	8.04

Fig. 22 Outcome of the two approaches

The outcome of both optimization approaches shows the following:

- The model successfully fulfilled both objectives, while abiding by the constraints set.
- In first approach (min total cost), the total cost was reached 10,551,220 EGP with repair cost of 3,720,220 EGP, along with average overall facility condition of 7.84.
- In second approach (max overall condition), the average overall condition of the hotel reached 8.04, along with total cost of 12,405,600 EGP and repair cost of 5,444,600 EGP.
- Although the average overall condition in second approach is improved, the total user cost slightly increased, and that can be related to repair user cost is slightly higher than the condition user cost, on the short-term analysis. This can be interpreted that the optimization model utilizes most of the repair allocated budget of 5.44 million in order to maximizes the overall condition of the hotel.

## References

1. Aryee S (2011) Hotel maintenance management. Department of Real Estate and Construction Management, pp 1–78
2. Beach Erosion Board Office of the Chief of Engineers (n.d.) Factors affecting durability of concrete in coastal structures, no 96. Department of the Army
3. Bonić Z, Čurčić GT, Davidović N, Savić J (2015) Damage of concrete and reinforcement of reinforced-concrete foundations caused by environmental effects. *Procedia Eng* 117(1):416–423. <https://doi.org/10.1016/j.proeng.2015.08.187>
4. Chan KT, Lee RHK, Burnett J (2001) Maintenance performance: a case study of hospitality engineering systems. *Facilities* 19:494–504. <https://doi.org/10.1108/02632770110409477>
5. Department of Housing and Community Development (2017) Expected lifespan guidelines, 2–5 July 2017
6. Edirisinghe R, Setunge S, Zhang G (2015) Markov model—based building deterioration prediction and ISO factor analysis for building management. *J Manag Eng* 31(6):04015009. [https://doi.org/10.1061/\(ASCE\)ME.1943-5479.0000359](https://doi.org/10.1061/(ASCE)ME.1943-5479.0000359)
7. Ghazi KM (2016) Hotel Maintenance Management Practices. *J Hotel Bus Manage* 5:136. <https://doi.org/10.4172/2169-0286.1000136>
8. Gordian Group (2008) Facilities maintenance and repair cost data. United States of America
9. Grussing MN (2015) Risk-based facility management approach for building components using a discrete Markov process—predicting condition, reliability, and remaining service life
10. Hassanien A, Losekoot E (2002) “The application of facilities management expertise to the hotel renovation process”, *Facilities*, 20(7/8), pp. 230–238
11. Ihsan B, Alshibani A (2018) Factors affecting operation and maintenance cost of hotels. *Prop Manag* 36(3):296–313. <https://doi.org/10.1108/PM-04-2017-0023>
12. Kaewunruen S, Wu L, Goto K, Najih Y (2018) Vulnerability of structural concrete to extreme climate variances. *Climate* 6(2):40. <https://doi.org/10.3390/cli6020040>
13. Lounis Z, Madanat SM (2002) Integrating mechanistic and statistical deterioration models for effective bridge management. In: 7th ASCE international conference on applications of advanced technology in transportation, pp 513–520. [https://doi.org/10.1061/40632\(245\)65](https://doi.org/10.1061/40632(245)65)
14. Marteinsson B (2003) Durability and the factor method of ISO 15686-1. *Build Res Inf* 31(6):416–426. <https://doi.org/10.1080/0961321032000105412>

15. Mohseni H (2012) Deterioration prediction of community buildings in Australia. School of Civil, Environmental and Chemical Engineering, RMIT University
16. Moncmanová A (2007) Environmental factors that influence the deterioration of materials. *Environ Deterior Mater* 28:1–25. <https://doi.org/10.2495/978-1-84564-032-3/01>
17. Riveros GA, Rosario-Pérez ME (2018) Deriving the transition probability matrix using computational mechanics. *Eng Comput (Swansea, Wales)* 35(2):692–709. <https://doi.org/10.1108/EC-02-2017-0051>
18. Seyedolshohadaie SR (2011) Modeling risks in infrastructure asset management. ProQuest dissertations and theses, Aug 2011, p 120. [http://search.proquest.com/docview/909528971?accountid=14745%5Cnhttp://sfx.fcla.edu/usf?url\\_ver=Z39.88-2004&rft\\_val\\_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+&+theses&sid=ProQ:ProQuest+Dissertations+&+Theses+Full+Text&atitle=&title=Modelin](http://search.proquest.com/docview/909528971?accountid=14745%5Cnhttp://sfx.fcla.edu/usf?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+&+theses&sid=ProQ:ProQuest+Dissertations+&+Theses+Full+Text&atitle=&title=Modelin)
19. Shi C (2004) Effect of mixing proportions of concrete on its electrical conductivity and the rapid chloride permeability test (ASTM C1202 or ASSHTO T277) results. *Cem Concr Res* 34(3):537–545. <https://doi.org/10.1016/j.cemconres.2003.09.007>
20. Silva HGC, Terradillos PG, Zornoza E, Mendoza-Rangel JM, Castro-Borges P, Alvarado CAJ (2018) Improving sustainability through corrosion resistance of reinforced concrete by using a manufactured blended cement and fly ash. *Sustainability (Switzerland)* 10(6):1–15. <https://doi.org/10.3390/su10062004>
21. Social Investment Forum (2009) 2009 annual report. Management
22. Wang S (2014) By, Dec 2014
23. Yiğiter H, Yazici H, Aydın S (2007) Effects of cement type, water/cement ratio and cement content on sea water resistance of concrete. *Build Environ* 42(4):1770–1776. <https://doi.org/10.1016/j.buildenv.2006.01.008>

# Protection of Concrete Surface from the Canadian Standard, ICRI, and ACI Perspectives



Claudiane M. Ouellet-Plamondon

**Abstract** The Canadian standard CSA A23.1:19 on concrete materials and methods of concrete construction addresses the concrete surface treatment in chapter 7 placing, finishing, and curing concrete and in chapter 8 concrete with special performance requirements. The factors affecting the abrasion resistance are specified. The surface defects are named as honeycombing, sand streaking, lift lines, variations in color, soft areas, and large surface void. Special finish of architectural formed concrete must minimize texture and color variations. The standard does not mention specific organic coating material. The standard is clear on specific forbidden action that can affect the surface finish. The International Concrete Repair Institute (ICRI) has more specific guidelines on the types of product to apply. The sealers and coating recommenders are classified as high molecular weight methacrylic sealing compounds, low viscosity epoxy sealing compounds, silane and siloxane sealing compounds, and coating compounds for concrete. The causes of damage to concrete are explained. The American Concrete Institute has a number of committees addressing surface finishing. The ACI 546.3R-14 Guide to materials selection for concrete repair has tables for sealers and anti-carbonatation coating selection based on durability factors.

**Keywords** Protection of concrete surface · Canadian standard · ICRI · ACI

## 1 Introduction

Sealers and coating can increase the durability of concrete infrastructure repairs, when well applied. From the Canadian perspective, the guidelines on the application of the coating and sealer are available in the CSA standard, the guides from the International Concrete Repair Institute (ICRI), and the American Concrete Institute. The objective of this extended abstract is to report the key point on the protection

---

C. M. Ouellet-Plamondon (✉)

École de Technologie Supérieure, Université du Québec, 1100 Notre-Dame Ouest, Montreal, QC H3C 1K3, Canada

e-mail: [Claudiane.Ouellet-Plamondon@etsmtl.ca](mailto:Claudiane.Ouellet-Plamondon@etsmtl.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_58](https://doi.org/10.1007/978-3-031-34593-7_58)

of surface concrete from the Canadian standard CSA 23.1 [2], ICRI, and the ACI guides. The best practices aim to prevent the degradation of the surface treatment. Limited information is available on the degradation mechanisms. There is worldwide interest in the best practices for coating application to prevent their degradation, as demonstrated by the Rilem technical committee on the degradation of protective surface treatment. The first step is to summarize the relevant information in standard and guides.

## 2 The Canadian Standard

The Canadian standard on concrete materials and methods of concrete construction addresses the concrete surface treatment in chapter 7 placing, finishing, and curing concrete and in chapter 8 concrete with special performance requirements. It does not mention specific organic coating material. The standard is clear in specific forbidden action that can affect the surface finish (Table 1). The conditions for underwater concreting are explained. The anti-washout admixtures are recommended to prevent mass loss (7.5.5). Regarding abrasion and wear resistance, the owner shall specify the surface treatment to the intended use of the surface (7.7.5). The factors affecting the abrasion resistance are the hardness of the aggregates, surface compressive strength, water-to-cementitious material ratio, quality and duration of curing, and the type of final finish. The finishing of formed surface must happen as soon after stripping the forms (7.10). The surface defects are named as honeycombing, sand streaking, lift lines, variations in color, soft areas, and large surface void. Surface void and color variations cannot be patched, unless they are too different from the reference sample or if there are special specifications that they must be patched. Architectural concrete is specified in the contract document (8.3). Special finish of architectural formed concrete must minimize texture and color variations. The specific finishes are specified in ACI guides. The current standard does not have any consideration regarding freeze–thaw exposure.

**Table 1** Forbidden action according to CSA 23.1

Actions	Example	Chapter section
Concreting underwater	Not under 5 °C, unless strength is sufficient; velocity of the current shall not exceed 3 m/min; maximum washout not greater than 8%	7.5.5
Finishing of formed surfaces	Surface void and color variations cannot be patched	7.10.2.4

### **3 The ICRI Perspective**

The section on the ICRI perspective is based on guide to concrete repair second edition [3]. More guide is available on surface preparation prior to sealers and coatings applications.

#### **3.1 Sealers and Coatings**

Sealers and coatings are used reduced deterioration from corrosion of rebar, freeze–thaw, carbonation, and sulfate damage. They protect concrete and repair surface damage and small cracks. They are applied on dry and cure concrete as a maintenance and repair procedure of infrastructure in overall good conditions. Sealing cracks are important to reduce the permeability of the concrete connected to its durability. The concrete treatment methods are the surface treatment with coating, treatments that can fill surface voids and cracks, and surface sealers. The main chemical groups are reviewed.

##### **3.1.1 High Molecular Weight Methacrylic (HMWM) Sealing Compounds**

The HMWM sealing is made up of methacrylic monomer and a polymerization catalyst comparable to monomer systems used in polymer concrete. It penetrates to a depth of 1.6 mm in concrete and it is more effective at sealing fine cracks in the concrete surface. It is affected by solar radiation and it can disappear after 1–2 years.

##### **3.1.2 Low Viscosity Epoxy (LVE) Sealing Compounds**

The LVE sealing compound is easier to use than HMWM, but they have a higher viscosity and they do not penetrate in cracks as much. The epoxy is made of epoxy resin with an appropriate catalyst of hardener. It better seals the cracks on the surface. LVE is also deteriorated with solar radiation in 1–2 years. In the typical Western United State climatic conditions, it is expected to last 10–20 years. Afterward, it must be reapplied.

##### **3.1.3 Silane and Siloxane (SS) Sealing Compounds**

The SS is effective and easy to use. They are used to seal concrete surfaces and cracks thinner than 0.005 mm, and they reduce water penetrated of treated concrete. For that efficiency, the sealing compound must contain 20% solid and the crack cannot



be larger than 0.25 mm. The SS is suspended in water, alcohol, or mineral spirits that evaporate after the application. Solvent-based tends to work better than water-based. Silanes cure at a high pH, while siloxanes do not need. The water beads at the surface when wet and the concrete does not change color. This sealing does not protect against prolonged inundation. They can be applied afterward. They have a limited service life of 5–7 years.

### **3.1.4 Coating Compounds for Concrete**

The surface treatment is usually with epoxy, polyurethane, or polyurea. There are four reasons to apply coating: (1) waterproofing to prevent water flow in the concrete, (2) damp proofing can seal the porosity of the concrete and prevent the absorption of the water, (3) decorative concrete is used to enhance the aesthetics, (4) barrier coating protects the concrete from the exposure to chemicals or from contamination.

## **3.2 The Damage to Concrete**

The cause of the damage to concrete is the excess concrete mix water, faulty design, construction defects, sulfate deterioration, alkali-aggregate reaction, freeze and thawing, abrasion–erosion damage, cavitation damage, and corrosion of reinforced steel, acid exposure, cracking, structural overload, and other multiple causes of damage. The organic coatings are used in crack and water leak repairs. Resin injection works with epoxy resins, polyurethanes, and methacrylic acrylates. Epoxy can rebound structural cracks still relatively dry. The polyurethanes and methacrylic acrylate can seal cracks and water leaks. To rebound a structural crack, it must be dry and exempt of dirt. Resin injection is used to seal deep cracks, while shallow cracks with sealers.

## **4 The ACI Perspective**

The ACI has a number of technical committees with mixed members from the industry, academia, and the government providing the best practices for concrete finish and repair. This article report Tables 2 and 3 adapted from the ACI 546.3R-14 Guide to materials selection for concrete repair [1]. Sealer and elastomer coatings prevent water penetration into the concrete, and elastomer coating is more efficient. They improve the durability of concrete repair. They also limit the ingress of water-soluble chemical like chloride ions into the concrete. Anti-carbonation coatings limit air and carbon dioxide penetration in the concrete and slow down the carbonation process. They also limit water penetration, but it is not their main purpose. They

**Table 2** Materials' selection guide for surface sealers (adapted from ACI 546.3R-14)

Durability factor	Silane	Siloxane	Acrylic	Epoxy	Linen seed oil
Limit water absorption	Good	Good	Good	Good	Bad
Water vapor transmission	Good	Good	Good	Fair	Fair
Penetrates concrete	Yes	Yes	No	No	No
Clear or tinted	Clear	Clear	Both	Tinted	Clear
UV resistance	Good	Good	Moderate	Moderate	No data
VOCs	Broad range	Broad range	Broad range	High	No data

**Table 3** Material selection guide for anti-carbonatation coatings (adapted from ACI 546.3R-14)

Durability factors	Acrylic	Methacrylate	Polymer-modified cementitious
CO <sub>2</sub> screening	Good	Good	Very good
Vapor transmission	Good	Good	Good
Water, chloride ion screening	Yes	Yes	Yes
Elasticity	Yes	No	No
Color	Tinted	Clear	Gray or white
Weathering resistance	Good	Good	Good
Primer	Yes	Yes	No

are used on vertical and horizontal surfaces not exposed to pedestrian and vehicular traffic.

## 5 Conclusion

This conference article presented the material selection of coatings and sealers, as well as the best practices to prevent their degradation. The ACI guide is clear on the durability factors of the sealers and coatings. The exposure to freeze–thaw is not yet considered clearly in the standard and guidelines.

## References

1. ACI (2014) 546.3R-14. Guide to materials selection for concrete repair. ACI, Farmington Hills, MI
2. CSA (2019) A23.1:19/CSA A23.2:19. Concrete materials and methods of concrete construction/ test methods and standard practices for concrete. CSA, Toronto, Canada
3. Fay KJV (2015) Guide to concrete repair, 2nd edn. U.S. Department of the Interior Bureau of Reclamation

# **Environmental Specialty: Water and Wastewater Treatment**

# Let Us Talk About Microplastic Pollution in Drinking Water Treatment



Jinkai Xue, Seyed Hesam-Aldin Samaei, and Jianfei Chen

**Abstract** Microplastic particles (MPs) have become interesting to the water research community and general public due to their global ubiquity and potential impacts on the environment and public health. MPs are plastic beads, fibers, and fragments in the micrometer size range ( $< 5$  mm), which can be ingested by aquatic fauna across a range of feeding guilds. MPs result from a wide variety of sources, including fragmentation of larger plastics and laundering of synthetic fabrics. Previous studies on animals have indicated that ingestion of MPs could lead to serious health issues, including liver stress response and tumor formation. Through literature analyses and laboratory experiments, we explored microplastic removal in various conventional and advanced water treatment processes, such as coagulation, granular media filtration, and membrane filtration. This presentation speaks to the major MP compositions (materials, sizes, shapes, and concentrations) in drinking water sources and the removal efficiency of MPs in these different drinking water treatment (DWT) processes. Current DWT processes that are purposed for particle removal are generally effective in reducing MPs in water. Various influential factors to MP removal are discussed, such as coagulant type and dose, MP material, shape and size, and water quality. It is anticipated that better MP removal can be achieved by optimizing the treatment conditions. Moreover, this talk will frame the major challenges and future research directions on MPs and nanoplastics (NPs) in DWT.

**Keywords** Microplastic pollution · Drinking water treatment

---

J. Xue (✉) · S. H.-A. Samaei · J. Chen

Environmental Systems Engineering, Faculty of Engineering and Applied Science, University of Regina, 3737 Wascana Parkway, Regina, SK S4S 0A2, Canada

e-mail: [jinkai.xue@uregina.ca](mailto:jinkai.xue@uregina.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

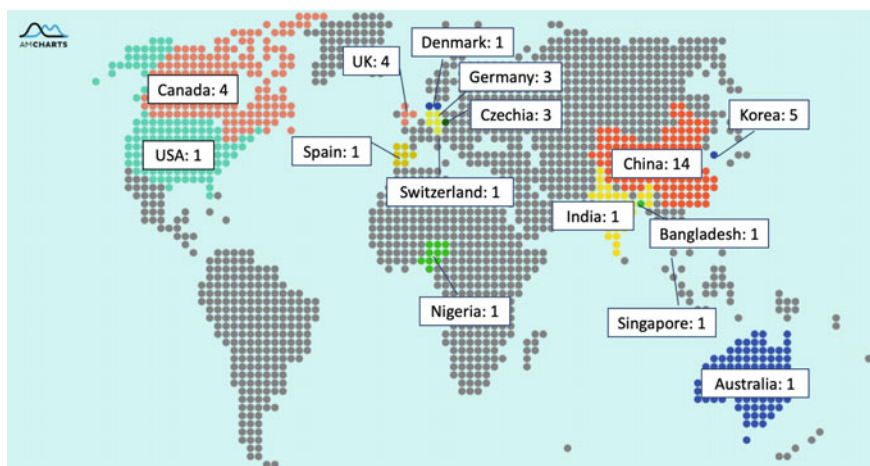
*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_59](https://doi.org/10.1007/978-3-031-34593-7_59)

## 1 Introduction

In 2019, the global plastic production was 368 million tons [21]. Over 40% of the total plastic production goes to single-use packaging purpose, leading to serious plastic pollution (Wright and Kelly, 2017). The discarded plastics go through various natural weathering processes, such as mechanical abrasion, ultraviolet (UV) radiation, leading to the creation of smaller fragments. Plastic particles that are smaller than 5 mm are defined microplastics [18, 26]. Recently, the contamination of microplastics (MPs) has become interesting to the research society due to their global ubiquity and potential adverse impacts to the environment and public health [22, 27]. In general, microplastics can cause dysfunction of the human organs and tissues and neurodegenerative diseases mostly through the translocation mechanism [9]. In addition, microplastics can induce local or systemic immune responses leading to autoimmune diseases or immunosuppression [22]. Small MPs (0.1–20  $\mu\text{m}$ ) can cross the cell membranes, the blood–brain barrier, and the placenta. On the other hand, microplastics within the size range of 20–150  $\mu\text{m}$  tend to be translocated from the gut cavity to the circulatory system and cause systematic exposure [3, 9]. Due to their hydrophobic nature and complex surface characteristics, MPs can also serve as ideal vectors of other pollutants, such as emerging contaminants of concerns (CECs), heavy metals, and even pathogens [10, 18], which further aggravates the environmental and health impacts of MPs. The most dominant originated additives in microplastics are inorganic aluminum, arsenic, bromine, copper, zinc, and phthalates [12]. The high abundance of these additives in microplastics can cause the dysfunction of the internal organs including kidneys, neurological system, and reproductive system. In fact, MPs have been detected in a broad variety of environments all over the world, including oceans, estuaries, bodies of fresh water, soils, and even the remote arctic ice [2], raising wide concerns about environmental and public health. In particular, the research community and general public are wondering whether our drinking water treatment plants (DWTPs) can adequately remove MPs (and their smaller counterparts, nanoplastics (NPs)).

To evaluate and improve DWTPs' performance in removing MPs, a number of laboratory and field studies have published [14, 26]. The country origins of the published studies on MP removal from drinking water are indicated in Fig. 1. Most of the studies on MPs in drinking water treatment were published by researchers from East Asia, Europe, and North America. Given that MPs are one of the hottest research topics in the world [28], the number of studies on MPs in drinking water treatment is surprisingly small, suggesting the infancy of direction and sheer magnitude of difficulty in studying MPs in drinking water treatment. The diverse chemical origins, surface properties, low concentrations (usually < 5000 MPs/L), and association with other substances (such as natural organic matter (NOM) and biofilms) make the extraction and detection of MPs in drinking water technical challenging, time consuming, and financially costly. It is thus convincible that only limited laboratories have sufficient research resource to perform MP studies. This literature review study aims at the following objectives: (1) to discuss the occurrence of MPs in drinking

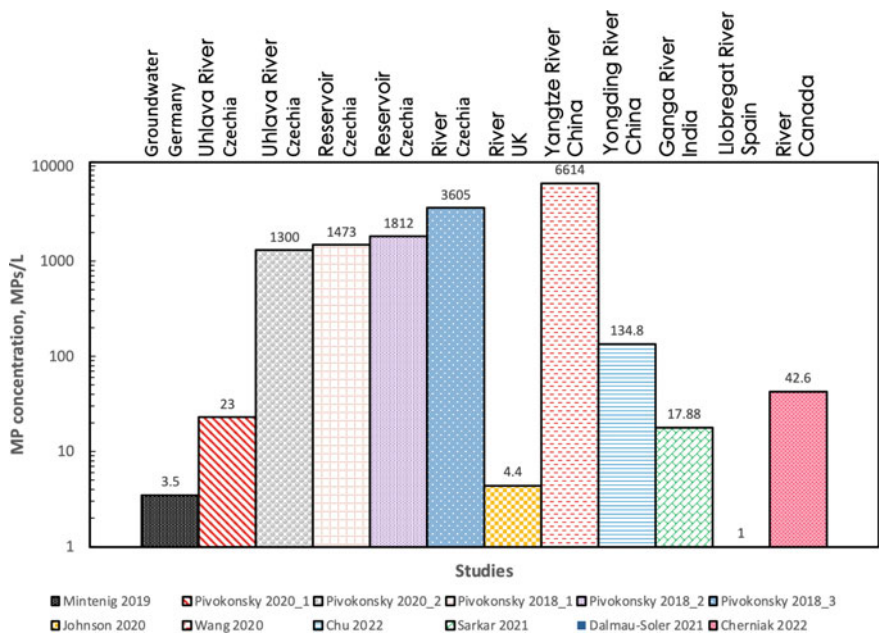


**Fig. 1** Geographic distribution of studies focused on MPs in drinking water treatment (up to March 6th, 2022)

water sources, (2) to evaluate the effectiveness of typical drinking water treatment processes in removing MPs, and (3) to identify future research directions on MPs in drinking water treatment.

## 2 MPs in Drinking Water Sources

Although MPs are ubiquitous, their occurrence is not consistent either globally or even locally. As is indicated in Fig. 2, the MP concentrations in drinking water sources significantly differ from location to location. The water bodies studied by Pivokonský et al. [20] (river water in the Czech Republic), Pivokonsky et al. [19] (reservoir and river water in the Czech Republic), and Wang et al. [25] (river water in China) had MP concentrations higher than 1000 MPs/L, with the Yangtze River in Wang et al. [25] showing 6 614 MPs/L. In contrast, the water sources in Europe exhibited the lowest MP concentrations, e.g., 1 MPs/L in the Llobregat River in Spain [7]. Although the detected MP concentrations are influenced by a number of factors, such as times, locations, and methods of sampling and analytical methods, Fig. 2 suggests that the severity of MP contamination in drinking water sources in different countries and water sheds are different. In addition, MP compositions differ significantly depending on source, sampling location, and time. Certain MP materials are more frequently detected than others in drinking water sources. The most commonly reported MPs in drinking water sources are polyethylene (PE), polypropylene (PP), polystyrene (PS), and poly(ethylene terephthalate) (PET), polyvinyl chloride (PVC), and nylon [2, 26]. As for shapes, the most dominant MP shapes in source water are fragments, films, fibers, and spheres [26]. It should be noted that most of these studies cited

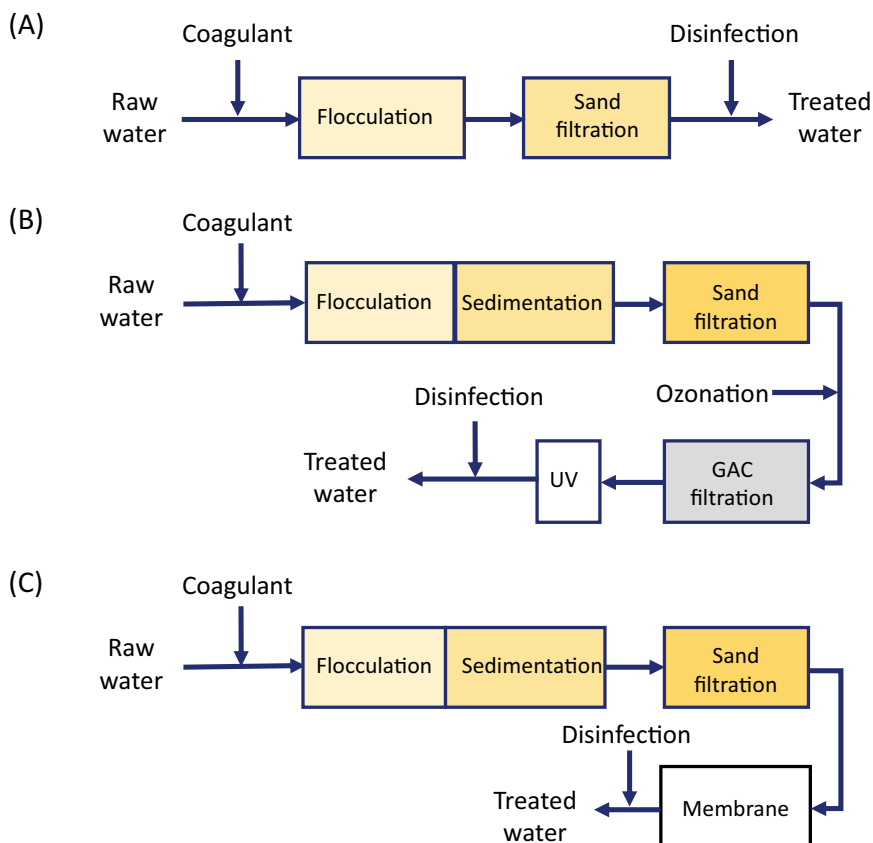


**Fig. 2** Reported MP concentrations in source water in different countries [5–7, 13, 17, 19, 20, 23, 25]

herein measured MPs that are larger than 1  $\mu\text{m}$ . Studies on NPs are limited due to the difficulty of sampling NPs and lack of analytical instruments presented in research laboratories [1, 24]. Given their possibly severer health impacts [11], NPs in drinking water sources are worth more research attention.

### 3 MP Removal in Drinking Water Treatment Processes

Drinking water treatment plants (DWTPs) can differ substantially in terms of the treatment train configuration. Figure 3 shows three typical DWTP systems that are widely used globally and have been investigated for MPs. Coagulation-based process (usually consisting of coagulation-flocculation-sedimentation) and sand filtration constitute a basic drinking water treatment train (Fig. 3a), whereas more advanced DWTP may have more treatment processes (such as ozonation, granular activated carbon (GAC) filtration, and membrane filtration) following sand filtration (Fig. 3b, c). Based on published literature, this overview will summarize MP removal by coagulation, media filtration (e.g., granular activated carbon (GAC) or sand filters), membrane filtration, ozonation, etc.

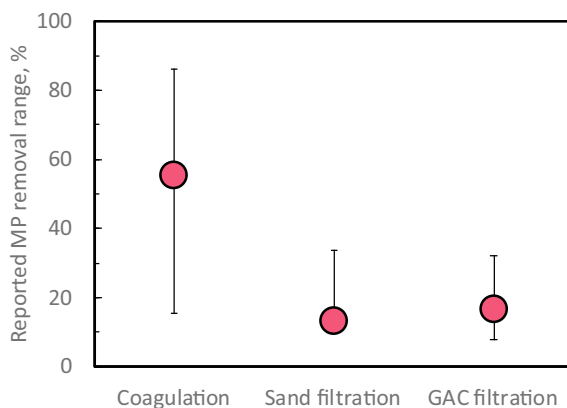


**Fig. 3** Representative drinking water treatment trains (adapted from Xue et al. [26])

According to a few full-scale studies that have been published, coagulation can remove 40–71%, sand filtration 3–24%, and GAC 9–25% of the total MPs (1–100  $\mu\text{m}$ ) (Fig. 4) [5, 19, 20, 25]. Hence, coagulation-sand filtration can remove ~ 70% of source water MPs, whereas coagulation-sand filtration-GAC can remove ~ 86% of source water MPs. Sand filtration is more effective in removing MPs > 10  $\mu\text{m}$  (~100% removal), but less efficacious for smaller MPs (< 10  $\mu\text{m}$ ) [19, 20, 25]). In contrast, GAC has been proved effective in removing smaller MPs (e.g., 1–5  $\mu\text{m}$ ). It is clear that most modern DWTPs can effectively remove a substantial fraction of source water MPs, with additional treatment steps generally enhancing the total treatment. When membrane filtration (especially ultrafiltration or finer membranes) is adopted at a plant, due to the physical size exclusion mechanism, near complete removal of MPs can be achieved [25]. However, there are more unknowns or issues to be addressed. Most of these DWTPs were focused on MPs. NPs that are smaller than MPs remain unknown and can be more problematic and more challenging to detect and study. Systematic understanding of MPs and NPs of different properties



**Fig. 4** Removal of MPs by water treatment processes (error bars indicate minimal and maximal values reported) [5, 19, 20, 25]



(chemical composition, shape, and size) in each treatment process remains lacking. Some treatment may actually contribute MPs and NPs into the treated water, such as polymers used in coagulation and polymer-based membrane modules [8, 25]. Even though ozonation demonstrated minimal influence on MPs in full-scale studies, some laboratory studies have indicated that ozonation can aggressively breaking down PS NPs (e.g., 99.9% molecular weight degradation and 42.7% mineralization) [15]. Thus, the treatment contribution of ozonation and chlorination in removing MPs and NPs should be further investigated. Moreover, the interactions between MPs (or NPs) and heavy metals and/or trace organic pollutants [4, 16] can result in compounded challenges, which requires more research.

## 4 Conclusions and Recommendations

The presentation overviewed the occurrence and removal of MPs in drinking water treatment. It is found that conventional and more advanced DWTPs can effectively remove a substantial fraction of MPs from water. Advanced DWTPs with more barriers achieved significantly better removal of MPs. However, a number of unknowns and challenges remain to be addressed in the future. The removal of MPs in drinking water treatment should be further improved, especially given that MPs contains various additives (very harmful to human health) and can be vectors of heavy metals and trace organic pollutants. In addition, NPs that can be more toxic but have seldom been studied in drinking water treatment warrant attention.

**Acknowledgements** This study was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Discovery Grants Program, NSERC Discovery Launch Supplement Grant, and NSERC Alliance Grant. In addition, we would thank the support by the University of Regina (UofR) through the New Faculty Start-Up Fund, the UofR President's

Seed Fund, and the UofR Faculty of Engineering & Applied Science Research Opportunities Fund. S. Samaei and J. Chen would like thank the Saskatchewan Innovation and Excellence Graduate Scholarship.

## References

1. Adhikar S, Kelkar V, Kumar R, Halden RU (2022) Methods and challenges in the detection of microplastics and nanoplastics: a mini-review. *Polym Int* 71:543–551
2. Andrady AL (2017) The plastic in microplastics: a review. *Marine Pollut Bull* 119(1):12–22
3. Barboza LGA, Vethaak AD, Lavorante BR, Lundebye A-K, Guilhermino LJM (2018) Marine microplastic debris: an emerging issue for food security, food safety and human health. *Mar Pollut Bull* 133:336–348
4. Chen CC, Zhu X, Xu H, Chen F, Ma J, Pan K (2021) Copper adsorption to microplastics and natural particles in seawater: a comparison of kinetics, isotherms, and bioavailability. *Environ Sci Technol* 55(20):13923–13931
5. Cherniak SL, Almuhtaram H, McKie MJ, Hermabessiere L, Yuan C, Rochman CM, Andrews RC (2022) Conventional and biological treatment for the removal of microplastics from drinking water. *Chemosphere* 288:132587
6. Chu X, Zheng B, Li Z, Cai C, Peng Z, Zhao P, Tian Y (2022) Occurrence and distribution of microplastics in water supply systems: in water and pipe scales. *Sci Total Environ* 803:150004
7. Dalmau-Soler J, Ballesteros-Cano R, Boleda MR, Paraira M, Ferrer N, Lacorte S (2021) Microplastics from headwaters to tap water: occurrence and removal in a drinking water treatment plant in Barcelona Metropolitan area (Catalonia, NE Spain). *Environ Sci Pollut Res* 1–11
8. Ding H, Zhang J, He H, Zhu Y, Dionysiou DD, Liu Z, Zhao C (2021) Do membrane filtration systems in drinking water treatment plants release nano/microplastics? *Sci Total Environ* 755:142658
9. Ebrahimi P, Abbasi S, Pashaei R, Bogusz A, Oleszczuk PJC (2022) Investigating impact of physicochemical properties of microplastics on human health: a short bibliometric analysis and review. *Chemosphere* 289:133146
10. Frère L, Maignien L, Chalopin M, Huvet A, Rinnert E, Morrison H, Kerninon S, Cassone A-L, Lambert C, Reveillaud J (2018) Microplastic bacterial communities in the Bay of Brest: influence of polymer type and size. *Environ Pollut* 242:614–625
11. Gigault J, El Hadri H, Nguyen B, Grassl B, Rowenczyk L, Tufenkji N, Feng S, Wiesner M (2021) Nanoplastics are neither microplastics nor engineered nanoparticles. *Nat Nanotechnol* 16(5):501–507
12. Hahladakis JN, Velis CA, Weber R, Iacovidou E, Purnell PJJ (2018) An overview of chemical additives present in plastics: migration, release, fate and environmental impact during their use, disposal and recycling. *J Hazard Mater* 344:179–199
13. Johnson AC, Ball H, Cross R, Horton AA, Jurgens MD, Read DS, Vollertsen J, Svendsen C (2020) Identification and quantification of microplastics in potable water and their sources within water treatment works in England and Wales. *Environ Sci Technol* 54(19):12326–12334
14. Lee J, Gibert Y, Dumée LF (2021) Nano and micro plastics in water processing—where are we at? *J Water Process Eng* 43:102281
15. Li Y, Li J, Ding J, Song Z, Yang B, Zhang C, Guan B (2022) Degradation of nano-sized polystyrene plastics by ozonation or chlorination in drinking water disinfection processes. *Chem Eng J* 427:131690
16. Liu L, Fokkink R, Koelmans AA (2016) Sorption of polycyclic aromatic hydrocarbons to polystyrene nanoplastic. *Environ Toxicol Chem* 35(7):1650–1655

17. Mintenig S, Löder M, Primpke S, Gerdts G (2019) Low numbers of microplastics detected in drinking water from ground water sources. *Sci Total Environ* 648:631–635
18. Pham DN, Clark L, Li M (2021) Microplastics as hubs enriching antibiotic-resistant bacteria and pathogens in municipal activated sludge. *J Hazard Mater Lett* 2:100014
19. Pivokonsky M, Cermakova L, Novotna K, Peer P, Cajthaml T, Janda V (2018) Occurrence of microplastics in raw and treated drinking water. *Sci Total Environ* 643:1644–1651
20. Pivokonský M, Pivokonská L, Novotná K, Čermáková L, Klimtová M (2020) Occurrence and fate of microplastics at two different drinking water treatment plants within a river catchment. *Sci Total Environ* 741:140236
21. PlasticsEurope (2020) Plastics—the facts 2020: an analysis of European plastics production, demand and waste data. Association of Plastics Manufacturers
22. Prata JC, da Costa JP, Lopes I, Duarte AC, Rocha-Santos T (2020) Environmental exposure to microplastics: an overview on possible human health effects. *Sci Total Environ* 702:134455
23. Sarkar DJ, Sarkar SD, Das BK, Praharaj JK, Mahajan DK, Purokait B, Mohanty TR, Mohanty D, Gogoi P, Kumar S (2021) Microplastics removal efficiency of drinking water treatment plant with pulse clarifier. *J Hazard Mater* 413:125347
24. Sullivan G, Gallardo JD, Jones E, Holliman P, Watson T, Sarp S (2020) Detection of trace sub-micron (nano) plastics in water samples using pyrolysis–gas chromatography time of flight mass spectrometry (PY-GCToF). *Chemosphere* 249:126179
25. Wang Z, Lin T, Chen W (2020) Occurrence and removal of microplastics in an advanced drinking water treatment plant (ADWTP). *Sci Total Environ* 700:134520
26. Xue J, Samaei SH-A, Chen J, Doucet A, Ng KTW (2022) What have we known so far about microplastics in drinking water treatment? A timely review. *Front Environ Sci Eng* 16(5):1–18
27. Zhang C, Wang J, Zhou A, Ye Q, Feng Y, Wang Z, Wang S, Xu G, Zou J (2021) Species-specific effect of microplastics on fish embryos and observation of toxicity kinetics in larvae. *J Hazard Mater* 403:123948
28. Zhu J-J, Dressel W, Pacion K, Ren ZJ (2021) ES&T in the 21st century: a data-driven analysis of research topics, interconnections, and trends in the past 20 years. *Environ Sci Technol* 55(6):3453–3464

# Selecting a Hybrid Treatment Technology for Upgrading a Lagoon-Based WWTP



Jeremy Enarson

**Abstract** For many small to mid-sized rural communities within Canada, municipal wastewater is treated within aerated lagoons, prior to being released back to the environment. Under changing provincial and federal regulations, these communities are required to treat their wastewater to a much higher standard than what lagoons can typically achieve. However, for many of these communities, the existing lagoons are still functioning well and represent a significant prior investment by the community. Communities are left with big decisions, such as abandoning their existing assets in favor of fully mechanical wastewater treatment plants (WWTPs), or connecting to a regional wastewater system (not feasible when the community is located a far distance away from an existing facility). In addition to significant capital costs, the ongoing operating costs for these systems can be significant as well, particularly when considering operator certification requirements for fully mechanical WWTPs. The City of Camrose (population of ~ 20,000) began initial design on its WWTP upgrade project in late 2008. After extensive testing of the City's wastewater and the downstream receiving environment, in 2012, the City received confirmation from the provincial and federal regulators regarding the treatment standards that would be applied to the upgraded WWTP. In addition to considering more traditional mechanical wastewater treatment options, in 2015, the City received approval from the regulators to consider "hybrid" systems, which could leverage the investment of the existing aerated lagoons while minimizing the complexity and operator requirements of the upgraded WWTP. The City's technology selection process involved many discussions with regulators, as some of the proposed technologies had not been used within Alberta. Ultimately, in 2019, the Province approved the City to proceed with detailed design on a hybrid WWTP that would see the existing aerated lagoons supplemented with a Moving Bed Biofilm Reactor (MBBR). The following report will highlight the process that the City followed to select the MBBR technology for its WWTP upgrades (currently under construction). Once complete, the Camrose WWTP will be the first facility in Alberta to use a full-scale MBBR treatment process within a hybrid WWTP.

---

J. Enarson (✉)  
City of Camrose, Canada  
e-mail: [jenarson@camrose.ca](mailto:jenarson@camrose.ca)

**Keywords** Hybrid treatment technology • Upgrading a lagoon-based WWTP

## 1 Background and Introduction

Similar to the situation for many small to mid-sized rural communities within Canada, municipal wastewater from the City of Camrose is treated within aerated lagoons prior to being released back to the environment. The City's wastewater system is regulated by the Province of Alberta (through the department of Alberta Environment and Parks or AEP). The Province grants a wastewater approval to operate to the City, typically lasting ten years at a time, which outlines the operating and capital upgrade requirements for the City's wastewater system over that ten-year period.

Currently, the City of Camrose (population of approximately 20,000 people) treats its wastewater using aerated lagoons, a process that has been in place since 1992 when the City's population was around 13,000 people. Aerated lagoons are a fairly simple process that involves adding air to the municipal wastewater to facilitate the breakdown of certain components within the wastewater through biological means. This process is relatively effective at reducing the carbonaceous biological oxygen demand (cBOD) of the wastewater to below 25 mg per liter (mg/L). However, aerated lagoons are not able to provide a tertiary level of treatment of municipal wastewater, and specifically for the removal of such parameters as ammonia (which can be acutely toxic to fish), phosphorus, and nitrogen (nutrients which can contribute to more chronic impacts to the environment through eutrophication of lakes and rivers).

## 2 Camrose Approval Renewal Proposal

In the mid-2000s, the City recognized that it would likely be required to complete a major upgrade to its wastewater treatment plant (WWTP) as a condition of a future provincial approval. Because of this, the City began initial work on its WWTP upgrade project in late 2008. The City retained Associated Engineering Alberta Ltd. (AE) to assist with:

- determining the required level of upgrading to the City's existing WWTP through environmental testing, and through discussions with provincial and/or federal regulators,
- preparing a formal proposal on the City's behalf, and submitting this proposal to the applicable regulatory agencies,
- completing a conceptual design for potential treatment technologies to achieve the level of upgrades outlined in the City's proposal, and

- continuing to develop the potential treatment technologies, and completing preliminary engineering design for a preferred treatment process for the Camrose WWTP upgrade project.

Through a separate contract following the end of preliminary design, the City later retained AE to also assist with detailed engineering design for the Camrose WWTP upgrade project, as well as tendering assistance, contract management and construction/post-construction administration on the City's behalf for this project. Construction recently began on the Camrose WWTP upgrade project in September 2021, and the upgrades are scheduled to be substantially complete by October 2023.

Early in the design process, the City and AE completed a review of nine potential end-use options for the Camrose municipal wastewater. This included such options as:

- continued treatment of municipal wastewater at the existing WWTP site, including continued discharge to the existing receiving watercourse, Camrose Creek (also referred to as Stoney Creek),
- the use of wastewater effluent (treated or untreated) for irrigation of agricultural lands,
- the use of treated wastewater effluent for urban non-potable uses, or for industrial process water needs,
- the pumping of untreated wastewater to a regional WWTP, such as the Alberta Capital Region Wastewater Commission, located just northeast of Edmonton (about 100 km away from Camrose), and
- the construction of a new regional WWTP to service additional communities within the larger region.

Through the evaluation of the nine different end-use options, the City and AE considered a number of key guiding principles, including:

1. Maximizing the use of existing assets, and especially those assets which might have capacity and/or life span available. (Many of the Camrose wastewater assets were less than 20 years old at that time, and were estimated to have used less than half of their anticipated life expectancy.)
2. Maintaining a strong focus on environmental stewardship, and sustainable watershed management. The City has a history of environmental stewardship within the larger region, and it wanted to promote improved wastewater treatment that served the community's needs for the foreseeable future, and which was both affordable to the community and environmentally sustainable.
3. Recognizing that wastewater is a resource that should be protected. The City's focus is that wastewater effluent should be seen in the context of "end-use" rather than "disposal," and the best solution will be the one that provides the most value to meet the needs of the community and the environment.
4. Recognizing the water flow needs of the Battle River, which is an overstressed "prairie-fed" river with low seasonal background flows, due in part to the fact that it does not receive any runoff from glaciers or mountain snow-pack. A potential loss of water (i.e., by diversion to another watershed or by beneficial re-use such

as agricultural irrigation) would effect a significant change on the water balance of the Battle River and could have detrimental effects on the environment and on downstream users.

Based on the above review, the City determined that the best option would be to continue treating its wastewater at the Camrose WWTP site, and discharging to the Battle River via the nearby Camrose Creek.

Another task that the City and AE completed early in the design process was to meet with the provincial regulator, Alberta Environment and Parks (AEP). During this meeting, which happened in February 2009, AEP confirmed that the City would be required to upgrade its WWTP as a condition of the City's next wastewater approval. However, rather than identifying a certain set of treatment parameters and limits, the Province noted that the City and AE would need to determine the treatment/discharge limits that would be appropriate for the receiving environment (Camrose Creek and the Battle River). AEP also noted that the City would need to consider "cumulative effects" within the larger watershed and to demonstrate that the upgraded WWTP would represent an acceptably low risk to human health.

To satisfy these requirements, the City and AE undertook an extensive seasonal monitoring program during the spring, summer, and fall of 2009. Samples were taken from six different sampling locations, which were then analyzed for almost 100 different parameters. The City and AE also reviewed the hydrology of Camrose Creek and the Battle River to determine whether there would be adequate flow available in order to achieve the 10-to-1 dilution that is generally desired when considering continuous or semi-continuous of treated wastewater to the environment.

Because of the general lack of natural flows within both Camrose Creek and the Battle River at certain times of the year, the City and AE found that the standard calculation methods outlined in the provincial "Water Quality-Based Effluent Limits" procedures manual could not be applied. Based on the results of the wastewater sampling, and the environmental and human health considerations that followed, the City and AE proposed that the terms of the City's next wastewater approval for the upgraded WWTP should provide for effluent quality at least as good as the existing lagoon discharge, and improved in the following aspects:

- Fish toxicity should be reduced by nitrification of ammonia to limits that are practically achievable,
- Nutrients, and particularly total phosphorus, should be reduced in a way that is practically achievable and justifiable in light of the watershed-based management approach (i.e., upgrades to the Camrose WWTP over time, at similar timeframes as upgrades to other regulated and non-regulated wastewater discharges further upstream).
- The WWTP effluent should be disinfected in a controlled way that preferably avoids the use of chlorine.

This initial review by AE proposed the suggested approval limits that the Province should apply to the discharge of the upgraded Camrose WWTP, as indicated in Table 1:

**Table 1** Initially proposed approval limits for the upgraded Camrose WWTP

Flow, (ML/d) <sup>a</sup> , maximum average daily volume	15
cBOD <sub>5</sub> (mg/L) <sup>b</sup>	20
TSS (mg/L) <sup>b</sup>	20
NH <sub>3</sub> -N (mg/L) <sup>b,d</sup>	5 summer/10 winter
Total P (mg/L) <sup>b</sup>	1.0
<i>E. coli</i> (MPN/100 ml) <sup>c</sup>	200

<sup>a</sup> Based on monthly arithmetic mean of daily flows—prorating annual flow into 7 months

<sup>b</sup> Based on monthly arithmetic mean of daily samples

<sup>c</sup> Based on monthly geometric mean of daily samples

<sup>d</sup> Limit may need to change to avoid toxicity of NH<sub>3</sub>-N at the prevailing effluent pH and temperature

AE finalized their review in February 2011, and the City submitted the AE report to the Province as its proposal for the next approval renewal. The Province of Alberta (through AEP) accepted the City's recommendations in August 2012 through the issuance of Approval 481-02-00. This became the basis for the following step in the design process, the conceptual engineering design.

Of note, the February 2011 Approval Renewal Proposal did not specify a preferred treatment technology for the Camrose WWTP upgrades. AE noted that a Biological Nutrient Removal (BNR) process could be considered as a "baseline" process option. However, AE noted that a number of other treatment options could and should be considered during conceptual design.

### 3 Conceptual Engineering Design

In the fall of 2012, the City and AE started on the conceptual design phase of this project. In addition to considering the recommendations of the 2011 Approval Renewal Proposal, the City and AE recognized that the Camrose WWTP upgrades would need to meet the newly announced federal requirements outlined in the Wastewater Systems Effluent Regulations (WSER), which introduced restrictions regarding un-ionized ammonia and chlorine, and which included a general requirement that treated municipal wastewater would not be acutely toxic to fish.

Based on the initial testing program completed in 2009, as well as ongoing sampling by the City since that time, the City recognized that the key parameters that the upgraded Camrose WWTP would need to address in order to meet the provincial and federal requirements would include ammonia (both total and un-ionized) and total phosphorus. Neither of these parameters are removed to any significant level using the City's existing aerated lagoons.

Whereas there are already a number of well-established technologies for the removal of phosphorus from municipal wastewater, the review of treatment technologies for the nitrification of ammonia would require additional effort. While AE



and the City continued to consider a more traditional, fully mechanized BNR treatment process, they also considered a number of other options that would capitalize on the ability of the existing aerated lagoons to achieve the proposed limits for cBOD and total suspended solids (TSS). A new ammonia removal (nitrification) process would be added after the existing aerated lagoons, prior to storage in one of the six existing post-treatment storage lagoons. (Because of the general lack of flow within the receiving Camrose Creek, the City of Camrose is currently required to store its treated wastewater for up to seven months per year, releasing the stored water back to the environment in the spring and in the fall, over a discharge period of up to six weeks at a time.)

Based on literature review and their own experience working with other clients, AE identified two potential nitrification processes that could be installed following the existing aerated lagoons to assist with the removal of ammonia:

- Submerged Attached Growth Reactor (SAGR)—a proprietary technology developed in Manitoba by Nelson Environmental (now NEXOM), and which had been successfully implemented at a number of sites within western and central Canada,
- Moving Bed Biofilm Reactor (MBBR)—a non-proprietary process used by Veolia Water Technologies Canada (Veolia) and others. While the MBBR technology was well-established in Europe and within Quebec, there were no installations of the MBBR technology (specifically for post-lagoon nitrification) anywhere within western Canada.

With initial input from Nelson Environmental and Veolia, AE developed process flow diagrams and conceptual site layouts of the SAGR and MBBR processes for the Camrose WWTP upgrade application. AE also identified the advantages and disadvantages of each technology at a high level, as well as conceptual level cost estimates for each technology that also included costs for the remaining components of the WWTP upgrade. These advantages, disadvantages, and conceptual level costs were then compared against corresponding information related to the more widely accepted BNR process.

In the end, AE and the City determined that both the SAGR and MBBR technologies, along with the associated upgrades for the remaining parameters such as phosphorus, should be able to provide a level of treatment that is similar to that found through the BNR process. Moreover, the SAGR and MBBR concepts were both significantly cheaper than the BNR treatment concept, both in terms of initial capital costs and of ongoing operating costs. As a result, AE and the City concluded that a “hybrid” WWTP (utilizing the existing aerated lagoons for carbon and solids removal, and new mechanical treatment technologies for ammonia and phosphorus) should be considered further through the preliminary design phase of the project.

AE concluded their review in October 2014, and the City submitted the AE conceptual design report to the Province at that time.

## 4 Preliminary Engineering Design

Immediately following the submission of the 2014 conceptual design report (CDR), the City and AE started working on the preliminary engineering phase of the Camrose WWTP upgrade project. The City and AE met with AEP staff in January 2015 to review the recommendations of the CDR and to confirm a go-forward plan for this project. Unfortunately, between 2009 and 2015, the Province had completed an internal shuffle of staff and workloads between their offices. As a result, a different AEP office was now managing the Camrose wastewater facilities. This also meant that most of the AEP staff who would be working with the City through the remaining steps of the WWTP upgrade project were different from those who had reviewed the 2011 Approval Renewal Proposal or who had issued the City's 2012 wastewater approval.

During the January 2015 meeting, AEP noted concerns with the hybrid WWTP concept that the City and AE were proposing. Since neither of the proposed ammonia removal technologies (SAGR or MBBR) had been approved to that time for use within Alberta, the City and AE would need to conduct additional work to prove that these nitrification technologies would be able to meet the limits outlined in the 2011 Approval Renewal Proposal and particularly in cold weather conditions (as low as 1 °C). AEP also noted concerns about the ability of the hybrid technologies to meet future treatment requirements that may be imposed by the Province, up to and including the approval limits currently imposed on many of the larger communities within Alberta, such as Edmonton or Calgary.

As a takeaway from that meeting, the City and AE committed to addressing AEP's concerns through the preliminary engineering design process. The City and AE compared the limits already agreed to by the Province for the initial upgrades (Stage 1, as outlined in the 2012 wastewater approval) with two potential future scenarios: "near future" limits (Stage 2) and "far future" limits (Stage 3). The limits that were considered as listed in Table 2.

The City and AE then re-engaged Nelson Environmental and Veolia and asked them to identify how their technologies would be able to achieve the Stage 2 and Stage 3 approval limits. (The ability of the SAGR and MBBR technologies to meet the Stage 1 limits was previously established during conceptual engineering design.)

**Table 2** Potential future operating approval limits

Parameter, (units)	2022 Limits (stage 1)	Potential near future limits (stage 2)	Potential far future limits (stage 3)
TSS, cBOD5, (mg/L)	20, 20	15, 15	5, 5
NH <sub>3</sub> -N (mg/L, summer, winter)	5, 10	3, 5	1, 3
Total N (mg/L)	N/A	15	5
Total Phosphorus (mg/L)	1.0	0.5	0.10
Fecal Coliform (CFU/100 mL)	200	20	2

Both companies were also asked to provide reference WWTPs within Canada and/or the USA where their technologies had been used to address similar approval limits.

Through this review, the City and AE were able to make a number of relevant conclusions:

- Both the SAGR and MBBR processes were sufficiently flexible to handle potential future changes in approval limits for ammonia, total nitrogen and total phosphorus.
- Being an older and more-established technology than the SAGR process (particularly within northern Europe countries and within Quebec), the MBBR process provided more operational experience to show that the post-lagoon nitrification processes could be a solid base on which to build tertiary denitrification and tertiary phosphorous removal facilities.

Given that both technologies had sufficient “future proofing” capabilities to adapt to the more stringent approval limits that might be applied by AEP in the future, the City and AE conducted a side-by-side comparison of the SAGR and MBBR technologies against eight evaluation criteria to determine the best option for the Camrose WWTP upgrades project. These evaluation criteria included:

1. track record of the technology under similar operating conditions (as low as 1 °C),
2. ability to reduce overall requirements for operating certification,
3. ease of operation for routine (day-to-day) maintenance,
4. ease of operation for major maintenance,
5. ease of adapting to tighter future effluent needs,
6. ease of conversion to a future mechanical WWTP,
7. ultimate foot print requirements, and
8. total life cycle costs (for the assumed 20-year design horizon for the project).

For each of the above evaluation criteria, a relative weighting was assigned between 1 (least critical) and 10 (most critical). The City and AE decided that of the eight criteria, the most critical criteria were the technology’s track record within cold operating conditions, as well as the ease of operation for routine and major maintenance. On the other hand, other criteria such as the ease of conversion to a future mechanical WWTP and the ultimate footprint requirements were determined to be less critical for this review. Once the evaluation criteria were identified and relative weightings were assigned, the City and AE assigned a relative ranking (between 1 and 5) for SAGR and for MBBR for each criteria. Weighted scores were calculated for each evaluation criteria, and then combined to provide an overall score for each technology.

Based on this review, the City and AE identified that the MBBR technology appeared to be better suited for the Camrose WWTP upgrade project. A sensitivity analysis was also conducted to determine what scenario would need to change in order to alter the outcome of this review in favor of SAGR. The City and AE concluded that the only scenario where MBBR was not preferred for the Camrose WWTP application was the scenario where the MBBR technology was assigned the lowest possible score under the criteria of its track record in performing under cold weather conditions. The evaluation team of the City and AE concluded that this scenario

was unlikely to be the case, based on the performance of this technology in various northern and cold-climate jurisdictions. However, the team agreed to conduct some further due diligence reviews on the viability of MBBR in cold weather conditions.

To help understand the MBBR technology better, the City and AE undertook reference checks of existing WWTPs within North America which were similar to Camrose in terms of size of facility and typical operating conditions. Two of those references were within the Province of Quebec, while one was located within a mountainous region of the State of Wyoming, USA. All of the reference WWTP operators noted that the MBBR technology was quite simple and easy to operate, that it had no significant operational or maintenance requirements, and that the upgraded WWTPs were able to consistently meet the treatment requirements imposed by their respective regulatory agencies.

To ensure that the MBBR technology could be successfully applied to the Camrose WWTP site, the City and AE were made aware of research being conducted by Dr. Robert Delatolla with the University of Ottawa, where Dr. Delatolla was studying the MBBR's ability for denitrification in very cold conditions (as low as 1 °C). The City and AE contacted Dr. Delatolla, who agreed to assist with a review of the suitability of this technology for the specific wastewater found at the Camrose WWTP site. The City collected approximately 1200 L of its post-lagoon (treated) wastewater and shipped the samples to Ontario, where Dr. Delatolla performed various tests over a two month period. From his research, Dr. Delatolla determined that MBBR should be able to consistently achieve a 10 mg/L limit for ammonia at temperatures as low as 1 °C.

Finally, Veolia was working with another municipality in Manitoba (the Town of Neepawa) on a pilot project to test the effectiveness of MBBR technology for post-lagoon removal of ammonia. This work happened during the winter of 2016/2017. Based on that pilot, Veolia reported that the Neepawa WWTP would be able to consistently achieve ammonia levels below 10 mg/L with influent temperatures as low as 1 °C.

Based on the above work, the City and AE concluded that the MBBR technology would be well suited for application as part of the Camrose WWTP upgrade project. On this basis, AE finalized their review in December 2017, and the City submitted the AE preliminary design report to the Province at that time.

## **5 Final Acceptance and Approval**

Following completion of preliminary design, the City did not receive immediate approval from AEP to proceed with detailed engineering design. During 2018 and the early part of 2019, the City and AE communicated with AEP on a regular basis to answer any of the outstanding questions raised by AEP staff. (By this point, the regulation of the Camrose wastewater facilities transferred back to the original AEP

office that the City was working with at the start of this project, resulting in the need for further discussions to re-familiarize AEP staff on the recent history of the project.)

As part of this final step in the approval process, AEP asked the City and AE to prepare and submit a final “due diligence” report to the Province. This due diligence report, which was submitted in March 2019, confirmed that the conclusions and recommendations of the 2017 Preliminary Design Report were consistent with the recommendations from the 2011 Approval Renewal Proposal and the 2014 Conceptual Design Report. As well, the due diligence report sought to provide additional certainty to the impact that the upgraded WWTP discharge would have on the receiving water bodies, as well as the strategies that the City could implement to mitigate any adverse effects that might arise after the upgraded WWTP was in operation (i.e., identifying potential upsets to the treatment processes within an upgraded WWTP, and develop appropriate measures to address those upsets).

Finally, in May 2019, the Province provided written authorization to the City to proceed with detailed engineering design for the upgrades being proposed by the City and AE, over ten years after the City and AE started initial work on this project.

The City and AE completed the detailed engineering design for this project by the spring of 2021, and by August 2021, the City had retained a general contractor to assist with completing the upgrades. Construction began on the Camrose WWTP upgrades in the following month, and work is expected to be substantially complete by the fall of 2023, with final project cleanup by the spring of 2024. Once complete, the Camrose WWTP will be the first facility of its kind within Alberta to use the MBBR technology for post-lagoon removal of ammonia. However, over the past few years, the City had been approached by a number of other municipalities and consultants who are looking to upgrade their wastewater treatment facilities in the near future. Most of these conversations have focused around the City’s decision to proceed with a hybrid WWTP (combination of lagoons and mechanical processes), as well as the steps taken to move this decision through the regulatory processes.

Over the past 10+ years, the City learned a number of lessons related to designing and implementing a hybrid WWTP, including:

- That out-of-the-box treatment options (such as hybrid treatment technologies for ammonia removal) can be a cost effective way of achieving a high level of treatment, sometimes to the same levels as what are provided through more traditional, well-established WWTP technologies.
- That introducing a new treatment technology into a regulatory jurisdiction can be challenging.
- That it is important to have early and regular/ongoing engagement with regulatory agencies.
- That it is important to have good documentation regarding direction provided by regulators.

**Acknowledgements** The City of Camrose wishes to acknowledge the assistance of its design consultant, Associated Engineering Alberta Ltd., as well as the regulatory staff with Alberta Environment and Parks and with Environment and Climate Change Canada in their assistance on this project.

# Treatment of Aqueous Arsenite Using Modified Biomass-Based Sorbent



Khaled Zoroufchi Benis, Kerry McPhedran, and Jafar Soltan

**Abstract** The occurrence of high concentrations of arsenic (As) in water has been recognized as a global health and environmental problem. Sorption is regarded as a promising As treatment method due to its simplicity and potential for high efficiency. Canada has a strong agricultural industry that produces waste products that can be converted to value-added products. Considering the availability of agricultural residue in Canada, the cost of the sorption process can be decreased by using agricultural residue-based sorbents (biosorbents) as an eco-friendly alternative for commercial sorbents. In this study, sorption of arsenite, As(III), from aqueous solutions onto Fe oxide-modified canola straw (MCS) was investigated. The results showed that the negligible As(III) sorption capacity of raw canola straw increased significantly to 791  $\mu\text{g/g}$  after modification in the removal of As(III) from a 1000  $\mu\text{g/L}$  solution. Studying the effect of solution pH showed that As(III) sorption capacity of MCS increased by increasing the solution pH from 3 to 10. A kinetic study showed that about 66% of the ultimate sorption capacity was reached within four hours. The sorption kinetic data was best represented by pseudo-second-order and Elovich models suggesting that chemisorption may be the rate-determining step of the sorption process. The isothermal data of As(III) sorption followed Freundlich and Redlich–Peterson models indicating a hybrid adsorption mechanism with a higher probability of a multilayer heterogeneous adsorption. Studying the effect of co-existing anions in the solution upon the As(III) removal efficiency of MCS indicated a significant antagonistic impact of selenate ( $\text{SeO}_4^{2-}$ ), selenite ( $\text{SeO}_3^{2-}$ ), and phosphate ( $\text{PO}_4^{3-}$ ). However, the effect of nitrate ( $\text{NO}_3^-$ ) and chloride ( $\text{Cl}^-$ ) on As(III)

---

K. Zoroufchi Benis · J. Soltan

Department of Chemical and Biological Engineering, University of Saskatchewan, Saskatoon, SK, Canada

K. McPhedran (✉)

Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, SK, Canada

e-mail: [kerry.mcphedran@usask.ca](mailto:kerry.mcphedran@usask.ca)

K. Zoroufchi Benis · K. McPhedran · J. Soltan

Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_61](https://doi.org/10.1007/978-3-031-34593-7_61)

removal efficiency was insignificant, indicating that inner-sphere complexation was the leading mechanism in As(III) sorption.

**Keywords** Arsenite · Biosorption · Canola straw · Co-existing ions

## 1 Introduction

Arsenic (As) is a toxic and carcinogenic metalloid and exists ubiquitously in nature. The World Health Organization (WHO) has acknowledged As contamination as a “major public health issue” [1]. The As poisoning from drinking contaminated waters is the main route of As exposure for humans [2] with source natural waters having a wide range of As concentrations from 0.5 to 5000  $\mu\text{g/L}$  [3]. WHO guidelines indicate a maximum As concentration of 10  $\mu\text{g/L}$  in drinking water [4], and it has been reported that over 200 million people worldwide are exposed to As concentrations higher than this permissible level [5]. Prolonged exposure to As causes a variety of potentially lethal health problems by creating risks of different diseases such as cancer, melanosis, gangrene, hyperkeratosis, enlargement of liver, and black foot disease [6]. As occurs in four oxidation states [As(V), As(III), As(0), and As(−III)] in the natural environment and its toxicity and mobility in the environment depends on its oxidation state [7]. Generally, inorganic As compounds are more toxic compared to organic arsenics. Among these, arsenite, As(III), is the most toxic and harmful to human beings due to its higher mobility which also makes it more difficult to remove from contaminated waters [8].

The common As removal technologies are based on precipitation techniques followed by a separation system to remove insoluble As-bearing precipitates from water [9]. However, initial As concentration, target treatment concentration, oxidation state, and regulatory requirements are the key factors in the selection of an effective As removal method [10]. Therefore, the potential of different technologies such as filtration, reverse osmosis, membrane separation, and sorption for arsenic removal has been investigated in recent years [11]. Sorption has gained much attention among these technologies due to its cost effectiveness and easy operation [12]. The sorption performance of different materials such as activated carbon, resins, metal oxides, gels, minerals, and biomasses has been investigated for As removal [11, 13]. However, using many of these materials as As sorbents is not economically feasible [14]. Therefore, the consideration of using abundantly available ubiquitous agricultural residues for As sorption is of increasing interest [10]. Biosorption is a term that describes the removal of contaminants from aqueous solutions using biomass (e.g., agricultural residues).

Despite the abundance and low cost of agricultural residues, using them for As biosorption is not effective due to their low sorption capacities [15]. The low As sorption capacities of agricultural residues is due to the lack of appropriate chelating functional groups on their surface to complex with As and uptake it from aqueous solution.



Therefore, agricultural residues should be modified via deposition of suitable functional groups on their surface to make their sorption performance comparable with commercially available (but expensive) sorbents [16]. Generally, the biomasses can be activated by immersion in acid or alkaline solutions or they can be modified by deposition of modifying agents on their surface [10]. For example, Abid et al. [17] reported that the As(V) sorption capacity of orange peel biomass increased two times after treating the biomass with  $\text{H}_2\text{SO}_4$  solution, which was attributed to the increased surface area of the treated biomass. Ebrahimi et al. [18] used  $\text{NaHCO}_3$  for the treatment of wheat straw biomass. They found that As(V) sorption capacity of biomass increased from 54 to 108  $\mu\text{g/g}$  after chemical treatment. Roy et al. [12] applied NaOH treatment on neem tree biomass followed by treatment in an  $\text{H}_2\text{SO}_4$  solution and used the treated biomass for As(III) sorption. The As(III) sorption capacity of the modified biomass was 31  $\mu\text{g/g}$ .

Although acid or alkali treatment of biomasses can improve their As sorption capacities, this improvement is limited because these treatments do not incorporate optimal functional groups for As sorption on the biomass surface [10]. However, Fe oxides have shown a high affinity for As ions [19], making them a popular agent in the modification of biomass for As biosorption [15]. For example, Tian et al. [20] modified wheat straw by deposition of  $\text{Fe}_3\text{O}_4$  on straw particles. They reported that As sorption capacity of the modified biomass increased by increasing the amount of deposited  $\text{Fe}_3\text{O}_4$ . They also found that while the straw particles could not sorb As, the sorption capacity of  $\text{Fe}_3\text{O}_4$ /straw composite was higher than  $\text{Fe}_3\text{O}_4$  alone. Hao et al. [21] developed a Fe-coated jute fiber biosorbent using a two-step process. First, the biomass surface was esterified with  $\text{C}_4\text{H}_4\text{O}_3$  to graft with surface carboxyl groups to enhance Fe deposition. Then,  $\text{C}_4\text{H}_4\text{O}_3$ -treated biomass was modified with  $\text{Fe}(\text{NO}_3)_3$  solution to form Fe oxyhydroxide on the surface of biomass. They observed an As(III) sorption capacity of 12.6 mg/g for the modified biosorbent. Meng et al. [22] reported that modification of orange peel using a mixed solution of  $\text{Fe}(\text{NO}_3)_3$  and  $\text{Fe}(\text{NO}_3)_2$  increased the surface area and As(V) sorption capacity of the biomass. However, As(V) reduced to more toxic As(III) during the sorption process by oxidation of deposited Fe(II) oxides on the surface to Fe(III) oxides.

Canola is a Canadian invention derived from rapeseed in the 1970s. Currently, Canada is the biggest global producer and exporter of canola, producing about 20 million tons of canola annually [23, 24]. Saskatchewan is the top province for canola production, contributing to about 55% of the total production in 2021 [23]. Considering the canola seeded area of 8.4 million ha in 2021 in Canada [23], and the average straw yield of 3 dry ton/ha [25], the total canola straw (CS) production can be estimated as 25.2 million ton/year. Using 60% of the produced CS for soil and livestock requirements [26], about 10 million ton/year would be available for other applications such as energy generation, chemical conversion, and production of other value-added products. Given the abundance of CS in Canada, it can be an appropriate precursor for the preparation of As biosorbent.

So far, the performance of chemically modified CS has been studied for the removal of As(V) from water [27], but there has yet to be investigation on its application for As(III) sorption. Therefore, this study aims to modify CS particles by deposition of Fe oxide using  $\text{FeCl}_3$  solution and investigate the As(III) sorption performance of the modified CS (MCS). Additionally, the effect of co-existing ions, namely selenate ( $\text{SeO}_4^{2-}$ ), selenite ( $\text{SeO}_3^{2-}$ ), phosphate ( $\text{PO}_4^{3-}$ ), nitrate ( $\text{NO}_3^-$ ), and chloride ( $\text{Cl}^-$ ), which could interfere with As(III) removal is investigated. Finally, a sorption mechanism has been proposed for the removal of As(III) by deposited Fe oxides on the surface of MCS.

## 2 Materials and Methods

### 2.1 Arsenic Solution

Arsenic was studied under the trivalent As(III) oxidation state as the more toxic form of As. As(III) stock solution (1 g/L) was prepared using sodium arsenite ( $\text{NaAsO}_2$ ) that was purchased from Fisher Scientific, USA. All solutions were prepared by using ultra-pure water (18.2 M $\Omega$  cm, Direct-Q UV, Millipore, USA). The concentration of As(III) in the solution was determined using an atomic absorption spectrometer coupled with a continuous flow vapor generator (VGA-AAS; VP100, Thermo Scientific, USA).

### 2.2 Biosorbent Preparation

The CS was collected from a local agricultural field in Saskatchewan, washed with tap water, dried at 60 °C, before being ground, and sieved (400–840  $\mu\text{m}$ ). The MCS was prepared based on the optimized procedure reported by [27]. Briefly, 5 g CS was immersed in a 0.15 M  $\text{FeCl}_3$  solution and sonicated for 30 min. Then the iron was precipitated by dropwise addition of 10 M NaOH solution and adjusting the final pH to 3 under magnetic stirring. The stirrer was switched off, and the suspension was allowed to age for a day. Afterward, the created Fe oxide-loaded particles were filtered, washed using deionized water, dried at 60 °C for 24 h, and used for As(III) sorption experiments.

### 2.3 Biosorbent Characterization

The morphology and structure of the CS and MCS were characterized by field emission scanning electron microscopy (FE-SEM; Hitachi SU8010, Japan). Fourier transform infrared spectroscopy (FTIR) was used to investigate the functional groups of CS and MCS (Smith's Detection IlluminatIR FTIR microscope, USA). The crystallinity of the biosorbent was determined by a Rigaku Ultima IV X-Ray Diffractometer (Rigaku Americas Corp., USA). Brunauer–Emmett–Teller (BET) surface area of the biosorbent was determined by N<sub>2</sub> adsorption using an ASAP 2020 (Micromeritics, USA). The point of zero charge (pH<sub>PZC</sub>) was determined using the pH drift method [28]. The Fe content of the MCS was determined using acid digestion followed by atomic absorption spectrometry (AAS, Thermo Scientific iCE 3000 series, USA) to verify the deposition of Fe oxide on the surface of MCS.

### 2.4 As(III) Sorption Experiments

#### 2.4.1 Sorption Isotherms and Kinetics

Adsorption isotherms were determined in order to study the relation between the amount of As(III) in solution and the sorbed amount on the MCS. MCS (dosage of 1 g/L) was placed in contact with As(III) solutions with different initial concentrations ranging from 500 to 40,000 µg/L at 25 °C and stirred for 72 h at 200 RPM. Initial experiments indicated that the 72 h duration was adequate to achieve equilibrium. Four sorption isotherm models were used to fit the experimental equilibrium data of As(III) sorption to MCS, namely the Langmuir, Freundlich, Temkin, and Redlich–Peterson models. For the sorption kinetics experiments, As(III) sorption was evaluated as a function of time to determine the influence of contact time on sorption capacity at an initial As(III) concentration of 2500 µg/L. Four kinetics models (pseudo-first-order, pseudo-second-order, intra-particle diffusion, Elovich) were used to investigate the sorption mechanism, characteristic constants, and solid-phase sorption. The amount of As(III) sorbed per unit mass ( $q_t$ , mg/g) at any time  $t$  was calculated as (Eq. 1):

$$q_t = \frac{C_0 - C_t}{m} V \quad (1)$$

where  $C_0$  is the initial concentration of As(III) in the solution (µg/L),  $C_t$  is the As(III) concentration in solution at any time ( $t$ ) (µg/L),  $V$  is the volume of the solution (L), and  $m$  is the mass of the biosorbent (g).

### 2.4.2 Effect of pH and Co-existing Ions

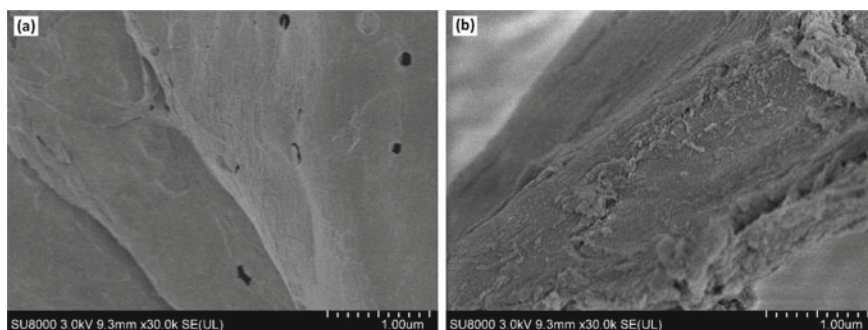
In order to study the effect of the initial pH of the solution, the sorption experiments were conducted at a pH range of 3–10 at an initial As(III) concentration of 2500  $\mu\text{g/L}$ , constant MCS suspension density of 1 g/L, and a temperature of 25 °C. Initial solution pH was adjusted by dropwise addition of 0.1 M HCl or NaOH solutions. The effect of different co-existing ions in water including selenite ( $\text{SeO}_3^{2-}$ ), selenate ( $\text{SeO}_4^{2-}$ ), phosphate ( $\text{PO}_4^{3-}$ ), nitrate ( $\text{NO}_3^-$ ), and chloride ( $\text{Cl}^-$ ) on the removal of As(III) was investigated by increasing the concentration of co-existing ions from 0 to 10 mg/L at a fixed As(III) concentration of 2500  $\mu\text{g/L}$ .

## 3 Results and Discussion

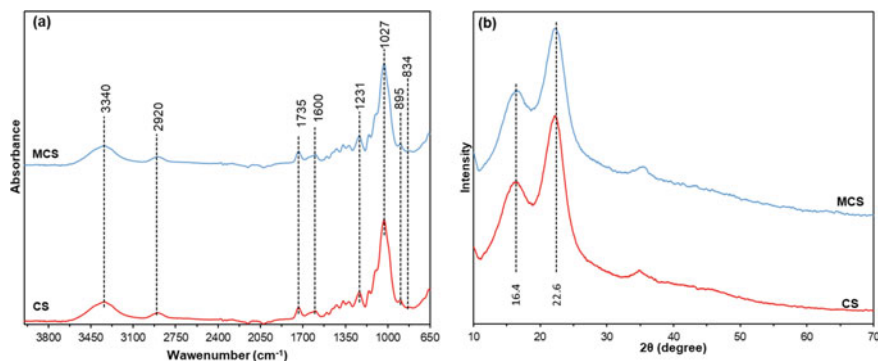
### 3.1 Biosorbent Characterization

The surface morphological characteristics of the CS and MCS are presented in the SEM images (Fig. 1). These images indicate the relatively rougher structure of MCS (Fig. 1b) as compared to CS (Fig. 1a). While the surface of CS is smooth, the surface of the MCS is rough, indicating deposition of a Fe oxide layer on the biomass surface. The Fe oxide deposition was confirmed by measuring the Fe content of the MCS, which indicated the deposition of 74.8 mg Fe per gram of MCS. In addition, the BET surface area of the CS increased from 2.0 to 3.0  $\text{m}^2/\text{g}$  after modification, which can increase the adsorptive surface area and subsequently provide more active As(III) sorption sites [29].

FTIR spectroscopy (wavenumber range of 4000–650  $\text{cm}^{-1}$ ) revealed no significant change in the functional groups upon modification of CS (Fig. 2a). However, both spectra showed the presence of cellulose, hemicellulose, and lignin in the materials [30]. The bands at  $\sim 834$  and  $\sim 895$   $\text{cm}^{-1}$  can be associated with aromatic C–H



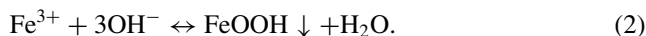
**Fig. 1** SEM images of, **a** canola straw (CS), **b** modified canola straw (MCS)



**Fig. 2** **a** FTIR spectra and **b** XRD pattern of canola straw (CS) and modified canola straw (MCS)

present in lignin [31], and C–O–C rings in cellulose [32], respectively. The band at  $\sim 1027\text{ cm}^{-1}$  can be attributed to the C–O in cellulose, and the band at  $\sim 1231\text{ cm}^{-1}$  can be assigned to the C–OH of the phenolic groups [33]. The spectral peak at  $\sim 1600\text{ cm}^{-1}$  can be related to aromatic skeletal vibrations and C=O stretches present in the aromatic structure of the lignin. The band at  $\sim 1735\text{ cm}^{-1}$  can be assigned to the acetyl groups in hemicellulose [34]. The band at  $\sim 2920\text{ cm}^{-1}$  can be attributed to CH<sub>2</sub> and CH<sub>3</sub> groups in cellulose, hemicellulose, and lignin [30], and the broad peak at  $\sim 3340$  can be due to –OH surface functional groups [35].

The XRD patterns were in good agreement with the FTIR spectra (Fig. 2b). The CS and MCS showed similar broad peaks at  $2\theta$  values of 16.4 and 22.6°, which can be attributed to cellulose [36]. However, the XRD pattern of the MCS lacks any diffraction peak indicating the amorphous nature of the deposited Fe oxides. These results are in agreement with the previous results indicating the formation of amorphous Fe oxide under low pH and drying conditions [37, 38]. FeOOH will be the dominant Fe oxide phase in the modification condition that can be deposited on the surface of CS (Eq. 2) [39]:



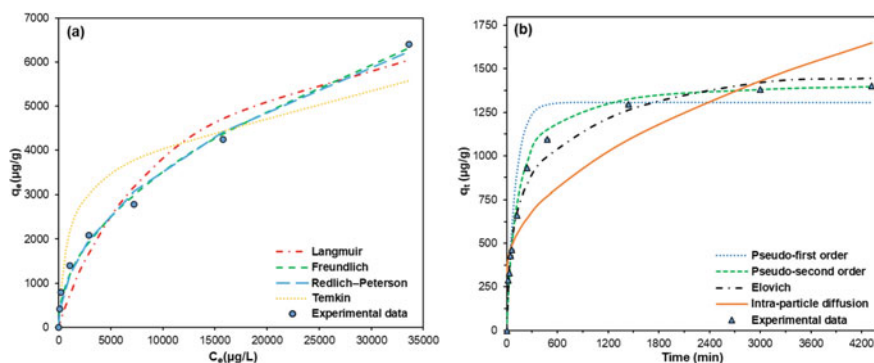
### 3.2 Sorption Isotherms and Kinetics

Sorption isotherms can be used to describe the interaction between the sorbate and the sorbent. Equilibrium sorption results indicated that the amount of sorbed As(III) increased from 407 to 6407  $\mu\text{g/g}$  by increasing the initial As(III) concentration from 500 to 20,000  $\mu\text{g/L}$ . The maximum As(III) sorption capacity of 6407  $\mu\text{g/g}$  is comparable with the reported values for other Fe treated biomasses such as 1400  $\mu\text{g/g}$  for *Melia azedarach* sawdust [40], 4370  $\mu\text{g/g}$  for hazelnut shells [41], 9740  $\mu\text{g/g}$

for pinecone [42], and 12,600 for jute fibers [21]. Overall, it is clear that the MCS developed in the current study performed well overall.

Four isotherm models were used to describe the sorption behavior of As(III) on MCS. The linearized isotherm models were fitted with the experimental data (Fig. 3a), and the estimated isotherm coefficients are shown in Table 1. It is observed from Fig. 3a and the  $R^2$  values that the data fitted better to the Freundlich (0.99) and Redlich–Peterson (0.99) isotherm models than the Langmuir (0.97) and Temkin (0.89) models. The Langmuir model assumes that sorption is monolayer and surface-active sites have uniform energy. In contrast to Langmuir, the Freundlich model assumes that the sorbent surface energy is heterogeneous and the sorption process is multilayer [43]. The Temkin model assumes sorption energy decreases linearly with increasing sorption quantity [44]. Lastly, the Redlich–Peterson model is a three-parameter model that incorporates features of both the Langmuir and Freundlich models. This model assumes a hybrid sorption mechanism and approaches the ideal Langmuir condition when the exponent  $\beta$  is close to 1 and resembles the Freundlich model if values of  $\beta$  are close to zero [45]. Therefore, considering the  $R^2$  values and the value of  $0 < \beta(0.56) < 1$  (Table 1), a hybrid sorption mechanism took place in biosorption of As(III) using MCS. In addition, the value of  $n$  in the Freundlich model ( $n_{fr} = 2.05$ ) indicated that the sorption process was favorable. Generally, when  $0 < 1/n < 1$ , the sorption is considered to be favorable, and when  $1/n > 1$ , the sorption is considered to be unfavorable [46].

The kinetic experiments of As(III) removal were carried out to understand the sorption behavior of MCS. The contact time was varied between 0 and 4320 min (72 h) to establish equilibrium. The As(III) sorption rate was fast, with 66% of the ultimate sorption occurring in the first 4 h, and the sorption capacity continued to increase for the next 72 h with a lower sorption rate to approach equilibrium (Fig. 3b). The sorption kinetics were best modeled by pseudo-second-order ( $R^2 = 0.99$ ) and Elovich ( $R^2 = 0.98$ ) models than the pseudo-first-order ( $R^2 = 0.95$ ) and intra-particle diffusion models ( $R^2 = 0.82$ ) (Table 1). The high correlation coefficient of the pseudo-second-order model suggested that the overall mechanism of sorption



**Fig. 3** a Biosorption isotherm, b biosorption kinetics for As(III) using MCS

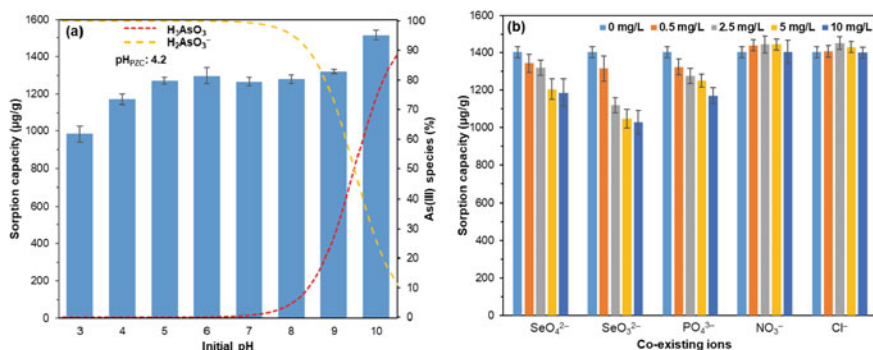
**Table 1** Results of the isotherm and kinetic models, equations, and estimated parameters for biosorption of As(III) by MCS

Model	Equation	Parameter	Value
<i>Isotherm</i>			
Langmuir	$q_e = \frac{K_L q_{\max} C_e}{1 + K_L C_e}$	$q_{\max}$ ( $\mu\text{g/g}$ ) $K_L$ ( $\text{L/mg}$ ) $R^2$	8077 0.089 0.97
Freundlich	$q_e = K_F C_e^{\frac{1}{n}}$	$K_{\text{Fr}}$ ( $\mu\text{g/g}$ )( $\text{L}/\mu\text{g}$ ) $^{1/n_{\text{Fr}}}$ $n_{\text{Fr}}$ $R^2$	39.0 2.05 0.99
Temkin	$q_e = \frac{RT}{b_T} \ln(K_T C_e)$	$K_T$ ( $\text{L/mg}$ ) $b_T$ $R^2$	12.1 845 0.89
Redlich–Peterson	$q_e = \frac{K_{\text{RP}} C_e}{1 + a_{\text{RP}} C_e^{\beta}}$	$K_{\text{RP}}$ ( $\text{L/g}$ ) $a_{\text{RP}}$ ( $\text{L}/\mu\text{g}$ ) $^{\beta}$ $\beta$ $R^2$	6.3 0.091 0.56 0.99
<i>Kinetics</i>			
Pseudo-first-order	$q_t = q_e(1 - e^{-k_1 t})$	$k_1$ ( $1/\text{min}$ ) $q_e$ exp. ( $\mu\text{g/g}$ ) $q_e$ cal. ( $\mu\text{g/g}$ ) $R^2$	0.01 1404 1307 0.95
Pseudo-second-order	$\frac{t}{q_t} = \frac{1}{k_2 q_e^2} + \left(\frac{1}{q_e}\right)t$	$k_2$ ( $\text{g}/\mu\text{g}\cdot\text{min}$ ) $q_e$ cal. ( $\mu\text{g/g}$ ) $R^2$	5.8e-6 1436.7 0.99
Intra-particle diffusion	$q_t = k_p t^{1/2} + C$	$k_p$ ( $\mu\text{g/g}\cdot\text{min}^{0.5}$ ) $C$ ( $\mu\text{g/g}$ ) $R^2$	56.3 905.3 0.82
Elovich	$q_t = \frac{1}{b} \ln(1 + abt)$	$a$ ( $\mu\text{g/g}\cdot\text{min}$ ) $b$ ( $\text{kg/m}$ ) $R^2$	35.2 4.3 0.98

of As(III) onto MCS was controlled by a chemisorption process [47]. The intra-particle diffusion model failed to describe the experimental kinetic data indicating that the intra-particle-diffusion was not the only rate-limiting step. The validity of the Elovich model suggested that the chemisorption mechanism (e.g., surface complexation) is likely the main rate-determining step for the sorption which is in agreement with the pseudo-second-order model.

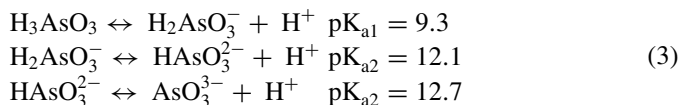
### 3.3 Effect of pH

The pH of solution is an important factor in the As(III) sorption process because both the speciation of As ions in an aqueous solution and the surface charge of the



**Fig. 4** Impacts to As(III) sorption capacity of MCS: **a** initial pH; **b** co-existing ions

biosorbent are pH dependent [16]. The speciation of As(III) species and equilibrium constants are shown below (Eq. 3) [48]:



Based on the  $\text{pK}_a$  values in Eq. 3, As(III) exists mainly as neutral  $\text{H}_3\text{AsO}_3$  at pH values lower than 9.2, and  $\text{H}_2\text{AsO}_3^-$  becomes dominant at pH values above 9.2 (Fig. 4a).

The effect of the initial pH of solution on the removal As(III) by MCS was investigated (Fig. 4a). The interaction between As(III) species and the surface of the MCS was influenced by solution pH, and the solution pH had a significant effect on As(III) sorption capacity of MCS ( $p < 0.01$ ). As the pH increased from 3 to 6, the As(III) sorption capacity increased from 984 to 1298  $\mu\text{g/g}$  and remained relatively constant between pH 6 and 9 (ranged between 1266 and 1320  $\mu\text{g/g}$ ) before reaching a maximum value of 1515  $\mu\text{g/g}$  at pH 10. Previously, similar sorption behaviors have been reported for As(III) on Fe oxides surfaces [49, 50]. The determined  $\text{pH}_{\text{pzc}}$  of the MCS was 4.2. Considering that at the pH values lower than 4.2 As(III) exists as neutral species ( $\text{H}_3\text{AsO}_3$ ), and the sorption capacity of MCS was the lowest at low pH values; thus, the electrostatic attraction was not responsible for As(III) uptake. The enhanced sorption capacity at  $5 < \text{pH} < 9$  can be attributed to outer-sphere complexation or inner-sphere complexation of As(III) with Fe oxides on the biosorbent surface [51]. The higher As(III) sorption capacity at pH 10 can be attributed to the stronger interaction between the deposited Fe oxides on the surface of MCS and As(III) ions. It has been reported that As(III) is sorbed more strongly at alkaline conditions, and in general, the maximum anion sorption occurs at pH values in the  $\text{pK}_a$  range of the conjugate acid (currently the  $\text{pK}_{\text{a}1} = 9.3$  for arsenious acid) (Eq. 3) [52, 53].



### 3.4 Effect of Co-existing Anions

Given that some anions in natural waters and anthropogenic wastewaters may compete with As(III) for the sorption sites on the MCS surface, investigating the possible competition between As(III) and examples of these anions is necessary. Therefore, the effect of different anions ( $\text{Cl}^-$ ,  $\text{NO}_3^-$ ,  $\text{PO}_4^{3-}$ ,  $\text{SeO}_3^{2-}$ ,  $\text{SeO}_4^{2-}$ ) on the biosorption of As(III) on MCS was investigated (Fig. 4b). The presence of  $\text{Cl}^-$  and  $\text{NO}_3^-$  did not have a significant effect ( $p > 0.05$ ) on the adsorption of As(III) when the concentration of the co-existing ions varied from 0 to 10 mg/L. However, the removal of As(III) was affected significantly ( $p < 0.01$ ) in the presence of  $\text{PO}_4^{3-}$ ,  $\text{SeO}_3^{2-}$ ,  $\text{SeO}_4^{2-}$ . With an increase in the concentration of ions from 0 to 10 mg/L, the As(III) sorption capacity decreased from 1404 to 1171, 1028, and 1186  $\mu\text{g/g}$  in the presence of  $\text{PO}_4^{3-}$ ,  $\text{SeO}_3^{2-}$ , and  $\text{SeO}_4^{2-}$ , respectively.

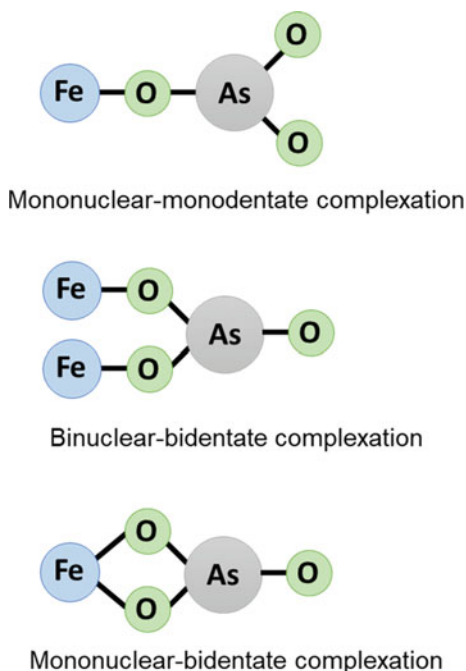
It has been reported that  $\text{Cl}^-$  and  $\text{NO}_3^-$  could only bind weakly to the surface of metal oxides by forming an outer-sphere complex [54, 55]. Generally, outer-sphere complexation is significantly affected by ionic strength and decreases with increasing ionic strength. Conversely, in the case of inner-sphere surface complexation, sorption capacity increases (or stays constant) with increasing ionic strength [56]. Considering that increase of  $\text{Cl}^-$  and  $\text{NO}_3^-$  concentrations did not influence the As(III) sorption capacity, the dominant mechanism of As(III) sorption onto the MCS can be attributed to inner-sphere surface complexation. On the other hand,  $\text{PO}_4^{3-}$  and  $\text{SeO}_3^{2-}$  bind strongly to metal oxides by forming inner-sphere complexes, and  $\text{SeO}_4^{2-}$  forms both relatively weaker inner- and outer-sphere complexes [55, 57]. Therefore, the impact of  $\text{PO}_4^{3-}$  and  $\text{SeO}_3^{2-}$  on As(III) sorption will be stronger than  $\text{SeO}_4^{2-}$ , which is in line with the amount of reduction in As(III) sorption capacity in the presence of  $\text{PO}_4^{3-}$  (233  $\mu\text{g/g}$ ) and  $\text{SeO}_3^{2-}$  (376  $\mu\text{g/g}$ ), and  $\text{SeO}_4^{2-}$  (218  $\mu\text{g/g}$ ).

Therefore, based on the results of sorption experiments and the effect of co-existing ions on the As(III) sorption capacity of MCS, it can be deduced that inner-sphere complexation is the main As(III) sorption mechanism. The inner-sphere complexation of As(III) with Fe oxides on the surface of MCS may take place by three complexation types, including mononuclear-monodentate, binuclear-bidentate, and mononuclear-bidentate complexes (Fig. 5).

### 3.5 Conclusion

The current study showed the viability of using Fe-modified canola straw (MCS) for the removal of As(III) from water. According to the results, the maximum As(III) sorption capacity of the MCS compares favorably to other similar sorbents in the literature. The Freundlich and Redlich–Peterson isotherm models were best-fitted to equilibrium data suggesting that a hybrid sorption mechanism took place in the removal of As(III) using MCS. Adsorption kinetics were well described by pseudo-second-order and Elovich models indicating the chemisorption nature of the process.

**Fig. 5** Possible inner-sphere surface complexes of As(III) formed on the surface of MCS



As(III) uptake by MCS increased with increasing pH from 3 to 5, remained constant in the pH range of 5–9, and reached a maximum of 1515  $\mu\text{g/g}$  at pH 10. Studying the inhibition effects of co-existing ions on As(III) sorption showed an insignificant effect of  $\text{Cl}^-$  and  $\text{NO}_3^-$ , while the effects of other ions were in the following order:  $\text{SeO}_3^{2-} > \text{PO}_4^{3-} > \text{SeO}_4^{2-}$ . Based on the sorption experiment results, inner-sphere complexation can be the main mechanism of As(III) sorption on MCS. While the batch adsorption experiments showed the promising potential of the MCS for As(III) sorption, further experiments are required to study the stability of the sorbent in the long-term adsorption process, identify a proper desorption agent, and investigate the regeneration-reuse capability of the adsorbent.

**Funding** The research was financially supported by the Saskatchewan Agriculture Development Fund and two NSERC Discovery Grants (K. McPhedran and J. Soltan). Kh. Zoroufchi Benis is supported by the Vanier Canada Graduate Scholarship and Saskatchewan Opportunity Scholarship.

## References

1. Chappells H, Parker L, Fernandez CV, Conrad C, Drage J, O'Toole G, Campbell N, Dummer TJB (2014) Arsenic in private drinking water wells: an assessment of jurisdictional regulations and guidelines for risk remediation in North America. *J Water Health* 12:372–392. <https://doi.org/10.2166/wh.2014.054>
2. Shakoor MB, Niazi NK, Bibi I, Shahid M, Sharif F, Bashir S, Shaheen SM, Wang H, Tsang DCW, Ok YS, Rinklebe J (2018) Arsenic removal by natural and chemically modified water melon rind in aqueous solutions and groundwater. *Sci Total Environ* 645:1444–1455. <https://doi.org/10.1016/J.SCITOTENV.2018.07.218>
3. Sadeghi F, Nasser S, Yunesian M, Nabizadeh R, Mosaferi M, Mesdaghinia A (2018) Carcinogenic and non-carcinogenic risk assessments of arsenic contamination in drinking water of Ardabil city in the Northwest of Iran. *J Environ Sci Heal Part A* 53:421–429. <https://doi.org/10.1080/10934529.2017.1410421>
4. Zhou Z, Liu Y, Liu S, Liu H, Zeng G, Tan X, Yang C, Ding Y, Yan Z, Cai X (2017) Sorption performance and mechanisms of arsenic(V) removal by magnetic gelatin-modified biochar. *Chem Eng J* 314:223–231. <https://doi.org/10.1016/J.CEJ.2016.12.113>
5. WHO (2011) Guidelines for drinking-water quality, 4th edn. World-Health-Organization, pp 315–318
6. Huq ME, Fahad S, Shao Z, Sarven MS, Al-Huqail AA, Siddiqui MH, Habib ur Rahman M, Khan IA, Alam M, Saeed M, Rauf A, Basir A, Jamal Y, Khan SU (2019) High arsenic contamination and presence of other trace metals in drinking water of Kushtia district, Bangladesh. *J Environ Manag* 242:199–209. <https://doi.org/10.1016/J.JENVMAN.2019.04.086>
7. Zoroufchi Benis K, Soltan J, McPhedran KN (2022) A novel method for fabrication of a binary oxide biochar composite for oxidative adsorption of arsenite: characterization, adsorption mechanism and mass transfer modeling. *J Clean Prod* 131832. <https://doi.org/10.1016/J.JCLEPRO.2022.131832>
8. Kumar ASK, Jiang S-J (2016) Chitosan-functionalized graphene oxide: a novel adsorbent an efficient adsorption of arsenic from aqueous solution. *J Environ Chem Eng* 4:1698–1713. <https://doi.org/10.1016/j.jece.2016.02.035>
9. Sogaard E (2014) Chemistry of advanced environmental purification processes of water: fundamentals and applications. Newnes
10. Zoroufchi Benis K, Motalebi Damuchali A, McPhedran KN, Soltan J (2020) Treatment of aqueous arsenic—a review of biosorbent preparation methods. *J Environ Manage* 273:111126. <https://doi.org/10.1016/j.jenvman.2020.111126>
11. Zoroufchi Benis K, Damuchali AM, Soltan J, McPhedran KN (2020) Treatment of aqueous arsenic—a review of biochar modification methods. *Sci Total Environ*. <https://doi.org/10.1016/j.scitotenv.2020.139750>
12. Roy P, Dey U, Chattoraj S, Mukhopadhyay D, Mondal NK (2017) Modeling of the adsorptive removal of arsenic(III) using plant biomass: a bioremedial approach. *Appl Water Sci* 7:1307–1321. <https://doi.org/10.1007/s13201-015-0339-2>
13. Siddiqui SI, Chaudhry SA (2017) Iron oxide and its modified forms as an adsorbent for arsenic removal: a comprehensive recent advancement. *Process Saf Environ Prot*. <https://doi.org/10.1016/j.psep.2017.08.009>
14. De D, Aniya V, Satyavathi B (2019) Application of an agro-industrial waste for the removal of As(III) in a counter-current multiphase fluidized bed. *Int J Environ Sci Technol* 16:279–294. <https://doi.org/10.1007/s13762-018-1651-9>
15. Vieira BRC, Pintor AMA, Boaventura RAR, Botelho CMS, Santos SCR (2017) Arsenic removal from water using iron-coated seaweeds. *J Environ Manag* 192:224–233. <https://doi.org/10.1016/J.JENVMAN.2017.01.054>
16. Zoroufchi Benis K, Soltan J, McPhedran KN (2021) Electrochemically modified adsorbents for treatment of aqueous arsenic: pore diffusion in modified biomass vs. biochar. *Chem Eng J* 423:130061. <https://doi.org/10.1016/j.cej.2021.130061>

17. Abid M, Niazi NK, Bibi I, Farooqi A, Ok YS, Kunhikrishnan A, Ali F, Ali S, Igalavithana AD, Arshad M (2016) Arsenic(V) biosorption by charred orange peel in aqueous environments. *Int J Phytorem* 18:442–449. <https://doi.org/10.1080/15226514.2015.1109604>
18. Ebrahimi R, Maleki A, Shahmoradi B, Daraei H, Mahvi AH, Barati AH, Eslami A (2013) Elimination of arsenic contamination from water using chemically modified wheat straw. *Desalin Water Treat* 51:2306–2316. <https://doi.org/10.1080/19443994.2012.734675>
19. Howell RJ, Alpers CN, Jamieson HE, Nordstrom DK, Majzlan J (2014) The environmental geochemistry of arsenic—an overview. *Rev Mineral Geochem* 79:1–16. <https://doi.org/10.2138/rmg.2014.79.1>
20. Tian Y, Wu M, Lin X, Huang P, Huang Y (2011) Synthesis of magnetic wheat straw for arsenic adsorption. *J Hazard Mater* 193:10–16. <https://doi.org/10.1016/J.JHAZMAT.2011.04.093>
21. Hao L, Zheng T, Jiang J, Hu Q, Li X, Wang P (2015) Removal of As(III) from water using modified jute fibres as a hybrid adsorbent. *RSC Adv* 5:10723–10732. <https://doi.org/10.1039/c4ra11901k>
22. Meng F, Yang B, Wang B, Duan S, Chen Z, Ma W (2017) Novel dendrimerlike magnetic biosorbent based on modified orange peel waste: adsorption-reduction behavior of arsenic. *ACS Sustain Chem Eng* 5:9692–9700. <https://doi.org/10.1021/acssuschemeng.7b01273>
23. Canola Council (2022) Grown on Canadian farms, consumed around the world [WWW Document]. <https://www.canolacouncil.org/about-canola/industry/>
24. Saskcanola (2022) The Canola Story [WWW Document]. <https://www.saskcanola.com/about/story.php>
25. Yousefi H (2009) Canola straw as a bio-waste resource for medium density fiberboard (MDF) manufacture. *Waste Manag* 29:2644–2648. <https://doi.org/10.1016/J.WASMAN.2009.06.018>
26. Pronyk C, Mazza G (2012) Fractionation of triticale, wheat, barley, oats, canola, and mustard straws for the production of carbohydrates and lignins. *Bioresour Technol* 106:117–124. <https://doi.org/10.1016/j.biortech.2011.11.071>
27. Zoroufchi Benis K, Shakouri M, McPhedran K, Soltan J (2020) Enhanced arsenate removal by Fe-impregnated canola straw: assessment of XANES solid-phase speciation, impacts of solution properties, sorption mechanisms, and evolutionary polynomial regression (EPR) models. *Environ Sci Pollut Res*. <https://doi.org/10.1007/s11356-020-11140-0>
28. Alchouron J, Navarathna C, Chludil HD, Dewage NB, Perez F, Hassan EB, Pittman CU Jr, Vega AS, Mlsna TE (2020) Assessing South American *Guadua chacoensis* bamboo biochar and Fe<sub>3</sub>O<sub>4</sub> nanoparticle dispersed analogues for aqueous arsenic(V) remediation. *Sci Total Environ* 706:135943. <https://doi.org/10.1016/j.scitotenv.2019.135943>
29. Jung K-W, Jeong T-U, Kang H-J, Chang J-S, Ahn K-H (2016) Preparation of modified-biochar from *Laminaria japonica*: Simultaneous optimization of aluminum electrode-based electro-modification and pyrolysis processes and its application for phosphate removal. *Bioresour Technol* 214:548–557. <https://doi.org/10.1016/j.biortech.2016.05.005>
30. Gautam SB, Alam MS, Kamsonlian S (2017) Adsorptive removal of As(III) from aqueous solution by raw coconut husk and iron impregnated coconut husk: kinetics and equilibrium analyses. *Int J Chem React Eng* 15. <https://doi.org/10.1515/ijcre-2016-0097>
31. Huang Z, Liang X, Hu H, Gao L, Chen Y, Tong Z (2009) Influence of mechanical activation on the graft copolymerization of sugarcane bagasse and acrylic acid. *Polym Degrad Stab* 94:1737–1745. <https://doi.org/10.1016/j.polymdegradstab.2009.06.023>
32. Qu G, Huang X, Yin Q, Ning P (2014) Dissolution of garlic stem in the 1-butylpyridinium bromide ionic liquid. *J Chem Eng Jpn* 47:435–441. <https://doi.org/10.1252/jcej.13we163>
33. Nadeem R, Manzoor Q, Iqbal M, Nisar J (2016) Biosorption of Pb(II) onto immobilized and native *Mangifera indica* waste biomass. *J Ind Eng Chem* 35:185–194. <https://doi.org/10.1016/j.jiec.2015.12.030>
34. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA, Hughes C, Lasch P, Martin-Hirsch PL, Obinaju B, Sockalingum GD, Sulé-Suso J, Strong RJ, Walsh MJ, Wood BR, Gardner P, Martin FL (2014) Using Fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc* 9:1771–1791. <https://doi.org/10.1038/nprot.2014.110>

35. Niazi NK, Bibi I, Shahid M, Ok YS, Shaheen SM, Rinklebe J, Wang H, Murtaza B, Islam E, Farrakh Nawaz M, Lüttge A (2018) Arsenic removal by Japanese oak wood biochar in aqueous solutions and well water: Investigating arsenic fate using integrated spectroscopic and microscopic techniques. *Sci Total Environ* 621:1642–1651. <https://doi.org/10.1016/J.SCITOTENV.2017.10.063>
36. Dymińska L, Gągor A, Hanuza J, Kulma A, Preisner M, Zuk M, Szatkowski M, Szopa J (2014) Spectroscopic characterization of genetically modified flax fibers. *J Mol Struct* 1074:321–329. <https://doi.org/10.1016/j.molstruc.2014.06.013>
37. El-Moselhy MM, Ates A, Çelebi A (2017) Synthesis and characterization of hybrid iron oxide silicates for selective removal of arsenic oxyanions from contaminated water. *J Colloid Interface Sci* 488:335–347. <https://doi.org/10.1016/J.JCIS.2016.11.003>
38. Zeng L (2004) Arsenic adsorption from aqueous solutions on an Fe(III)-Si binary oxide adsorbent. *Water Qual Res J Can* 39:267–275. <https://doi.org/10.2166/wqrj.2004.037>
39. Zeng L (2003) A method for preparing silica-containing iron(III) oxide adsorbents for arsenic removal. *Water Res* 37:4351–4358. [https://doi.org/10.1016/S0043-1354\(03\)00402-0](https://doi.org/10.1016/S0043-1354(03)00402-0)
40. Davodia M, Alidadib H, Ramezanib A, Jamali-Behnamc F, Bonyadib Z (2019) Study of the removal efficiency of arsenic from aqueous solutions using Melia azedarach sawdust modified with FeO: isotherm and kinetic studies. *Desalin Water Treat* 137:292–299
41. Sert S, Celik A, Tirtom VN (2017) Removal of arsenic (III) ions from aqueous solutions by modified hazelnut shell. *Desalin Water Treat* 75:115–123
42. Pholosi A, Naidoo BE, Ofomaja AE (2018) Clean application of magnetic biomaterial for the removal of As(III) from water. *Environ Sci Pollut Res* 25:30348–30365. <https://doi.org/10.1007/s11356-018-2990-2>
43. Singh P, Pal P, Mondal P, Saravanan G, Nagababu P, Majumdar S, Labhsetwar N, Bhowmick S (2021) Kinetics and mechanism of arsenic removal using sulfide-modified nanoscale zerovalent iron. *Chem Eng J* 128667. <https://doi.org/10.1016/j.cej.2021.128667>
44. Zhuang H, Zhong Y, Yang L (2020) Adsorption equilibrium and kinetics studies of divalent manganese from phosphoric acid solution by using cationic exchange resin. *Chin J Chem Eng* 28:2758–2770. <https://doi.org/10.1016/j.cjche.2020.07.029>
45. Kim S, Gholamirad F, Yu M, Park CM, Jang A, Jang M, Taheri-Qazvini N, Yoon Y (2021) Enhanced adsorption performance for selected pharmaceutical compounds by sonicated Ti3C2TX MXene. *Chem Eng J* 406:126789. <https://doi.org/10.1016/j.cej.2020.126789>
46. Al-Ghouti MA, Da'ana DA (2020) Guidelines for the use and interpretation of adsorption isotherm models: a review. *J Hazard Mater*. <https://doi.org/10.1016/j.jhazmat.2020.122383>
47. Xiang L, Niu CG, Tang N, Lv XX, Guo H, Li ZW, Liu HY, Lin LS, Yang YY, Liang C (2020) Polypyrrole coated molybdenum disulfide composites as adsorbent for enhanced removal of Cr(VI) in aqueous solutions by adsorption combined with reduction. *Chem Eng J* 408:127281. <https://doi.org/10.1016/j.cej.2020.127281>
48. Yu X, Tong S, Ge M, Zuo J, Cao C, Song W (2012) One-step synthesis of magnetic composites of cellulose@iron oxide nanoparticles for arsenic removal. *J Mater Chem A* 1:959–965. <https://doi.org/10.1039/C2TA00315E>
49. Pervez MN, Fu D, Wang X, Bao Q, Yu T, Naddeo V, Tian H, Cao C, Zhao Y (2021) A bifunctional  $\alpha$ -FeOOH@GCA nanocomposite for enhanced adsorption of arsenic and photo Fenton-like catalytic conversion of As(III). *Environ Technol Innov* 22:101437. <https://doi.org/10.1016/J.ETI.2021.101437>
50. Yu F, Sun S, Ma J, Han S (2015) Enhanced removal performance of arsenate and arsenite by magnetic graphene oxide with high iron oxide loading. *Phys Chem Chem Phys* 17:4388–4397. <https://doi.org/10.1039/C4CP04835K>
51. Zubair YO, Fuchida S, Tokoro C (2020) Insight into the mechanism of arsenic(III/V) uptake on mesoporous zerovalent iron-magnetite nanocomposites: adsorption and microscopic studies. *ACS Appl Mater Interfaces* 12:49755–49767. [https://doi.org/10.1021/ACSAMI.0C14088/SUPPL\\_FILE/AM0C14088\\_SI\\_001.PDF](https://doi.org/10.1021/ACSAMI.0C14088/SUPPL_FILE/AM0C14088_SI_001.PDF)
52. Brechbühl Y, Christl I, Elzinga EJ, Kretzschmar R (2012) Competitive sorption of carbonate and arsenic to hematite: combined ATR-FTIR and batch experiments. *J Colloid Interface Sci* 377:313–321. <https://doi.org/10.1016/J.JCIS.2012.03.025>

53. Manning BA, Fendorf SE, Goldberg S (1998) Surface structures and stability of arsenic(III) on goethite: spectroscopic evidence for inner-sphere complexes. *Environ Sci Technol* 32:2383–2388. <https://doi.org/10.1021/ES9802201>
54. Ma Z, Shan C, Liang J, Tong M (2018) Efficient adsorption of Selenium(IV) from water by hematite modified magnetic nanoparticles. *Chemosphere* 193:134–141. <https://doi.org/10.1016/j.chemosphere.2017.11.005>
55. Xin Y, Gu P, Long H, Meng M, Yaseen M, Su H (2021) Fabrication of ferrihydrite-loaded magnetic sugar cane bagasse charcoal adsorbent for the adsorptive removal of selenite from aqueous solution. *Colloids Surf A: Physicochem Eng Aspects* 614:126131. <https://doi.org/10.1016/j.colsurfa.2020.126131>
56. Zoroufchi Benis K, McPhedran KN, Soltan J (2022) Selenium removal from water using adsorbents: a critical review. *J Hazard Mater* 424:127603. <https://doi.org/10.1016/J.JHAZMAT.2021.127603>
57. Zhang N, Lin LS, Gang D (2008) Adsorptive selenite removal from water using iron-coated GAC adsorbents. *Water Res* 42:3809–3816. <https://doi.org/10.1016/J.WATRES.2008.07.025>

# A Framework for the Economic Assessment of a More Sustainable Wastewater Management System



Bibhas B. Tanmoy and M. Abdel-Raheem

**Abstract** Blackwater and greywater are transported through the same sewage pipe in the traditional wastewater network in a city. This mixture turns both the blackwater and greywater into blackwater, which needs to go through extensive treatment processes before being discharged. On the other hand, greywater can be reused for numerous purposes, even after basic filtration. Fortunately, there are varieties of household greywater reuse systems (GWS) available in the market that can make good use of this resource by reusing this massive amount of generated greywater with its internal filtration or UV disinfection unit. This stops this massive amount of greywater from leaving the household, eventually reducing the sewage bill drastically. Furthermore, the lower volume of blackwater also allows for smaller sewer pipes and reduces the number of wastewater treatment plants required in a city. These reduced wastewater treatment plants will result in a significant drop in cost per capita for the city's population, as well as less environmental impact. Altogether, it can be affirmed that the outputs of this wastewater management system will satisfy the economic, environmental, and social concerns for sustainable cities. However, dual plumbing and greywater system installation, operation, maintenance, and replacement (OMR) costs will initially add new expenses to the bills. Therefore, an economic assessment is necessary to verify that the assumed savings can significantly outweigh this cost, resulting in a net benefit. This study provides the framework for conducting this economic assessment of implementing this system. The framework will act as a guideline to conduct a thorough economic assessment of various similar sustainable design concepts for sustainable cities.

**Keywords** Life cycle cost analysis • Greywater reuse • Wastewater management • Sustainable construction • Sustainable cities

---

B. B. Tanmoy · M. Abdel-Raheem (✉)  
University of Texas Rio Grande Valley, Edinburg, TX, USA  
e-mail: [mohamed.abdelraheem@utrgv.edu](mailto:mohamed.abdelraheem@utrgv.edu)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_62](https://doi.org/10.1007/978-3-031-34593-7_62)

977

# 1 Introduction

Water is the most valuable natural resource in the world. By 2050, global water use is expected to increase by 55%, resulting in scarcity and competition among water users [1]. As a result, many sustainable water reuse schemes are being developed and implemented worldwide, particularly in urban areas. The reason for this is that cities are seen as producers of secondary resources (their own), such as greywater. These resources are undeniably crucial in any future sustainable city design that attempts to make cities more resilient and self-sufficient [2].

Greywater comprises the major share of the household wastewater amount. Not only it can be discharged into the waterbodies directly [3], but it also possesses a high potential for reuse and reducing the water footprint. Therefore, it necessitates thorough research on how the maximum potential of greywater can be achieved. To solve this challenge in a sustainable manner, a new wastewater management system is proposed in this study, and its economic assessment framework is developed. The new system proposes the generalization of greywater systems in every household. These systems will treat all the greywater generated in each household in its treatment unit and make it available for reuse almost instantly. As a result, this vast amount of greywater will not leave the houses, rather it will be reused for multiple purposes such as irrigation, car washing, toilet flushing, laundry, dishwashing. The framework will first demonstrate how the implementation of this system will alter the wastewater management scenario in a city, followed by a demonstration of the significant cost and non-cost elements, necessary tools, and all the calculation procedures required to determine this system's feasibility in practical life. The framework will support as a guideline for researchers, water engineers, environmentalists, town planners, and policymakers who are willing to determine the economic viability of a sustainable wastewater management system.

## 2 Background Study

### 2.1 Sustainable Strategies

In 2015, the United Nations (UN) established 17 Sustainable Development Goals (SDG) for 2030, including access to clean water (Goal 6), sustainable cities and communities (Goal 11), and actions to mitigate the effects of climate change (Goal 12) (THE 17 GOALS | Sustainable Development n.d.). Various water demand control strategies have recently been implemented to achieve these goals. The importance of establishing solutions that contribute to more sustainable use of reclaimed water is highlighted in current water and wastewater management issues. This type of water reuse is promoted because it can (i) restore freshwater supplies, (ii) limit sewage discharge into bodies of water, and (iii) reduce rising water treatment costs [4].



Numerous studies have been conducted in recent decades to suggest alternative designs and strategies for reducing the amount of water wasted in homes. Almost the majority of them are centred on decentralized wastewater management. Otterpohl et al. [5] suggested the separation of domestic wastewater from stormwater and the fractionation of domestic wastewater into several streams for a more effective treatment [5]. Kjerstadius et al. [6] analysed the sustainability factors of separating food waste (FW) from domestic wastewaters through five urban systems for collection, transport, treatment, and nutrient recovery from blackwater, greywater, and food waste using data from implementations in Sweden or northern Europe [6]. Walker and Duquette [7] noticed that the wastewater stream that goes away unutilized could be used as a source of energy that is required to pump water in that same building. The research evaluates the possible techno-economic benefits of recovering energy from building wastewater systems utilizing a tank and turbine generator system [7]. Chen et al. [8] also used a similar concept regarding heat generation. This research aimed to establish a system for evaluating the possibility of utilizing sewage heat based on a suggested urban sewage state prediction model (USSPM). In this work, a prediction model was built and then deployed to a real-world situation to examine the possibilities for sewage heat utilization utilizing case studies [8].

## 2.2 *Greywater Systems*

Greywater systems, also known as whole-house greywater systems, central water recycling systems, or GWS, are a recent technological advancement in the utilization of greywater generated in standard houses. These systems use various treatment methods, resulting in a wide range of system scopes and pricing. Product variants range from greywater diversion systems, which direct generated greywater to irrigation sprinklers, to more advanced variants equipped with filtration and/or UV disinfection, which provides quality water comparable to freshwater but not necessarily drinkable [9]. Almost all of these systems have two main components: a surge tank that retains the produced wastewater and a filtering chamber that filters the wastewater through and surges out as clean water, which is then connected to the inflow pipes for use in activities like garden sprinklers [10], toilet flushing, car washing, laundry, dishwashing, etc. (The Complete Beginner's Guide to Greywater Systems n.d.). Surge tank storage, filtration rates, outflow capacities, and pumping or no pumping systems are used to identify model variants. The water intended for reuse is drawn from the system's tank rather than the municipal water supply because these systems have their own tank. Consequently, the usage is not charged to the water metre, resulting in a lower utility bill at the end of the month, saving up to 50% [9].

### 2.3 Life Cycle Cost Analysis

Life cycle costing (LCC), also known as life cycle cost analysis (LCCA), establishes economic viability, a significant component in deciding whether or not to use this innovative wastewater management method. In essence, economic analysis using the LCC technique necessitates not only a cost item (i.e. the initial cost, energy cost, and maintenance and replacement costs) for the target component over its lifetime but also several assumptions, such as the discount rate, analysis period, and maintenance and replacement information [11]. This technique is commonly used in the water and energy industries. Following that, the results are displayed in a variety of forms. Financial indicators such as net present value (NPV), return on investment (ROI), benefit-cost ratio (BCR), and payback period (PP) were employed by [12] in their model [12]. Similar measures, as well as an internal rate of return (IRR), were used by Matos et al. [13]. Khastagir and Jaysuriya [14] employed the Levelized cost (LC), also known as “amortized cost”, strategy defined for Australia’s water and energy sectors [14].

## 3 Methodology

This study applied the life cycle cost analysis (LCCA) approach to conduct the economic assessment of this new system. The Life Cycle Costing Manual for the Federal Energy Management Program (FEMP), 2020 Edition, from the National Institute of Standards and Technology (NIST) [15] was a beneficial source in the creation of the life cycle costing model utilized in this study.

After reviewing available literature, a base case (BC) and the alternative case (A) were defined. A benchmark against which the alternative case can be compared is established by defining a base case. Here, the alternative case represents the implementation of the proposed system in a hypothetical manner.

Upon completing these phases, the parameters that would be required to conduct the assessment were defined. These parameters mostly involved the financial metrics, such as construction costs, maintenance costs, and utility bills. Government websites were researched to gather the cost indices and escalation rates, and globally established databases and tools were used to estimate the cost schemes, along with necessary equations and procedures collected from the literature. Different programs and applications were researched to determine the best option to conduct the economic assessment based on these parameters and within the scope of the study.

4 Study Parameters

The economic assessment requires two significant parameters to conduct the assessment: the wastewater data and the cost data. Other parameters such as time and rate are used for the calculations. Some minor assumptions are also taken into consideration to avoid extensive complexity.

4.1 Wastewater Data

According to Tchobanoglous et al. [16], every household exerts approximately 65 Gallons of wastewater per capita per day (GPCD). Table 1 shows the distribution of the sources of this volume, considering the standard water-efficient fixtures installed in the household [16].

Table 1 US average indoor water use distribution

Use	WW flow, GPCD
Bath	1.2
Laundry	15
Dishwashing	0.9
Faucet	10.4
Shower	11.7
Toilet flushing	18.2
Other domestic	1.4
Leakage	6.2
Total	65

Among this distribution, toilet flushing and dishwashing are the main sources of the blackwater portion, leaving 43.4 GPCD of greywater in the total wastewater flow volume.

Table 2 Wastewater outflow comparison

Settings	BW outflow, GPCD	Reusable GW, GPCD	Total WW outflow, GPCD
Conventional	65	0	65
Proposed	21.6	43.4	21.6

4.2 Cost Data

4.2.1 Cost Elements

The base case and the alternative case will account for a specific set of cost elements, which will need to be calculated and then compared. The following table shows these set of cost elements for both cases.

Table 3 Cost elements

Base case, conventional, BC	Alternative case, proposed system, A
Wastewater network costs	Wastewater network costs
Wastewater treatment plant costs	Greywater system purchase, installation, operation, maintenance, and replacement costs
Water costs	Wastewater treatment plant costs
	Household dual plumbing costs
	Water costs

4.2.2 Cost Factors

The implementation of this system will generate substantial changes in the costing schemes over the course of time. These changes include both increases and decreases. The following table demonstrates these factors:

Table 4 Cost factors

Case	Cost increasing factors	Cost decreasing factors
Conventional settings	None	None
Proposed system	Greywater system purchase	Lateral sewer pipe materials
	Greywater system installation, operation, maintenance, and replacement	Lateral sewer pipe construction
	Household dual plumbing	Lateral sewer pipe maintenance
	WWTP chemical costs	Number of WWTP
		Water costs

### **4.3 Rates**

This framework uses multiple sets of rates for an accurate assessment. However, the application of the computer programs will automatically compute most of the region-dependent rates such as labour rates, overhead, and profits (O&P) costs once the region is specified. However, the nominal discount rate obtained from the FEMP publication will be used as the interest rate in the equations for calculating the present values of the life cycle costs.

### **4.4 Time Data**

Because LCCA deals exclusively with money, the study must also verify that the time value of money is accounted for. There are numerous parts of time that are connected with an LCCA such as (1) base date, (2) study term, and (3) service period. The base date is the commencement of the project and is the point in time that all project expenses are discounted to [15]. The research period is the amount of time that expenses connected to a project and its alternatives are of interest to the decision-maker [15]. For a non-energy LCCA, the recommended study length is 40 years.

### **4.5 Assumptions**

This study has a set of assumptions. First of all, stormwater runoff and infiltration are also not considered in the assessment. Secondly, this study considers all the payments occurring at the end of the time cycle, including capital replacement. It is also assumed that when a greywater system reaches the end of its functional life, in this case after ten years (The Complete Beginner's Guide to Greywater Systems n.d.), it is discarded since it has no value. It should also be noted that this study uses the US standard sewer piping size for the wastewater network [17], and the greywater systems are installed in residential buildings only, not commercial or industrial buildings.

## **5 Economic Assessment Framework**

### **5.1 Assessment Baseline**

Before initiating the cost assessment, a city is defined for the case study, and the population of a city is extracted from its respective census bureau. Later, its water footprint is calculated by multiplying this population with the wastewater flow amount

discussed previously. These two parameters will be the base of all the estimations required for this assessment. Life cycle costs of 40 years will be calculated for all the major sections required to do the comparative assessment. The generic equation for computing life cycle cost can be written as follows [15]:

$$LCC = I - \text{Res} + \text{Repl} + W + E + \text{OMR} + X, \quad (1)$$

where LCC represents life cycle cost,  $I$  stands for initial investment, which is the purchase and installation costs, Res is residual value, Repl equals replacement cost,  $W$  is the water cost,  $E$  is the energy cost, OMR represents the operation and maintenance cost, and  $X$  is any other relevant costs.

This equation will be modified in accordance with the elements involved in each set of calculations. All the life cycle costs will be divided later by the responsible population amount to get the cost per capita.

## 5.2 Calculation Procedures

### 5.2.1 Wastewater Network Costs

Two tools would be used to calculate this parameter. Firstly, CIGMAT-LCC, a spreadsheet-based model developed by the Center for Innovative Grouting Materials and Technology (CIGMAT), University of Houston [17]. This model estimates the life cycle cost (LCC) for constructing, operating, and maintaining a wastewater sewer network for a specific US city when provided the region-specific values. For instance, in the case of Houston, the population would be 2,304,580 and the population increase will be 2.3% (Houston Metro Area Population 1950–2022 n.d.). The model comes with all the parameters such as housing units, pipe diameters, pipe maintenance costs according to the US average as default. To estimate the cost differences, these parameters will be changed in accordance to the design specifications.

Another parameter that varies between cities is the cost indices. For these values, pricings from RSMeans Data would be used. The acquired data will then be updated in the CIGMAT-LCC model to get a more accurate cost estimation for any specific city.

The following equation would be used to calculate the LCC for the wastewater network:

$$LCC_{BC-WN} = (\text{Pipe Material Cost} + \text{Pipe Installation Cost} + \text{Manholes Cost} + \text{Pipe Maintenance Cost})_{BC}, \quad (2)$$

where BC represents the conventional settings (base case) and WN stands for wastewater network costs. All the costs are discounted to present values, which will later be divided by the population ( $P$ ) of the city to acquire the cost/capita.

**Table 5** Street lateral piping distribution scenario

Pipe diameters	US standard amount in the whole WW network (%)	Amount in proposed system in the whole WW network (%)
4 inches	34	70
6 inches	33	30
8 and above inches	33	0

$$WN_{BC}(\$/Capita) = LCC_{BC-WN} / P.$$

(3)

For the calculation of the proposed system, as the wastewater flow volume will be reduced, the piping classifications will be changed in the model as well. Table 5 demonstrates the change pipe diameter scenario of the two cases.

The major alteration occurs in the street lateral pipe class, as these are mainly the pipes that transport the wastewater from the residential buildings to the sewer main lines. As stated in Table 2, implementation of this system will result in a 65% reduction in wastewater outflow, which can be transported with pipes even smaller than 4 inches, as the pipe flows half-full [16]. However, the 4-inch diameter pipes are the smallest standard and construction-friendly options; therefore, this class amount is increased to 70%, as it will serve the purpose easily. The 30% of 6-inch pipes are also proposed as a safety consideration, especially for larger and multipurpose homes. The life cycle cost calculation will be a similar process as in the base case, with different values.

$$LCC_{A-WN} = (\text{Pipe Material Cost} + \text{Pipe Installation Cost} + \text{Manholes Cost} + \text{Pipe Maintenance Cost})_A$$

(4)

$$WN_A(\$/Capita) = LCC_{A-WW} / P,$$

(5)

where A represents the proposed system.

5.2.2 Wastewater Treatment Plant Costs

The computer program CapdetWorks would be used to determine this cost. In the first stage, a standard wastewater treatment plant process chain is programmed into the software accordingly. To estimate the life cycle cost of a plant, the program requires a wastewater inflow amount, average plant capacity, the influent wastewater characteristics, and the region-specific cost parameters such as discount rate, labour costs, operation costs, land costs. The US standards of these values are preset on the program initially according to the EPA database. An important point to note here is that although the assessment focuses on the residential wastewater scheme, the treatment plant accounts for all the wastewater in the city, which includes the commercial,

recreational, and industrial wastewater. The US standard gallon wastewater flow per capita for each sector was collected from the CIGMAT-LCC model [17]. The other cost parameters will be changed according to the city that is being worked on. For instance, an assessment for Houston will require the land cost of \$21,000/acre (Seidel n.d.).

A treatment plant of a standard capacity will be chosen at first, which will be the same in both cases, the life cycle cost will be computed for one plant and will be later multiplied based on the population of the city, determining how many of these plants would be required to treat this amount of wastewater for the whole city. The key player in the cost differences will be the number of plants required, and the cost of treating that quality of wastewater with each of the plants. For the base case, the US standard wastewater characteristics will be input into the program [16]. The region-specific parameters would be adjusted accordingly, and finally, the LCC of one plant will be computed ( $LCC_{Plant}$ ). The WW flow will be multiplied by the population and then divided by the capacity of the plant to calculate the required number of plants. This number will be again multiplied with the LCC of one plant for the total LCC responsible for the treatment plants for the whole city, and then dividing this cost by the population will compute the cost per capita.

$$\text{Number of Plants Required}_{BC} = (\text{Avg. per capita WW flow}_{BC} * P) / \text{Avg. Plant Capacity} \quad (6)$$

$$LCC_{BC - TP} = LCC_{Plant - BC} * \text{Number of Plants required}_{BC} \quad (7)$$

$$TP_{BC} (\$/\text{Capita}) = LCC_{BC - TP} / P, \quad (8)$$

where TP stands for treatment plant costs. The operation and periodic maintenance costs are discounted automatically in this program; therefore, no further discounting would be required.

For the calculation of the proposed system, the wastewater characteristics will change. It can be seen from Table 6 that wastewater volume has decreased by 24.11%, but that is the amount of greywater being absent, which means even though wastewater volume is lower now, the blackwater concentration is higher, meaning the constituents in this volume is higher in mg/L scale. This will change the wastewater characteristics input in the program, which needs to be adjusted before running the costing simulation. The rest calculation procedure will remain the same.

$$\text{Number of Plants Required}_A = (\text{Avg. per capita WW flow}_A * P) / \text{Avg. Plant Capacity} \quad (9)$$

$$LCC_{A - TP} = LCC_{Plant - A} * \text{Number of Plants required}_A \quad (10)$$

$$TP_A (\$/\text{Capita}) = LCC_{A - TP} / P. \quad (11)$$



**Table 6** Wastewater outflow from different sectors

Sector	Conventional, GPCD	Proposed system, GPCD
Residential	65	21.6
Industrial	60	60
Business	25	25
Educational	20	20
Recreational	10	10
Total	180	136.6

### 5.2.3 Water Cost Calculation

Typically, the water cost includes both the water and sewage charges. This cost breakdown varies from city to city. For instance, in Houston, the water section of the utility bill includes a base price and a constant rate for subsequent kilogallons consumed. The sewage section likewise consists of a base price but is followed by two blocks accounting for subsequent consumption. These charges are then added together to create a comprehensive utility bill (Houston Public Works—Utility Rates 2021 n.d.). The following equation can be a generic representation of the water cost calculation:

$$w_{\text{flat}} = [\text{Water Base Charge} + (\text{Water Rate} * \text{Consumption})] + [\text{Sewage Base Charge} + (\text{Sewage Rate} * \text{Consumption})_{\text{First Block}} + (\text{Sewage Rate} * \text{Consumption})_{\text{Second Block}}] \quad (12)$$

Which can be later discounted using the following equation [15]:

$$W (\$/\text{Capita}) = \frac{w_{\text{flat}} [1 - (1 + g_w / (1 + i))^n]}{i - g_w} + \frac{s_{\text{flat}} [1 - (1 + g_s / (1 + i))^n]}{i - g_s}, \quad (13)$$

where  $W$  represents the total water cost per capita, and on the other side,  $w$  is the Water portion,  $s$  is the Sewage portion,  $i$  is the nominal discount rate,  $n$  is the study period, and  $g_w$ ,  $g_s$  are the expected increase of water and sewage rate every year, respectively.

The implementation of the new system will change the consumption amount in the equation, hence altering the overall water cost.

### 5.2.4 Greywater System Cost Calculation

This calculation will only be applied to the costs for the proposed system (Alternative Case, A). The LCC for a greywater system involves multiple parameters, as can be seen from the equation:

$$LCC_{GWS} = I_{GWS} - Res_{GWS} + Repl_{GWS} + W_{GWS} + E_{GWS} + OMR_{GWS}, \quad (14)$$

where GWS represents the Greywater system[15].

A point to note here is that the installation of this system adds to the energy cost only if a pump-operated greywater system is installed. This energy consumption can be determined using the power of the pump and the number of hours per month required to treat the volume of greywater generated during the same period. Finally, the energy cost is determined using the equation below:

$$e_{\text{flat}} = \text{Power (kW)} * \frac{\text{Water Amount (gal)}}{\text{Capacity} \left( \frac{\text{gal}}{h} \right)} * \text{energy rate} \left( \frac{\$}{\text{kWh}} \right), \quad (15)$$

which can be discounted using the following equation:

$$E_{GWS} = \frac{e_{\text{flat}} \left[ 1 - \left( 1 + g_e / (1 + i) \right)^n \right]}{i - g_e}, \quad (16)$$

where  $g_s$  is the expected increase of energy rate every year[15].

All the other single payments such as maintenance costs, replacement costs, residual value can be discounted to present value using the following formula [15]:

$$PV = F(1 + i)^{-n}, \quad (17)$$

where  $F$  is the payment that will occur at a point in future.

Annuities, excluding water and energy bills, such as the annual maintenance costs can be discounted to present value using the following formula [15]:

$$P = A \left[ \frac{(1 + i)^n - 1}{i(1 + i)^n} \right], \quad (18)$$

where  $A$  is the payment occurring annually, and  $P$  is the discounted value of the amount.

After the life cycle cost of a greywater system has been calculated, it will be divided by the number of residents one system accounts for. This portion of the calculation will be determined based on what sort of greywater system model is being used. For instance, if the system model is a residential variant and is installed in a four-person household, the LCC will be divided by 4, whereas for commercial variants, which have higher capacity, one system would count for the number of

families the system is serving. This is the only calculation in the study, that does not count the whole population of the city, but rather the population in one household.

$$GWS_A (\$/Capita) = LCC_{GWS} / (\text{Residents in the house}). \quad (19)$$

### 5.3 Total Cost Comparison

After all the individual costs per capita in each case set have been calculated, it can be summed up to determine the total cost per capita (CPC) in each case.

$$CPC_{BC} (\$/Capita) = WN_{BC} + TP_{BC} + W_{BC} \quad (20)$$

$$CPC_A (\$/Capita) = WN_A + TP_A + GWS_A. \quad (21)$$

A point to note here is that the water cost is a part of the greywater system's total cost ( $GWS_A$ ); therefore, for the alternative case, the water cost ( $W_A$ ) is not present in the equation for  $CPC_A$ .

### 5.4 Decision Factor: Net Savings

The study uses two cases to compare and determine the viability of this new system, by calculating the cost per capita (CPC) for each system. Multiplying the CPC values for both cases will compute the total life cycle cost (LCC) of both cases.

$$LCC_{BC} = CPC_{BC} \times \text{Population} \quad (22)$$

$$LCC_A = CPC_A \times \text{Population}. \quad (23)$$

These two LCC values will be compared using the following equation, also known as Net Savings [NS] [15]:

$$NS = LCC_{BC} - LCC_A. \quad (24)$$

If the NS is positive, it indicates that the new system will result in savings; if it is negative, it suggests that the choice will result in a loss project. The system will be regarded as neither advantageous nor detrimental if the NS is zero [15, 18].

## 6 Discussion

The framework developed in this study was built to perform a comprehensive life cycle cost analysis for a new wastewater management system. This system is assumed to result in a reduction of wastewater treatment plants, a major threat to the environment, deeming itself an environment-friendly concept. Furthermore, promoting the reuse of wastewater reduces pollution and pressure on natural resources, resulting in a healthy society. Altogether, implementation of this benefits the community socially, environmentally, and economically, satisfying the three criteria for sustainability. However, apart from the economic part, this framework does not analyse the other two criteria. Applying life cycle analysis (LCA) can solve this issue, which will address the environmental benefits of this system. Community surveys can help regarding the social impact as well.

## 7 Future Work

The future work of this study would be the results of the assessment of this system, using practical data sets of different sustainable cities. Furthermore, outputs from the different analyses can also involve comparison between cities, as well as sensitivity analyses on the key parameters.

## 8 Conclusion

Based on technical, quantitative, and financial parameters, the study provides a guideline to conduct an economic assessment of a new wastewater management system. Economic assessment is a very important metric for a new system since it is a firm commitment to society, the economy, and the environment. With concerns about sustainability growing among the general public, the decrease in natural resource use is one means of saving and protecting the environment as well as saving money for homeowners. LCCA is a very beneficial approach to evaluating designs, and projects. By employing cost data that is significant to investors and decision-makers, the computation of life cycle costs may clearly depict the overall cost associated with a given project and its potential alternatives. Additionally, LCCA has established a niche in the examination of sustainable designs in urban landscapes.

## References

1. Dumit Gómez Y, Teixeira LG (2017) Residential rainwater harvesting: Effects of incentive policies and water consumption over economic feasibility. *Resour Conserv Recycl* 127:56–67. <https://doi.org/10.1016/j.resconrec.2017.08.015>
2. Burszta-Adamiak E, Spychalski P (2021) Water savings and reduction of costs through the use of a dual water supply system in a sports facility. *Sustain Cities Soc* 66:102620. <https://doi.org/10.1016/j.scs.2020.102620>
3. US EPA O (2015) Wastewater technology fact sheets [overviews and factsheets]. US EPA. <https://www.epa.gov/septic/wastewater-technology-fact-sheets>
4. Carden K (2006) Understanding the use and disposal of greywater in the non-sewered areas of South Africa. <https://open.uct.ac.za/handle/11427/14591>
5. Otterpohl R, Braun U, Oldenburg M (2004) Innovative technologies for decentralised water-, wastewater and biowaste management in urban and peri-urban areas. *Water Sci Technol* 48(11–12):23–32. <https://doi.org/10.2166/wst.2004.0795>
6. Kjerstadius H, Haghighatafshar S, Davidsson Å (2015) Potential for nutrient recovery and biogas production from blackwater, food waste and greywater in urban source control systems. *Environ Technol* 36(13):1707–1720. <https://doi.org/10.1080/09593330.2015.1007089>
7. Walker T, Duquette J (2022) Performance evaluation of a residential building-based hydroelectric system driven by wastewater. *Sustain Cities Soc* 79:103694. <https://doi.org/10.1016/j.scs.2022.103694>
8. Chen W-A, Lim J, Miyata S, Akashi Y (2022) Methodology of evaluating the sewage heat utilization potential by modelling the urban sewage state prediction model. *Sustain Cities Soc* 80:103751. <https://doi.org/10.1016/j.scs.2022.103751>
9. Ferguson D (2014) Greywater systems: can they really reduce your bills? *The Guardian*. <https://www.theguardian.com/lifeandstyle/2014/jul/21/greywater-systems-can-they-really-reduce-your-bills>
10. Simple Greywater Systems for Your Home (2019) The tiny life. <https://thetinylife.com/greywater-systems/>
11. Cho K, Chang H, Jung Y, Yoon Y (2017) Economic analysis of data center cooling strategies. *Sustain Cities Soc* 31:234–243. <https://doi.org/10.1016/j.scs.2017.03.008>
12. Morales-Pinzón T, Lurueña R, Gabarrell X, Gasol CM, Rieradevall J (2014) Financial and environmental modelling of water hardness—implications for utilising harvested rainwater in washing machines. *Sci Total Environ* 470–471:1257–1271. <https://doi.org/10.1016/j.scitotenv.2013.10.101>
13. Matos C, Bentes I, Santos C, Imteaz M, Pereira S (2015) Economic analysis of a rainwater harvesting system in a commercial building. *Water Resour Manag: Int J* 29(11):3971–3986. Published for the European Water Resources Association (EWRA)
14. Khastagir A, Jayasuriya N (2011) Investment evaluation of rainwater tanks. *Water Resour Manag* 25(14):3769–3784. <https://doi.org/10.1007/s11269-011-9883-1>
15. Kneifel J, Webb D (2020) Life cycle cost manual for the federal energy management program (NIST HB 135-2020; p. NIST HB 135-2020). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.HB.135-2020>
16. Tchobanoglous G, Stensel HD, Tsuchihashi R, Burton FL, Abu-Orf M, Bowden G, Pfrang W, Metcalf and Eddy (eds) (2014) *Wastewater engineering: treatment and resource recovery*, 5th edn. McGraw-Hill Education
17. Vipulanandan C, Pasari G (2012) Life cycle cost model (LCC-CIGMAT) for wastewater systems, pp 740–751. [https://doi.org/10.1061/40800\(180\)59](https://doi.org/10.1061/40800(180)59)
18. Krarti M (2018) Life-cycle cost and energy productivity analyses (chapter 5). In: Krarti M (ed) *Optimal design and retrofit of energy efficient buildings, communities, and urban centers*. Butterworth-Heinemann, pp 247–312. <https://doi.org/10.1016/B978-0-12-849869-9.00005-3>
19. Houston Metro Area Population 1950–2022 (n.d.) Retrieved from <https://www.macrotrends.net/cities/23014/houston/population> Accessed on 14 Apr 2022

20. Houston Public Works—Utility Rates 2021 (n.d.) Retrieved from <https://www.houstonsecured.org/NewRates.html> Accessed on 29 Jan 2022
21. Seidel S (n.d.) How much does an acre of land cost in Texas in 2022? Retrieved from <https://info.southerngreenbuilders.com/blog/acre-of-land-cost-in-texas> Accessed on 6 Feb 2022
22. THE 17 GOALS | Sustainable Development (n.d.) Retrieved from <https://sdgs.un.org/goals> Accessed on 1 Jan 2022
23. The Complete Beginner's Guide to Greywater Systems (n.d.) Retrieved from <https://elemental.green/complete-beginner-guide-to-greywater-systems/> Accessed on 27 Sept 2021

# Economic Analysis of the Utilization of a Greywater System in Residential Dwellings



Bibhas B. Tanmoy and M. Abdel-Raheem

**Abstract** Greywater accounts for most of the wastewater generated in a residence, making it an excellent resource for wastewater reuse, especially in this age of water scarcity. Fortunately, a household greywater system can make the best application of this potential by reusing the generated greywater. Although it may appear to be a convenient solution, it is necessary first to examine the economic feasibility of the system. This is because they demand a significant initial investment and frequent maintenance and replacement schemes. This study examines the economic impact of installing one of the greywater systems currently available in the US market, in a typical Houston residence. The life cycle costs (LCC) comparison shows a reduction in the overall utility bill and economic indicators such as net savings (NS), SIR or AIRR further justifies that it is an economically beneficial solution to be considered for the long term.

**Keywords** Greywater · Life cycle cost analysis · Economic analysis · Financial feasibility · Water savings

## 1 Introduction

Water consumption is increasing in tandem with population growth and urbanization, transforming water into a scarce resource in the future years [1, 2]. Along with this increasing depletion, the United Nations predicts that by 2025, 2.7 billion people, or half to one-third of the global population, would confront water shortages or shortage concerns [3]. However, water waste reduction opportunities may be heavily promoted in the residential sector, mainly through greywater recycling. Greywater recycling can be used as a decentralized primary or auxiliary supply of water on-site and for a

---

B. B. Tanmoy · M. Abdel-Raheem (✉)  
University of Texas Rio Grande Valley, Edinburg, TX, USA  
e-mail: [mohamed.abdelraheem@utrgv.edu](mailto:mohamed.abdelraheem@utrgv.edu)

variety of non-potable uses, resulting in the conservation of both raw and processed water and the reduction of resources used in water treatment and distribution, namely power and chemical products.

The most significant barrier to a household adopting a greywater recycling system is that, while this system will minimize utility bills by using little or no energy, the initial high purchase price and recurring maintenance expenses appear to be a deterrent to the consumer's approach. Therefore, in this framework, an economic assessment of a greywater system is carried out from a residential perspective of Houston, TX. This specific city was chosen to conform to a fixed energy and water price and the per capita consumption amount required for the study.

The objectives of this study are to assess the economic impact of a greywater system in residence and to determine its viability. The term 'residence' refers to a four-person household, with an average US water consumption level [4]. This work will serve as an essential economic guideline on adopting water reusing projects for researchers, water engineers, environmentalists, town planners, and policymakers dealing with water challenges in urban and peri-urban areas.

## **2 Background Study**

### ***2.1 Previous Instances and Studies on Greywater Reuse***

Recently, water quality has been the focus of greywater reuse research. Maintenance systems, breakdowns, smells, flushing procedures, and, most crucially, costs have all been left unexplored. Nowadays, the water used for toilet flushing is of drinking water quality; thus, utilizing greywater for urinal and toilet flushing is a viable alternative. Greywater has been demonstrated to support the amount of water required for toilet flushing and outdoor uses like car-washing and garden watering in a single household [5]. Water savings of 36% in houses and 42% in multi-story residential buildings were found in studies that looked at the potential for potable water savings in homes and multi-story residential buildings using rainwater and greywater [6]. Several countries have actively created greywater technology and regulations and adapted them to agricultural and home applications. Due to water shortages, growing awareness about water conservation, and a lack of centralized treatment facilities, greywater reuse is a common practice in areas with low population densities, such as parts of North America and Australia [7]. Greywater from toilets has been successfully reused in Germany, indicating that it is technically feasible and health-safe [8]. It is also a frequent occurrence for GW to be reused without being treated. Bathwater reuse for garden irrigation has been performed for ages in nations such as Australia, Syria, and South Africa, Israel for land irrigation, and Jordan for fruit tree irrigation [9].



## 2.2 Life Cycle Costing

Life cycle cost analysis (LCCA), sometimes called life cycle costing (LCC), can be considered the ‘economic cousin’ of the life cycle analysis (LCA) [10]. Currently, LCC is used in purchasing for more than only source selection. It has been used for design compromises, optimization, planning, budgeting, repair level analysis, and marketing [11]. The method is commonly employed in the economic examination of numerous aspects of construction projects. In essence, a cost item (i.e., the original cost, energy cost, and maintenance and replacement costs) for the target component during its life term is required, as well as a variety of assumptions, such as the discount rate, analysis period, and maintenance and replacement information [12]. This technique is commonly used in the water and energy industries. Following that, the findings are displayed in several forms. Financial metrics such as net present value (NPV), return on investment (ROI), benefit-cost ratio (BCR), and payback period (PP) were employed by Morales-Pinzón et al. in their model [13]. Similar measures, as well as an internal rate of return (IRR), were used by Matos et al. [14].

## 3 Greywater Systems

Greywater systems, also known as whole-house greywater systems, central water recycling systems, or shortened as GWS, are a modern technical improvement in utilizing greywater generated in typical residences. These systems employ several treatment procedures, resulting in a wide range of system scopes and prices. Models range from greywater diversion systems, which channel generated greywater straight to irrigation sprinklers, to advanced ones with combination of filtration and/or UV disinfection, which produce quality water equal to freshwater, but not necessarily drinking water [15]. Almost all of these systems have two key components: a surge tank that holds the generated wastewater and a filtering chamber that passes the wastewater through and surges out as clean water, which is linked to the inflow pipes for needed purposes like garden sprinklers [16]. Model variants are categorized mainly on surge tank storage, filtration rates, and outflow capacities, along with pumping or no pumping mechanisms. Because these systems have their own tank, the water required for reuse purposes is drawn from the system’s tank rather than the municipal water supply. As a result, the consumption is not credited towards the water metre, resulting in a lower utility bill at the end of the month cutting down up to 50% [17].

## 4 Methodology

The economic assessment in this study is conducted using the life cycle cost analysis (LCCA) approach. The initial stage in this research was to focus on various articles on the subject of this approach. The life cycle costing manual for the Federal Energy Management Program (FEMP), 2020 Edition, from the National Institute of Standards and Technology (NIST) [18] was a beneficial source in the creation of the life cycle costing model utilized in this study. Many of the criteria and procedures used to evaluate the life cycle costs of household appliances were prescribed by this book.

Following the background study, a base case was established. This is to create a benchmark against which the alternative scenario will be assessed. To determine the amount of cost required and energy consumed every month, it is necessary to know the system's relevant expenses and energy consumption and the length of time it will be used every day. The estimated life expectancy will be required to evaluate the cash flow and understand the appliances' consumption habits.

## 5 Model Development

This model estimates a predefined set of financial indicators for the system. Even though a greywater system usually has a usable life of ten years, a minimum LCCA research length of 40 years was determined in compliance with NIST LCCA parameters for the study period [18].

### 5.1 Study Parameters

#### 5.1.1 Product Selection

The US market was investigated to list various greywater systems. Each system was assessed to see if it met the study scope and if it was accessible to residents. Among the shortlisted greywater systems, the Aqua2Use Pro was chosen as it satisfies the scope of the study, as shown in Table 1. This system collects the greywater from different sources via dual plumbing, treats the water through its State of The Art 4 Stage Filtration unit, and finally diverts the treated water to specified units using a 200W auto controlled submersible pump. The system has a treatment capacity of 6.8 GPM and a storage capacity of 50 gallons [19].

**Table 1** Selected greywater system [19]

Name	Variant	Treatment mechanism	Flow mechanism
Aqua2Use Pro	Residential	Filter and divert	Submersible pump

### 5.1.2 Usage Distribution, Sources of Generation, and Scopes of Reuse

The Water Calculator developed by Home Water Works was used to calculate the amount of water utilized in a typical Houston household [20]. Based on location and factors such as residents, fixtures, usage frequency, and efficiency status, this calculator generates an approximate estimate of water consumption in a house (Table 2).

The base case is represented by the total water (before system) quantity in the table. The installation of a greywater system alters the distribution of consumption. In residential dwellings, although toilets are commonly thought to be the major source of blackwater, [6, 9, 21, 22], faucet and dishwasher wastewater are also blackwater due to the presence of food waste [23, 24] in them. According to this hypothesis, all remaining sources may be classified as greywater generators. The quantity of greywater generated and routed to the system is shown in the total GW available for system portion of the table.

According to the Texas Administrative Code [25], Texas Commission on Environmental Quality (TCEQ) [26], as well as product brochure, manufacturer website, and available literature, the system treated water can be used for irrigation, toilet flushing, washing machine, dishwasher, and other miscellaneous uses [27, 19]. The amount of water that would be pulled from the system tank is represented in the total water reused part of the table. Another factor to consider is that, because the total

**Table 2** A Typical efficient water consumption scenario in a Houston residence [20]

Category	gal/Year	gal/day	gal/person/day
Faucet	5096	14	3.50
Toilet	6665	18.30	4.60
Shower	21,556	59.10	14.80
Bathtub	20,630	56.50	14.10
Dishwasher	413	1.10	0.30
Washing machine	709	1.90	0.50
Other	544	1.50	0.40
Irrigation	31,583	86.50	21.60
Total water (before system)	93,544	256.30	64.10
Total water (after system)	51,726	142.78	35.41
Total GW available for system	45,343	124.22	31.09
Total water reused	39,914	109.30	27.4

GW available for the system is more than the total water reused, there should be no shortage of reusable water in the system tank, as it is a simultaneous operation.

### 5.1.3 Energy Rates

Energy Information Administration publications were reviewed to obtain data on the anticipated growth in energy prices over time. The Houston average price per kilowatt-hour was provided by the US Bureau of Labor Statistics.

### 5.1.4 Costs and Economic Indicators

Two parameters were considered for the basic economic estimates for the system. The first is the purchase and installation costs, which are the costs of the product based on the pricing supplied by the manufacturer of this system. The other is the maintenance and replacement cost, which, unlike installation expenses, refers to future expenditures paid throughout the system's operation. The economic indicators chosen to conduct the assessment are life cycle cost (LCC), net savings (NS), savings-to-investment ratio (SIR), adjusted internal rate of return (AIRR), and payback period (PB).

Below are the formulae for these indicators. All cost parameters utilized in this analysis are discounted to present value using FEMP's nominal discount rate. The equations were gathered from the literature and converted to fit the study's scope.

$$LCC = I - \text{Res} + \text{Repl} + W + E + \text{OMR} \quad (1)$$

$$NS = LCC_{\text{BaseCase}} - LCC_{\text{GWS}} \quad (2)$$

$$SIR = \frac{\Delta E + \Delta W + \Delta \text{OMR}}{\Delta I + \Delta \text{Repl} - \Delta \text{Res}} \quad (3)$$

$$\text{AIRR} = (1 + i) * \text{SIR}^{\frac{1}{n}} - 1, \quad (4)$$

where LCC represents life cycle cost,  $I$  stands for initial investment, Res is residual value, Repl equals replacement cost,  $W$  is the utility bill,  $E$  is the energy cost, OMR represents the operation and maintenance cost,  $i$  is the nominal discount rate,  $n$  is the study period, and  $\Delta$  denotes the difference of cost between base and alternative case [9, 18, 23–29].

Typically, the utility bill includes both the water and sewage charges. To calculate this bill, the City of Houston's 2021 residential water and wastewater rates were used [30]. The water section of the utility bill includes a base price of \$6.46 and a constant \$5.50 rate for subsequent kilogallons consumed. The sewage section likewise consists of a base price of \$10.10 but is followed by two blocks. The first block

accounts for the first three kilogallons consumption, minus the irrigation or sprinkler amount, and costs \$4.00 per kilogallons. The second block activates when it exceeds the first three kilogallons, which charges \$10.50 per kilogallons. These charges are then added together to create a comprehensive utility bill. The following equation demonstrates this calculation [30]. All the charges are in USD, rates are in USD/month, and consumption amounts are in gallons/month.

$$w_{\text{flat}} = [\text{Water Base Charge} + (\text{Water Rate} * \text{Consumption})] \\ + [\text{Sewage Base Charge} + ((\text{Sewage Rate} * \text{Consumption})_{\text{First Block}}) \\ + (\text{Sewage Rate} * \text{Consumption})_{\text{Second Block}}]. \quad (5)$$

The energy consumption was determined using the power of the pumping systems and the number of hours per month required to treat the volume of greywater generated during the same period. The energy cost was determined using the equation below, where the rate is \$0.15/kWh:

$$e_{\text{flat}} = \text{Power (kW)} * \frac{\text{Water Amount (gal)}}{\text{Capacity} \left( \frac{\text{gal}}{h} \right)} * \text{energy rate} \left( \frac{\$}{\text{kWh}} \right). \quad (6)$$

The subscript ‘flat’ implies that these costs have not been discounted to their current value. The geometric gradient formula was used to discount both the monthly utility bill and energy expenses to present value [18, 31].

$$E = \frac{e_{\text{flat}} \left[ 1 - \left( 1 + g_e / (1 + i) \right)^n \right]}{i - g_e} \quad (7)$$

$$W = \frac{w_{\text{flat}} \left[ 1 - \left( 1 + g_w / (1 + i) \right)^n \right]}{i - g_w} + \frac{s_{\text{flat}} \left[ 1 - \left( 1 + g_s / (1 + i) \right)^n \right]}{i - g_s}, \quad (8)$$

where  $W$  represents the total utility bill and on the other side  $w$  is the water portion,  $s$  is the sewage portion, and  $g_e$ ,  $g_w$ ,  $g_s$  are the expected increase of energy, water, and sewage rate every year, respectively.

All the other single payments, such as maintenance costs, replacement costs, residual value were discounted to present value using the following formula [18]:

$$PV = F(1 + i)^{-n}, \quad (9)$$

where  $F$  is the payment that will occur at a point in the future.

Annuities, excluding water and energy bills, such as the annual maintenance costs were discounted to present value using the following formula [18]:

$$P = A \left[ \frac{(1+i)^n - 1}{i(1+i)^n} \right], \quad (10)$$

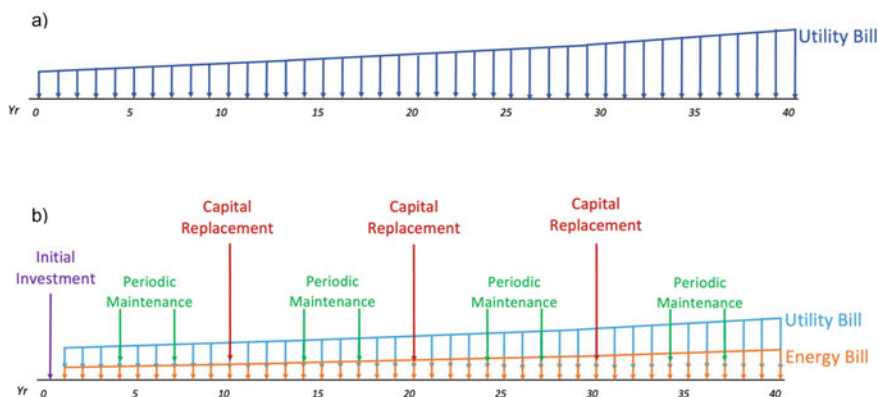
where  $A$  is the payment occurring annually, and  $P$  is the discounted value of the amount.

## 5.2 Assumptions

This study considers all the payments occurring at the end of the year, including capital replacement. It is also assumed that when a greywater system reaches the end of its functional life, in this case after ten years [32], it is discarded since it has no value. This signifies that the appliance depreciates ‘straight-line’. The residual values are calculated using this straight-line depreciation assumption; the residual value is \$0 if the device reaches the end of its useful life at the end of the study period, or it is some value greater than \$0 if the appliance’s useful life beyond the end of the study period.

## 5.3 Cash Flow Diagrams

Before initiating the research, cash flow diagrams for both cases were created. The cash flow diagram illustrates the occurrence time, kind, and amount of the associated payments over the study period. The diagrams illustrate that in the base case, where the only cost is the utility bill or  $W$  in the equations, and in the alternative case, various types of payments occur at their respective times, such as periodic maintenance and energy costs.



**Fig. 1** a Cash flow of base case, b cash flow of Aqua2Use Pro

## 6 Economic Assessment

### 6.1 Basic Data

The previously discussed equations were applied to the data collected. The following table displays the present values of all the cost parameters used in the economic assessment.

**Table 3** Basic cost data

System Name	Initial Investment, $I$	Replacement, Repl	Residual Value, Res	Energy Cost, $E$	Water Cost, $W$	OMR
Aqua2Use Pro	\$1199.00	\$2997.00	\$0	\$127.07	\$ 72,748.57	\$1672.00

### 6.2 Life Cycle Cost (LCC)

The system's life cycle cost (LCC) included the initial investment, operation and maintenance, sewage and electricity costs, replacement costs, and residual values. It is important to note that the base case's life cycle cost is just the utility bill paid over 40 years. The system's life cycle cost of 40 years was lower than the base case. This substantiates the claim that installing this system in a home will result in cost savings. Appendix A provides a detailed explanation of the calculation.

**Table 4** Life cycle costs

Settings	Life cycle costs
Conventional	\$106,784
A2U GWDD Pro	\$78,744

### 6.3 Net Savings (NS)

When deciding between alternatives, such as whether to install a system or not, the alternatives' net savings (NS) can be easily observed. If the NS of an alternative is positive, it indicates that it will result in savings; if it is negative, it suggests that the choice will result in a loss project. The option will be regarded neither advantageous nor detrimental if the NS is zero [29, 33]. The LCC value from the previous section shows a clear decrease in cost, whereas the NS value numerically expands the argument. The system saves a total of \$24,040 during the whole study period.

6.4 Savings-to-Investment Ratio (SIR)

The savings-to-investment ratio (SIR) provides a quick ‘grading’ of a project. A ratio larger than one shows that overall savings exceed total investments [12, 28], hence a beneficial option. This analysis refreshes the previous decision and reveals an anomaly that the previous LCC and NS assessment could not identify. Analysis shows that the system yields a SIR value of 7.68, which indicates that the overall savings are sufficient to compensate for the investment required for this option.

6.5 Adjusted Internal Rate of Return (AIRR)

The adjusted internal rate of return (AIRR) is a project performance metric. The AIRR is a metric that measures an investment’s annual rate of return. The yields outweigh the losses if the AIRR is higher than the discount rate [28]. In line with the results of SIR, the system resulted in an AIRR of 7.02%, which is substantially higher than the discount rate of 1.7% used in the study.

6.6 Payback Period

The payback period is the time it takes for a complete economic return to equal the cost of an alternative’s initial investment, and it is essential in terms of practicality. A beneficial project should ideally have a payback time in between the research period. The faster an option recoups its investments and recurring expenditures, the better it is in the long term [28, 34]. Analysing the system’s payback period, it was seen that this system takes 13 years before the savings start exceeding the investment amount.

6.7 Summary of Results

The following table summarizes all the parameters calculated in the assessment:

Table 5 Summary of results

System	LCC	NS	SIR	AIRR	PB
A2U GWDD Pro	\$78,744	\$28,040	7.68	7.02%	13



## 7 Discussion and Limitations

The model developed in this study was built to perform a comprehensive life cycle cost analysis for a greywater system installed in a residential dwelling. The output data shows that, although it yields a net benefit in a mathematical calculation, the inherent costs, such as maintenance and replacement costs, often hinder this financial benefit in the long run. Therefore, it is necessary to do deep research before investing significant money into one of these systems. This is important because the cost calculations, such as life cycle cost or net savings, show positive results, indicating benefits, and the SIR and AIRR values justifies that the benefit remains steady in the long run.

However, this study only assesses the impact of just one system, it cannot determine what would be the perfect choice for a consumer among the many variations. For instance, in homes where residents are usually out of the house most of the day, choosing a high storage system would be unnecessary. On the other hand, a system with higher pump speed would be more beneficial in areas that face water scarcity frequently. All these considerations will also play a vital role in the economic factors. On the assessment side, adding more parameters such as return on investment (ROI), along with the increased sample size, can further expand the results. A sensitivity analysis to the computed results can also demonstrate the fluctuation factors.

## 8 Conclusion

With public awareness of sustainability growing, reducing water use and reusing is one strategy to save and preserve natural resources while lowering monthly utility expenses for homes. The LCCA approach provides a strong value in this regard since it provides a thorough analysis of the economic impact before making a decision, highlighting the overall cost associated with a given project and its potential alternatives by using cost data essential to investors and decision-makers. Furthermore, reusing wastewater reduces the demand for natural resources, which is beneficial to the environment as well, and the reduction in pollution benefits the society. Economic savings, environmental safety, and social benefit, altogether create the pathway for a sustainable tomorrow.

## References

1. Oron G, Adel M, Agmon V, Friedler E, Halperin R, Leshem E, Weinberg D (2014) Greywater use in Israel and worldwide: standards and prospects. *Water Res* 58:92–101. <https://doi.org/10.1016/j.watres.2014.03.032>
2. Rodríguez C, Sánchez R, Rebolledo N, Schneider N, Serrano J, Leiva E (2021) Life cycle assessment of greywater treatment systems for water-reuse management in rural areas. *Sci Total Environ* 795:148687. <https://doi.org/10.1016/j.scitotenv.2021.148687>

3. Juan Y-K, Chen Y, Lin J-M (2016) Greywater reuse system design and economic analysis for residential buildings in Taiwan. *Water* 8:546. <https://doi.org/10.3390/w8110546>
4. Tchobanoglous G, Stensel HD, Tsuchihashi R, Burton FL, Abu-Orf M, Bowden G, Pfrang W, Metcalf and Eddy (eds) (2014) *Wastewater engineering: treatment and resource recovery*, 5th edn. McGraw-Hill Education, New York
5. Fountoulakis MS, Markakis N, Petousi I, Manios T (2016) Single house on-site grey water treatment using a submerged membrane bioreactor for toilet flushing. *Sci Total Environ* 551–552:706–711. <https://doi.org/10.1016/j.scitotenv.2016.02.057>
6. López Zavala MA, Castillo Vega R, López Miranda RA (2016) Potential of rainwater harvesting and greywater reuse for water consumption reduction and wastewater minimization. *Water* 8:264. <https://doi.org/10.3390/w8060264>
7. Mandal D, Labhasetwar P, Dhone S, Dubey AS, Shinde G, Wate S (2011) Water conservation due to greywater treatment and reuse in urban setting with specific context to developing countries. *Resour Conserv Recycl* 55:356–361. <https://doi.org/10.1016/j.resconrec.2010.11.001>
8. Eriksson E, Auffarth K, Henze M, Ledin A (2002) Characteristics of grey wastewater. *Urban Water* 4:85–104. [https://doi.org/10.1016/S1462-0758\(01\)00064-4](https://doi.org/10.1016/S1462-0758(01)00064-4)
9. Boyjoo Y, Pareek VK, Ang M (2013) A review of greywater characteristics and treatment processes. *Water Sci Technol* 67:1403–1424. <https://doi.org/10.2166/wst.2013.675>
10. Rebitzer G, Hunkeler D, Joliet O (2003) LCC—the economic pillar of sustainability: methodology and application to wastewater treatment. *Environ Prog* 22:241–249. <https://doi.org/10.1002/ep.670220412>
11. Ilyas M, Kassa FM, Darun MR (2021) Life cycle cost analysis of wastewater treatment: a systematic review of literature. *J Clean Prod* 310:127549. <https://doi.org/10.1016/j.jclepro.2021.127549>
12. Cho K, Chang H, Jung Y, Yoon Y (2017) Economic analysis of data center cooling strategies. *Sustain Cities Soc* 31:234–243. <https://doi.org/10.1016/j.scs.2017.03.008>
13. Morales-Pinzón T, Lurueña R, Gabarrell X, Gasol CM, Rieradevall J (2014) Financial and environmental modelling of water hardness—Implications for utilising harvested rainwater in washing machines. *Sci Total Environ* 470–471:1257–1271. <https://doi.org/10.1016/j.scitotenv.2013.10.101>
14. Matos C, Bentes I, Santos C, Imteaz M, Pereira S (2015) Economic analysis of a rainwater harvesting system in a commercial building. *Water Resour Manag Int J* 29:3971–3986. Published for the European Water Resources Association (EWRA)
15. *Graywater Systems* [WWW Document] (2016) Beachapedia. [https://beachapedia.org/Graywater\\_Systems](https://beachapedia.org/Graywater_Systems). Accessed 21 Jan 2022
16. *Simple Greywater Systems For Your Home* [WWW Document] (2019) Tiny life. <https://thetinylife.com/greywater-systems/>. Accessed 27 Nov 21
17. Ferguson D (2014) Greywater systems: can they really reduce your bills? *The Guardian*
18. Kneifel J, Webb D (2020) Life cycle cost manual for the federal energy management program (No. NIST HB 135-2020). National Institute of Standards and Technology, Gaithersburg, MD. <https://doi.org/10.6028/NIST.HB.135-2020>
19. *Greywater System Aqua2use Pro* (n.d.) Water wise group. <https://waterwisegroup.com/greywater-systems-sale/aqua2use-pro/>. Accessed 30 Nov 2021
20. *Water Calculator | Home Water Works* [WWW Document] (n.d.) <https://home-water-works.org/calculator>. Accessed 5 Nov 21
21. Christova-Boal D, Eden RE, McFarlane S (1996) An investigation into greywater reuse for urban residential properties. *Desalination* 106:391–397. [https://doi.org/10.1016/S0011-9164\(96\)00134-8](https://doi.org/10.1016/S0011-9164(96)00134-8)
22. Ren X, Zhang Y, Chen H (2020) Graywater treatment technologies and reuse of reclaimed water for toilet flushing. *Environ Sci Pollut Res* 27:34653–34663. <https://doi.org/10.1007/s11356-019-05154-6>
23. Gao M, Guo B, Zhang L, Zhang Y, Yu N, Liu Y (2020) Biomethane recovery from source-diverted household blackwater: Impacts from feed sulfate. *Process Saf Environ Prot* 136:28–38. <https://doi.org/10.1016/j.psep.2020.01.010>

24. Hernández Leal L, Temmink H, Zeeman G, Buisman CJN (2010) Comparison of three systems for biological greywater treatment. *Water* 2:155–169. <https://doi.org/10.3390/w2020155>
25. Texas Administrative Code [WWW Document] (n.d.) [https://texreg.sos.state.tx.us/public/readtac\\$ext.TacPage?sl=T&app=9&p\\_dir=F&p\\_rloc=174662&p\\_tloc=14842&p\\_ploc=1&pg=2&p\\_tac=&ti=30&pt=1&ch=317&rl=2](https://texreg.sos.state.tx.us/public/readtac$ext.TacPage?sl=T&app=9&p_dir=F&p_rloc=174662&p_tloc=14842&p_ploc=1&pg=2&p_tac=&ti=30&pt=1&ch=317&rl=2). Accessed 25 Apr 2021
26. Beneficial re-use of graywater and alternative onsite water—Texas Commission on Environmental Quality—[www.tceq.texas.gov](http://www.tceq.texas.gov) [WWW Document] (n.d.) <https://www.tceq.texas.gov/permitting/wastewater/graywater>. Accessed 29 Jan 2022
27. Ecoviv (n.d.) Aqualoop | The only NSF350 greywater recycling solution. Ecoviv Water Management. <https://www.ecovivwater.com/products/aqualoop/>. Accessed 29 Nov 2021
28. Amini Toosi H, Lavagna M, Leonforte F, Del Pero C, Aste N (2020) Life cycle sustainability assessment in building energy retrofitting: a review. *Sustain Cities Soc* 60:102248. <https://doi.org/10.1016/j.scs.2020.102248>
29. Eltamaly AM, Mohamed MA (2018) Optimal sizing and designing of hybrid renewable energy systems in smart grid applications (chapter 8). Elsevier Enhanced Reader [WWW Document]. <https://doi.org/10.1016/B978-0-12-813185-5.00011-5>
30. Houston Public Works—Utility Rates 2021 [WWW Document] (n.d.) <https://www.houstonsewage.org/NewRates.html>. Accessed 29 Jan 2022
31. Ruegg R, Marshall H (2013) Building economics: theory and practice. Softcover reprint of the original 1st edn. (1990 edn). Springer
32. The Complete Beginner's Guide to Greywater Systems [WWW Document] (n.d.) <https://elemental.green/complete-beginner-guide-to-greywater-systems/>. Accessed 27 Sep 2021
33. Project Management for Construction: Economic Evaluation of Facility Investments [WWW Document] (n.d.) [https://www.cmu.edu/cee/projects/PMbook/06\\_Economic\\_Evaluation\\_of\\_Facility\\_Investments.html](https://www.cmu.edu/cee/projects/PMbook/06_Economic_Evaluation_of_Facility_Investments.html). Accessed 17 Dec 2021
34. Leong JYC, Balan P, Chong MN, Poh PE (2019) Life-cycle assessment and life-cycle cost analysis of decentralised rainwater harvesting, greywater recycling and hybrid rainwater-greywater systems. *J Clean Prod* 229:1211–1224. <https://doi.org/10.1016/j.jclepro.2019.05.046>

# Adsorption of Sulfamethoxazole by Dried Biomass of Activated Sludge Collected from Biological Nutrient Removal (BNR) Systems



S. Minaei, K. N. McPhedran, and J. Soltan

**Abstract** Widespread use of human and veterinary antibiotics has drawn attention to the occurrence and fate of antibiotics in the environment. Despite very low concentrations of these chemicals ranging from  $< 1$  ng/L to hundreds  $\mu\text{g/L}$ , they are bioaccumulating in the environment and potentially spreading antibiotic resistance genes into ecosystems. The focus of current research interest is on municipal wastewater treatment plant (MWTP) effluents which are the main point sources for introducing these chemicals into surface waters. This study investigated the feasibility of using MWTP activated sludge as a cost-effective and sustainable adsorbent for sulfamethoxazole (SMX) removal. The activated sludge was collected from the aerobic basin of a biological nutrient removal (BNR) MWTP system. The sludge was oven-dried and used as an adsorption material without further modification. Langmuir and Freundlich isotherms were used to clarify the SMX adsorption parameters on dried activated sludge (DAS). The specific surface area of DAS correlates with the adsorption capacity. The specific surface area is reported as  $10.5 \text{ m}^2/\text{g}$  for DAS. The data reported a rise in SMX adsorbed, from 0.02 to 11.07 mg/L, with an increase in the SMX initial concentration, increasing the interaction with adsorbent. SMX adsorption on DAS satisfactorily fits with the Freundlich isotherm, with  $n = 1.87$ , which implies a high adsorption affinity of SMX on DAS. Further modification could improve the adsorption capacity by adding functional groups with higher adsorption capacity of DAS toward pharmaceutical removal from water and wastewater matrices.

---

S. Minaei · J. Soltan

Department of Chemical and Biological Engineering, University of Saskatchewan, Saskatoon, SK, Canada

K. N. McPhedran (✉)

Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, SK, Canada

e-mail: [kerry.mcphedran@usask.ca](mailto:kerry.mcphedran@usask.ca)

K. N. McPhedran · J. Soltan

Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_64](https://doi.org/10.1007/978-3-031-34593-7_64)

1007

**Keywords** Wastewater treatment · Biological nutrient removal (BNR) · Dried activated sludge (DAS) · Sulfamethoxazole · Adsorption isotherm

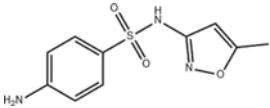
## 1 Introduction

Recently, it has been increasingly crucial to meet the need for safe drinking water in communities worldwide to help to safeguard public health. The waste and effluent released from manufacturing industries include a large and diverse group of chemicals such as antibiotics, synthetic musks, pharmaceuticals, and personal care products (PPCPs) [16, 18]. These chemicals are used to improve human health while also enhancing quality of life. Overall, a significant amount of PPCPs is being released into the sewage network in metabolized or unmetabolized forms through urine, fecal matter, and medical wastewater discharge. Given that pharmaceuticals generally have both high polarity and water solubility, it is challenging to treat these chemicals through conventional wastewater treatment processes [9, 10, 36]. Eventually, all these chemicals make their way into receiving waterbodies and connected ecosystems, either indirectly through municipal wastewater treatment plant (WWTP) effluent or directly through raw sewage release into receiving environments [2, 9].

The presence of antibiotics, an important group of PPCPs, is a growing global environmental concern in water resources. Antibiotics are persistent and easily accumulate in aquatic systems, causing toxicity and ecotoxicological risks for many species throughout the water environment [2]. Currently, the continuous release of these pollutants persists while environmental hazards of these pollutants has not been thoroughly understood and established. For example, sulfamethoxazole (SMX) (Table 1) is used to treat various bacterial infections but has been shown to typically not be completely removed/treated during typical WWTP processes. For SMX, the unmetabolized fraction entering into municipal wastewater after being subject to human metabolism is reported to be 15–30% [2, 5, 11, 20]. Exposure to this chemical can cause hepatic cancer and altered genetic traits, while its environmental behavior is still unclear [3, 13, 30, 37]. Given that WWTP effluents have been identified as the primary point sources for the entry of these chemicals into surface waters, the existing treatment system processes need to be modified or upgraded in order to adequately treat antibiotics to prevent further releases into the environment.

Different approaches have been investigated previously for the removal of PPCPs from wastewaters including coagulation, membrane filtration, bioremediation, ozonation, flocculation, advanced oxidation photocatalysis, sedimentation, microbial degradation, electrochemical processes, and adsorption [8, 17, 25, 26, 31, 33, 37]. Despite the various benefits and drawbacks of each of these wastewater treatment processes, adsorption, especially using biosorbents, has recently become recognized as one of the most promising treatment processes. This is due to the benefits of adsorption that include cost-effectiveness, energy efficiency, and simple operation [6, 24]. In addition, biosorbents developed from a variety of different

**Table 1** Molecular structure and chemical properties of sulfamethoxazole (SMX)

Molecular structure	Molecular weight (g/mol)	Water solubility (mg/L)	pK <sub>a1</sub>	pK <sub>a2</sub>	References
Sulfamethoxazole $C_{10}H_{11}N_3O_3S$ 	253.28	0.5–0.6	1.39–1.97	5.81–6.16	[4, 19]

source biomasses (e.g., agricultural wastes) are currently being developed world-wide to replace the need for expensive, commercially prepared activated carbons that are currently the dominant adsorbents used worldwide.

Agricultural residues, various species of algae, and bacterial and industrial wastes (e.g., wastewater activated sludges), are sustainable and inexpensive primary sources for the synthesis of biosorbents for treatment of PPCPs, among other pollutants [15, 22]. Interestingly, the biological nature of activated sludge leads to the development of different surface functional groups on the sludges, which can help to facilitate hydrophobic surface interactions with the target pollutant(s) such as the currently considered SMX when used as biosorbents. Both live and dead microbial cells can adsorb chemicals such as SMX; however, living cells require nutritional resources, which could cause an increase in the biological oxygen demand (BOD) and chemical oxygen demand (COD) of the wastewater if used directly as biosorbents. On the other hand, dead microbial cells require no additional nutritional resources, less maintenance to keep viable, and are readily available for use as biosorbents given vast and continuous volumes of sludges continually created within MWTPs. Previously activated sludge has been reported as an effective biomaterial for creation of biosorbents used for biosorption or bioleaching of metal ions from wastewater [1, 23, 32, 35]. Adsorbents could be used in the filtration beds prior to discharge to surface waters. Due to challenges with reintroducing the waste activated sludge to the treated wastewater, synthesized biosorbents require to be modified in terms of mechanical stability. In this regard, pelletization is suggested as a convenient immobilization technique which facilitates the separation and multicycle application of biosorbents. Pelletized biosorbent also reduces clogging of the column used for treatment purposes compared to the powdered biosorbent. However, further research is needed to optimize and develop sludge-based biosorbents for use in the adsorption of current and newly created chemicals of environmental concern.

In this study, the feasibility of employing WWTP activated sludge waste as a biosorbent for the removal of SMX was investigated. In this regard, aerobic sludge was collected from the Saskatoon Wastewater Treatment Plant (SWTP) biological nutrient removal (BNR) system and processed in the University of Saskatchewan Environmental Engineering laboratories prior to assessment of their adsorption potential for SMX. The adsorption isotherm of SMX adsorption on dried activated

sludge (DAS) was analyzed using Langmuir and Freundlich models to calculate thermodynamic parameters used to clarify the feasibility and characteristics of the adsorption process.

## 2 Materials and Methods

### 2.1 Sampling and Sludge Processing

Secondary activated sludge biomass was collected using sampling buckets directly from the aerobic tank of the biological nutrient removal (BNR) system of a full-scale MWTP in Saskatoon, SK, Canada (Fig. 1a). The SWTP is a Class 4 treatment facility, which is the highest degree of accreditation available in Canada (City of Saskatoon). The sampling location used for each sampling event is indicated by the star shown in Fig. 1b. During the steady-state operation, the mixed liquor suspended solids (MLSS), total protein, total carbohydrate, and pH value were 6.9 g/L, 4.98 mg COD/L, 4.3 mg COD/L, and  $7.0 \pm 0.5$ , respectively. The samples from the aerobic basin were transferred in 20 L containers, transported to the lab, and processed immediately. The sludge was filled into 50 mL centrifuge tubes (Millipore Sigma, ON, Canada) and centrifuged at 5000 rpm for five minutes. The supernatant was removed and replaced with distilled water to 50 mL volume two times to thoroughly wash the sludge from wastewater. The precipitate was collected and dried in an oven at 60 °C to a constant weight. The obtained DAS was directly used in biosorption experiments.

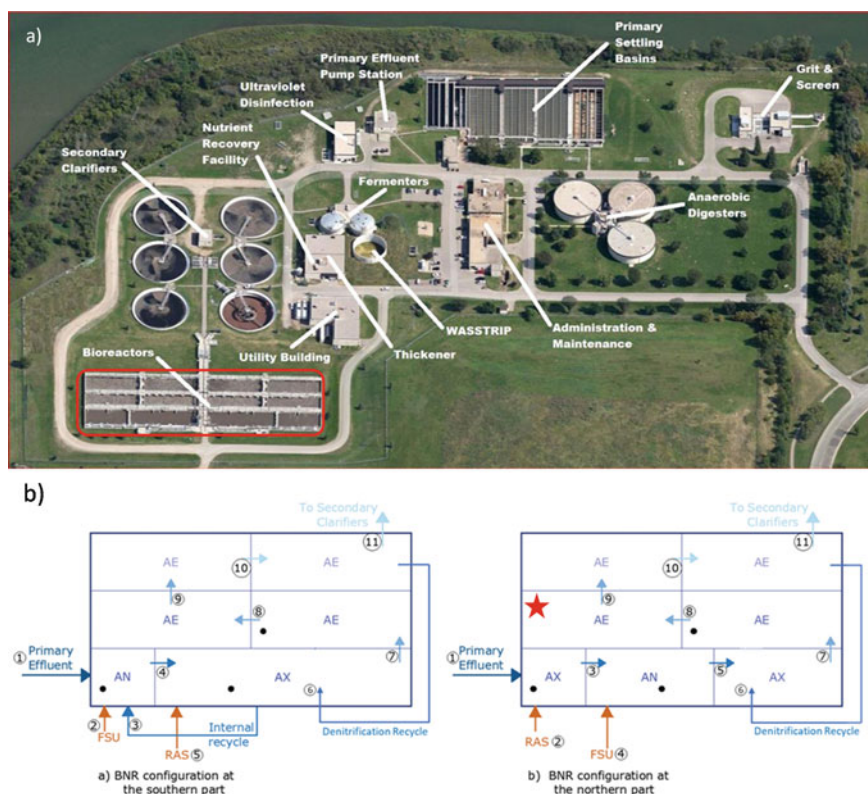
### 2.2 Materials

Sulfamethoxazole (SMX,  $C_{10}H_{11}N_3O_3S$ ) and methanol (HPLC grade,  $\geq 99.9\%$ ) were purchased from Thermo Fisher Scientific (Thermo Fisher Scientific Inc., CA, USA). The physicochemical properties and molecular structure of SMX are summarized in Table 1. A standard 500 ppm SMX stock solution was prepared by weighing of SMX and dissolving it in 10 mL of methanol and then diluting to a final volume of 1000 mL using distilled water.

### 2.3 Adsorption Experiments

The adsorption experiments were carried out using the batch method. For the equilibrium adsorption experiments, DAS and SMX were added to 50 mL glass vials. The concentration of DAS was 2 g/L which has been reported as an optimum dosage





**Fig. 1** a) Saskatoon wastewater treatment plant (SWTP), b) biological and nutrient removal configuration of the SWTP. AN, AX, AE, FSU, and RAS represent anaerobic, anoxic, and aerobic basins, fermenter supernatant, and return activated sludge

in the literature [2]. The adsorption isotherm tests were performed at a varying initial SMX concentration in the range of 0–200 mg/L. The pH value of the mixture remained almost constant at pH 6.9 during the experiment as measured using a Thermo Scientific pH probe (Cole-Parmer Canada, QC, Canada). The vials were placed on a rotary shaker (Orbi-shaker CO2, Benchmark, USA) operated at 250 rpm and 25 °C for 24 h in order to achieve equilibrium adsorption. The vial solutions were then filtered in order to remove any adsorbent particles. The freely dissolved concentrations of SMX before and after adsorption were characterized using ultra-high performance liquid chromatography (UltiMate 3000 UHPLC, ThermoFisher, MA, USA) with a UV/Vis detector at 270 nm, and the flow rate and injection volume were set at 1 mL.min<sup>-1</sup> (0.1% Formic acid/Acetonitrile (ACN) 50:50 v/v) and 10 µL, respectively. The adsorption of SMX on the DAS was calculated according to Eq. (1):



$$q_e = \frac{(C_0 - C_e)V}{M}, \quad (1)$$

where  $q_e$  (mg/g) is the amount of SMX adsorbed at equilibrium;  $C_0$  and  $C_e$  (mg/L) are the initial and equilibrium SMX concentrations, respectively;  $V$  (L) is the mixture volume; and  $M$  (g) is the mass of activated sludge adsorbent. The BET surface area ( $S_{\text{BET}}$ ) and pore size distribution of the activated sludge were determined by  $N_2$  adsorption and desorption isotherms at 77 K, using a pore size analyzer (Quantachrome, NOVA 2200e, USA). The  $S_{\text{BET}}$  value was determined by the Brunauer–Emmett–Teller (BET) multipoint technique in a relative pressure ( $p/p_0$ ) range from 0.048 to 0.30.

## 2.4 Adsorption Isotherms

The Langmuir isotherm model describes the monolayer sorption of an adsorbate (SMX currently) on an adsorbent (DAS currently) surface having a finite number of identical sites without surface diffusion as:

$$q_e = \frac{QbC_e}{1 + bC_e}, \quad (2)$$

where  $Q$  (mg/g) indicates the binding strength,  $b$  (L/mg) the maximum adsorption capacity,  $C_e$  is the equilibrium concentration of the adsorbate (mg/L), and  $q_e$  (mg/g) is the amount of adsorbed solute per unit adsorbent mass. Equation (2) could be rewritten in a linear format and plotted to calculate the various parameters using experimental data.

The Freundlich isotherm model correlates the sorption density of adsorbate on the adsorbent surface and the concentration in the liquid phase empirically:

$$q_e = K_f C_e^{\frac{1}{n}}, \quad (3)$$

where  $K_f$  ( $\text{mg}^{1-n} \text{mL}^n/\text{gr}$ ) is the Freundlich adsorption coefficient and  $1/n$  indicates the isotherm nonlinearity. The isotherm fits linearly when explaining the adsorption of a chemical at low mass loading or missing specific bonding between the target compound and the adsorbent.

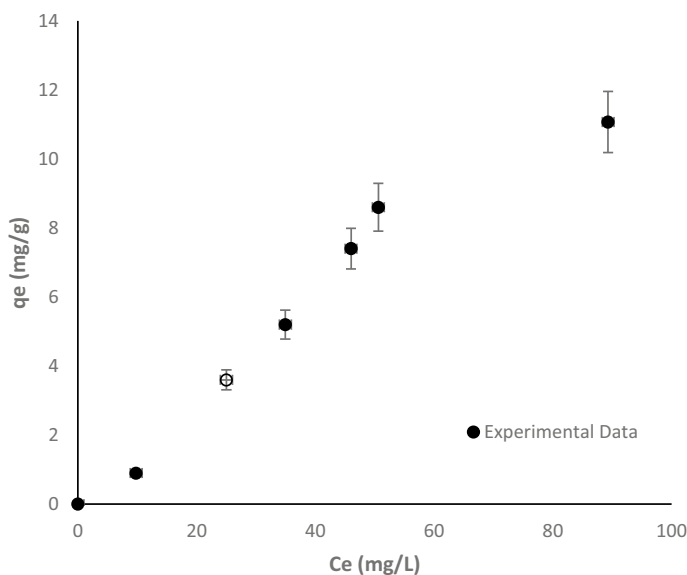
In general, the Langmuir model is based on the monolayer adsorption mechanism, and better describes a homogenous sorption system. In contrast, Freundlich model fits better with multilayer heterogeneous adsorption [34].

### 3 Results and Discussion

#### 3.1 SMX Adsorption

Adsorption isotherm developed from data obtained at equilibrium state can be used to describe the correlation between the amount of target pollutant adsorbed on the adsorbent and the remaining amount of the pollutant in the solution [34]. Equilibrium data from the batch adsorption experiment shows SMX sorption increased from 0.02 to 11.07 mg/g with increasing initial SMX concentrations up to 100 mg/L (Fig. 2). The obtained adsorption capacity for DAS (11.07 mg/g) is higher than some raw biosorbent materials such as giant reed, rice, and wheat straw (Table 2), but it is still lower than chemically treated biomasses and biochars [29, 38]. Therefore, treatment of these agricultural biomasses using different chemicals may be useful to increase their porosities and surface areas by extracting soluble organic compounds, changing or introducing new functional groups [21].

The specific surface area analysis of the thermal treated activated sludge (DAS) exhibited an  $S_{\text{BET}}$  of 10.5 m<sup>2</sup>/g (Table 2). This area is higher than other raw biomasses such as raw alfalfa (*Medicago sativa* L.) crops ( $S_{\text{BET}}$  = 0.6 m<sup>2</sup>/g) [7]. However, it has been reported that high-temperature pyrolysis drastically increases the surface area, leading to much higher adsorption of SMX [7]. However, the processed DAS presented higher adsorption capacity for SMX than some biochars made from giant reed, rice, and wheat straw [29, 38] (Table 2).



**Fig. 2** SMX adsorption data on dried activated sludge (DAS), sorption isotherms including experimental data points,  $q_e$  mg/g versus  $C_e$  mg/L (sorbent dosage 2 g/L; temperature 25 °C)

**Table 2** Specific surface area and SMX adsorption capacity using different biochars

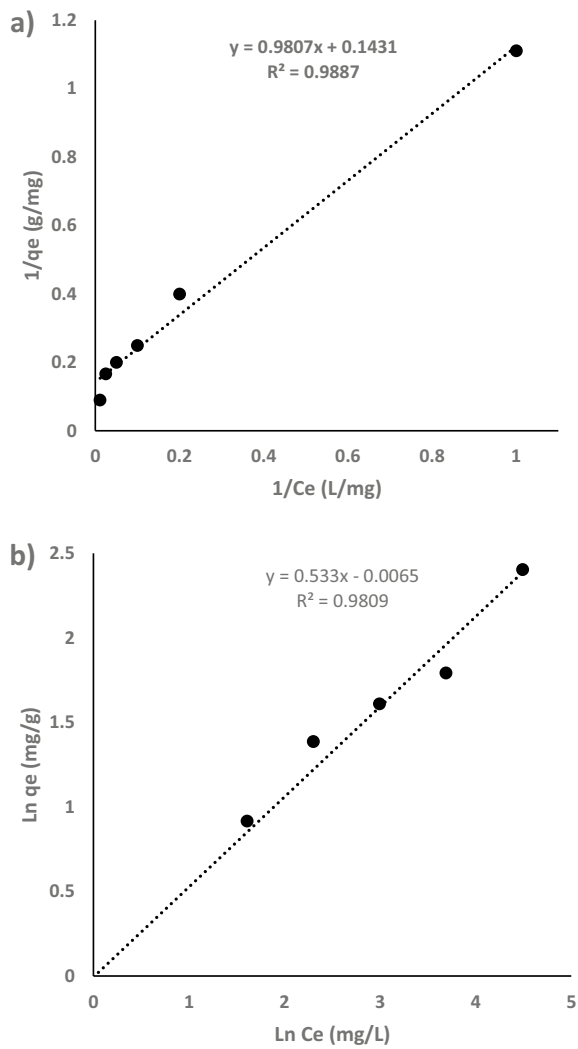
Adsorbent	$S_{\text{BET}}$ ( $\text{m}^2/\text{g}$ )	$q_{\text{max}}$ ( $\text{mg}/\text{g}$ )	References
Giant reed (untreated)	2	5	[38]
Biochar (giant reed-300 °C)	59	1.9	
Rice Straw (untreated)	6	4.2	[29]
Biochar (rice straw-300 °C)	27	7.4	
Wheat Straw (untreated)	8	6.8 38	[29]
Biochar (wheat straw-300 °C)	38	38	
Dried activated sludge	10.5	6.98	This Study

### 3.2 Sorption Isotherms

The Langmuir and Freundlich isotherm models are depicted in Fig. 3, and the results of the model parameters are presented in Table 3 with comparison of Langmuir  $q_{\text{max}}$  data shown in Table 2. Comparing the  $R^2$  values for the investigated isotherms shows that these values are close to unity for both the Langmuir ( $R^2 = 0.9887$ ) and Freundlich isotherms ( $R^2 = 0.9809$ ). Given these values, it appears that both Langmuir and Freundlich isotherm models may be suitable for modeling the sorption of SMX on DAS. The current Langmuir  $q_{\text{max}}$  result of 6.98 mg/g is similar in comparison to biochars presented in Table 2.

Alternatively, the experimental data is well fitted with the Freundlich model which assumes multilayer heterogeneous sorption [12, 27, 28]. Additionally, the Freundlich model constant,  $n$ , is associated with the linearity of SMX adsorption data. The  $n$  value obtained from the model fitting is not much higher than 1 ( $n = 1.87$ ), implying linear sorption and a high affinity of DAS for SMX adsorption. Linear isotherms occur when the adsorbate concentration is relatively low compared to the adsorption capacity of the solid, which implies the adsorption condition is far below saturation. Based on the fitting data, comparing the  $R^2$  values, in these cases, hybrid adsorption mechanism is assumed. At lower concentrations, the ideal homogenous monolayer adsorption is occurring, while at higher concentrations, multilayer heterogeneous adsorption is suggested [14]. Given small surface area of DAS, there is not high enough active sites available for SMX adsorption; hence, the multilayer adsorption has not started at lower concentrations. The resulting data indicates that the DAS can be an appropriate sorbent and beneficial for the SMX removal from the wastewater and further investigation is required to optimize the DAS surface to achieve higher SMX removal rates.

**Fig. 3** **a** Langmuir sorption isotherm, **b** Freundlich sorption isotherm, plots for SMX removal using DAS (sorbent dosage: 2 g/L; temperature: 25 °C)



**Table 3** Isotherm models, equations, and estimated parameters for SMX adsorption using DAS: sorbent dosage 2 g/L; Temperature: 25 °C

Model	Parameter	Value
Langmuir	$q_{\max}$ (mg/g)	6.9881
	$K_L$ (L/mg)	0.145
	$R^2$	0.9887
Freundlich	$K_f$ (mgr/gr)	0.9935
	$n$	1.8761
	$R^2$	0.9809

## 4 Conclusion

The processed aerobic activated sludge collected from the biological nutrient removal treatment system is a potential biosorbent for removing SMX from the wastewater. The samples were thermally modified to achieve a dead microbial population as the biosorbent. Given the resulting data, chemical treatment and pyrolysis could be additional modifications to improve adsorption capacity. In this study, the possibility of using activated sludge as a source of biosorbents to remove pharmaceuticals was investigated. Overall, the study suggests activated sludge as a sustainable, easy-to-apply source for pharmaceutical removal from wastewater. The sorption capacity of DAS indicates the need for further research into these biomasses to further improve the efficiency to achieve higher removal rates. Chemical treatment could improve adsorption capacity by providing appropriate surface functional groups for DAS toward pharmaceutical removal from water and wastewater matrices.

**Acknowledgements** The authors would like to acknowledge the University of Saskatchewan for awarding Devolved and Dean's Scholarships to Mr. Shahab Minaei and Natural Sciences and Engineering Research Council (NSERC) Discovery Grants (Profs. Jafar Soltan and Kerry McPhedran) for supporting the funding for this research.

## References

1. Ahmad A, Ghufuran R, Faizal WM (2010) Cd(II), Pb(II) and Zn(II) removal from contaminated water by biosorption using activated sludge biomass. *CLEAN Soil Air Water* 38(2):153–158. <https://doi.org/10.1002/clen.200900202>
2. Akhtar J, Amin NS, Aris A (2011) Combined adsorption and catalytic ozonation for removal of sulfamethoxazole using Fe<sub>2</sub>O<sub>3</sub>/CeO<sub>2</sub> loaded activated carbon. *Chem Eng J* 170(1):136–144. <https://doi.org/10.1016/j.cej.2011.03.043>
3. Ania CO, Parra JB, Menéndez JA, Pis JJ (2007) Microwave-assisted regeneration of activated carbons loaded with pharmaceuticals. *Water Res* 41(15):3299–3306. <https://doi.org/10.1016/j.watres.2007.05.006>
4. Boreen AL, Arnold WA, McNeill K (2004) Photochemical fate of sulfa drugs in the aquatic environment: sulfa drugs containing five-membered heterocyclic groups. *Environ Sci Technol* 38(14):3933–3940. <https://doi.org/10.1021/es0353053>
5. Bound JP, Voulvoulis N (2005) Household disposal of pharmaceuticals as a pathway for aquatic contamination in the United Kingdom. *Environ Health Perspect* 113(12):1705–1711. <https://doi.org/10.1289/ehp.8315>
6. Chaukura N, Gwenzi W, Tavengwa N, Manyuchi MM (2016) Biosorbents for the removal of synthetic organics and emerging pollutants: opportunities and challenges for developing countries. *Environ Dev* 19:84–89. <https://doi.org/10.1016/j.envdev.2016.05.002>
7. Choi YK, Kan E (2019) Effects of pyrolysis temperature on the physicochemical properties of alfalfa-derived biochar for the adsorption of bisphenol A and sulfamethoxazole in water. *Chemosphere* 218:741–748. <https://doi.org/10.1016/j.chemosphere.2018.11.151>
8. Dantas RF, Contreras S, Sans C, Esplugas S (2008) Sulfamethoxazole abatement by means of ozonation. *J Hazard Mater* 150(3):790–794. <https://doi.org/10.1016/j.jhazmat.2007.05.034>

9. Ebele AJ, Abou-Elwafa Abdallah M, Harrad S (2017) Pharmaceuticals and personal care products (PPCPs) in the freshwater aquatic environment. *Emerg Contam* 3(1):1–16. <https://doi.org/10.1016/j.emcon.2016.12.004>
10. Evgenidou EN, Konstantinou IK, Lambropoulou DA (2015) Occurrence and removal of transformation products of PPCPs and illicit drugs in wastewaters: a review. *Sci Total Environ* 505:905–926. <https://doi.org/10.1016/j.scitotenv.2014.10.021>
11. Hirsch R, Ternes T, Haberer K, Kratz K-L (1999) Occurrence of antibiotics in the aquatic environment. In: *The science of the total environment*, vol 225
12. Jang HM, Yoo S, Park S, Kan E (2019) Engineered biochar from pine wood: characterization and potential application for removal of sulfamethoxazole in water. *Environ Eng Res* 24(4):608–617. <https://doi.org/10.4491/eer.2018.358>
13. Ji L, Chen W, Zheng S, Xu Z, Zhu D (2009) Adsorption of sulfonamide antibiotics to multiwalled carbon nanotubes. *Langmuir* 25(19):11608–11613. <https://doi.org/10.1021/la9015838>
14. Benis KZ, Soltan J, McPhedran KN (2021) Electrochemically modified adsorbents for treatment of aqueous arsenic: pore diffusion in modified biomass vs. biochar. *Chem Eng J* 423:130061
15. Laurent J, Casellas M, Pons MN, Dagot C (2009) Flocs surface functionality assessment of sonicated activated sludge in relation with physico-chemical properties. *Ultrason Sonochem* 16(4):488–494. <https://doi.org/10.1016/j.ultsonch.2008.12.012>
16. Lee I-S, Lee S-H, Oh J-E (2010) Occurrence and fate of synthetic musk compounds in water environment. *Water Res* 44(1):214–222. <https://doi.org/10.1016/j.watres.2009.08.049>
17. Lima DRS, Baêta BEL, Aquino SF, Libânio M, Afonso RJCF (2014) Removal of pharmaceuticals and endocrine disruptor compounds from natural waters by clarification associated with powdered activated carbon. *Water Air Soil Pollut* 225(11):2170. <https://doi.org/10.1007/s11270-014-2170-z>
18. Liu J-L, Wong M-H (2013) Pharmaceuticals and personal care products (PPCPs): a review on environmental contamination in China. *Environ Int* 59:208–224. <https://doi.org/10.1016/j.envint.2013.06.012>
19. Manallack DT (2009) The acid–base profile of a contemporary set of drugs: implications for drug discovery. *SAR QSAR Environ Res* 20(7–8):611–655. <https://doi.org/10.1080/10629360903438313>
20. Mompelat S, le Bot B, Thomas O (2009) Occurrence and fate of pharmaceutical products and by-products, from resource to drinking water. *Environ Int* 35(5):803–814. <https://doi.org/10.1016/j.envint.2008.10.008>
21. Ngah WW, Hanafiah MM (2008) Removal of heavy metal ions from wastewater by chemically modified plant wastes as adsorbents: a review. *Biores Technol* 99(10):3935–3948
22. Oliveira FR, Patel AK, Jaisi DP, Adhikari S, Lu H, Khanal SK (2017) Environmental application of biochar: current status and perspectives. *Biores Technol* 246:110–122. <https://doi.org/10.1016/j.biortech.2017.08.122>
23. Salihi IU, Kutty SRM, Ismail HHM (2018) Copper metal removal using sludge activated carbon derived from wastewater treatment sludge. *MATEC Web Conf* 203:03009. <https://doi.org/10.1051/mateconf/201820303009>
24. Sardar M, Manna M, Maharana M, Sen S (2021) Remediation of dyes from industrial wastewater using low-cost adsorbents, pp 377–403. [https://doi.org/10.1007/978-3-030-47400-3\\_15](https://doi.org/10.1007/978-3-030-47400-3_15)
25. Saritha V, Srinivas N, Srikanth Vuppala NV (2017) Analysis and optimization of coagulation and flocculation process. *Appl Water Sci* 7(1):451–460. <https://doi.org/10.1007/s13201-014-0262-y>
26. Sheng J, Yin H, Qian F, Huang H, Gao S, Wang J (2020) Reduced graphene oxide-based composite membranes for in-situ catalytic oxidation of sulfamethoxazole operated in membrane filtration. *Sep Purif Technol* 236:116275. <https://doi.org/10.1016/j.seppur.2019.116275>

27. Shi Y, Liu G, Wang L, Zhang H (2019) Activated carbons derived from hydrothermal impregnation of sucrose with phosphoric acid: remarkable adsorbents for sulfamethoxazole removal. *RSC Adv* 9(31):17841–17851. <https://doi.org/10.1039/c9ra02610j>
28. Shi Y, Liu G, Wang L, Zhang H (2019) Heteroatom-doped porous carbons from sucrose and phytic acid for adsorptive desulfurization and sulfamethoxazole removal: a comparison between aqueous and non-aqueous adsorption. *J Colloid Interface Sci* 557:336–348. <https://doi.org/10.1016/j.jcis.2019.09.032>
29. Sun B, Lian F, Bao Q, Liu Z, Song Z, Zhu L (2016) Impact of low molecular weight organic acids (LMWOAs) on biochar micropores and sorption properties for sulfamethoxazole. *Environ Pollut* 214:142–148. <https://doi.org/10.1016/j.envpol.2016.04.017>
30. Tonucci MC, Gurgel LVA, de Aquino SF (2015) Activated carbons from agricultural byproducts (pine tree and coconut shell), coal, and carbon nanotubes as adsorbents for removal of sulfamethoxazole from spiked aqueous solutions: kinetic and thermodynamic studies. *Ind Crops Prod* 74:111–121. <https://doi.org/10.1016/j.indcrop.2015.05.003>
31. Vieno N, Tuhkanen T, Kronberg L (2006) Removal of pharmaceuticals in drinking water treatment: effect of chemical coagulation. *Environ Technol* 27(2):183–192. <https://doi.org/10.1080/0959332708618632>
32. Wang J, Chen C (2009) Biosorbents for heavy metals removal and their future. *Biotechnol Adv* 27(2):195–226. <https://doi.org/10.1016/j.biotechadv.2008.11.002>
33. Yuan R, Zhu Y, Zhou B, Hu J (2019) Photocatalytic oxidation of sulfamethoxazole in the presence of TiO<sub>2</sub>: effect of matrix in aqueous solution on decomposition mechanisms. *Chem Eng J* 359:1527–1536. <https://doi.org/10.1016/j.cej.2018.11.019>
34. Zama EF, Zhu Y-G, Reid BJ, Sun G-X (2017) The role of biochar properties in influencing the sorption and desorption of Pb(II), Cd(II) and As(III) in aqueous solution. *J Clean Prod* 148:127–136. <https://doi.org/10.1016/j.jclepro.2017.01.125>
35. Zare H, Heydarzade H, Rahimnejad M, Tardast A, Seyfi M, Peyghambarzadeh SM (2015) Dried activated sludge as an appropriate biosorbent for removal of copper (II) ions. *Arab J Chem* 8(6):858–864. <https://doi.org/10.1016/j.arabjc.2012.11.019>
36. Zhang D, Gersberg RM, Ng WJ, Tan SK (2014) Removal of pharmaceuticals and personal care products in aquatic plant-based systems: a review. *Environ Pollut* 184:620–639. <https://doi.org/10.1016/j.envpol.2013.09.009>
37. Zhang D, Pan B, Zhang H, Ning P, Xing B (2010) Contribution of different sulfamethoxazole species to their overall adsorption on functionalized carbon nanotubes. *Environ Sci Technol* 44(10):3806–3811. <https://doi.org/10.1021/es903851q>
38. Zheng H, Wang Z, Zhao J, Herbert S, Xing B (2013) Sorption of antibiotic sulfamethoxazole varies with biochars produced at different temperatures. *Environ Pollut* 181:60–67. <https://doi.org/10.1016/j.envpol.2013.05.056>

# H<sub>3</sub>PO<sub>4</sub> and NaOH Treated Canola Straw Biochar for Arsenic Adsorption



Julia Norberto, Khaled Zoroufchi Benis, Jafar Soltan,  
and Kerry McPhedran

**Abstract** It is known that arsenic is remarkably toxic and can be found in high concentrations in a variety of natural waters and wastewaters both in Canada and worldwide. Thus, it is important to adequately treat these waters, with removal using adsorption methods being promising due to their simplicity, effectiveness, and relatively low costs. Agricultural residues are abundant in Saskatchewan, Canada, and can be utilized as an inexpensive biomass to produce adsorbents. In this study, biochar prepared from a raw canola straw biomass was investigated for its arsenate, As(V), and arsenite, As(III) adsorption capacities after being modified using phosphoric acid (H<sub>3</sub>PO<sub>4</sub>) and sodium hydroxide (NaOH). The biomass was treated prior to being converted into biochar using a prepyrolysis method. Canola straw biochar (CSB) was made through both conventional and microwave pyrolysis methods and a variety of different pretreatment parameters were explored including varying H<sub>3</sub>PO<sub>4</sub> and NaOH to biomass ratios. Studying the effect of acid or base to biomass ratio suggested that the highest H<sub>3</sub>PO<sub>4</sub> and NaOH ratio led to higher adsorption for both As(III) (10.59 µg/g and 12.46 µg/g, respectively) and As(V) (12.65 µg/g and 16.69 µg/g, respectively). Furthermore, varying solution pH at values of 3, 7, and 10 demonstrated that H<sub>3</sub>PO<sub>4</sub> and NaOH CSB had markedly increased adsorption capacity at pH 7 for As(V). Overall, H<sub>3</sub>PO<sub>4</sub> and NaOH CSB showed marginal adsorption enhancements for As(III) and As(V).

**Keywords** Acid · Adsorption · Base · Biochar · Canola straw

---

J. Norberto · K. McPhedran (✉)

Department of Civil, Geological and Environmental Engineering, University of Saskatchewan,  
Saskatoon, Canada

e-mail: [kerry.mcphedran@usask.ca](mailto:kerry.mcphedran@usask.ca)

K. Z. Benis · J. Soltan

Department of Chemical and Biological Engineering, University of Saskatchewan, Saskatoon,  
Canada

J. Soltan · K. McPhedran

Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_65](https://doi.org/10.1007/978-3-031-34593-7_65)



## 1 Introduction

Heavy metals found in natural waters and anthropogenic wastewaters have become an increasingly critical threat to humans, animals, and the environment overall. A common heavy metal or metalloid (referred to as a heavy metal herein) that has been studied in the past, and continues to be relevant for present research, is arsenic. It is known that arsenic is remarkably toxic and can be found in high concentrations in a variety of aqueous environments worldwide [29]. Arsenic has an atomic number of 33 and holds different positions concerning its presence, including being 20<sup>th</sup> in natural abundance, 14<sup>th</sup> in seawater, and 12<sup>th</sup> in the human body. It has an atomic weight of 74.9, specific gravity of 5.73, and boiling point of 613 °C. Additionally, this heavy metal can be found in four different oxidation states in a variety of compounds: As(-III), As(0), As(III), and As(V) [22]. In the environment, arsenic occurs mainly as the arsenic oxides including arsenite, As(III), and arsenate, As(V), [17] which are the commonly found species in waters [32].

Arsenic contamination can occur by ingestion of this heavy metal through drinking water [1], foodstuffs, use of groundwater, and in contaminated soils [33]. Both forms of arsenic, As(III) and As(V), are currently considered important for assessment of treatment methodologies due to their high toxicity and capacity to be human carcinogens and/or ability to lead to chronic illnesses [44]. For example, distinct diseases such as lung, skin, and bladder cancer can be caused by elevated levels of arsenic exposure to humans [27]. Therefore, due to its abundance and critical effects on humans, animals, and the environment, further research must be done to address and remediate this worldwide issue of arsenic contaminated waters and wastewaters.

Various treatment methodologies and techniques have been considered for arsenic removal from both water and wastewater matrices. Among the most commonly used are microfiltration and ultrafiltration [31], coagulation and filtration [32], adsorption [10], flocculation [25], and membrane filtration [44]. Each of these techniques have their own advantages and disadvantages. For example, [9] studied an ultrafiltration membrane combined with adsorption process for arsenic removal, which showed that this hybrid method was able to remove arsenic along with other dissolved pollutants. This technique displays a promising separation refinement and modest energy necessities. On the other hand, coagulation and flocculation can be used to absorb finer particles, but this technique has a high cost and generally requires further treatment of the created sludges [35]. Hoang et al. [11] highlights that membrane filtration is an effective process, however it has an increased cost due to membrane fouling and pump usage. Overall, adsorption can be an economic and highly efficient method when compared to other techniques [32], thus, it is the focus of the current study.

While adsorption has been an increasingly established process, utilizing biochar as an adsorbent has become a more effective option and the focus of a variety of recent research studies. Biochar is an environmentally friendly material that can be generated from different types of biomasses and often results in the creation of high adsorption capacity adsorbents particularly for target heavy metals [41]. Biochar is made from pyrolysis processes which are a method of biomass decomposition at

high temperatures and in the absence, or significantly small quantity of, oxygen [23]. These biochar materials present a high specific surface area and a porous structure which are both contributing factors for the adsorption of pollutants [41]. Previous studies have shown that biochar modification and/or activation can be done to improve adsorption capacities. However, different chemicals considered in the literature have presented distinct outcomes and results. For example, potassium hydroxide (KOH) [12], phosphoric acid (H<sub>3</sub>PO<sub>4</sub>) [13, 46], iron (Fe) [19], sodium hydroxide (NaOH) [40], among other agents, have been reported to be able to successfully increase biochar adsorption efficiency for various heavy metals. Thus, these agents may be useful in enhancing biochar adsorption for arsenic removal from water.

H<sub>3</sub>PO<sub>4</sub> and NaOH have previously been used for biomass activation in different heavy metal adsorption scenarios. Additionally, H<sub>3</sub>PO<sub>4</sub> is also attractive for its environmentally-friendliness due to it being less polluting and readily washed away with water [7, 28]. These activation agents can enhance the biochar by surface cell modification and unmasking supplementary metal binding sites which can be used for heavy metal attraction [24]. Furthermore, these chemical modifications can also increase surface areas, micropores and stimulate the appearance of new functional groups in the biochar [7]. Peng et al. [28] reported of new phosphorus functional groups on pine saw biochar after activation with H<sub>3</sub>PO<sub>4</sub>. These modified functional groups and increased surface area resulted in higher cadmium and copper adsorption. Moreover, [15] used NaOH to activate rice husk for potential cadmium adsorption and found that the modified adsorbent had enhanced sorption capacity from 8.58 mg/g to 20.24 mg/g when compared to raw biomass. Other studies have also reported increased adsorption using chemical activation such as As(III) removal using H<sub>3</sub>PO<sub>4</sub>-activated sugarcane bagasse [13], nickel and copper adsorption using NaOH pretreated *Magnifera indica* biomass [24], and As(V) sorption using cotton stalk biochar activated with H<sub>3</sub>PO<sub>4</sub> [12]. Chemical pretreatment efficiency will be influenced by the biomass used and heavy metal to be adsorbed.

Therefore, the overall objective of this study is to investigate the impacts of H<sub>3</sub>PO<sub>4</sub> and NaOH pretreatments on the effectiveness of canola straw biochar (CSB) for the sorption of both As(III) and As(V). Previous studies have reported the use of these activating agents [7, 28, 40] for other heavy metals and metalloids on biochars generated from distinct biomasses. However, most studies have prioritized the usage of conventional pyrolysis only. Thus, this current research aims to compare and analyze the effect of pyrolysis types, including conventional and microwave methodologies, on the biochar adsorption capacity and to determine how the activating agents may enhance the biochar arsenic adsorption capacity. Canola straw was used as the biomass due to its abundance in Saskatchewan from agricultural residues [6] and the project being funded through the Saskatchewan Agricultural Development Fund (ADF). Furthermore, variables including acid/base ratio and pH were examined to investigate their effects on biochar adsorption to further enhance the adsorption capacity of the created biochar.

## 2 Materials and Methods

### 2.1 Biomass Preparation

Raw canola straw was gathered by the industrial partner from an agricultural field in Saskatchewan, Canada, before being delivered to the Environmental Engineering Laboratories, Engineering Building, University of Saskatchewan. For biomass preparation, the canola straw was washed using deionized water and then dried at 60 °C for approximately 24 h. Next, the canola straw was ground and sieved to the appropriate size range of 425–850 µm which has been shown to be a reasonable size fraction for use in the production of biochar by our research group. Once this process was complete, the biomass was ready for biochar production based on the following sections.

### 2.2 $H_3PO_4$ and NaOH Biomass Pretreatment

The  $H_3PO_4$  and NaOH were used to treat the canola straw biomass (CSB) prior to undergoing pyrolysis, or a prepyrolysis modification. Three sets of different ratios were used for both  $H_3PO_4$  and NaOH for investigating the effect of acid and base quantity on the arsenic adsorption efficiency. The acid/base to biomass ratios are shown for the three biochar samples produced in Table 1. An 85%  $H_3PO_4$  stock solution (Fisher Scientific, USA) was used in different ratios in relation to the initial biomass amount to be pyrolyzed, which was 5 g. Since  $H_3PO_4$  was a liquid solution, a w/v ratio was calculated to develop an equivalent proportion to the biomass. For these acid solutions, the volumes equivalent to 10 g, 20 g, and 40 g (5.93 mL; 11.86 mL; and 23.72 mL, respectively) were used, and deionized water was added to reach a final volume of 100 mL for mixing with the canola straw. The NaOH (Fisher Scientific, USA) solid was weighed according to the distinct ratios and added to 100 mL of deionized water for mixing as well. All samples were mixed for 30 min which was adequate for the modification process based on preliminary experiments by our research team.

**Table 1**  $H_3PO_4$  and NaOH to biomass (BM) ratios for biochar production

BC	Ratio (ACID/BASE:BM)
$H_3PO_4$ BC—1	2:1
$H_3PO_4$ BC—2	4:1
$H_3PO_4$ BC—3	8:1
NaOH BC—1	0.5:1
NaOH BC—2	1:1
NaOH BC—3	2:1

### **2.3 Microwave Pyrolysis Biochar**

For microwave pyrolysis, a common microwave (Goldstar, GSC-4006-M, Canada) was modified to allow for nitrogen flow directly to the sample and other necessary adjustments for the pyrolysis procedure to proceed. After being treated with H<sub>3</sub>PO<sub>4</sub> or NaOH, the biomass samples were washed, and each were pyrolyzed in the microwave for 6 min total (two 3-min intervals with 30 s in between). These intervals were performed in an attempt to produce optimal biochar yields following [20]. During pyrolysis, a constant nitrogen flow rate of 25 mL/min was being supplied to sample to eliminate oxygen presence. Once completed, the biochar samples were allowed to cool to approximately room temperature, removed from the microwave, washed with 300 mL of deionized water, and then dried at 100 °C for 24 h before being used in any experiments.

### **2.4 Conventional Pyrolysis Biochar**

For this conventional pyrolysis biochar process, the biomass samples were dried after being treated with acid or base prior to being subjected to pyrolysis. Conventional pyrolysis was carried out in a tubular furnace (Lindberg/Blue M, Model TF55030A-1, Thermo Scientific, USA) with a nitrogen flow rate of 40 mL/min used to minimize oxygen in the pyrolysis process. The heating rate of the furnace was 8 °C/min which was constant up to reaching a maximum value of 400 °C, and then the sample was maintained at 400 °C for 1 h to complete the pyrolysis process. Once created, the biochar samples were allowed to cool to approximately room temperature, removed from the furnace, washed with approximately 300 mL of deionized water, and dried at 100 °C for 24 h before being used in any experiments in a process analogous to the microwave pyrolysis treatment.

### **2.5 As(III) and As(V) Adsorption Experiments**

Several experiments were performed in order to better understand the effects of the acid and base pretreatments coupled with the two different pyrolysis techniques. For these experiments, the produced adsorbents were placed into volumetric flasks containing 80 mL solutions of 1 mg/L of As(III) or As(V) with 0.08 g of the prepared biochar resulting in an adsorbent dosage of 1 g/L. These solutions were placed on an orbital shaker (Orbishaker CO<sub>2</sub>, Benchmark, USA) for approximately 24 h at 200 rpm (25 °C). The 24 h interval has been previously determined to be adequate for attainment of equilibrium for adsorption experiments. Samples were then filtered using 6 µm filter papers (Whatman 1003–055, GE, USA), and the aqueous arsenic concentrations were determined using an atomic absorption spectroscopy (AAS,

iCE 3000, Thermo Scientific, USA) coupled with vapor generation assembly (VGA, VP100, Thermo Scientific, USA) methodology.

The biochar sample from each pretreatment method (both  $\text{H}_3\text{PO}_4$  and NaOH) resulting in the highest arsenic adsorption capacity was chosen for further investigation including the development of isotherm models and kinetic models, and variable pH experiments. Isotherm experiments were completed to determine how the two selected optimal adsorbents reacted to different arsenic concentrations. These experiments were performed at five different arsenic concentrations (1, 5, 10, 20 and 40 mg/L) for both As(III) and As(V). The isotherm models being considered currently included Freundlich, Langmuir, Temkin, and Redlich-Peterson. The kinetic experiments were carried out for the two selected optimal adsorbents at the optimized arsenic concentration found in each of the isotherm experiments, and samples were collected at different time intervals (30 min, 1.5 h, 3 h, 17 h, 24 h and 48 h). Lastly, pH experiments were assessed at three levels including pH 3 (acidic), 7 (neutral), and 10 (basic) for both As(III) and As(V). The pH was adjusted by dropwise addition using 0.1 M HCl and NaOH solutions. All experiments were run in triplicates, and the software package Origin Pro 2021 was used to best fit the isotherm and kinetic data to their respective models.

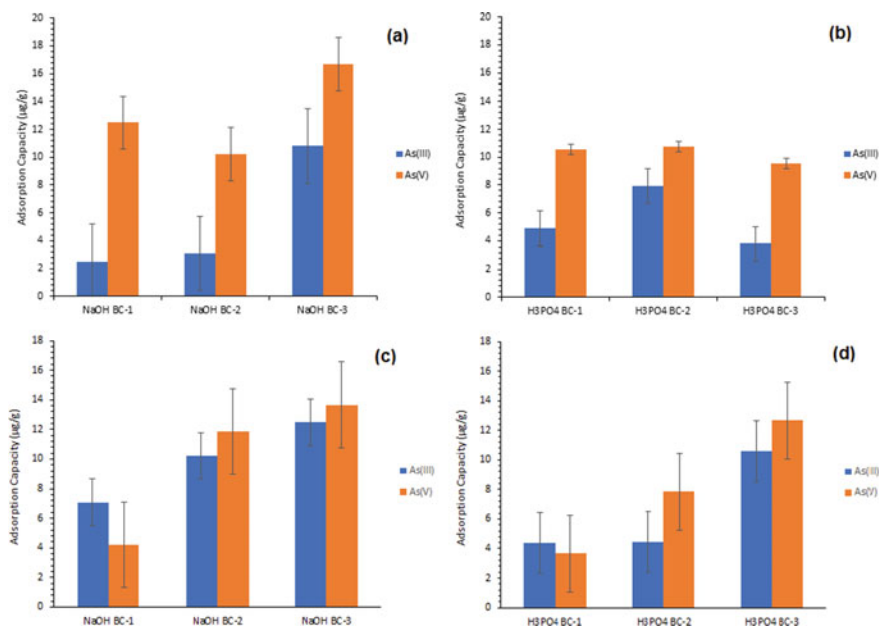
### 3 Results and Discussion

#### 3.1 As(III) and As(V) Pyrolysis Technique Effects

General adsorption tests were first completed to determine the effect of varying acid and base ratios and pyrolysis techniques, as was shown in Table 1. Figure 1 shows the resulting adsorption capacities of each pretreated CSB made in microwave pyrolysis for As(III) and As(V) in (a) and (b), with results of conventional pyrolysis shown in (c) and (d).

Overall, both the  $\text{H}_3\text{PO}_4$  and NaOH treated CSBs had higher adsorption capacities for As(V) as compared to As(III), which can be attributed to the As(V) having a greater stationary state versus As(III) which is more mobile [32]. For example, the NaOH treated CSB had a moderately higher adsorption capacity (16.69  $\mu\text{g/g}$ ) than  $\text{H}_3\text{PO}_4$  (10.76  $\mu\text{g/g}$ ) for As(V) for the NaOH BC-3 treatment. Additionally, NaOH CSB showed increasing sorption as the NaOH ratio increased resulting in the highest arsenic adsorption for both As(III) and As(V) for the NaOH BC-3 (base to biomass ratio of 2:1) treatment. In contrast, the  $\text{H}_3\text{PO}_4$  adsorption capacity did not increase linearly with  $\text{H}_3\text{PO}_4$  ratio. Overall, the  $\text{H}_3\text{PO}_4$  CSB adsorption capacities were similar for all treatments for As(V), while for As(III), the  $\text{H}_3\text{PO}_4$  BC-2 treatment (acid to biomass ratio of 4:1) was slightly higher than the other two treatments.

Acid and base treatments each generate distinct functional groups on the biomass that can influence the resultant adsorption capacities of the created biochar. Generally, the highest adsorption capacities currently were obtained for the NaOH treated CSB



**Fig. 1** Results for H<sub>3</sub>PO<sub>4</sub> and NaOH treated biochar sorption experiments using microwave pyrolysis (a) and b and conventional pyrolysis (c) and d. Error bars represent standard deviations with  $n = 3$

which showed higher values for both As(III) and As(V) when compared to the H<sub>3</sub>PO<sub>4</sub> treated CSB. This may be attributed to decreased particle size and greater surface area and porosity as a result of the NaOH base treatment, as [12] reported when activating a biochar with KOH for As(V) adsorption experiments. Biochar base pretreatment can also lead to organic matter dissolution and the creation of additional oxygenated groups which may assist in activating biochar resulting in improved adsorption capacities [2].

For comparison, Fig. 1 panels (c) and (d) show the adsorption capacities for the H<sub>3</sub>PO<sub>4</sub> and NaOH CSBs produced using the conventional pyrolysis process. Similar to the microwave pyrolysis results, most of the biochars adsorbed As(V) better than As(III). However, this excludes the two biochars with the lowest ratios of both H<sub>3</sub>PO<sub>4</sub> and NaOH which had similar adsorption capacities for both As(III) and As(V). Conventional pyrolysis shows a more linear trend in the relationship between acid/base ratio and adsorption efficiency, with the highest adsorption capacities being for NaOH BC-3 (13.65 µg/g) and H<sub>3</sub>PO<sub>4</sub> BC-3 (12.65 µg/g). Thus, in conventional pyrolysis, the two biochar pretreatments resulted in more similar adsorption capacities in comparison to the microwave pyrolysis experiments.

Although the microwave pyrolysis CSB showed the highest NaOH sorption ability overall for NaOH BC-3 at 16.69 µg/g, the conventional pyrolysis CSB proved to create a more stable biochar in solution while still producing a similar range of

adsorption efficiencies to the microwave treatments. Furthermore, the conventional pyrolysis process had a greater biochar yield for both  $\text{H}_3\text{PO}_4$  (average yield of 76% vs 55%) and NaOH (average yield of 28% vs. 26%) when compared to microwave pyrolysis. Therefore, the conventional pyrolysis treatments including  $\text{H}_3\text{PO}_4$  BC-3 and NaOH BC-3 were considered for further investigations.

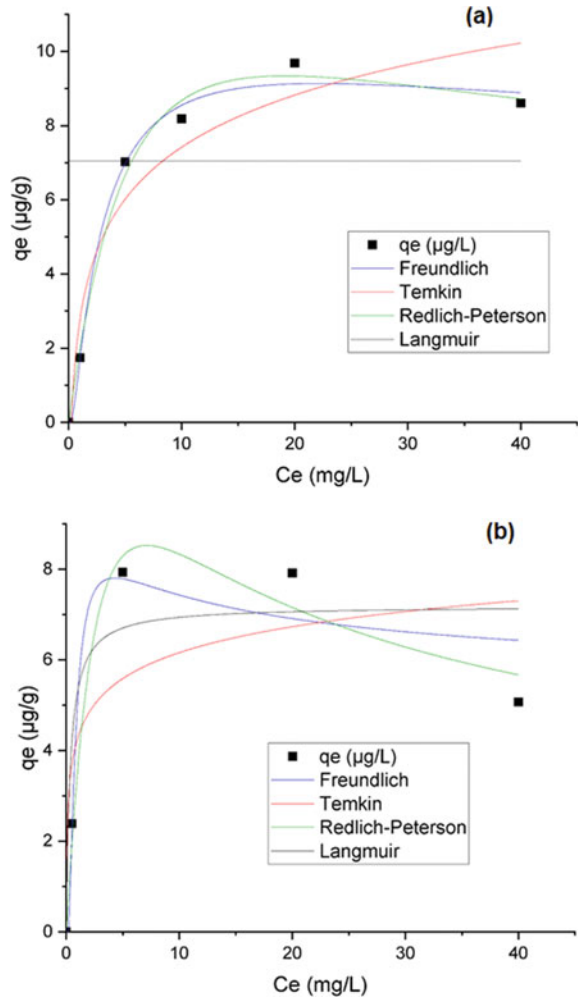
### 3.2 Arsenic Adsorption Isotherm and Kinetic Models

Adsorption isotherms describe equilibrium relationships between the contaminant (e.g., arsenic) and the adsorbent (e.g., biochar). Thus, they are essential for use in determining surface features and technique alternatives that can improve adsorption system outcomes [8]. A wide variety of isotherm models exist including Freundlich, Langmuir, Temkin, and Redlich-Peterson, which were each used to fit the present isotherm data. Each model describes a unique adsorption mechanism and can help to further comprehend the process that occurs for the adsorption phenomenon.

In Fig. 2 and Table 2, the isotherm data for As(III) is presented for both the  $\text{H}_3\text{PO}_4$  and NaOH pretreated CSBs. The data was tested against the different models in order to determine the model which produces the best fit to the experimental data. An ANOVA one-way test was done and used to determine if the data was statistically significant. For  $\text{H}_3\text{PO}_4$ , all three models were deemed to have an acceptable fit with  $R^2 > 0.9$ . However, the best fit was for the Redlich-Peterson model with an  $R^2$  of 0.99. For NaOH, the Temkin and Freundlich models were deemed to be unacceptable with  $R^2$  values of 0.72 and 0.86, respectively. Similar to  $\text{H}_3\text{PO}_4$ , the Redlich-Peterson model was the best fit for NaOH with an  $R^2$  of 0.97. The Redlich-Peterson isotherm model is essentially a three-parameter isotherm that associates both Langmuir and Freundlich isotherms [4]. Thus, this model suggests that the adsorption mechanism that occurred for both  $\text{H}_3\text{PO}_4$  and NaOH treatments are unique and do not adhere to monolayer adsorption alone [16]. The flexibility of this model allows applicability to homogeneous and heterogeneous systems [4]. For both  $\text{H}_3\text{PO}_4$  and NaOH CSB, the Freundlich model was a better match than Langmuir and Temkin, indicating a greater feasibility that the adsorption mechanism that transpired was a heterogeneous multilayer process. This data is in agreement with past studies where both acid ( $\text{H}_3\text{PO}_4$ ) and base (NaOH) pretreatments showed similar adsorption mechanism processes [7, 38]. Additionally, previous studies have shown comparable experimental data fits with similar number of data points [5, 26, 34, 45], further corroborating the observed results.

Given the isotherm data showed higher adsorption capacities at higher arsenic concentrations, the kinetic experiments were completed at 40 mg/L As(V) concentrations. The data collected was tested against kinetic models that are widely used including: pseudo-first-order, pseudo-second-order, Elovich, and intra-particle diffusion models. Each of these models can be useful in the understanding of the rate of adsorption equilibrium.

**Fig. 2** Isotherm model fits including Freundlich, Temkin, Redlich-Peterson, and Langmuir for **a** H<sub>3</sub>PO<sub>4</sub> CSB and **b** NaOH CSB. See Table 2 for further information



**Table 2** Equations, constant, and  $R^2$  values for isotherm model fits for H<sub>3</sub>PO<sub>4</sub> and NaOH biochars

Model	Equation	Constant values		$R^2$ Values	
		H <sub>3</sub> PO <sub>4</sub>	NaOH	H <sub>3</sub> PO <sub>4</sub>	NaOH
Freundlich	$q_e = K_F C_e^{\frac{1}{n_F}}$	$K_F = 1.681$ $n_F = 0.323$	$K_F = 5.078$ $n_F = 0.685$	0.986	0.859
Temkin	$q_e = \frac{RT}{b_T} \ln(K_T C_e)$	$K_T = 3.905$ $b_T = 1234.75$	$K_T = 175.187$ $b_T = 3023.27$	0.925	0.719
Redlich-Peterson	$q_e = \frac{K_{RP} C_e}{1 + a_{RP} C_e^{\beta}}$	$K_{RP} = 2.089$ $a_{RP} = 0.069$	$K_{RP} = 4.06$ $a_{RP} = 0.148$	0.993	0.973
Langmuir	$q_e = \frac{K_L q_{max} C_e}{1 + K_L C_e}$	$q_{max} = 7.047$ $K_L = 7.02E^{46}$	$q_{max} = 7.187$ $K_L = 2.768$	0.515	0.331



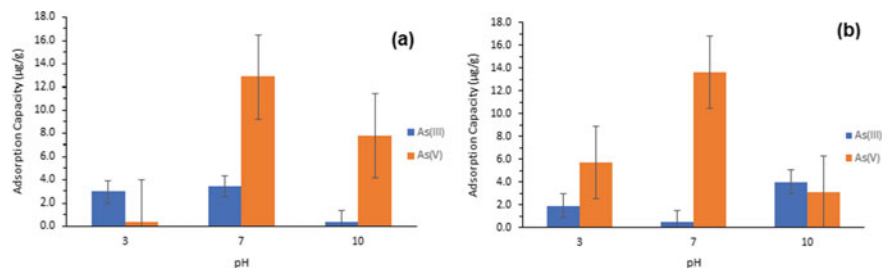
**Table 3** Equations, constant, and  $R^2$  values for kinetic model fits for  $\text{H}_3\text{PO}_4$  and NaOH biochars

Model	Equation	Constant values		$R^2$ values	
		$\text{H}_3\text{PO}_4$	NaOH	$\text{H}_3\text{PO}_4$	NaOH
Pseudo-first-order	$q_t = q_e \times (1 - e^{-k_1 t})$	$k_1 = 8.426$ $q_e = 3008$	$k_1 = 0.0077$ $q_e = 9.719$	0.94	0.547
Pseudo-second-order	$\frac{t}{q_t} = \frac{1}{k_2 \times q_e^2} + \left(\frac{1}{q_e}\right) \times t$	$k_2 = 0.0011$ $q_e = 9.719$	$k_1 = 0.0011$ $q_e = 10.341$	0.99	0.626
Elovich	$q_t = \frac{1}{b} \times \ln(1 + abt)$	$b = 7.641$ $a = 7.565\text{E}^{23}$	$b = 0.678$ $a = 0.609$	0.759	0.732
Intra-particle diffusion	$q_t = k_p \times t^{1/2} + C$	$k_p = 6.515\text{E}^{-4}$ $C = 6.562$	$k_p = 0.001$ $C = 7.33$	0.113	0.753

Table 3 presents the kinetic model results used to fit  $\text{H}_3\text{PO}_4$  and NaOH CSB experiments. For the  $\text{H}_3\text{PO}_4$  CSB, there were two satisfactory model fits for both the pseudo-first-order and pseudo-second-order models ( $R^2$  of 0.94 and 0.99, respectively), while the Elovich and intra-particle diffusion models did not fit as well ( $R^2$  of 0.76 and 0.11, respectively). This model fit indicates that the arsenic adsorption occurred through chemisorption in many active sites [36], which is in agreement with the isotherm conclusions presented previously. For the NaOH CSB, there was no clear model that was highly acceptable for the experimental data. However, the highest  $R^2$  value of 0.75 was found for the intra-particle diffusion model which suggests that the arsenic adsorption occurred in a homogeneous sphere [30]. For both cases, the adsorption capacities maintained similar values to those that were previously seen in the initial conventional pyrolysis experiments presented in Sect. 3.1. Additionally, the kinetic model fitting is in agreement with previous research using  $\text{H}_3\text{PO}_4$  and NaOH treated biochars for other heavy metals [14, 45, 47].

### 3.3 pH Effect on Adsorption

The pH of a solution has been reported to have a significant effect on the removal of arsenic due to the impact it can have on the adsorbent, as well as the arsenic ion species, in the water being treated [48]. The impacts of the three levels of pH (pH 3, 7, and 10) on the  $\text{H}_3\text{PO}_4$  and NaOH CSB adsorption capacities are presented in Fig. 3. Clearly, both  $\text{H}_3\text{PO}_4$  and NaOH CSBs showed an optimal As(V) adsorption at pH 7.



**Fig. 3** pH effects for As(III) and As(V) for: **a** H<sub>3</sub>PO<sub>4</sub> CSB and **b** NaOH CSB

In fact, the H<sub>3</sub>PO<sub>4</sub> CSB only showed meaningful adsorption at pH 7 (12.65 μg/g) for As(V) with the adsorptions at pH 3 and pH 10 being less than 6.0 μg/g. Similarly, the NaOH CSB showed significantly higher adsorption for pH 7 (13.6 μg/g) as opposed to pH 3 (5.75 μg/g) and 10 (3.12 μg/g) for As(V). However, As(III) did not present considerable differences in adsorption at different pH values for either the H<sub>3</sub>PO<sub>4</sub> or NaOH CSB (Fig. 3). Thus, the As(III) adsorption currently appears to be independent of pH, while the As(V) adsorption presents an optimal adsorption at pH 7. This may be due to the fact that at pH 7 As(V) exists as H<sub>2</sub>AsO<sub>4</sub><sup>-</sup> or, HAsO<sub>4</sub><sup>2-</sup>, which can be attracted to the positively charged adsorbent surface [21].

## 4 Conclusions

Both the H<sub>3</sub>PO<sub>4</sub> and NaOH pretreated CSBs were prepared by two distinct pyrolysis techniques, conventional and microwave, for assessment of their abilities to remove As(III) and As(V) from water. The different pyrolysis techniques exhibited minor differences on the overall adsorption capacities for both the acid and base treated biochars. However, the biochar yield from conventional pyrolysis is considerably greater than that from microwave pyrolysis making it the preferred method for biochar preparation. The CSBs with highest ratios for both H<sub>3</sub>PO<sub>4</sub> and NaOH pretreatments showed the best adsorption capacities in initial experiments, thus, were utilized for further experiments. Isotherm, kinetic, and pH experiments were performed to investigate the adsorption mechanisms that occurred for As(III) and As(V). The results of the kinetic and isotherm models for the H<sub>3</sub>PO<sub>4</sub> CSB were in agreement and were used to determine that arsenic was adsorbed in a heterogeneous and multilayer adsorption process. The NaOH CSB showed similar kinetic model results to the H<sub>3</sub>PO<sub>4</sub> CSB; however, it did not show a good fit with the isotherm models under consideration. Moreover, former studies have reported biochar desorption capacities using NaOH, NaNO<sub>3</sub>, and HCl solutions [37, 39, 42, 43]. Further experiments are needed to determine optimum desorption agents and concentrations for these biochars. The pH experiments suggested that pH 7 was optimal for both acid and base treated biochars for As(V) adsorption, but As(III) adsorption was not significantly influenced by

changes in the solution pH. Thus,  $\text{H}_3\text{PO}_4$  and NaOH CSB showed minor sorption enhancements for As(III) and As(V) adsorption when compared to previously treated acid/base adsorbents for adsorbing heavy metals. Previous research showed adsorption capacities of 4.48 mg/g and 4.32 mg/g using  $\text{H}_3\text{PO}_4$  and KOH treated biochar for As(V) removal, respectively [12], 423 mg/g with an  $\text{H}_3\text{PO}_4$  treated biochar for cadmium sorption [18], 93  $\mu\text{g/g}$  using  $\text{H}_3\text{PO}_4$ -activated jute stick for As(III) sorption [3]. Therefore, additional study and research should be done to further improve this adsorbent.

## References

1. Ahmad A, van der Wens P, Baken K, de Waal L, Bhattacharya P, Stuyfzand P (2020) Arsenic reduction to  $< 1 \mu\text{g/L}$  in Dutch drinking water. *Environ Int* 134(September 2019):105253. <https://doi.org/10.1016/j.envint.2019.105253>
2. Amen R, Bashir H, Bibi I, Shaheen SM, Niazi NK, Shahid M, Hussain MM, Antoniadis V, Shakoor MB, Al-Solaimani SG, Wang H, Bundschuh J, Rinklebe J (2020) A critical review on arsenic removal from water using biochar-based sorbents: the significance of modification and redox reactions. *Chem Eng J* 396(November 2019). <https://doi.org/10.1016/j.cej.2020.125195>
3. Asadullah M, Jahan I, Ahmed MB, Adawiyah P, Malek NH, Rahman MS (2014) Preparation of microporous activated carbon and its modification for arsenic removal from water. *J Ind Eng Chem* 20(3):887–896. The Korean Society of Industrial and Engineering Chemistry. <https://doi.org/10.1016/j.jiec.2013.06.019>
4. Ayawei N, Ebelegi AN, Wankasi D (2017) Modelling and Interpretation of adsorption isotherms. *J Chem* 2017. <https://doi.org/10.1155/2017/3039817>
5. Bakshi S, Banik C, Rathke SJ, Laird DA (2018) Arsenic sorption on zero-valent iron-biochar complexes. *Water Res* 137:153–163. <https://doi.org/10.1016/j.watres.2018.03.021>
6. Canola Council of Canada (2021) Canola industry in Canada, from farm to global markets. Available from <https://www.canolacouncil.org/about-canola/industry/>. Accessed 17 Dec 2021
7. Chen H, Li W, Wang J, Xu H, Liu Y, Zhang Z, Li Y, Zhang Y (2019) Adsorption of cadmium and lead ions by phosphoric acid-modified biochar generated from chicken feather: selective adsorption and influence of dissolved organic matter. *Bioresour Technol* 292. <https://doi.org/10.1016/j.biortech.2019.121948>
8. Foo KY, Hameed BH (2010) Insights into the modeling of adsorption isotherm systems. *Chem Eng J* 156(1):2–10. <https://doi.org/10.1016/j.cej.2009.09.013>
9. Hao L, Wang N, Wang C, Li G (2018) Arsenic removal from water and river water by the combined adsorption—UF membrane process. *Chemosphere* 202:768–776. <https://doi.org/10.1016/j.chemosphere.2018.03.159>
10. He R, Peng Z, Lyu H, Huang H, Nan Q, Tang J (2018) Synthesis and characterization of an iron-impregnated biochar for aqueous arsenic removal. *Sci Total Environ* 612:1177–1186. <https://doi.org/10.1016/j.scitotenv.2017.09.016>
11. Hoang AT, Nižetić S, Cheng CK, Luque R, Thomas S, Banh TL, Pham VV, Nguyen XP (2022) Heavy metal removal by biomass-derived carbon nanotubes as a greener environmental remediation: a comprehensive review. In: *Chemosphere*
12. Hussain M, Imran M, Abbas G, Shahid M, Iqbal M, Naeem MA, Murtaza B, Amjad M, Shah NS, Ul Haq Khan Z, Ul Islam A (2020) A new biochar from cotton stalks for As (V) removal from aqueous solutions: its improvement with  $\text{H}_3\text{PO}_4$  and KOH. *Environ Geochem Health* 42(8):2519–2534. <https://doi.org/10.1007/s10653-019-00431-2>
13. Joshi S, Sharma M, Kumari A, Shrestha S, Shrestha B (2019) Arsenic removal fromwater by adsorption onto iron oxide/nano-porous carbon magnetic composite. *Appl Sci (Switzerland)* 9(18). <https://doi.org/10.3390/app9183732>

14. Kim H, Ko RA, Lee S, Chon K (2020) Removal efficiencies of manganese and iron using pristine and phosphoric acid pre-treated biochars made from banana peels. *Water (Switzerland)* 12(4):1–13. <https://doi.org/10.3390/W12041173>
15. Kumar U, Bandyopadhyay M (2006) Sorption of cadmium from aqueous solution using pretreated rice husk. *Biores Technol* 97(1):104–109. <https://doi.org/10.1016/j.biortech.2005.02.027>
16. Kumara NTRN, Hamdan N, Petra MI, Tennakoon KU, Ekanayake P (2014) Equilibrium isotherm studies of adsorption of pigments extracted from Kuduk-kuduk (*Melastoma malabathricum* L.) pulp onto TiO<sub>2</sub> nanoparticles. *J Chem* 2014. <https://doi.org/10.1155/2014/468975>
17. Leist M, Casey RJ, Caridi D (2000) The management of arsenic wastes: Problems and prospects. *J Hazard Mater* 76(1):125–138. [https://doi.org/10.1016/S0304-3894\(00\)00188-6](https://doi.org/10.1016/S0304-3894(00)00188-6)
18. Li X, Wang C, Tian J, Liu J, Chen G (2020) Comparison of adsorption properties for cadmium removal from aqueous solution by *Enteromorpha prolifera* biochar modified with different chemical reagents. *Environ Res* 186(March):109502. <https://doi.org/10.1016/j.envres.2020.109502>
19. Lin L, Qiu W, Wang D, Huang Q, Song Z, Chau HW (2017) Arsenic removal in aqueous solution by a novel Fe–Mn modified biochar composite: characterization and mechanism. *Ecotoxicol Environ Saf* 144(June):514–521. <https://doi.org/10.1016/j.ecoenv.2017.06.063>
20. Menéndez JA, Domínguez A, Inganzo M, Pis JJ (2004) Microwave pyrolysis of sewage sludge: analysis of the gas fraction. *J Anal Appl Pyrol* 71(2):657–667. <https://doi.org/10.1016/j.jaap.2003.09.003>
21. Mishra SP, Mohapatra D, Mishra D, Chattopadhyay P, Roy Chaudhury G, Das RP (2014) Arsenic adsorption on natural minerals. *J Mater Environ Sci* 5(2):350–359
22. Mohan D, Pittman CU (2007) Arsenic removal from water/wastewater using adsorbents-A critical review. *J Hazard Mater* 142(1–2):1–53. <https://doi.org/10.1016/j.jhazmat.2007.01.006>
23. Mohan D, Sarswat A, Ok YS, Pittman CU (2014) Organic and inorganic contaminants removal from water with biochar, a renewable, low cost and sustainable adsorbent—a critical review. *Bioresour Technol* 160:191–202. <https://doi.org/10.1016/j.biortech.2014.01.120>
24. Nadeem R, Zafar MN, Afzal A, Hanif MA, Saeed R (2014) Potential of NaOH pretreated *Mangifera indica* waste biomass for the mitigation of Ni(II) and Co(II) from aqueous solutions. *J Taiwan Inst Chem Eng* 45(3):967–972. <https://doi.org/10.1016/j.jtice.2013.09.012>
25. Nicomel NR, Leus K, Folens K, Van Der Voort P, Du Laing G (2015) Technologies for arsenic removal from water: current status and future perspectives. *Int J Environ Res Public Health* 13(1):1–24. <https://doi.org/10.3390/ijerph13010062>
26. Olatunji MA, Khandaker MU, Amin YM, Mahmud HNME (2016) Cadmium-109 radioisotope adsorption onto polypyrrole coated sawdust of dryobalanops aromatic: kinetics and adsorption isotherms modelling. *PLoS ONE* 11(10):1–14. <https://doi.org/10.1371/journal.pone.0164119>
27. Palma-Lara I, Martínez-Castillo M, Quintana-Pérez JC, Arellano-Mendoza MG, Tamay-Cach F, Valenzuela-Limón OL, García-Montalvo EA, Hernández-Zavala A (2020) Arsenic exposure: a public health problem leading to several cancers. *Regul Toxicol Pharmacol* 110(November 2019):104539. <https://doi.org/10.1016/j.yrtph.2019.104539>
28. Peng H, Gao P, Chu G, Pan B, Peng J, Xing B (2017) Enhanced adsorption of Cu(II) and Cd(II) by phosphoric acid-modified biochars. *Environ Pollut* 229:846–853. <https://doi.org/10.1016/j.envpol.2017.07.004>
29. Qiu B, Tao X, Wang H, Li W, Ding X, Chu H (2021) Biochar as a low-cost adsorbent for aqueous heavy metal removal: a review. *J Anal Appl Pyrol* 155(December 2020). <https://doi.org/10.1016/j.jaap.2021.105081>
30. Qiu H, Lv L, Pan BC, Zhang QJ, Zhang WM, Zhang QX (2009) Critical review in adsorption kinetic models. *J Zhejiang Univ Sci A* 10(5):716–724. <https://doi.org/10.1631/jzus.A0820524>
31. Rath BS, Kumar PS (2021) A review on sources, identification and treatment strategies for the removal of toxic Arsenic from water system. *J Hazard Mater* 418(June):126299. <https://doi.org/10.1016/j.jhazmat.2021.126299>

32. Singh P, Borthakur A, Singh R, Bhadouria R, Singh VK, Devi P (2021) A critical review on the research trends and emerging technologies for arsenic decontamination from water. *Groundwater Sustain Dev* 14(May):100607. <https://doi.org/10.1016/j.gsd.2021.100607>
33. Upadhyay MK, Shukla A, Yadav P, Srivastava S (2019) A review of arsenic in crops, vegetables, animals and food products. *Food Chem* 276(October 2018):608–618. <https://doi.org/10.1016/j.foodchem.2018.10.069>
34. Vandenbruwane J, De Neve S, Qualls RG, Sleutel S, Hofman G (2007) Comparison of different isotherm models for dissolved organic carbon (DOC) and nitrogen (DON) sorption to mineral soil. *Geoderma* 139(1–2):144–153. <https://doi.org/10.1016/j.geoderma.2007.01.012>
35. Vardhan KH, Kumar PS, Panda RC (2019) A review on heavy metal pollution, toxicity and remedial measures: current trends and future perspectives. *J Mol Liquids* 290:111197. <https://doi.org/10.1016/j.molliq.2019.111197>
36. Wang J, Guo X (2020) Adsorption kinetic models: physical meanings, applications, and solving methods. *J Hazard Mater* 390(November 2019). <https://doi.org/10.1016/j.jhazmat.2020.122156>
37. Wang S, Gao B, Li Y (2016a) Enhanced arsenic removal by biochar modified with nickel (Ni) and manganese (Mn) oxyhydroxides. *J Ind Eng Chem* 37:361–365. <https://doi.org/10.1016/j.jiec.2016.03.048>
38. Wang S, Gao B, Li Y, Zimmerman AR, Cao X (2016b) Sorption of arsenic onto Ni/Fe layered double hydroxide (LDH)-biochar composites. *RSC Adv* 6(22):17792–17799. <https://doi.org/10.1039/c5ra17490b>
39. Wang S, Gao B, Li Y, Zimmerman AR, Cao X (2016c) Sorption of arsenic onto Ni/Fe layered double hydroxide (LDH)-biochar composites. *RSC Adv* 6(22):17792–17799. <https://doi.org/10.1039/c5ra17490b>
40. Xu Y, Bai T, Yan Y, Ma K (2020) Influence of sodium hydroxide addition on characteristics and environmental risk of heavy metals in biochars derived from swine manure. *Waste Manage* 105:511–519. <https://doi.org/10.1016/j.wasman.2020.02.035>
41. Yin Q, Zhang B, Wang R, Zhao Z (2017) Biochar as an adsorbent for inorganic nitrogen and phosphorus removal from water: a review. *Environ Sci Pollut Res* 24(34):26297–26309. <https://doi.org/10.1007/s11356-017-0338-y>
42. Yoon K, Cho DW, Tsang DCW, Bolan N, Rinklebe J, Song H (2017) Fabrication of engineered biochar from paper mill sludge and its application into removal of arsenic and cadmium in acidic water. *Bioresour Technol* 246:69–75. <https://doi.org/10.1016/j.biortech.2017.07.020>
43. Yu Z, Zhou L, Huang Y, Song Z, Qiu W (2015) Effects of a manganese oxide-modified biochar composite on adsorption of arsenic in red soil. *J Environ Manage* 163:155–162. <https://doi.org/10.1016/j.jenvman.2015.08.020>
44. Zakhar R, Derco J, Čácho F (2018) An overview of main arsenic removal technologies. *Acta Chimica Slovaca* 11(2):107–113. <https://doi.org/10.2478/acs-2018-0016>
45. Zeng H, Zeng H, Zhang H, Shahab A, Zhang K, Lu Y, Nabi I, Naseem F, Ullah H (2021) Efficient adsorption of Cr (VI) from aqueous environments by phosphoric acid activated eucalyptus biochar. *J Clean Prod* 286:124964. <https://doi.org/10.1016/j.jclepro.2020.124964>
46. Zhao L, Zheng W, Mašek O, Chen X, Gu B, Sharma BK, Cao X (2017) Roles of phosphoric acid in biochar formation: synchronously improving carbon retention and sorption capacity. *J Environ Qual* 46(2):393–401. <https://doi.org/10.2134/jeq2016.09.0344>
47. Zhao N, Li B, Huang H, Lv X, Zhang M, Cao L (2020) Modification of kelp and sludge biochar by TMT-102 and NaOH for cadmium adsorption. *J Taiwan Inst Chem Eng* 116:101–111. <https://doi.org/10.1016/j.jtice.2020.10.036>
48. Zoroufchi Benis K, Soltan J, McPhedran KN (2021) Electrochemically modified adsorbents for treatment of aqueous arsenic: pore diffusion in modified biomass vs. biochar. *Chem Eng J* 423(April):130061. <https://doi.org/10.1016/j.cej.2021.130061>

# Evaluation of Performance of Pilot-Scale Engineered Permeable Bio-barriers for Removal of Nitrogenous Compounds from Waters Contaminated with Manure Slurry



Ali Ekhlasi Nia, Kharazm Khaledi, Bernardo Predicala, Terry Fonstad, and Mehdi Nemati

**Abstract** Agricultural and livestock run off are major sources of contamination of surface and ground waters by nitrogenous compounds, resulting in eutrophication and disruption of the ecosystem in these water bodies. Ammonium fertilizers and urea can be transformed into nitrate, which is highly soluble in water and can leach out from the soils and cause water pollution. Nitrate and nitrite have been known to cause health issues and thus should be prevented from entering water resources. The present study evaluated the effectiveness of pilot-scale permeable bio-barriers for the removal of nitrate, nitrite, and a mixture of nitrate and nitrite from contaminated water. The designed bio-barriers that are operated in continuous flow mode are comprised of parallel channels that can be run independently. Each channel has a bio-reactive compartment that houses an inexpensive carbon source and the biofilm. As part of this study, following the development of biofilm in the bio-reactive compartment, bio-barrier channels were fed with an aqueous medium containing either nitrite, nitrate, or both, to assess the impacts of nitrogenous compounds concentrations ( $10\text{--}250\text{ mg L}^{-1}$ ) and the presence of manure slurry supernatant on the removal of nitrogenous compounds at an average hydraulic retention time of 20 days. Apart from evaluating the performance of the bio-barrier in removal of nitrogenous compounds, composition of microbial community developed under various conditions has been

---

A. E. Nia (✉) · K. Khaledi · B. Predicala · M. Nemati  
Department of Chemical and Biological Engineering, University of Saskatchewan, Saskatoon, Canada  
e-mail: [Ali.Nia@usask.ca](mailto:Ali.Nia@usask.ca)

M. Nemati  
e-mail: [Mehdi.Nemati@usask.ca](mailto:Mehdi.Nemati@usask.ca)

B. Predicala  
Prairie Swine Centre Inc., Saskatoon, SK, Canada

T. Fonstad  
Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, Canada

examined. Results obtained to date show the effectiveness of engineered bio-barriers and their potential use as a solution for removing nitrite and nitrate from agricultural and livestock run off, hence preventing the contamination of ground waters by these compounds.

**Keywords** Pilot-scale engineered permeable bio-barriers • Removal of nitrogenous compounds • Waters contaminated

## 1 Introduction

Focus on maximizing the yield of various crops has resulted in intensive agricultural activities, which led to increased use of manure and chemical fertilizers. Consequently, nitrogen-containing chemicals could accumulate in the root zone of crop lands and subsequently leach to sub-surface drainage systems. Agricultural drainage waters have long been considered an important factor in polluting water bodies with nitrogenous compounds such as nitrates and ammonium [1]. Apart from agricultural activities, animal wastes are also a major contributor to the release of nitrogenous compounds in water bodies through inadvertent spills, improper discharge, and seepage of wastes in lagoons of dairies and open feedlots [2]. While manure is a source of many valuable nutrients for crop production, the excess nutrients can become a source of pollution by promoting extreme plant and algae growth in water bodies. After the death of these plants, their bacterial decomposition consumes dissolved oxygen from the water, thus severely affecting the health and sustainability of aquatic ecosystems through eutrophication [2, 3].

Among the nitrogenous compounds, ammonia, nitrate, and nitrite are of greater concern due to their higher availability in aquatic environments and potentially harmful effects on human health and the environment [4]. Ammonia is widely used in agricultural production to provide the required nitrogen for plant growth, thus making it one of the most common pollutants discharged into water bodies. Under oxidizing conditions, ammonia can be transformed into nitrate and nitrite, which are highly soluble in water and can leach out from the soils. Nitrate and nitrite have been known to cause health problems, such as inducing blue-baby syndrome in infants and forming potentially carcinogenic N-nitrosamines [5].

Various treatment technologies, mainly *ex situ*, have been developed to prevent these nitrogenous compounds from entering water bodies. Among these, biological treatments are commonly used as they can remove nitrogen compounds by utilizing bacteria in the absence of harsh physicochemical conditions, which makes this process relatively inexpensive compared to other methods [6]. Biological permeable reactive barriers (PRBs), also known as bio-barriers, are novel biological technologies capable of *in situ* remediation of waters contaminated by nitrogenous compounds. Bio-barriers are passive systems which involve positioning a permeable bio-reactive media to intercept the contaminated water plume and transforming the pollutants into less harmful compounds or adsorb them into the reactive media

[7]. Since most of the contaminant removal (i.e., denitrification) happens in the bio-reactive media, it is important to choose the proper materials to provide the conditions favorable to denitrification process. Proper media provides an optimum environment for growth and bacterial activity and has a hydraulic conductivity higher than the adjacent aquifer devoid of contaminated plume [8]. Wood in various forms is commonly used as bio-reactive media, especially in the denitrification processes, as they can serve as a source of carbon and energy for bacteria, while providing a solid matrix for formation of biofilm [9]. In this study, a pilot-scale bio-barrier has been designed and evaluated for the potential use of this treatment technology in large scale in situ removal of nitrate and nitrite from livestock and agricultural runoff.

## 2 Methodology

### 2.1 Materials

Spruce wood chips (Sta-Green, 0.75-in) were used as media. Wood chips served as a matrix for the development of biofilm in the denitrifying bio-barrier, as well as an inexpensive carbon and energy source. Manure slurry was collected from the manure pit of a production room at Prairie Swine Centre Inc., Saskatoon, SK. The collected manure was passed through a commercial wire mesh (pore size 120  $\mu\text{m}$ ) and kept at 4 °C in an environment chamber at the University of Saskatchewan. Modified McKinney's medium was used for culture maintenance and in denitrification experiments. The medium contained the following ingredients:  $\text{KH}_2\text{PO}_4$  (840  $\text{mg L}^{-1}$ ),  $\text{K}_2\text{HPO}_4$  (750  $\text{mg L}^{-1}$ ),  $(\text{NH}_4)_2\text{SO}_4$  (474  $\text{mg L}^{-1}$ ),  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$  (60  $\text{mg L}^{-1}$ ),  $\text{NaCl}$  (60  $\text{mg L}^{-1}$ ),  $\text{CaCl}_2$  (60  $\text{mg L}^{-1}$ ) and  $\text{Fe}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$  (20  $\text{mg L}^{-1}$ ).

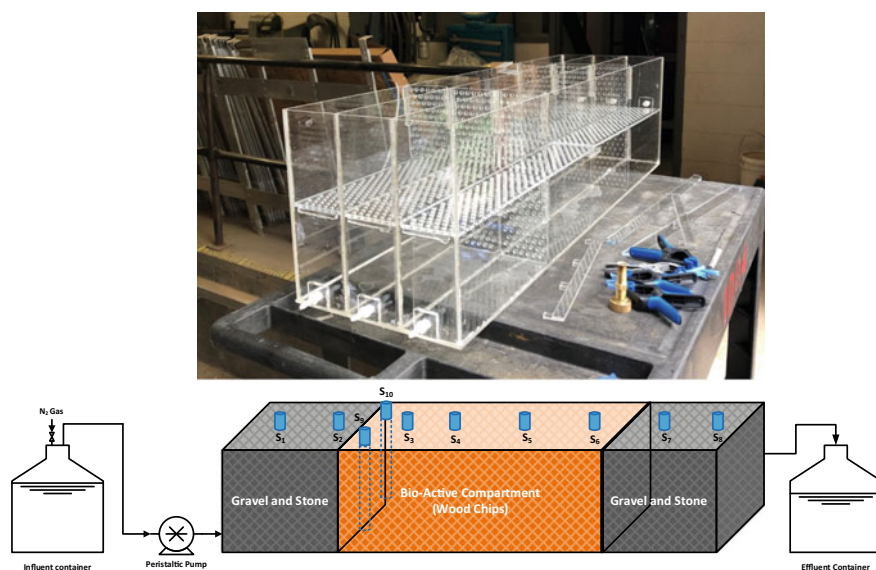
### 2.2 Denitrifying Culture

The denitrifying culture was derived from an anaerobic culture originally developed for the removal of naphthenic acid under denitrifying conditions [9]. The culture was adapted to nitrite and nitrate removal using modified McKinney's medium with 300  $\text{mg L}^{-1}$  of acetate and 620  $\text{mg L}^{-1}$  of either nitrite, nitrate, or both (310  $\text{mg L}^{-1}$  each), with the original culture developed with naphthenic acid used as the initial inoculum. The developed culture was then used as inoculum in subsequent sub-culturing which was carried out regularly.



### 2.3 Bio-barrier Experimental Setup

Figure 1 presents a photo of the designed pilot-scale bio-barrier and a schematic diagram of the experimental setup used to study the removal of nitrite, nitrate, and their mixtures in a system simulating a field-scale bio-barrier. The designed bio-barrier is comprised of three independent rectangular cuboid channels (without any interconnection) made of plexiglass, with each individual channel partitioned into upper and lower layers, using a perforated plexiglass panel. Additionally, each layer is divided longitudinally into three flow compartments, allowing various configurations of the entire system in terms of channel width and depth, as well as the operation of the three longitudinal flow channels independently and under various conditions. The lower layers of each channel that were used in the current study have a liquid capacity of 9 L (length:110 cm, width:30 cm and depth:15 cm), with inlet and outlet ports fitted to the front and end compartments filled with stone and gravel, and the middle compartment being filled with wood chips (bio-reactive chamber). The use of Spruce wood chips and a mixture of stone and gravel provided the system with porosities of 35% for the bio-reactive chamber and 55% for the front and end compartments. The gravel and stone were used to fill the part of channel that was not used as bio-reactive region. It should be pointed out that the upper layer of each channel is packed with soil on top of thin foam sheets in order to create an anoxic condition in the lower layer. In each channel, the feed is introduced into the gravel and stone compartment



**Fig. 1** Photo of the designed pilot-scale bio-barrier (top) and a schematic diagram of the experimental setup (one channel) used to evaluate anoxic denitrification

in the bottom layer using a peristaltic pump. The feed then passes through the bio-reactive chamber filled with wood chips and finally exits the system through the last compartment filled with gravel and stone. In each channel, ten sampling ports (S1-S10 in Fig. 1) have been devised in the longitudinal direction, with two ports (S9-S10) allowing sampling near the bottom of the bio-reactive chamber.

## 2.4 Biofilm Development in Bio-barrier

The biofilm development in all three channels was carried out initially batch-wise and then contiguously at a low flow rate of feed. These channels were designated for the removal of nitrate (channel 1), nitrite (channel 2), and simultaneous removal of nitrate and nitrite (channel 3). During the batch-wise operation, each channel was filled with 9 L sterilized modified McKinney's medium containing  $300 \text{ mg L}^{-1}$  acetate with  $200 \text{ mg L}^{-1}$  nitrite or nitrate, or a mixture of  $100 \text{ mg L}^{-1}$  of each nitrate and nitrite. The medium was purged with nitrogen gas for 30 min to remove dissolved oxygen from the solution prior to addition to the bio-barrier. The bio-reactive chamber of each channel was then inoculated with 1.35 L inoculum (15% v/v), then sampled and analyzed regularly until complete removal of nitrite and nitrate. Once the concentrations of nitrogenous compounds reached zero, a concentrated solution of nitrite, or nitrate, or their mixture and acetate was injected into the bio-reactive section to achieve final concentrations of 200 and 300 for nitrogenous compounds and acetate, respectively. After five successful batch cycles, the system was switched to continuous mode and fed with modified McKinney's medium containing nitrate or nitrite or their mixtures at a flow rate of  $10 \text{ mL h}^{-1}$  for 12 weeks. Concentration of nitrogenous compound in each channel was progressively increased from 10 to  $500 \text{ mg L}^{-1}$  to promote the formation of denitrifying cultures. Figure 2 shows the bio-barrier during the batch-wise development of biofilm.



**Fig. 2** Bio-barrier during the batch-wise development of biofilm

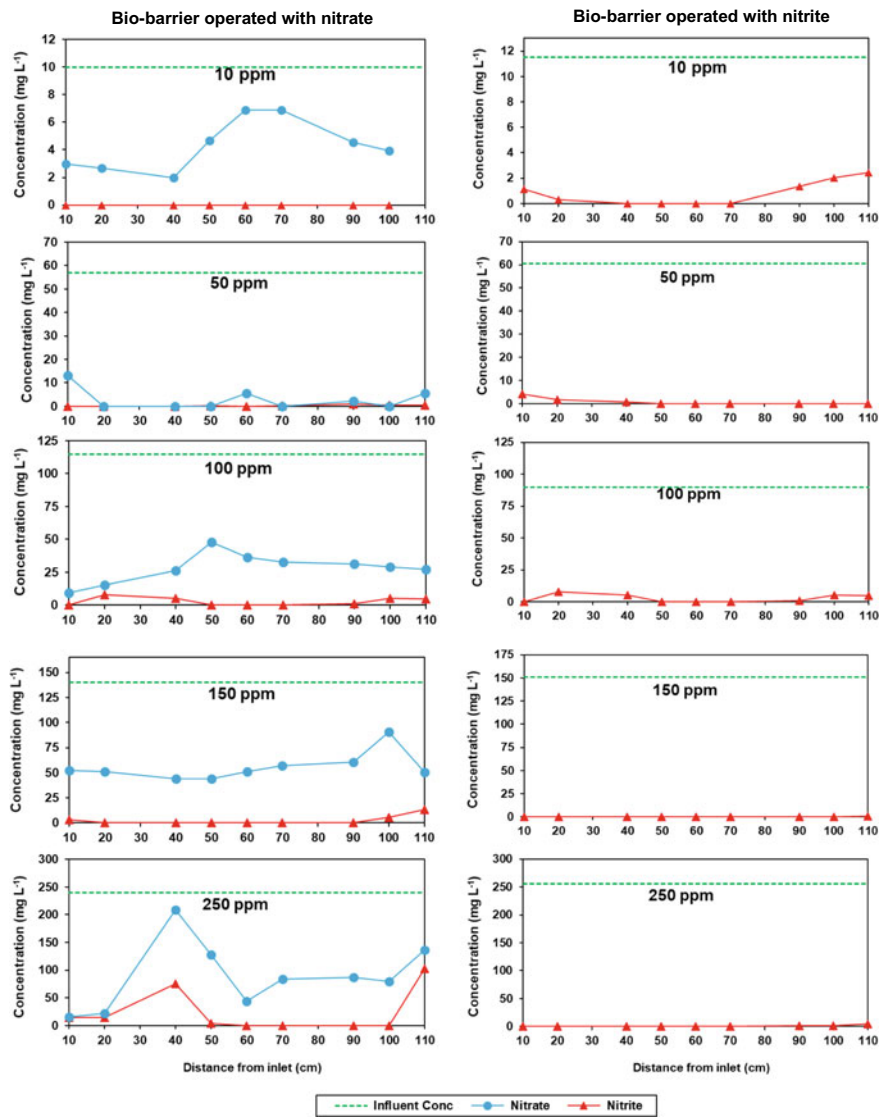
## 2.5 Experimental Procedure

After the development of biofilm in the bio-reactive chamber, the effects of nitrite and nitrate concentrations in the feed and their loading rates in presence of various levels of manure supernatant (0–10%) on the nitrite and nitrate removal percentages and rates were investigated. In each channel of bio-barrier, removal performances under nitrite and/or nitrate concentrations of 10, 50, 100, 150, and 250 mg L<sup>-1</sup> were evaluated. Equal nitrate and nitrite concentrations of 5, 25, 50, 75, and 125 mg L<sup>-1</sup> were added to the medium for the mixture experiments. These experiments were conducted in the absence of manure supernatant and in the presence of 2.5, 5, and 10% manure supernatant. All the experiments were conducted at room temperature (24 ± 2 °C) and a flow rate of 20 mL h<sup>-1</sup> that represented a linear velocity corresponding to those observed for ground water. The channels were regularly sampled and monitored for concentrations of nitrate and nitrite. Samples were taken from the sampling ports, filtered with 0.45 µm nylon syringe filters and analyzed for nitrite and nitrate concentrations by an ion chromatograph (ICS-2500, Dionex Corporation, Sunnyvale, USA). Analysis of developed microbial communities developed with nitrate, nitrite, and their mixture is also being conducted.

## 3 Results and Discussion

### 3.1 Removal of Nitrate and Nitrite in the Bio-barrier

Figure 3, a typical set of results, shows the effect of nitrite and nitrate initial concentrations on the removal efficiency of each compound during their treatment in the bio-barriers. Profiles of nitrate and nitrite concentrations are depicted as a function of distances from the inlet of the bio-barrier. In this set of experiment, the bio-barriers were run at a flow rate of 20 mL h<sup>-1</sup> and fed with modified McKinney's medium containing 2.5% manure supernatant with nitrate (channel 1) or nitrite (channel 2) at concentrations of 10 mg L<sup>-1</sup> (a), 50 mg L<sup>-1</sup> (b), 100 mg L<sup>-1</sup> (c), 150 mg L<sup>-1</sup> (c), and 250 mg L<sup>-1</sup> (d). Each data point is the average value of two to five measurements at steady state conditions in the bio-barrier, with the standard deviation of each data point being in the range of 0 to 18.79%. Steady state was assumed when there was less than 10% variation in concentration of nitrate in the effluent. In general, the nitrate concentrations decreased along the length of bio-barrier (blue line) and were substantially lower than that in the influent (green line). This was the case for all influent concentrations, indicating the capability of bio-barrier in the removal of nitrate. Nitrite was detected in the nitrate-removing channel, though at a low level



**Fig. 3** Profiles of nitrate and nitrite concentrations in the bio-barriers fed with either nitrate (left) or nitrite (right). Influent contained 2.5% manure supernatant and: **a** 10  $\text{mg L}^{-1}$ , **b** 50  $\text{mg L}^{-1}$ , **c** 100  $\text{mg L}^{-1}$ , **d** 150  $\text{mg L}^{-1}$ , and **e** 250  $\text{mg L}^{-1}$  nitrate or nitrite

(red line) which indicates that the nitrate was converted to nitrite and then other nitrogenous compounds, likely nitrogen. The nitrite concentrations in the nitrate-removing channel were found to follow the same trends as nitrate (i.e., decreasing trend along the bio-barrier length), especially at higher concentrations. In general, the nitrate removal efficiencies decreased with the increase of nitrate initial concentration. As an example, the nitrate removal efficiency was about 88% with 50 mg L<sup>-1</sup> nitrate, while a removal of 44% was observed with 250 mg L<sup>-1</sup> nitrate. Increase of loading rates from 0.53 to 11.65 mg L<sup>-1</sup> day<sup>-1</sup>, as a result of increase in the influent concentration, enhanced the nitrate removal rate from 0.33 mg L<sup>-1</sup> day<sup>-1</sup> to a maximum value of 5.48 mg L<sup>-1</sup> day<sup>-1</sup>. The removal of nitrite in the bio-barrier was more effective when compared to nitrate. In fact, a close look at nitrite concentration profiles shows that nitrite residual concentrations were lower than 5 mg L<sup>-1</sup> along the bio-barrier, regardless of nitrite concentration in the influent. This demonstrates the high capability of the bio-barrier in nitrite removal. The nitrite removal efficiencies were found to slightly decrease from 100% at 50 mg L<sup>-1</sup> to about 98% at 250 mg L<sup>-1</sup> nitrite. For nitrite, an increase in loading rates from 0.53 to 11.65 mg L<sup>-1</sup> day<sup>-1</sup>, enhanced the removal rate of nitrite from 0.46 to 11.43 mg L<sup>-1</sup> day<sup>-1</sup>, which is significantly higher than maximum removal rate obtained with nitrate.

## 4 Conclusion

Nitrate and nitrite removal from contaminated medium containing 2.5% manure supernatant and 10, 50, 100, 150, or 250 mg L<sup>-1</sup> of nitrogenous compounds was evaluated in a pilot-scale bio-barrier. The bio-barriers performance revealed the capability of this system for large scale removal of nitrate and nitrite. The results indicated that the removal efficiency was lower at high nitrate concentrations (61% at 10 mg L<sup>-1</sup> versus 44% at 250 mg L<sup>-1</sup>) but nitrite concentration did not have a marked impact. The removal percentage and rate in case of nitrite were higher than nitrate. These findings are based on the results obtained to date and work on the effects of manure slurry level, composition of microbial communities developed under various conditions, as well as simultaneous removal of nitrate and nitrite are ongoing.

**Acknowledgements** This work was made possible by an Agriculture Development Fund grant from the Saskatchewan Ministry of Agriculture. Technical assistance of D. Jayasinghe and R. Prokopishyn from the Department of Chemical and Biological Engineering, University of Saskatchewan, is gratefully acknowledged.

## References

1. Xia Y, Zhang M, Tsang DC, Geng N, Lu D, Zhu L, Ok YS (2020) Recent advances in control technologies for non-point source pollution with nitrogen and phosphorous from agricultural runoff: current practices and future prospects. *Appl Biol Chem* 63(1):1–13
2. Karr JD, Showers WJ, Gilliam JW, Andres AS (2001) Tracing nitrate transport and environmental impact from intensive swine farming using delta nitrogen-15. *J Environ Qual* 30(4):1163–1175
3. Salameh E, Harahsheh S (2010) Eutrophication processes in arid climates. *Eutrophication: Causes, Consequences and Control*. Springer, Dordrecht, Netherlands, pp 69–90
4. Lopez-Ponnada EV, Lynn TJ, Peterson M, Ergas SJ, Mihelcic JR (2017) Application of denitrifying wood chip bioreactors for management of residential non-point sources of nitrogen. *J Biol Eng* 11(1):1–14
5. Bhatnagar A, Sillanpää M (2011) A review of emerging adsorbents for nitrate removal from water. *Chem Eng J* 168(2):493–504
6. Karri RR, Sahu JN, Chimmiri V (2018) Critical review of abatement of ammonia from wastewater. *J Mol Liq* 261:21–31
7. Obiri-Nyarko F, Grajales-Mesa SJ, Malina G (2014) An overview of permeable reactive barriers for in situ sustainable groundwater remediation. *Chemosphere* 111:243–259
8. Careghini A, Saponaro S, Sezenna E (2013) Biobarriers for groundwater treatment: a review. *Water Sci Technol* 67(3):453–468
9. Nordström A, Herbert RB (2018) Determination of major biogeochemical processes in a denitrifying woodchip bioreactor for treating mine drainage. *Ecol Eng* 110:54–66
10. Valdes Labrada GM, Nemati M (2018) Biodegradation of surrogate naphthenic acids and electricity generation in microbial fuel cells: bioelectrochemical and microbial characterizations. *Bioprocess Biosyst Eng* 41(11):1635–1649

# Wetland Water Discharge Remediation Using On-Site Non-woven Geotextile Filtration



Antonio C. Pereira, Dileep Palakkeel Veetil, Mathew Cotton,  
Catherine N. Mulligan, and Sam Bhat

**Abstract** Lake Johanne, a shallow mesotrophic lake located in the *Sainte-Anne-des-Lacs* municipality in Quebec, is receiving water from a wetland discharge. This wetland water could be possibly contributing to this lake's ageing. Thus, as an initial investigation, an on-site potential remediation of this wetland water discharge using a non-woven geotextile filtration technique is proposed. The method was based on a tank located near the lakeshore with a floating geotextile filtration system in continuous mode (i.e., 0.5-day retention time). On deployment, wetland water was pumped into the tank using a calibrated peristaltic pump (inlet), filtered by a selected set of non-woven geotextiles with distinct apparent opening sizes, and then returned to the lake (outlet) by gravity. This experiment ran throughout the summer and mid-fall of 2021. Samples from both the inlet and tank were taken every 2 days, and filter layers were changed upon clogging. The objective of this study was to improve the quality of this discharge before entering the lake by filtration. The water quality parameters monitored during the experiment from both inlet and tank samples were total phosphorus, total nitrogen, nitrate, chemical oxygen demand, total suspended solids, particle size, and turbidity. The non-woven geotextile filters removed nutrients, organic matter, and suspended particles at levels of 32%, 17%, and 29% for total phosphorus, COD, and turbidity removals, respectively, when the tank inlet and outlet were compared during the entire experiment. In addition, it was proven that the proposed method is easy to install, deploy, and quickly adapt to water quality changes. Even though the results have pointed out that additional investigation is needed, the treatment improved the quality of a portion of this wetland water. Thus, the feasibility of treatment has been shown for this possible remediation.

**Keywords** Wetland water discharge remediation · On-site non-woven geotextile filtration

---

A. C. Pereira · D. P. Veetil · M. Cotton · C. N. Mulligan (✉) · S. Bhat  
Concordia University, Montreal, Canada  
e-mail: [mulligan@civil.concordia.ca](mailto:mulligan@civil.concordia.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_67](https://doi.org/10.1007/978-3-031-34593-7_67)

1043

## 1 Introduction

Ageing of waters is intensified around the globe due to climate change scenarios synergically associated with well-known human-made variables such as land modifications [1, 2], watershed pollution [3, 4], and consumption behavior [5, 6]. Triggered by organic matter and nutrient input increase from allochthonous sources (i.e., catchment-derived sources also known as external loads) and autochthonous sources (i.e., internal loads), this aging is hastening trophic level changes and increasing eutrophication scenarios in water resources.

In these possible eutrophication scenarios, benthic primary productivity (macrophytes and periphyton) will be shifted to pelagic primary production (phytoplankton) [7], causing plankton biomass to increase. Generally, this will cause a shift to a cyanobacteria-dominated phytoplankton community [8] in the water column. Therefore, with the increase of cyanobacteria-dominated phytoplankton communities in waters, issues of recreational and drinking advisories due to harmful toxin production [9, 10], as well as water anoxia, obnoxious scum and smells, will occur more frequently.

In this view, for definition purposes, allochthonous sources are conveyed by watershed runoff or incomplete treated effluent discharge. Autochthonous sources, on the other hand, are associated with past catchment-derived emissions, internal releases from sediment disturbances (e.g., natural summer lake turnover), and/or organic matter decomposition into the water. It should be emphasized that whereas external nutrient loads are predominantly in the particulate form (i.e., phosphorus), and not directly available for plankton communities, internal loads are mostly in the dissolved form and openly available for cyanobacteria growth (i.e., blue-green algae) [11]. Both of the sources are composed of macronutrients like nitrogen and phosphorus, organic matter characterized by carbon components, and some micronutrients in low concentrations. They are extremely important for further plankton development.

On the elements presented, phosphorus could be highlighted as a limiting nutrient, which can trigger the uncontrolled growth of photosynthetic aquatic microorganisms in natural waters [12]. Therefore, remediation methods to inhibit possible trophic changes or attenuate eutrophication scenarios on water resources are focused on the removal or reduction of this element. In the specialized literature, the following methods are highlighted as go-to approaches: sediment dredging, hypolimnetic water withdrawal, sediment capping by inert elements, and phosphorus inactivation in the water column and bottom sediments via chemical addition. Those practices are highly technical, drastic, energy-demanding, expensive, and could be harmful to water organisms and end-water users.

Conversely, methods that take advantage of a lake ecosystem's natural response to changes made within it, called ecological methods, are in development and study. One of those methods is been investigated by our research group. Through the use of in situ/on-site filtration units with geotextiles as filter membranes, removal of particulate nutrients, organic matter, and total suspended solids are being achieved [13–17]. In other words, the lake/pond water geotextile filtration has been extensively



investigated, indicating the potential of this technique as flexible, responsive, and environmentally-safe remediation that can be adjusted to other surface water types.

To continue this investigation, Lake Johanne (LJ) has been chosen as the study area. A shallow mesotrophic lake, located in *Sainte-Anne-des-Lacs* Quebec municipality, is being possibly degraded and aged throughout the slow increase of organic matter and nutrients on it. Our studies have been emphasizing that the wetland discharge, located on the northeast lake corner, is one of the possible ageing factors [14–16]. This discharge presents higher particle size, organic matter, and nutrient content than the overall lake quality. On this understanding, the on-site non-woven geotextile filtration treatment, for this lake water remediation, was applied in 2021. However, a different methodology was taken, as the system inlet was modified from the lakeshore to the wetland discharge. Therefore, the objectives of this present study are to assess the lake water quality and evaluate the usefulness of the geotextile on-site filtration for nutrient and suspended solid removal on the discharge in a continuous experiment with a retention time of 0.5 day.

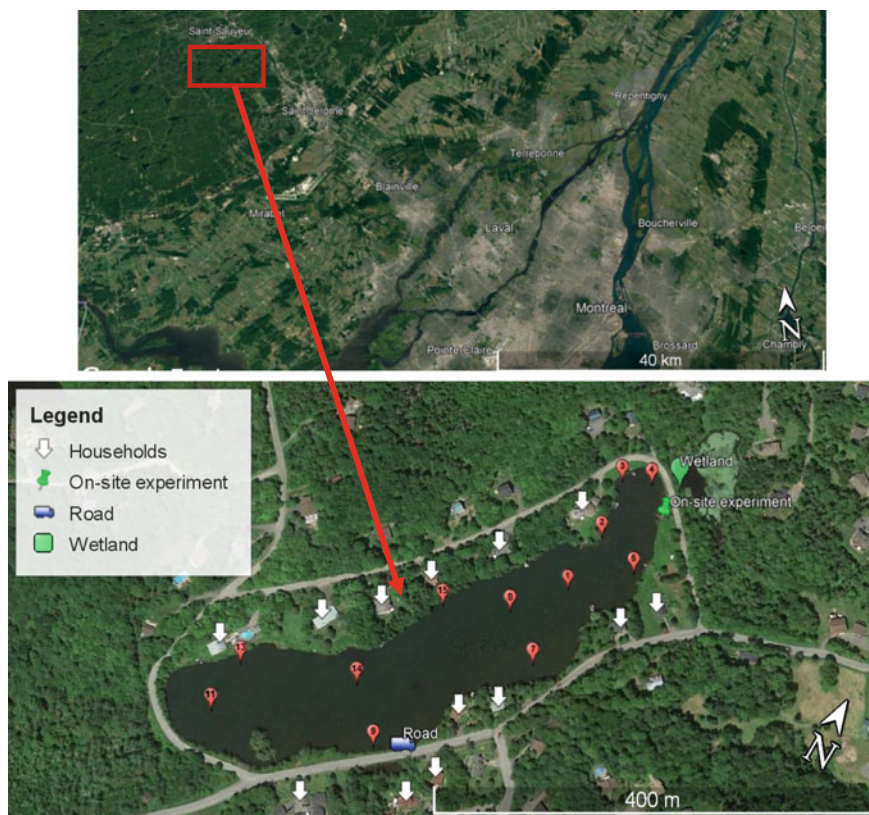
## 2 Materials and Methods

### 2.1 Influent Study Area

Lake Johanne (45°50'23"N; 74°08'19"W), a shallow mesotrophic lake in *Sainte-Anne des-Lacs* municipality, Quebec (situated around 75 km north of downtown Montreal near to *Laurentian* Mountains), was chosen for the study area. Located in the *Masse* watershed (i.e., one of the contributors to the *Rivière du Nord*), which is mainly composed of vegetation and few households. This mesotrophic Head Lake had passed through some recreational advisories in the past associated with excessive algae/cyanobacteria growth.

This lake is an artificial water body with maximum and average depths of 3.5 and 1.7 m and approximately surface area and water volume of 44,910 m<sup>2</sup> and 74,900 m<sup>3</sup>, respectively (ABVLacs Org., 2021). Also, the main sources of water renewal are associated with wetland discharge, precipitation, surface runoff, and snow melting and its renewal time is 0.48 years. Figure 1 shows Lake Johanne's location with sampling stations, wetland, and other details.

Related to the nutrients, external phosphorus loads occurrence on this lake comprises runoff from a nearby road (i.e., near Station 9) and the wooded land around it, as well as plant decay in the water. Internal loads are, however, attributed to phosphorus sediment release, wetland discharge at the lake inlet, and possibly diffuse pollution from septic tanks. Also, the sediment phosphorus content in this lake is not homogenous and according to [18] ranges from 1186 to 1451 mg/kg.



**Fig. 1** Lake Johanne location with sampling stations as shown by the numbered indicators. White arrows indicate the nearest lakeshore household

## 2.2 Geotextile Filtration Setup—On-Site Experiment

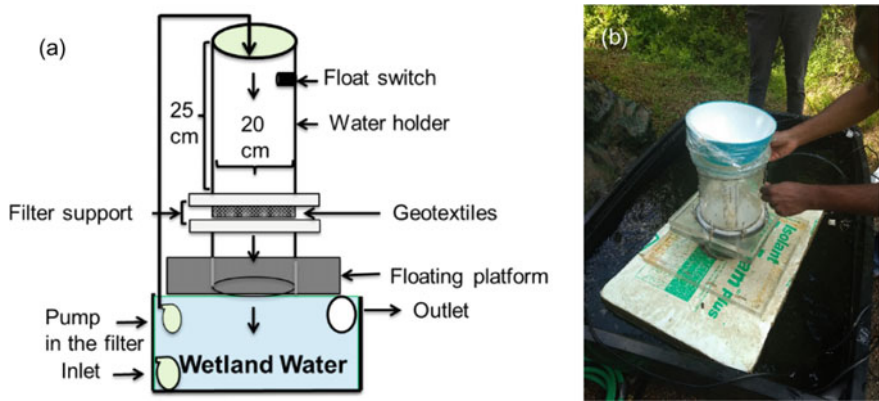
For the on-site geotextile filtering deployment, a 543 L plastic tank (35.6 cm in height and 97.8 cm in width) was installed near the lakeshore. Different from previous studies where the system inlet was located near station 2, this was altered to be near the wetland discharge on the lake's upper northeast corner, almost 50 m from the tank, as presented in Fig. 2. For the experiment start, a submersible pump was used to fill up the tank until 300 L. After that, a peristaltic calibrated pump was used to continuously feed the tank with the wetland water (i.e., considered the system input), with a retention time of 0.5 days throughout the 72 experiment days. Additionally, the treated wetland discharge water was returned to the lake using gravity (i.e., system output). For system protection from external influences, a tarpaulin was used as a cover.

Regarding the floating filtration unit, where the geotextile layers were placed, this was made as a cylindrical Plexiglas column with an internal diameter of 20 cm



**Fig. 2** Wetland discharge location and system placement on the lakeshore

and a height of 25 cm. In addition to a square-shaped base, which serves as a filter holder and has a circular hole in the center with the exact filtration column diameter (20 cm) shown in Fig. 3a, capable to sustain a maximum hydraulic head of 18 cm above it. The filtration unit with the sandwiched geotextile combination was mounted on a polystyrene foam sheet with a center 20 cm circular hole. (Fig. 3b). With this combination, the filtration unit was able to float over the tank water and allow the filtered water to return to the tank. Every 2–3 days, the experiment was evaluated for geotextile clogging, pumping status, output and input clogging, sampling, and any other external impacts.



**Fig. 3** a Schematic of the on-site filtration setup b on-site filtration unit deployed

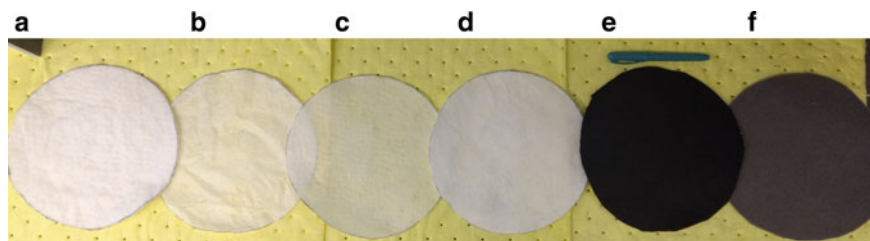
### 2.3 Filter Media

Five custom-made geotextiles were employed as filter media to capture suspended particles and particulate nutrients in this on-site remediation of this mesotrophic lake water. The filter selection and combination were based on earlier on-site research conducted in 2017–2018, which were validated by project data from 2019 and 2020 [14–16]. Titan Environmental Containment manufactured the geotextiles based on the particle size of 90 percent of the particles in this lake water (D90). Table 1 shows the parameters of the six non-woven geotextile membranes applied (TE-GTX300, TE-GTT100, TE-GTT120, TE-GTT200, TE-GTN350B, and TE-GTP250), and Fig. 4, the non-woven geotextiles before the filtration process in the apparent opening size (AOS) descending order.

The four geotextile membranes (i.e., from 4b to 4f) were comprised of polypropylene (PP) fibers, except for the non-woven geotextile TE-GTX300 (i.e., 4a), which was produced using PET fibers. They are all exceedingly flexible, with dimensionally stable fabrics that are ideal for use as filtering membranes. For experiment deployment, these geotextile layers were cut to 22 cm diameter and arranged in decreasing order of their AOS (110, 100, 90, 70, 65, and 60  $\mu\text{m}$ ) using one layer of each. The

**Table 1** Non-woven geotextile characteristics used in this study

Filters	Material	Apparent opening size (AOS) ( $\mu\text{m}$ )	Flow rate ( $\text{L}/\text{s}/\text{m}^2$ )	Permittivity ( $\text{sec}^{-1}$ )	Mass per unit area ( $\text{g}/\text{m}^2$ )	Thickness (mm)
TE-GTX300	Polyester	110	65	1.62	300	3.1
TE-GTT100	Polypropylene	100	75	–	150	0.8
TE-GTT120	Polypropylene	90	70	–	120	0.8
TE-GTT200	Polypropylene	70	50	–	200	1.5
TE-GTN350B	Polypropylene	65	45	0.56	350	2.1
TE-GTP250	Polypropylene	60	41	0.83	200	1.7



**Fig. 4** Non-woven geotextiles before the filtration process according to the AOS: **a** 110  $\mu\text{m}$  **b** 100  $\mu\text{m}$  **c** 90  $\mu\text{m}$  **d** 70  $\mu\text{m}$  **e** 65  $\mu\text{m}$ , and **f** 60  $\mu\text{m}$

overall thickness of the five layers when combined was roughly 10.0 mm. The combination was changed on the deployment depending on how fast geotextile clogging was experienced. The ideal filter run was adjusted to every sampling day (i.e., every 2–3 days). The base geotextile layer combination was composed of AOS from 110 to 70  $\mu\text{m}$  arranged in descending order. The 65  $\mu\text{m}$  geotextile was added and removed from the combination when found necessary. Also when colloidal particles accumulation was noticed in the tank water, in the middle of the experiment, a further layer of 60  $\mu\text{m}$  was then added, making up from 4 to 6 layers as the used combination.

## 2.4 Water Quality Analysis

Every 2–3 days, water samples were collected from the tank and the inlet. Additionally, overall lake quality was assessed by sampling in the lake (St. 1, St. 4, St. 7, St. 9, and St. 11) from the summer to mid-fall of 2021. All samples were taken in 1L amber bottles (high-density polyethylene (HDPE)) and 50 mL sterilized polypropylene test tubes, kept at 4 °C in the dark before any physicochemical study, and acidified when necessary (i.e., in the case of phosphorus analysis) with analysis completed within 48 h.

Particle size distribution (PSA), turbidity, total suspended solids (TSS), total phosphorus (TP), total nitrogen (TN), nitrate ( $\text{NO}_3^-$ ), and Chemical Oxygen Demand (COD) were all measured in the water samples collected. For TSS, the specific APHA method (SM 2540D) was used technique, and turbidity was assessed using an Oakton turbidity meter. A laser diffraction particle analyzer (LA-960 Horiba laser particle size analyzer) was used to perform particle size analysis (PSA). Thus, Hach chemicals test kits were used to analyze TN (TNT 826, Method 10,208, persulfate digestion) and COD (TNT 820, Method 10,221, reactor digestion method).

Phosphorus was evaluated by elemental analysis performed by ICP-MS with a quadruple mass analyzer following partial acid-peroxide digestion ( $\text{HNO}_3\text{-H}_2\text{O}_2$ ) of water samples (USEPA 3050B). Nitrate, on the other hand, was measured using a Metrohm Ion Chromatography under isocratic conditions with a Metrosep A Supp 5 - 150/4.0 analytical column (150  $\times$  4 mm), suppressed conductivity detection, and 3.2 mM  $\text{Na}_2\text{CO}_3$ –1.0 mM  $\text{NaHCO}_3$  as the eluent. The injection volume was 100 L/ml, and the eluent flow rate was 7.0 ml/min.

**Table 2** Lake Johanne overall water quality from July-Sep, 2020 and July-Sep, 2021

Parameters	2020 <sup>a</sup>	2021 <sup>b</sup>
TP (µg/L)	15.1 ± 1.5	17.1 ± 2.5
COD (mg/L)	21.0 ± 2.5	25.0 ± 3.5
NO <sub>3</sub> <sup>-</sup> (mg/L)	0.4 ± 0.3	0.8 ± 0.7
TN (mg/L)	1.0 ± 0.7	1.3 ± 0.8
TSS (mg/L)	4.6 ± 1.1	5.2 ± 1.1

<sup>a</sup> Average of 7 samplings at 5 lake stations; <sup>b</sup> Average of 5 samplings at 5 lake stations

### 3 Results and Discussion

#### 3.1 Overall Lake Water Quality Assessment

The *Réseau de surveillance volontaire des lacs* [18] by the Minister of the Environment and the Fight Against Climate Change (MELCC) in Quebec has a specific trophic status designation for lakes. Lake Johanne (LJ) on this classification is classified as a mesotrophic lake (13–20 g/L) from the 2017 to 2020 reports. Our findings, when based only on total phosphorus content, pointed to the same classification. Also, the same report recommended the adoption of measures to limit nutrient sources to avoid further degradation and further loss of use. Table 2 shows the overall lake water quality for the years 2020 and 2021.

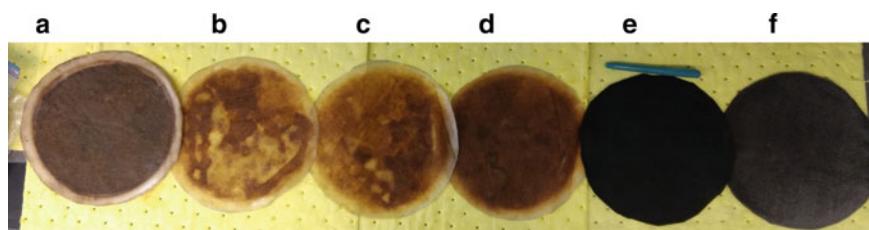
Particle size analysis from the overall lake quality was D90 in the range of 60–79 µm and the diameter of 50% particles (D50) under the 8–16 µm range over the 5 samplings of 2021. The COD concentration shown for this lake water is slightly higher than the average found in surface water of 20 mg/L, according to Chapman and Kimstach [19]. Additionally, total suspended solids (TSS), even though not too representative in the overall lake water quality, they have particulate nutrients associated with them and thus need to be attenuated.

#### 3.2 Filtration Deployment

The remediation experiment on this wetland discharge by the on-site floating geotextile filtration proposed was executed shortly after the start of summer until the mid-fall of 2021. More specifically from July 23, 2021, to October 2, 2021, a total of 72 experimental days were done. In this deployment, a retention time of 0.5 day (12 h) was maintained when possible, filtering a volume of 51 m<sup>3</sup> using 30 geotextile filter layer combinations (a total of 1.15 m<sup>2</sup> for each AOS). Thus, the cost related only to geotextile layers was 5.81 CAD.

Additionally, geotextile layer fouling was observed after some hours of the filter run and increased until it needed to be changed. It was observed that this has aided in





**Fig. 5** Non-woven geotextiles after the filtration process in the AOS order: **a** 110  $\mu\text{m}$  **b** 100  $\mu\text{m}$  **c** 90  $\mu\text{m}$  **d** 70  $\mu\text{m}$  **e** 65  $\mu\text{m}$ , and **f** 60  $\mu\text{m}$

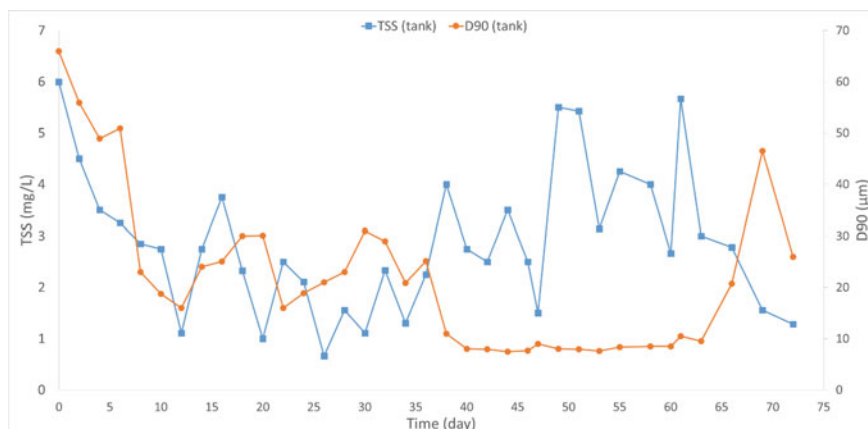
the decrease of AOS in the geotextile combination employed, ensuring the removal of suspended solids and nutrients in the tank water being remediated. The cake layer formed after one week of filtration is presented in Fig. 5. For this on-site filtration process, two filtration mechanisms have occurred. It was noted that the first filtration mechanism that occurred was straining or surface filtration. This filtration mechanism happens when the porous surface retains the particles solely based on particle size [20]. After that, solids then penetrate the surface of the media to block the open pores, thus forming a filter cake on the surface of the media called depth filtration [19]. Thus, by particle accumulation, on the top of each geotextile layer, the surface filtration mechanism was improved and combined with depth filtration.

### 3.3 Continuous Deployment

To assess the system's responsiveness and versatility for the on-site geotextile filtration application, a continuous experiment was piloted near the LJ lakeshore. The filtered water by the on-site method has shown a lower average particle size than the lake overall. The same behavior was captured for turbidity removal as well as for nutrients that were attenuated in the process. This assured that the on-site filtration was able to remove any particulate endogenous phosphorus and organic matter entering the lake.

#### 3.3.1 Particle Follow-Up

By particle size determination in the tank water, a slow reduction was observed in both the TSS amount and particle size from the start of the experiment until 35 days as shown in Fig. 6. This could be justified as the system was performing well in this scenario, and any pumped water was being constantly filtered. After the decrease, a small increase was noticed in the tank TSS concentration but with particle sizes below 10  $\mu\text{m}$ . This might be explained as the fine colloidal particles, which had passed through the geotextile combination, had accumulated in the first 35 days into



**Fig. 6** TSS concentration (mg/L) and D90 diameter ( $\mu\text{m}$ ) follow-up in the tank water

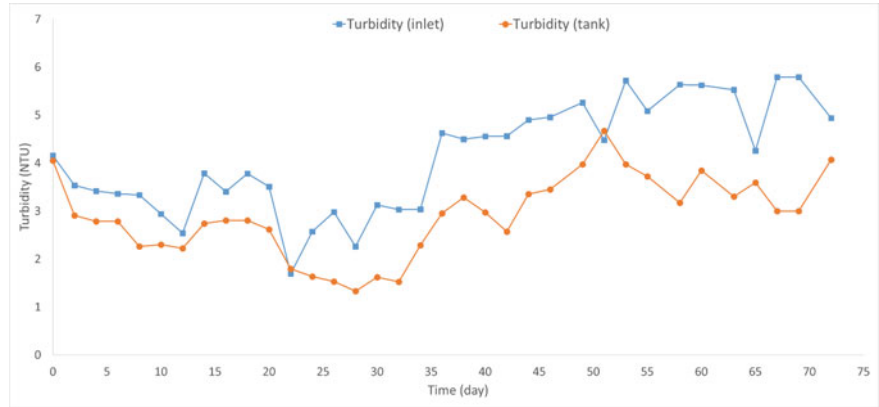
the tank and the system needed some days to adapt to this modification. Therefore, as the tank water was well mixed in the following days, TSS concentration decreased, but particle size increased. The average TSS concentration on the entire experiment on the tank was  $2.87 \pm 1.36$  mg/L and D90 (diameter of 90% of particles) was  $21.84 \pm 15.07$   $\mu\text{m}$ . Both were lower than the lake's overall quality. It was understood that by removing TSS quickly and consistently, a continuous D90 reduction in the tank water occurred, demonstrating that geotextile AOS reduction due to cake formation enhanced the filtering process.

### 3.3.2 Turbidity and Organic Matter Removal

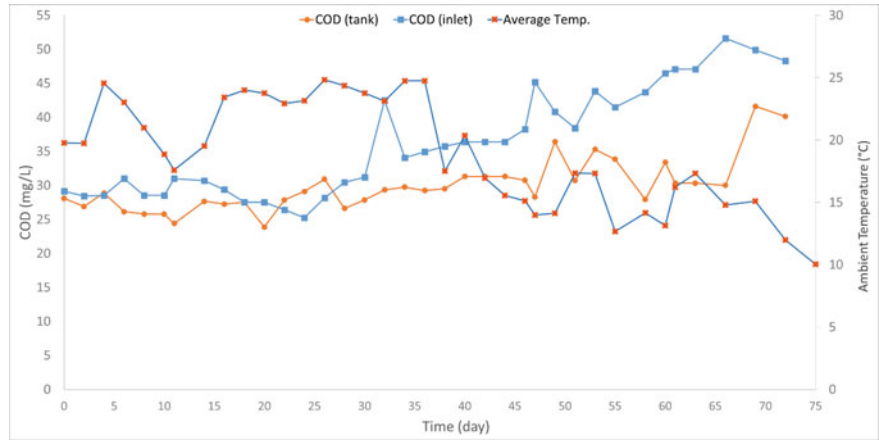
Following the same behavior of TSS decrease in the tank followed by an increase after 35 days, it was observed that even with constant replenishment of higher turbidity water from the system input, the floating filtration unit was able to ensure particle removal. As presented in Fig. 7, an average of 29% removal was obtained. In this view, less turbid water was being returned to the lake throughout the whole deployment. Average tank turbidity (same as the effluent) was  $2.88 \pm 0.82$  NTU, and the inlet was  $4.07 \pm 1.11$  NTU.

As mentioned, the wetland discharge was suggested as one of the contributors to this lake's ageing. After the follow-up on the organic matter (OM) concentration from the system inlet, the hypothesis was corroborated, presenting itself as an OM pollution source to the lake. This OM concentration from the inlet, presented as COD concentration, was always higher ( $36.12 \pm 7.68$  mg/L) than the overall lake water quality ( $25.0 \pm 3.50$  mg/L). Also, on the experiment deployment, shown in Fig. 8, it can be perceived as a constant replenishment of OM in the tank by the inlet, where some removal has been achieved (i.e., average removal of 17%). This slight attenuation can be clarified as the only OM attenuated was particulate and the





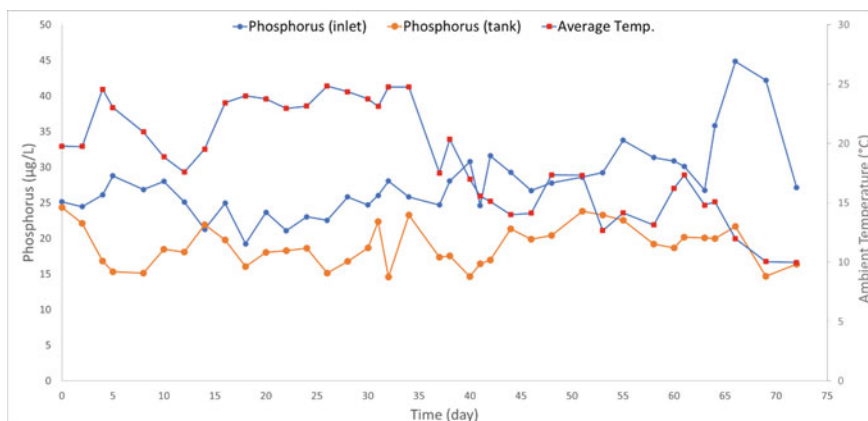
**Fig. 7** Turbidity removal (NTU) for 0.5-day retention time



**Fig. 8** COD concentrations (mg/L) for a 0.5-day retention time. Ambient temperatures were obtained from Almanac average extremes Montreal Mirabel intl A. — 7,034,900 (2022)

dissolved OM present was not removed. Tank water presented an average COD of  $29.87 \pm 3.80$  mg/L.

As in any filtration process, the assumption was that the removal of suspended solids in the study water is related to the reduction of particle magnitude and cloudiness in a well-mixed system. The proposed on-site filtration has also prevented new OM from being inputted in the lake and possible settling and accumulation on sediments, which can be transformed into an internal source in the near future. In this way, water with less TSS will be inputted into the lake, which will allow the lake to naturally respond to this change for a better ecosystem.



**Fig. 9** TP concentrations ( $\mu\text{g/L}$ ) for 0.5-day retention time. Ambient temperatures were obtained from Almanac average extremes Montreal Mirabel intl A. – 7,034,900 (2022)

### 3.3.3 Nutrient Removal

Likewise, as for the OM, the total phosphorus (TP) was always higher ( $27.75 \pm 5.03 \mu\text{g/L}$ ) than the overall lake water quality ( $17.1 \pm 2.50 \mu\text{g/L}$ ) supporting, even more, the suggestion of wetland discharge as one of the strong contributors to this lake ageing. It was observed that the geotextile filtering system, in addition to maintaining continuous removal of turbidity, suspended solids, and some organic material, also ensured that the water returning to the lake had a phosphorus concentration near the average lake's overall quality. In other words, the system was capable of removing some of the endogenous phosphorus being inputted into the lake by the wetland discharge, an average of 32%. Then, filtered water returned to the lake and had a phosphorus concentration of  $18.9 \pm 2.78 \mu\text{g/L}$  (Fig. 9).

Throughout the experiments, no significant change in the concentrations of TN and nitrate was observed as they were mainly in the dissolved form. The average value was kept below the values present in the influent and within the Quebec regulated values. For TN, the value was kept at an average of  $1.25 \pm 0.54 \text{ mg/L}$ , and for nitrate, its average was  $0.28 \pm 0.3 \text{ mg/L}$ .

## 4 Conclusions

Wetland discharge is further contributing to this lake aging by inputting water with higher particle sizes, higher organic matter concentration, and higher particulate phosphorus levels. The flexible, reactive, and environmentally friendly technology, on-site geotextile filtration, was able to enhance the water quality of a representative amount of lake water. By using the geotextile layer combination not only the straining

filtration mechanisms had occurred but also depth filtration, which was able to reduce the AOS of the filter layers and increase filtration performance. Results showed that any endogenous suspended solids and particulate phosphorus being inputted by the wetland were removed by the geotextile filtration and could prevent any further degradation or trophic level change of the lake. Regarding the geotextiles used, methods to reuse, reduce, and recycle are under study to assure waste reduction and promote a circular economy and sustainability in the process. Investigation of in-lake filtering testing, dissolved COD characterization, attenuation methods, and the reuse of clogged geotextile filters can be considered for future work.

**Acknowledgements** The authors thank NSERC and Concordia University for the financial support of this project. The authors are also grateful to their industrial partner, Titan Environmental Containment Ltd., for supplying geotextiles and providing technical and financial support for this project.

## References

1. Meyer-Jacob C, Michelutti N, Paterson AM, Cumming BF, Keller WB, Smol JP (2019) The browning and re-browning of lakes: divergent lake-water organic carbon trends linked to acid deposition and climate change. *Sci Rep* 9(1):1–10
2. Zheng X, Liu G, Yang W, Peng X, Liu H, Li H, Li W (2021) Dominant contribution of a lake's internal pollution to eutrophication during rapid urbanization. *Bull Environ Contam Toxicol* 107(5):904–910
3. Cooperrider MC, Davenport L, Goodwin S, Ryden L, Way N, Korstad J (2020) Case studies on cultural eutrophication on watersheds around lakes that contribute to toxic blue-green algal blooms. In: *Ecological and practical applications for sustainable agriculture*, pp 357–372. Springer, Singapore
4. Morabito G, Rogora M, Austoni M, Ciampittiello M (2018) Could the extreme meteorological events in Lake Maggiore watershed determine a climate-driven eutrophication process? *Hydrobiologia* 824(1):163–175
5. Belgacem W, Mattas K, Arampatzis G, Baourakis G (2021) Changing dietary behavior for better biodiversity preservation: a preliminary study. *Nutrients* 13(6):2076
6. Hamilton HA, Ivanova D, Stadler K, Merciai S, Schmidt J, Van Zelm R, Moran D, Wood R (2018) Trade and the role of non-food commodities for global eutrophication. *Nature Sustain* 1(6):314–321
7. Alexander TJ, Vonlanthen P, Seehausen O (2017) Does eutrophication-driven evolution change aquatic ecosystems? *Philos Trans Royal Soc B Biol Sci* 372(1712):20160041
8. Senar OE, Creed IF, Trick CG (2021) Lake browning may fuel phytoplankton biomass and trigger shifts in phytoplankton communities in temperate lakes. *Aquat Sci* 83(2):1–15
9. Li H, Gu X, Chen H, Mao Z, Shen R, Zeng Q, Ge Y (2022) Co-occurrence of multiple cyanotoxins and taste-and-odor compounds in the large eutrophic Lake Taihu, China: dynamics, driving factors, and challenges for risk assessment. *Environ Pollut* 294:118594
10. Yindong T, Xiwen X, Miao Q, Jingjing S, Yiyang Z, Wei Z, Mengzhu W, Xuejun W, Yang Z (2021) Lake warming intensifies the seasonal pattern of internal nutrient cycling in the eutrophic lake and potential impacts on algal blooms. *Water Res* 188:116570
11. Bormans M, Maršálek B, Jančula D (2015) Controlling internal phosphorus loading in lakes by physical methods to reduce cyanobacterial blooms: a review. *Aquat Ecol* 50(3):407–422
12. Zheng Y, Wang B, Wester AE, Chen J, He F, Chen H, Gao B (2019) Reclaiming phosphorus from secondary treated municipal wastewater with engineered biochar. *Chem Eng J* 362:460–468

13. Mulligan CN, Davarpanath N, Fukue M, Inoue T (2009) Filtration of contaminated suspended solids for the treatment of surface water. *Chemosphere* 74:779–786
14. Pereira AC, Veetil DP, Mulligan CN, Bhat S (2020) On-site non-woven geotextile filtration method for remediation of lake water. In: 2020 CSCE annual conference. Paper accepted but not presented due to conference cancellation
15. Pereira AC, Veetil DP, Mulligan CN, Bhat S (2021) Environmental remediation of a shallow mesotrophic lake water using on-site non-woven geotextile filtration treatment. In: 2021 CSCE annual conference. The conference was held online in May 2021
16. Pereira AC (2021) Novel sustainable in-situ geotextile filtration method for eco-remediation of eutrophic lake waters (Master's Thesis, Concordia University)
17. Veetil DP, Arriagada E, Mulligan C, Bhat S (2021) Filtration for improving surface water quality of a eutrophic lake. *J Environ Manage* 279:111766
18. Veetil DP, Mulligan CN, Bhat S (2018) Phosphorus Speciation of Sediments of a Mesoeutrophic Lake in Quebec, Canada. In: *The international congress on environmental geotechnics* Springer, Singapore, 1:780–787
19. Ebnesaajad S (2016) Chapter 10—Filtration in *Expanded PTFE applications handbook: technology, manufacturing and applications*, p 213–231 William Andrew Applied Science Publishers
20. Berk Z (2018) Chapter 8—Filtration in *food process engineering and technology*, 3rd edn, p 195–216, Academic Press
21. ABVLACS (Agence des Bassin Versants de Sainte-Anne-des-Lacs) (2021) Retrieved from <http://abvlacs.org/lac-johanne>
22. Almanac Average Extremes Montreal Mirabel intl. A. – 7034900, 2022. Retrieved from [https://climate.weather.gc.ca/climate\\_data/almanac\\_e.html?StationID=49608&month=8&day=17&timeframe=4&year=2022&month=8&day=17](https://climate.weather.gc.ca/climate_data/almanac_e.html?StationID=49608&month=8&day=17&timeframe=4&year=2022&month=8&day=17)
23. Chapman D, Kimstach V (1996) *Water quality assessments: a guide to the use of biota, sediments, and water in environmental monitoring*. 2nd ed. CRC Press. London
24. EPA, U.S. (1996) Method 3050B: acid digestion of sediments, sludges, and soils. Revision 2, p 12
25. RSVL (Réseau de surveillance volontaire des lacs), Lac Johanne—No RSVL: 497, 2021. Retrieved from [https://www.environnement.gouv.qc.ca/eau/rsvl/relais/rsvl\\_details.asp?fiche=497](https://www.environnement.gouv.qc.ca/eau/rsvl/relais/rsvl_details.asp?fiche=497)

# **Environmental Specialty: Ecohydrology and Environmental Hydraulics**

# On Developing Extreme Rainfall Intensity–Duration–Frequency Relations for Canada: A Comparative Study of Different Estimation Methods



Van-Thanh-Van Nguyen and Truong-Huy Nguyen

**Abstract** Extreme rainfall intensity–duration–frequency (IDF) relations are commonly used for estimating the design storm for the design of various urban hydraulic structures. Traditionally, these IDF relations were obtained by fitting the two-parameter Gumbel distribution to the annual maximum (AM) rainfalls for each rainfall duration independently using the method of moments (MOM). However, it has been widely known that this Gumbel/MOM-based traditional approach may not produce accurate estimates of extreme rainfalls as compared to those given by, for instance, the generalized extreme value (GEV)/L-Moment method as suggested in some recent studies. Consequently, there are several new IDF estimation procedures and products that are recently developed in Canada in an attempt to provide some improvements in the estimation of design rainfalls. This study proposed therefore new approaches for developing IDF relations based on the scale-invariance behaviour of extreme rainfall processes using the GEV distribution. A detailed comparative study was then carried out to compare the performance of traditional IDF estimation methods and the proposed new approaches using the available IDF data from 39 stations located across Canada with at least 50 years of record. Results of this comparative study have indicated that the new scale-invariance GEV approaches can provide the most accurate and most robust estimates of design rainfalls for all locations in Canada as compared to existing traditional methods.

**Keywords** Extreme rainfall intensity · Duration · Frequency relations

---

V.-T.-V. Nguyen (✉) · T.-H. Nguyen  
Department of Civil Engineering, McGill University, Montréal, Canada  
e-mail: [van.tv.nguyen@mcgill.ca](mailto:van.tv.nguyen@mcgill.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_68](https://doi.org/10.1007/978-3-031-34593-7_68)

1059

## 1 Introduction

Information on the spatio-temporal variability of extreme rainfall characteristics is of critical importance for planning, design, and management of various water systems [1]. In particular, for urban and small rural watersheds that are generally characterized by fast response, the design of hydraulic structures such as small dams, culverts, storm sewers, detention basins, and so on, requires extreme rainfall input for very short time durations (e.g. few minutes or hours) for runoff simulation models. More specifically, this extreme rainfall information is extracted from the “intensity–duration–frequency” (IDF) relations for various durations and return periods at a given site of interest [1–3].

In current engineering practice, the IDF relations are commonly derived based on statistical frequency analyses of annual maximum (AM) rainfall series for different durations. In Canada, Environment and Climate Change Canada (ECC) provides short-duration extreme rainfall data for nine different rainfall durations ( $D = 5, 10, 15, 30, 60, 120, 360, 720, \text{ and } 1440 \text{ min}$ ) and the IDF relations for approximately 650 stations across Canada with at least 10-year rainfall record [4]. Traditionally, the extreme rainfalls for six different return periods ( $T = 2, 5, 10, 25, 50, \text{ and } 100 \text{ years}$ ) were computed by fitting the two-parameter Gumbel distribution to the AM series for each rainfall duration independently using the method of moments (MOM). However, it has been widely known that this Gumbel/MOM-based traditional approach may not produce accurate and robust extreme rainfall estimates as compared to those given by, for instance, the generalized extreme value (GEV)/L-Moment method [5]. Consequently, there are several recently developed IDF products in Canada (as summarized in Table 1) to provide some improvements in the estimation of extreme rainfalls for both gaged or ungaged locations. Hence, there is an urgent need to carry out a critical review of existing rainfall estimation methods and to perform a detailed comparative study to assess their performance in order to identify the best estimation method for deriving IDF relations for Canada.

## 2 Extreme Rainfall Estimation Methods

As mentioned previously, Table 1 provides a summary of existing methods for extreme rainfall estimation and for developing the IDF relations for Canada. These include the traditional EC-IDF product by ECC [4], the Metro Vancouver IDF tool [6]; the IDF design values for Atlantic Provinces by Shephard [7], the Ontario Ministry of Transportation (MTO-IDF) tool by Soulis et al. [8], the IDF-CC tool by Simonovic et al. [3], and the SMExRain tool by Nguyen and Nguyen [2]. The key differences between these tools are: (i) the different probability model that was selected for describing the distribution of extreme rainfalls; (ii) the different estimation method that was used for estimating the probability model parameters; (iii) the different regression model that was chosen to represent the IDF curves; (iv) the different

**Table 1** Existing IDF products and tools in Canada

Product	Provider	Latest version (Year)	Region	Brief Description	Reference
EC-IDF	Environment Canada	EC_IDF v-3.1 (2020)	Canada	IDF graphs and tables for 651 stations across Canada	[4]
IDF-CC tool	University of Western Ontario	IDF_CC Tool 4.5 (2021)	Canada	Web-based application, IDF graphs and tables for gaged and ungaged stations using (EC_IDF v3.1 data)	[3]
SMEsRain tool	McGill University	SMEsRain 1.0 (2019)	Canada	Standalone software, IDF graphs and tables for any site with IDF data	[2]
MTO IDF Curves Finder	Ontario Ministry of Transportation	MTO IDF 3.0 (2016)	Ontario	Web-based application for gaged (using EC_IDF v2.3 data) and ungaged stations	[8]
Metro Vancouver	Metro Vancouver	IDF curves (2009)	Vancouver	IDF graphs and tables for different homogeneous zones	[6]
IDF climate design values	Atlantic Climate Adaptation Solutions Association	IDF curves (2011)	Atlantic Provinces	IDF curves for different Atlantic Provinces	[7]

estimation method that was used for estimating the regression model parameters; (v) the different spatial interpolation technique that was used for transferring IDF information from gaged sites to the target ungaged location; and (vi) the different consideration of the scale-invariance property of the extreme rainfall processes for different rainfall durations.

More specifically, for the whole Canada, only the EC-IDF and IDF-CC are readily available, and the data are maintained up to date. The MTO IDF tool is mainly developed for the province of Ontario using data primarily from EC-IDF in combination with USGS digital elevation data to derive physiographic characteristics [8]. The Metro Vancouver IDF curves are primarily developed for the City of Vancouver and its neighbouring regions based on Hosking and Wallis [9]’s regional approach [6]. The IDF curves for Atlantic Provinces use the same approach as in the EC-IDF and have been recently integrated into the EC-IDF product [4]. The SMEsRain tool is a standalone application that can be used to generate IDF curves for any location in Canada with available IDF data [2]. Therefore, only the EC-IDF product and



the IDF-CC tool were selected for this comparative study. A summary of the basic features of these two products is provided in the following sections.

The EC-IDF tool provides the historical IDF relations for approximately 650 stations across Canada with at least 10-year record [4]. The extreme rainfalls for these IDF relations were computed for nine different rainfall durations ( $D = 5\text{--}1440$  min) and for six different return periods ( $T = 2\text{--}100$  years) using the two-parameter Gumbel distribution. In particular, the Gumbel distribution is fitted to the historical AM series for each rainfall duration independently using the method of moments (MOM). In addition, the IDF relations are described by a simple power-form relation as follows:

$$I = aD^b \quad (1)$$

in which  $I$  is the rainfall intensity;  $D$  is the rainfall duration; and  $a$ ,  $b$  are coefficients that are computed for each return period by the least-square technique in the log-data space. These two coefficients are also computed for locations without data using a simple linear spatial interpolation technique [4].

The IDF-CC tool uses the generalized extreme value (GEV) distribution as the parent distribution for representing the distribution of AM rainfall series [3]. The L-moment (LMOM) estimation method is used to estimate the three parameters of the GEV model. This tool also provides the IDF curves derived from the Gumbel distribution for the purpose of comparison. The IDF curves are described by the following mathematical relation:

$$I = a(D + c)^b \quad (2)$$

in which  $a$ ,  $b$ ,  $c$  are coefficients that are computed for each return period using the differential evolution optimization algorithm [3] in the real data space.

In summary, these two traditional approaches are based on the regression of the computed rainfall intensities (or depths) over durations using Eq. (1) with two coefficients (QR2C method) or Eq. (2) with three coefficients (QR3C method). In the present study, a new approach is introduced based on the regression of the empirical statistical moments of observed rainfall amounts over durations. As compared to the traditional approaches, this new method can account for the observed scale-invariance property of the empirical statistical moments. In general, the proposed new method consists of the following four steps:

- (i) Firstly, compute the statistical moments of the AM rainfalls for the first few orders for each rainfall duration (e.g. from 5-min to 24-h intervals) for a given location of interest. More specifically, only the first two non-central moments (NCMs) are required for the GUM/MOM model, and the first three probability weighted moments (PWMs) are necessary for the GEV/LMOM model;
- (ii) Secondly, construct regression models to describe the relationships between the computed NCMs (or PWMs) of rainfall amounts and the rainfall durations. These relationships indicate the scale-invariance property of the AM processes.

- For instance, some previous studies have found that extreme rainfall processes for durations ranging from a few minutes to several days could display one or two different scaling regimes [10, 11]. These rainfall-statistical-moment-based regression models are quite useful for the estimation of the NCMs (or PWMs) of sub-hourly or sub-daily rainfall amounts from those of the daily amounts;
- (iii) Thirdly, estimate the parameters of a selected theoretical distribution using the method of moments (or the method of L-moments) in consideration of the regression relationships established in step (ii). More specifically, there are two different approaches are proposed in this study depending on the application of these regression relationships: (a) the rainfall-statistical-moment-based method (referred hereafter as MR method) if the parameters of the GUM/MOM or GEV/LMOM model are computed based on the direct regression of the statistical moments of AM rainfalls for different durations; and (b) the scaling-statistical-moment-based method (referred herein as MRS method) if the GUM/MOM and GEV/LMOM parameters are computed based on the scale-invariance relations between daily and sub-daily statistical moments; and
  - (iv) Finally, estimate the design rainfall intensity (or depth) for a given duration and return period of interest using the quantile function of the Gumbel distribution or GEV distribution [10].

In this study, two common dimensionless goodness-of-fit (GOF) criteria were selected to compare the performance of the four methods (QR2C, QR3C, MR, and MRS). These criteria include the mean absolute relative difference (MADr) and the root mean square relative error (RMSEr) as defined below:

$$\text{MADr} = \frac{1}{n} \sum \left\{ \frac{|x_{ij} - y_{ij}|}{x_{ij}} \right\} \quad (3)$$

$$\text{RMSEr} = \left[ \frac{1}{n} \sum \left\{ \frac{(x_i - y_i)}{x_i} \right\}^2 \right]^{1/2} \quad (4)$$

in which  $x_{ij}$  and  $y_{ij}$  are the rainfall quantiles corresponding to duration  $D_i$  ( $i = 1, 2, \dots, 9$ ) and return period  $T_j$  ( $j = 1, 2, \dots, 6$ ) estimated based on the observed data and based on one of the four estimation methods (i.e. QR2C, QR3C, MR, and MRS), respectively;  $n = i \cdot j$  is the number of design rainfall values at each station.

### 3 Study Sites and Data

As mentioned previously, the IDF network consists of approximately 650 raingages located across Canada [4]. For each site, ECC provides the AM series for nine different rainfall durations ranging from 5 min to 24 h. The record length for these AM series varies from 10 to 82 years, and most of the data are recorded after 1960 and updated to 2017. For this comparative study, only 39 stations with very long

record (containing at least 50 years) were selected since the estimation of rainfall quantiles for high return periods (e.g. for  $T = 50$  and 100 years) can be considered as reliable for these long datasets [2]. The details of the 39 chosen stations are presented in Table 2. The comparison of the performance of the two traditional (QR2C and QR3C) approaches and the two new MR and MRS methods will be carried out using the available IDF data for these 39 stations.

## 4 Results and Discussions

In the present study, the rainfall quantiles for each station were estimated for all six return periods ( $T = 2$ –100 years) and for all nine durations ( $D = 5$ –1440 min) using the four selected approaches (QR2C, QR3C, MR, and MRS). The IDF relations are then constructed for each station. For purposes of illustration, Fig. 1 shows the IDF relations for St Thomas WPCP Station for all methods. It can be observed that the QR2C method produced the least accurate results. As expected, the QR3C, MR, and MRS methods, considered as the general form of the QR2C method, provided more accurate results. Similar results were found for other stations.

More generally, Fig. 2 shows the comparison of these four methods based on the MADr (%) and RMSEr (%) for both Gumbel and GEV models and for all stations. The slope ratio between the first and second scaling regimes of AM series is also plotted. Notice that the first scaling regime is identified as from the 5-min duration to the breakpoint on the graph between the empirical statistical moments and the rainfall durations; the second scaling regime is defined from the breakpoint to the 24-h duration for each station. It can be seen that the QR2C method yields high error for most of stations, especially stations with large difference in the slope ratio. The QR3C method produces good results for most stations with MADr less than 5%, except the first two stations with the slope ratio less than 1 indicating the convex pattern for the IDF curves. On the other hand, both MR and MRS approaches can capture very well this convex pattern for these two stations and can also provide the MADr less than 5% for all other stations. The MADr and RMSEr results for all stations are also summarized in the form of boxplots as shown in Fig. 2. It can be seen that the QR2C is the least accurate method.

A more detailed investigation of the slope ratios of the first and second scaling regimes is shown in Fig. 3. These slopes are computed based on the first NCMs (or first PWMs) of rainfall depths over durations for all available IDF stations (approximately 650 stations). In general, it can be seen that the pattern of the IDF curves varies over different regions of Canada. However, for a large number of stations in the Pacific region the convex pattern (i.e. the slope ratio value is less than one) occurs more frequent. This could be due to the orographic effect in this region which causes a very different behaviour of short-duration extreme rainfall processes. Hence, the QR3C method should not be applied in this region.

**Table 2** Details of the selected 39 stations, including the province (PR), station identification (ID), station name, latitude (lat, degrees), longitude (long, degrees), and record period and length (No. of years)

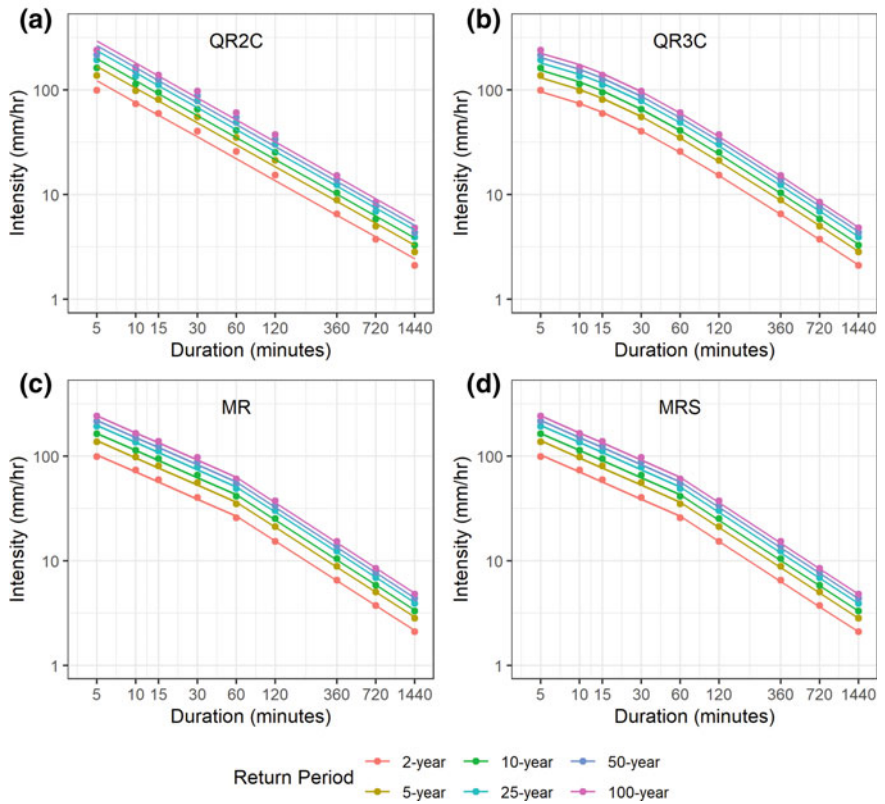
No.	PR	ID	Station Name	Lat (degree)	Long (degree)	Period	No. of years
1	BC	1018611	Victoria Gonzales CS	48.42	123.32	1925–2017	65
2	BC	1018621	Victoria Intl A	48.65	123.43	1965–2017	50
3	BC	1105192	Mission West Abbey	49.15	122.27	1963–2017	54
4	BC	1108395	Vancouver Intl A	49.18	123.18	1953–2017	63
5	AB	3012206	Edmonton Intl CS	53.32	113.62	1961–2017	52
6	AB	3012209	Edmonton Blatchford	53.57	113.52	1914–2017	71
7	AB	3031094	Calgary Int L CS	51.12	114.00	1947–2017	62
8	SK	4012410	Estevan	49.22	102.97	1964–2017	53
9	SK	4015322	Moose Jaw CS	50.33	105.53	1960–2017	50
10	SK	4016699	Regina R CS	50.43	104.67	1941–2017	62
11	SK	4043901	Kindersley A	51.52	109.18	1966–2016	50
12	MB	502S001	Winnipeg A CS	49.92	97.25	1944–2017	58
13	ON	6012199	Ear Falls (Aut)	50.63	93.22	1952–2017	56
14	ON	6016525	Pickle Lake (Aut)	51.45	90.22	1953–2017	51
15	ON	6042716	Geraldton A	49.78	86.93	1952–2016	54
16	ON	6048268	Thunder Bay CS	48.37	89.33	1952–2012	53
17	ON	6078285	Timmins Victor P. A	48.57	81.38	1952–2016	51
18	ON	6104175	Kingston Pumping Stn	44.23	76.48	1914–2007	63
19	ON	6105978	Ottawa Cda R CS	45.38	75.72	1905–2017	57
20	ON	6127519	Sarnia Climate	43.00	82.30	1962–2017	50
21	ON	6131983	Delhi CS	42.87	80.55	1962–2015	50
22	ON	6137362	St Thomas WPCP	42.77	81.22	1926–2016	82
23	ON	6139525	Windsor A	42.28	82.97	1946–2016	66
24	ON	6143089	Guelph Turfgrass	43.55	80.22	1954–2017	52
25	ON	6144478	London CS	43.03	81.15	1943–2017	66
26	ON	6153301	Hamilton RBG CS	43.28	79.92	1962–2017	53
27	ON	6158355	Toronto City	43.67	79.40	1940–2017	67
28	ON	6158731	Toronto Intl A	43.68	79.63	1950–2017	64
29	QC	701S001	Quebec Jean L. Intl	46.80	71.38	1961–2017	57
30	QC	7014160	L Assomption	45.82	73.43	1963–2017	55
31	QC	702S006	Montreal P.E.T. Intl	45.47	73.73	1943–2017	72
32	QC	7024280	Lennoxville	45.37	71.82	1960–2017	54
33	NB	8100885	Charlo Auto	47.98	66.33	1959–2017	55
34	NB	8103201	Moncton Intl A	46.12	64.68	1946–2016	67

(continued)

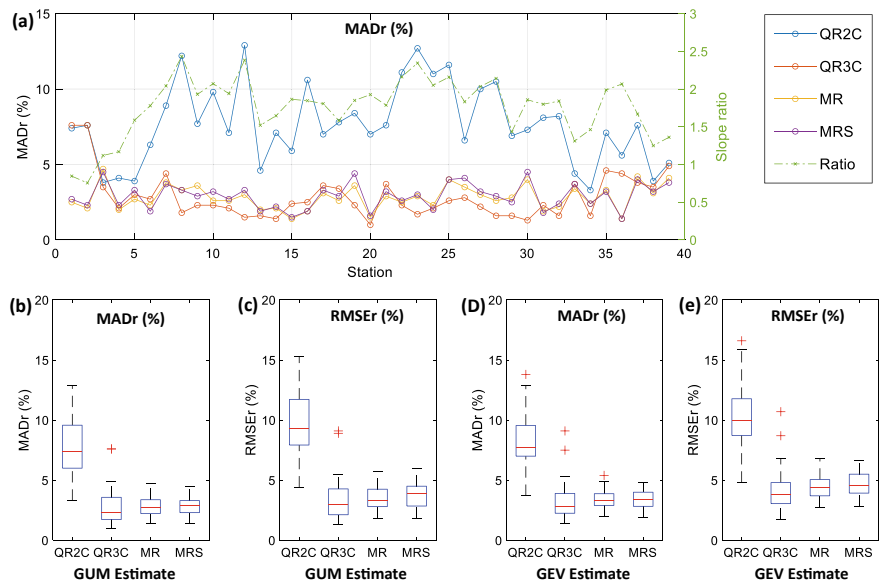
**Table 2** (continued)

No.	PR	ID	Station Name	Lat (degree)	Long (degree)	Period	No. of years
35	NS	8204700	Sable Island	43.93	60.02	1962–2013	51
36	NS	8205092	Shearwater R CS	44.63	63.52	1955–2017	60
37	NS	8205702	Sydney CS	46.17	60.03	1961–2016	53
38	NL	8401705	Gander Airport CS	48.95	54.57	1939–2017	70
39	NL	8501900	Goose A	53.32	60.42	1961–2016	53

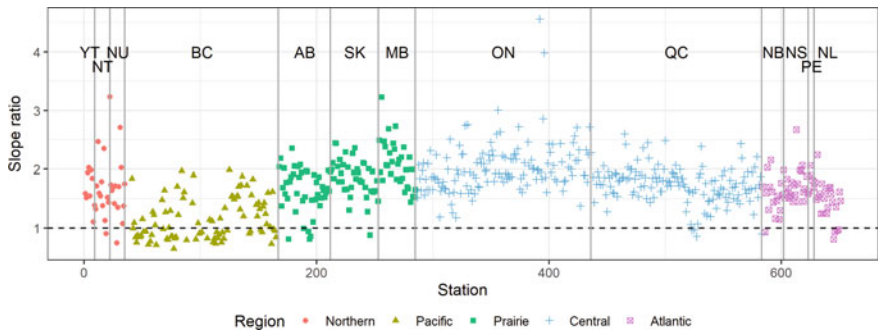
**Station 22 ST\_THOMAS\_WPCP**



**Fig. 1** IDF relations at St Thomas WPCP Station based on different methods: **a** and **b** quantile-based regression methods (QR2C and QR3C); **c** and **d** moment-based regression without and with scaling (MR and MRS). Markers represent the observed GUM/MOM rainfall quantiles



**Fig. 2** Plots of the scaling slope ratio, MADr (%), and RMSEr (%) for different methods (QR2C, QR3C, MR, and MRS) for Gumbel and GEV models



**Fig. 3** Slope ratios of the first and second scaling regimes based on the first NCMs (or PWMs) of rainfall depths for different regions of Canada

## 5 Summary and Conclusions

In Canada, there exist several different methods for developing IDF relations. Hence, in this research, a comparative study was carried out to assess the performance of these methods in order to identify the best approach for use in practice. More specifically, two traditional approaches (QR2C and QR3C) were compared with two new proposed procedures (MR and MRS). In general, the traditional methods were based on the relationships between the extreme rainfall quantiles with the rainfall

durations, while the new approaches were relied on the relationships between the statistical moments of extreme rainfalls and the rainfall durations. The comparison was performed using the available historical AM data for nine different durations (from 5 min to 24 h) for 39 stations with long record length (at least 50 years) across Canada to represent the diverse climatic conditions of Canada.

Results of this comparative study have indicated that the QR2C based on the GUM/MOM model is the least accurate method as compared with the other three methods. In addition, it was found that different patterns of the IDF curves for different locations in Canada can be linked to the scaling behaviour of the extreme AM processes. This scaling behaviour can be identified based on the relationships between the empirical statistical moments (NCMs or PWMs) of rainfall amounts and the rainfall durations. Consequently, the QR3C method can be used for IDF relations with linear and concave scaling patterns but cannot provide an accurate rainfall estimation for IDF curves with the convex pattern that is a common pattern observed in the Pacific region. On the other hand, it was found that the new MR and MRS methods proposed in this study can be used for all regions of Canada.

Furthermore, the MR method produces slightly better results than the MRS since it was relied on the best fit of the regression model to the empirical statistical moments over all different rainfall durations, while the MRS method was based on the approximate fit of the derived sub-daily rainfall statistical moments from the daily statistical moments. However, this difference is not significant (on average, less than 2% between the two methods). The MRS method, however, offers a key advantage for cases of ungaged or partially-gaged sites where the derivation of the distributions of sub-hourly and sub-daily extreme rainfalls from that of daily amounts is required.

**Acknowledgements** Financial support by the Natural Science and Engineering Research Council Canada (NSERC) (Discovery Grant Program) for this research work is gratefully acknowledged.

## References

1. CSA (2019) Technical guide: development, interpretation, and use of rainfall intensity-duration-frequency (IDF) information: guideline for Canadian water resources practitioners, CSA PLUS 4013:19. Canadian Standard Association, Toronto, Ontario, Canada, p 126
2. Nguyen T-H, Nguyen V-T-V (2019) Decision-support tool for constructing robust rainfall IDF relations in consideration of model uncertainty. *J Hydrol Eng ASCE* 24(7):06019004. [https://doi.org/10.1061/\(asce\)he.1943-5584.0001802](https://doi.org/10.1061/(asce)he.1943-5584.0001802)
3. Simonovic SP, Schardong A, Sandink D, Srivastav R (2016) A web-based tool for the development of intensity duration frequency curves under changing climate. *Environ Model Softw* 81:136–153. <https://doi.org/10.1016/j.envsoft.2016.03.016>
4. Environment Canada (2020) Rainfall intensity-duration-frequency (IDF) tables and graphs. Version 3-10. Released date: March 2020. Retrieved from [https://climate.weather.gc.ca/prods\\_servs/engineering\\_e.html](https://climate.weather.gc.ca/prods_servs/engineering_e.html). Accessed on 1 Oct 2020
5. Nguyen T-H, El Outayek S, Lim SH, Nguyen V-T-V (2017) A systematic approach to selecting the best probability models for annual maximum rainfalls—a case study using data in Ontario (Canada). *J Hydrol* 553:49–58. <https://doi.org/10.1016/j.jhydrol.2017.07.052>

6. MetroVancouver (2009). Regional IDF curves. <http://www.metrovancouver.org/services/liquid-waste/drainage/Pages/default.aspx>
7. Shephard MW (2011) Updating IDF climate design values for the Atlantic Provinces. Atmospheric and Climate Applications, Inc., Ontario, p 41
8. Soulis E, Princz D, Wong J (2014) Renewal and update of MTO IDF curves: defining the uncertainty. J Water Manag Model. <https://doi.org/10.14796/jwmm.c386>
9. Hosking JRM, Wallis JR (1997) Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge, p 224
10. Nguyen T-H, Nguyen V-T-V (2020) Linking climate change to urban storm drainage system design: an innovative approach to modeling of extreme rainfall processes over different spatial and temporal scales. J Hydro-environ Res 29:80–95. <https://doi.org/10.1016/j.jher.2020.01.006>
11. Yeo MH, Nguyen VTV, Kpodonu TA (2020) Characterizing extreme rainfalls and constructing confidence intervals for IDF curves using Scaling-GEV distribution model. Int J Climatol 41(1):456–468. <https://doi.org/10.1002/joc.6631>



# Artificial Neural Networks and Extended Kalman Filter for Easy-to-Implement Runoff Estimation Models



Arash Yoosefdoost, Syeda Manjia Tahsien, S. Andrew Gadsden,  
William David Lubitz, and Mitra Kaviani

**Abstract** Determination of suitable sites for small hydropower projects could offer new opportunities for sustainable developments. However, the non-scalable initial investigation costs are one of the biggest burdens when planning small projects. Moreover, solving a complex problem with only a few available parameters is almost impossible with many traditional models, and lack of data may make many design studies infeasible for remote, hard to access or developing areas. Artificial neural networks (ANNs) could help reduce investigation costs and make many projects feasible to study by acting as input–output mapping algorithms. This study provides an easy to understand and implement method to develop fast ANN-based estimation models using the multilayer perceptron (MLP) neural network and extended Kalman filter (EKF) or gradient descent (GD) as the training algorithm. Also, three approaches to feeding training data to the models were studied. Estimating runoff is an important challenge in water resources engineering, especially for development and operation plans. Therefore, the proposed method is applied for a runoff estimating problem using only easily measured precipitation and temperature. Results of this case study indicate that for a relatively similar performance, ANN models using EKF required a fewer number of neurons and training epochs than GD. Compared to the prior research in this study area, the methods in this study are much easier to understand and implement and are not dependent on data mining techniques or continuous long-term time series. Based on the results, a combination of the proposed data feeding methods and the EKF training algorithm improved estimation models by reducing the number of training epochs and the size of the network.

---

A. Yoosefdoost (✉) · S. M. Tahsien · W. D. Lubitz  
School of Engineering, University of Guelph, Guelph, Canada  
e-mail: [Yoosefdoost@gmail.com](mailto:Yoosefdoost@gmail.com)

S. A. Gadsden  
Department of Mechanical Engineering, McMaster University, Guelph, Canada

M. Kaviani  
School of Computer Science, University of Guelph, Guelph, Canada

**Keywords** Artificial neural network • ANN • Extended Kalman filter • EKF • Gradient descent • Multilayer perceptron • MLP • Machine learning • Runoff • Taleghan basin

## Nomenclature

$B$	Bias
$B^L$	Biases matrix of the layer $L$
$D_n$	Desired outputs (observations/measurements)
$E$	Error function
$EB$	The epoch of the best
$EKF$	Extended Kalman filter
$f, F$	Nonlinear, linear state function/matrix
$GD$	Gradient descent
$h, H$	Nonlinear, linear measurement function/matrix
$K$	Kalman gain matrix (i.e., EKF)
$k + 1 k$	A priori time step (i.e., before applied gain)
$k + 1 k + 1$	A posteriori time step (i.e., after update)
$L$	Layer number
$m$	The number of the neuron connected to the current neuron ( $n$ )
$MEB$	The minimum epoch of the best
$MSE$	Mean square error
$n$	Number of samples/neuron
$N(0, Q)$	The noise of the process with the covariance of $Q$
$N(0, R)$	The noise of the measurement with the covariance of $R$
$O_i$	Observed data
$\bar{O}$	Average of observational data
$P$	State error covariance matrix
$P_i$	Estimated value
$\bar{P}$	Average of estimated data
$Q$	System noise covariance matrix
$R$	Measurement noise covariance matrix
$RMSE$	Root mean square error
$S$	Innovation covariance matrix
$T$	Transpose of vector or matrix
$u$	Input to the system
$w$	Weight
$W^L$	Weight's matrix of the layer $L$
$x$	State vector
$X^0$	Inputs' matrix
$X^L$	Inputs' matrix of the layer $L$
$x_m$	Neuron's inputs
$z$	Measurement (system output) vector

$\Gamma$	Hard-limit (unit step) function
$\Sigma$	Neuron's sum function output
$\Sigma^L$	Sigma function matrix of the layer $L$
$\Psi_n$	Model's outputs (estimations/predictions)
$\mathfrak{N}$	Neuron's ultimate output $f(\Sigma)$
$\mathfrak{N}^L$	Neuron output
$\wedge$	Estimated vector or values

## 1 Introduction

Solving a complex problem with a few available variables could be almost impossible by many traditional models. Therefore, the lack of enough data may make many studies infeasible. Furthermore, studying new problems or even new cases usually requires a systematic approach that could be costly or time-consuming. Using a black box could be beneficial to speed up studies or make them feasible where utilizing traditional methods is hard, expensive, time-consuming or technically infeasible.

A black box could be defended as a model that maps the inputs of a system to its output without the need for knowing the governing rules and phenomena in the system, or details such as the dynamics of, or interactions between, the system's components. Based on this definition, many examples could be considered or utilized as a block box: from simple methods such as logistic regression to non-parametric supervised learning methods such as decision tree, evolutionary algorithms such as gene expression programming (GEP) and machine learning models (ML) as support vector machine (SVM), and artificial intelligence (AI) solutions such as artificial neural network (ANN).

Analytical solutions and conceptual models are preferable where the required data is available. However, recent studies indicate that using proper data for black-box models in some cases led to achieving even more accuracy. For example, a comparison of three black-box models (SVM, multi-layer perceptron (MLP), and GEP) with the HYMOD ("a lumped rainfall-runoff model based on conceptually simplified physical processes model in rainfall-runoff simulation" [1, 2]) indicated a more reasonable performance for black-box models compared to conceptual models in simulating rainfall-runoff in the Karaj basin [3].

In the modern world, widely applicable black-box models could offer new, faster, or more effective solutions or assist in solving complex problems where it is not possible using traditional models. For example, ANNs are widely applied for solving complex problems in areas that there are a few or no alternatives for them, such as problems such as image processing, and natural language processing.

Generally, using more samples for training the ANNs could improve the results. For example, a long-term dataset of effective variables in the precipitation-runoff phenomenon is required to train ANN-based runoff estimation models to achieve reasonable results. However, more training data comes at the cost of more training

time, particularly for large neural networks. Therefore, optimizing the size of neural networks is very important. In addition, efficient training algorithms are essential to reduce the training time of ANNs. Such algorithms could also benefit the computational costs by reducing the size of the ANN. Due to the proven applicability of ANNs in many areas in modern investigations, this research studied solutions to develop an easy-to-implement and understand ANN-based method to develop estimation models when the minimum number of the available variables limits the feasibility of the application of traditional models. For this purpose, the effects of using the extended Kalman filter (EKF) as the training algorithm for MLP neural network were evaluated by solving a runoff estimation problem and comparing the developed models with MLP networks using gradient descent (GD) as the training algorithm.

## ***1.1 Artificial Neural Networks***

Artificial intelligence (AI) is the intelligence exhibited by machines taught by humans. This term is applied when a machine mirrors cognitive functions such as learning and problem-solving [4]. AI research has been divided into subfields [5] that include reasoning, planning, and learning [4, 6–8]. Machine learning investigates the study and development of algorithms to learn from and make predictions on data [9]. Machine learning can accelerate and improve investigations by building a model from example inputs to make data-driven predictions instead of time-consuming and intensive deterministic analytical or computational approaches. Therefore, the quality of data is critical to the success of most AI-based or machine learning models.

Artificial neural networks (ANN) are biologically inspired neural networks used to estimate or approximate functions that can depend on a large number of inputs, which are generally unknown. ANNs usually make a correlation between inputs and outputs in a system, effectively acting as an input–output mapping algorithm with largely unknown properties [10]. Artificial neural networks (ANNs) date back to the idea of threshold logic by Warren McCulloch and Walter Pitts in 1943 [11], which is a generalization of the logic gates that work by comparing the inputs to a threshold [12]. In the 1940s, Hebb developed the so-called Hebbian learning, a learning hypothesis based on the mechanism of neural plasticity [13]. Hebbian learning was applied to computational models with Turing's B-type machines in 1948. In 1954, Farley and Clark simulated a Hebbian network in computational machines [14]. In 1958, Rosenblatt developed the concept of perceptron as a pattern recognition algorithm [15]. Group method of data handling is the first functional network with many layers, which Litvinenko and Lapa introduced in 1965 [16–18]. However, publishing the Minsky and Papert research result in 1969 caused stagnation in ANN research [19]. The backpropagation algorithm introduced by Werbos in 1975 effectively solved the 'exclusive or' problem by making the training of multilayer networks feasible and efficient. It distributed the error-term back through the layers and modified the

weights at each node [20]. The resilient backpropagation (Rprop) algorithm introduced by Riedmiller and Braun in 1993 is an effective learning algorithm for multilayer feedforward networks [21]. Today, many other algorithms and new approaches are developed for the optimization of the training process or outputs of ANNs.

## **1.2 Extended Kalman Filter and ANNs**

In signal processing, the Kalman filter is known as the most popular method used to estimate the states of a linear system in the presence of white noise [22]. Different versions of the KF have been developed for nonlinear systems and measurements and the most popular include the extended Kalman filter (EKF) and the unscented Kalman filter (UKF) [23]. The EKF estimation method has been found to be a fast training technique [24].

Many investigations studied the effectiveness of utilizing the recursive least squares technique (RLS) [25, 26] and EKFs [27, 28] in neural networks. In the 1990s, several studies autonomously used an EKF in training a multilayer perceptron and demonstrated that it performs better than utilizing a traditional backpropagation training approach [29–32]. In 1992, Ruck et al. studied the utilization of the EKF and backpropagation techniques as training algorithms [33]. In 1994, Puskorius and Feldkamp expanded the idea of training the neural network by applying an extended Kalman filter for trained recurrent networks in nonlinear dynamical systems [34]. In 1995 Plumer used Kalman filtering techniques to train multilayer perceptron neural networks (MLPs). Results demonstrated that the sequential-dual extended Kalman filter (DEKF) and batch-DEKF have fundamentally the same convergence properties and lead to a similar performance of the trained networks [35]. In 1998, Pui-Fai SUM [36] studied the possibility of applying the EKF in combination with pruning to speed up the learning process and specify the size of a trained neural network. In 1999, Wan and Nelson proposed a dual Kalman filtering strategy for feedforward neural network training [37].

Williams [38] and Suykens [39] used the EKF to train a recurrent neural network as a state estimation problem. In 2002, Puskorius and Feldkamp proposed a decoupled strategy for the EKF to accelerate the training rates [40]. Also, in 2002, Leung Chi and Chan Lai studied using a dual-EKF method in recurrent neural networks to improve the detection and identification of an aerospace problem [41]. In 2003, they proposed using a dual-EKF strategy in recurrent neural networks. The simulation results demonstrated that the proposed approach is an effective joint-learning-pruning method for RNNs under online operation [42].

Pietruszkiewicz [43] studied the application of nonlinear Kalman filtering to feedforward neural networks as the learning algorithm. This study showed that these filters have a high learning ability in the presence of noisy measurements when applied to a popular backpropagation neural network. A comparison of results showed that nonlinear Kalman filters outperformed the classical error backpropagation learning

algorithm [44]. In 2013, Kurylyak et al. reported that blood pressure could be evaluated accurately by combining Kalman filtering and ANNs [45]. The authors proposed a regulated KF approach to overcome the drawback of non-stationary impacts of breathing and measurement noise. The additional filtering was applied to remove low- and high-frequency noise and accurate blood pressure estimates [45]. Krok (2013) applied KF for ANN learning. This study demonstrated that implementing a KF reduced the calculation time by using fewer parameters for learning the network [46]. In 2019, Jondhale proposed a KF framework based on real-time target tracking in wireless sensor networks (WSN) in an application with generalized regression neural networks (GRNNs). The experimental results demonstrated that the proposed GRNN-UKF approach outperformed all other strategies for target tracking.

### ***1.3 ANN-EKF-Based Runoff Estimation Model***

Ahmat Ruslan et al. [47] studied developing flood prediction models for the Sungai Kelang basin by comparing a backpropagation (BPNN) and Elman (ENN) neural networks that use the extended Kalman filter (EKF). The rainfall at flood location and water level at three upstream rivers were defined as model inputs, and the flood location at the downstream river is defined as outputs. This study's dataset resolution is for four days with 10 min time intervals, leading to 493 data points for training and validation. The developed networks had a single hidden layer and applied gradient descent (GD) or EKF with tangent sigmoid as a transfer function in the hidden and output layers and evaluated for a maximum of 1000 epochs of training. This study stated that using GD leads to reasonable results for the BPNN and ENN models with 10 and 15 neurons in the hidden layer, respectively. The researchers reported that using EKF in BPNN led to some improvements in tracking and filtering the nonlinearity of the BPN output that helped in reducing the RMSE. They also stated that using EKF led to further improvement in ENN and smoother flood water level prediction results by filtering out the nonlinearity of output data.

De Vos [48] compared a variety of ANNs for forecasting twelve river basins in the Eastern United States. This study used time series of daily precipitation, potential evaporation, discharge, and the 20-day simple moving average of the precipitation time series as the inputs to the ANN models. EKF was applied to three ANNs of Elman recurrent ANN (EL), Williams–Zipser fully recurrent ANN (WZ), and Williams–Zipser fully recurrent and fully feedforward ANN (WZFF) as training methods. This study indicated that for the EKF models, four neurons in the hidden layer were optimal. It also stated that the EKF training algorithm shows better performance on most basins.

Ahmat Ruslan et al. [49] studied developing flood prediction models for Klang River using neural network autoregressive with exogenous input (NNARX) and EKF. In these models, rainfall and water level were the inputs, and the water level was the output: 120, 78, and 170 data samples for training, validation, and testing of developed models, respectively. This study does not provide information about the details of the

models' neural networks. The NNARX converged at 83 training epochs. The authors stated that cascading the NNARX outputs to EKF helps filter the nonlinearity of the output data that led to improvements in the performance of NNARX. The author noted that the performance of the NNARX-EKF model is much better.

Karunasingha et al. [50] incorporated ANN models in EKF to enhance the prediction of chaotic flow river time series. The proposed model was applied to the Lorenz and Mackey Glass series's benchmark time series and to the five daily flow time series of the rivers: Mississippi River, Wabash River, Ciliwung River at Katulampa, Ciliwung River at Ratujaya, and Ciliwung River at Sugutamu. The ANN model used in this study was previously described in [51]. This single-layer MLP was used to develop models of up to 100 neurons and 50 training epochs with five sets of initial weights. Nguyen–Widrow initialization algorithm was applied to initialize weights and biases. The logistic sigmoid transfer function was used for all hidden neurons. Levenberg–Marquardt optimization method was used as the training algorithm to update weights and biases. This study adopted the ANN models as the state-space models in EKF to time delays different from 1 unit. The researchers reported that EKF produces an improved prediction for the benchmark time series, and for the river flow time series of Ciliwung with low average flows, however, prediction for three other flow series with large average flows did not improve [50].

Hosseini et al. [52] compared the extended Kalman filter-based neural network (EKFNN) and the gene expression planning (GEP) in estimating the runoff of the Malayer basin. The dataset used for this study had a daily resolution for runoff and rainfall from 2001 to 2013. This study does not provide information about the type, structure, and topology of the developed EKFNNs. This study stated that the EKFNN model was superior to GEP for this case study.

A well-designed neural network can learn and break down complex relationships with adequate training [53]. However, designing an efficient ANN is challenging. A faster and/or smaller ANN with acceptable accuracy can reduce both the length and costs of scientific investigations. Therefore, the training algorithm, the number of neurons in each layer, and the type of topology are some of the most important design parameters. Finding the optimal size of a neural network for a particular problem is an important concern in neural networks. If this size is too small, it may not be possible to train the neural network well to solve the problem. Likewise, overfitting may occur if it is too large [54], which may waste resources. Pruning is an approach used to determine the optimal size of a neural network [55]. Great generalization capacity and quick training speed are two other essential criteria for evaluating the performance of the learning techniques in ANNs [36]. Model accuracy and computational costs are two important factors in model development.

Many different parameters may be included in real phenomena, which increases the complexity of the problem. Using traditional methods could be expensive or technically infeasible. Solving a complex problem with a few parameters is almost impossible by many common models. The lack of data may make many studies infeasible, particularly for remote, hard to access or developing regions. To deal with such limitations, artificial neural networks (ANNs) could be used as a black box that relates the system inputs to the output(s). This approach could help in reducing costs

and making many projects feasible to study. ANNs have been utilized to solve a wide range of problems that are challenging in traditional programming [56]. However, ANNs sometimes treated as tools rather than modern, flexible mathematical models. This study provides an easy to understand and implement method to develop ANN-based estimation models using the multilayer perceptron (MLP) neural network and extended Kalman filter (EKF) or gradient descent (GD) as the training algorithms. Moreover, three methods to feed training data to the models were evaluated. Then, the feasibility of this method is evaluated by solving a runoff estimation problem that uses the minimum number of input parameters.

The basin used in this study was formerly studied using different ANNs [57–59]. Applying a different method for the same case study is a contribution to any study. The value of additional studies is to provide additional materials for future works. Moreover, it helps to compare this study's method with the previous methods. Previous studies on this basin used advanced temporal neural networks such as focused time-delay neural networks (FTDNNs) and FGam [57–59]. This study aims to develop an easy-to-implement and fast method to estimate runoff with a minimized, simple yet effective model using a minimum number of effective parameters: precipitation and temperature. These parameters are easy-to-measure and so basic that they are widely available in nearly all climatology or hydrology stations as well as other records. In this study, it is assumed that the developed models could estimate runoff efficiently if they could be trained well. EKF and GD are applied as the training algorithms, and three approaches are used to feed the training data to models.

In this study, an artificial neural network (ANNs) is used to model the basin as a black box that relates the system inputs (temperature and precipitation) to its output (runoff). The study area information is used for training and testing the developed MLP-based models that apply EKF or GD as training algorithms. The results are compared using several evaluation criteria to measure the performance of each model and approach. This study also investigates the effects of three approaches to feed the training data to the developed ANN-based models. Finally, the results of the most efficient and effective models have been summarized as well as the models with a minimum requirement for training.

The ANNs and methods used in former studies in this basin [57–59] were difficult to understand and implement due to their complex structures. Additionally, advanced data mining techniques were applied to achieve effective results. Moreover, the developed models were highly dependent on the time-series continuity as inputs which is a strict assumption for reliable results [57–59]. The proposed approach in this study is much easier to understand and implement, and it does not need data mining techniques or complex temporal ANNs such as FTDNN or FGam to solve the same problem. In addition, the final model is not dependent on a continuous long-term time series and estimates the runoff with the current data.

In this study, the developed runoff estimation models use a minimum number of required parameters: precipitation and temperature, i.e., just two easy-to-measure and the most accessible parameters in nearly all climatology or hydrology stations. Estimating the runoff is among the most important and challenging issues in water resources management/engineering, not only for development plans but also for



operation plans. There are many effective factors in this phenomenon, and the lack of adequate data makes many common models impractical. The non-scalable initial investigation costs are one of the small projects' biggest burdens [60, 61]. This study helps to reduce costs and make it possible to study possible developments in many new potential sites in remote, hard to access or developing regions.

## 2 Materials and Methods

### 2.1 Multilayer Perceptron

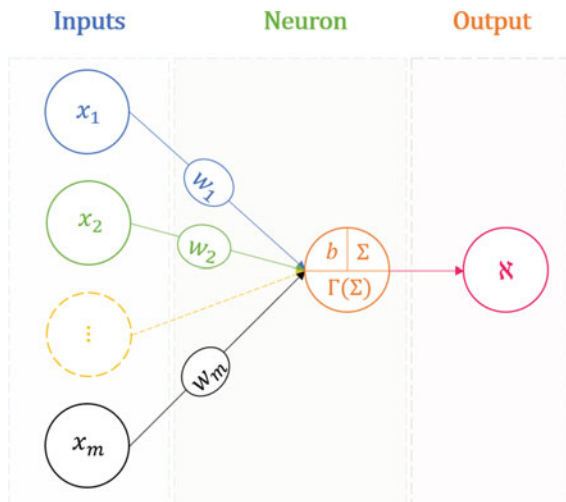
The perceptron is a simplified model of a biological neuron and is a linear model that can mimic some behavior seen in real neurons [62, 63]. This artificial neuron uses the Heaviside step function as the activation function. A perceptron neuron is represented in Fig. 1.

The single-layer perceptron is a linear classifier and the simplest feedforward neural network. It is a binary classifier learning algorithm known as the threshold (unit step) function. This function maps a vector of inputs ( $x$ ) to an output with a single binary value ( $\Gamma$ ). A simple perceptron is defined as [64, 65]:

$$\Sigma = \sum_{n=1}^m w_n x_n + b \quad (1)$$

$$\Gamma = \begin{cases} 1 & \Sigma \geq 0 \\ 0 & \Sigma < 0 \end{cases} \quad (2)$$

**Fig. 1** Perceptron neuron



where  $x$ ,  $\Gamma$ , and  $w$  have real values,  $m$  is the number of inputs,  $b$  is the bias, and  $w$  is the vector of weights. The bias does not depend on input values and allows for shifting the sigma function  $\Sigma$  (decision boundary) up or down [64]. The weights could describe the effectiveness of each node (here input).

The perceptron is a linear classifier and cannot distinguish data that is not linearly separable [66]; however, a multilayer perceptron (MLP) is able to do this [67]. The MLP is a feedforward ANN that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses an activation function [68, 69]. An MLP is distinguished from a linear perceptron by the multiple layers as well as a nonlinear activation function [67]. In perceptron,  $\Gamma$  is an activation (transfer) function called the ‘hard-limit’ or ‘unit step’ function that maps the  $\Sigma$  values into desired values (here 0 or 1). By replacing  $\Gamma$  with other activation functions,  $f(\Sigma)$  can map the neuron’s output into desired ranges. Therefore, in general form, the output of a neuron ( $\aleph$ ) with any activation function  $f$  could be calculated as

$$\aleph = f(\Sigma) = f\left(\sum_{n=1}^m w_n x_n + b\right) \quad (3)$$

Figure 2 shows the topology of a simple MLP. In this figure,  $x_m$ ,  $w_{m,n}^L$ ,  $b_n^L$ ,  $\Sigma_n^L$ , and  $f$  are the input, weight, bias, sigma function, and activation function, respectively;  $n$  is the number of the current neuron,  $m$  is the number of the neuron connected to the neuron  $n$ , and  $L$  is the number of layers. This figure could be described mathematically in the matrix form if for layer  $L$ , the matrixes of inputs ( $X^0$ ), weights ( $W^L$ ), biases ( $B^L$ ), sigma function ( $\Sigma^L$ ), and the neuron’s output ( $\aleph^L$ ) are defined as

$$X^0 = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix} \quad (4)$$

$$W^L = \begin{bmatrix} w_{1,1}^L & w_{1,2}^L & \cdots & w_{1,n}^L \\ w_{2,1}^L & w_{2,2}^L & \cdots & w_{2,n}^L \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1}^L & w_{m,2}^L & \cdots & w_{m,n}^L \end{bmatrix} \quad (5)$$

$$B^L = \begin{bmatrix} b_1^L \\ b_2^L \\ \vdots \\ b_n^L \end{bmatrix} \quad (6)$$

$$\Sigma^L = X^L \cdot W^L + B^L \quad (7)$$

$$\aleph^L = f(\Sigma^L) \quad (8)$$

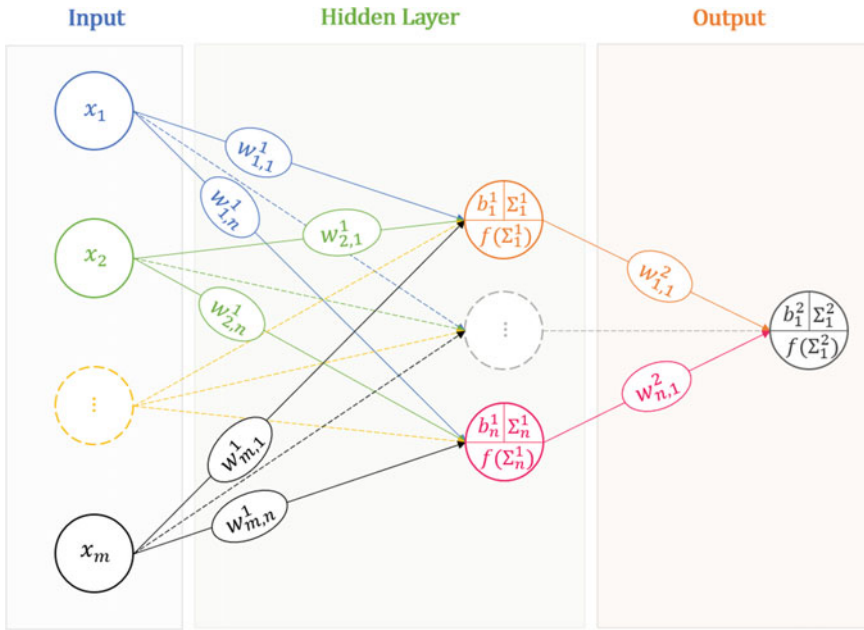


Fig. 2 Single-layer MLP

In a feedforward neural network, the output of each layer is the input of the next layer. Therefore, for layers after the first hidden layer, the neuron outputs of the former layer will be the input and

$$X^{L+1} = (\mathbb{N}^L)' \quad (9)$$

## 2.2 Learning Algorithms

For the neural network's output  $\Psi_i$  and the desired output  $D_i$ , the learning algorithm could be defined as a method for optimizing the weights in order to make  $\Psi_i \approx D_i$ . A supervised learning process of MLPs occurs by changing the connection weights after each training epoch. This is typically based on the output error (the difference between the output and the desired result).

### 2.2.1 Gradient Descent

In machine learning, backpropagation (backprop, BP) is a popular algorithm for supervised learning widely used to train feedforward neural networks [21, 70]. The

basic concept behind this approach is to use the chain rule repeatedly to calculate the effect of each weight in the network on an arbitrary error function  $E$  [71].

$$\frac{\partial E}{\partial w_{m,n}} = \frac{\partial E}{\partial \mathfrak{s}_n} \frac{\partial \mathfrak{s}_n}{\partial w_{m,n}} = \frac{\partial E}{\partial \mathfrak{s}_n} \frac{\partial \mathfrak{s}_n}{\partial \Sigma_n} \frac{\partial \Sigma_n}{\partial w_{m,n}} \quad (10)$$

By knowing the partial derivative for each weight,  $E$  could be minimized by using gradient descent [21, 72].

$$(w_{m,n})_{t+1} = (w_{m,n})_t - \eta \left( \frac{\partial E}{\partial w_{m,n}} \right)_t \quad (11)$$

where  $\eta$  is the learning rate, which scales the derivative and is an important factor in the required time for the convergence and needs to be determined reasonably: too large  $\eta$  values may prevent the error from falling below a certain value, and too small  $\in$  leads to too many iterations to reach to acceptable results [71]. It can be shown that

$$\frac{\partial E}{\partial w_{m,n}} = -(D_n - \Psi_n) f'(\Sigma_n) x_n \quad (12)$$

Combining Eqs. (12) and (13) leads to a special form of the backpropagation algorithm called the delta rule, which is a gradient descent (GD) learning algorithm (rule) [73] for updating the weights:

$$(w_{m,n})_{t+1} = (w_{m,n})_t + \eta (D_n - \Psi_n) f'(\Sigma_n) x_n \quad (13)$$

For linear activation functions, the derivative is equal to one, and it makes the delta rule similar to the perceptron's update rule (the delta rule cannot be used directly for the perceptron since its activation function's (Heaviside step function) derivative does not exist at zero and is equal to zero elsewhere.) [74, 75]:

$$(w_{m,n})_{t+1} = (w_{m,n})_t + \eta (D_n - \Psi_n) x_n \quad (14)$$

## 2.2.2 Extended Kalman Filter

Filters are a type of estimation strategy that is used to accurately extract knowledge of the states in the presence of noisy measurements and modeling uncertainties [76]. Kalman-based filters are the most well-known filters in estimation theory [77–79]. The Kalman filter (KF) is applicable for linear systems and yields the optimal estimation solution in the presence of known systems and white noise [77, 80]. The extended Kalman Filter (EKF) is formulated the same as the KF, except that first-order Taylor series expansions (Jacobian matrices) are used to linearize the nonlinearities.

The state function  $f$  and measurement function  $h$  are linearized to approximate the state and measurement error covariance matrices. Equations (15) and (16) are used to approximate the nonlinear systems [78].

$$F_k = \left. \frac{\partial f(x)}{\partial x} \right|_{x=\hat{x}_{k|k}, u_k} \quad (15)$$

$$H_{k+1} = \left. \frac{\partial h(x)}{\partial x} \right|_{x=\hat{x}_{k+1|k}} \quad (16)$$

Equations (17) and (18) are used to predict the state estimates and state error covariances, respectively [22]. The state error covariance found in Eq. (18) is used to calculate the innovation covariance in Eq. (19) and calculate the corresponding EKF gain in Eq. (20). The predicted state estimates in Eq. (17) are used in conjunction with the EKF gain found in Eq. (20) to update the state estimates as per Eq. (21). Note that the measurement error is also used in Eq. (21). Finally, the state error covariance matrix is updated as per Eq. (22).

$$\hat{x}_{k+1|k} = f(\hat{x}_{k|k}, u_k) \quad (17)$$

$$P_{k+1|k} = F_k P_{k|k} F_k^T + Q_k \quad (18)$$

$$S_{k+1} = H_{k+1} P_{k+1|k} H_{k+1}^T + R_{k+1} \quad (19)$$

$$K_{k+1} = P_{k+1|k} H_{k+1}^T S_{k+1}^{-1} \quad (20)$$

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}(z_{k+1} - h(\hat{x}_{k+1|k})) \quad (21)$$

$$P_{k+1|k+1} = (I - K_{k+1} H_{k+1}) P_{k+1|k} (I - K_{k+1} H_{k+1})^T + K_{k+1} R_{k+1} K_{k+1}^T \quad (22)$$

Equations (15) through (22) represent the EKF estimation process. The process is iterative, meaning that the values are found in Eqs. (21) and (22) are used again in Eqs. (17) and (18). Note that other nonlinear KF variants have been developed, such as the unscented Kalman filter (UKF). However, the EKF was found to provide accurate results for this case study [22].

In order to use EKF as the training algorithm of the neural network, it is needed to define the system's dynamic relationships in the form of state-space by defining the weights as system states and the neuron output as the system output as:

$$(w_{m,n})_{t+1} = (w_{m,n})_t + N(0, Q) \quad (23)$$

$$\mathfrak{N}_{t+1} = \mathfrak{N}(x_{t+1}, w_{t+1}) + N(0, R) = f(y)|_{y=\Sigma_{t+1}} + N(0, R) \quad (24)$$

where  $N(0, Q)$  is the process noise with the covariance of  $Q$ , and  $N(0, R)$  is the measurement noise with the covariance of  $R$ . The weights could be updated by knowing the Kalman gain:

$$K_{t+1} = P_t H' (H P_t H' + R)^{-1} \quad (25)$$

where  $P$  and  $H$  are the covariance matrix and Jacobian vector, respectively, and:

$$P_{t+1} = P_t (I - K_{t+1} H') + Q_t \quad (26)$$

where  $I$  is the identity matrix, and elements of vector  $H$  could be calculated as:

$$(H_n)_{t+1} = f(y) \frac{\partial f(y)}{\partial y} \big|_{y=\Sigma_{t+1}} \quad (27)$$

Therefore, the weights could be updated by the following equation:

$$(w_{m,n})_{t+1} = (w_{m,n})_t + K_{t+1} (D_n - \Psi_n) \quad (28)$$

### 2.3 Evaluation Criteria

A correlation refers to any statistically significant relationship between two variables [81]. According to the definition, to compare the results of model outputs (predictions) with observed data (observations or measurements) for a statistical sample with  $n$  couples  $(O_i, P_i)$ , the correlation could be calculated by the following equation:

$$R = \frac{\sum_{i=1}^n (O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^n (O_i - \overline{O})^2} \sqrt{\sum_{i=1}^n (P_i - \overline{P})^2}} \quad (29)$$

Note that  $O_i$  is the observed data,  $P_i$  is the estimated value,  $\overline{O}$  is the average of observational data,  $\overline{P}$  is the average of estimated data, and  $n$  is the number of data. The correlation coefficient values could be in a range between  $-1$  and  $+1$ . Correlation close to  $+1$  means a good and direct correlation between two datasets. Correlations close to  $-1$  mean a good but reverse relation between datasets. A correlation close to zero implies a lack of correlation. The range of  $R^2$  is between 0 and 1. Therefore, a higher value indicates better relation between datasets.

The mean absolute error (MAE) is a criterion that compares the predicted results to the desired or actual results. This criterion is calculated using the following equation [82]:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \quad (30)$$

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator measures the average of the squares of the errors. It is the average squared difference between the estimated values and desired or actual values. The MSE is a measure of the quality of an estimator. It is always non-negative, and smaller values indicate better performance [83].

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2 \quad (31)$$

Root mean square deviation (RMSD) or root mean square error (RMSE) is a common measuring criterion calculated from the difference between the predicted values by a model or estimator and the observed data. It indicates the sample standard deviation from the predicted values and the experimental data. These differences are called residual when calculations are estimated from samples and are called forecast error when they are predicted out of sample. RMSE is an acceptable measure to compare the prediction errors of a special variable and is calculated using the following equation [83]:

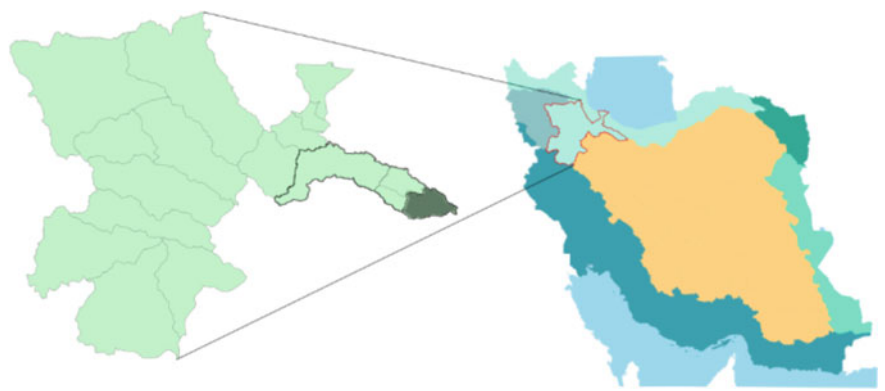
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2} \quad (32)$$

## 2.4 Study Area and Parameters

According to the water master plan of Iran, the great SefidRood (White River) basin is divided into 17 sub-basins. The Taleghan upper basin is located in the center of the Alborz Mountains and the east of the SefidRood basin. Figure 3 shows the Taleghan upper basin location. The basin's area is 960 Km<sup>2</sup>, representing 2.5% of this basin [59, 84]. The most important feature of this basin is the high altitude and steep slope. The slope of about half of this basin is greater than 40%, and its general direction is east–west. Its maximum and minimum heights are 4300 m and 1390 m above sea level, respectively. The average elevation is 2665 m [59, 85]. The length of the Taleghan River is about 180 km [59, 86].

Table 1 represents the longitude and latitude of the Taleghan basin. The Taleghan earth dam is located at the end with a capacity of 420 CMC and an area of 12 km<sup>2</sup> [59, 87, 88].

Over 40 years of hydrometry and climatology records were used for this study. The original records were provided by Iran Water Resources Management Company



**Fig. 3** Taleghan upper basin location [57–59, 85]

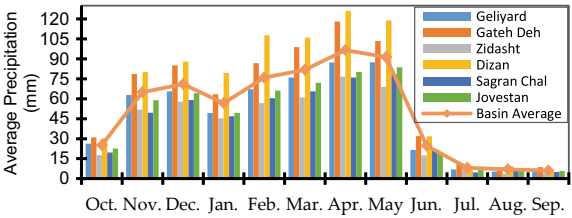
**Table 1** Specifications of stations in the upstream basin of Taleghan [57–59, 85]

Station name	Station type	Longitude	Latitude	Elevation (m)	Period
Galidar	Rain gage	50°:51′	36°:08′	2150	1966–2008
Gateh deh	Rain gage	51°:04′	36°:10′	2600	1966–2011
Zidasht	Climatology	51°:18′	31°:45′	2000	1969–2011
Dizan	Rain gage	50°:50′	36°:16′	3200	1967–2011
Sakratchal	Rain gage	50°:44′	36°:17′	2200	1966–2011
Jovestan	Rain gage	50°:41′	36°:10′	1850	1966–2011
Galinak	Hydrometry	50°:40′	36°:10′	1783	1958–2011

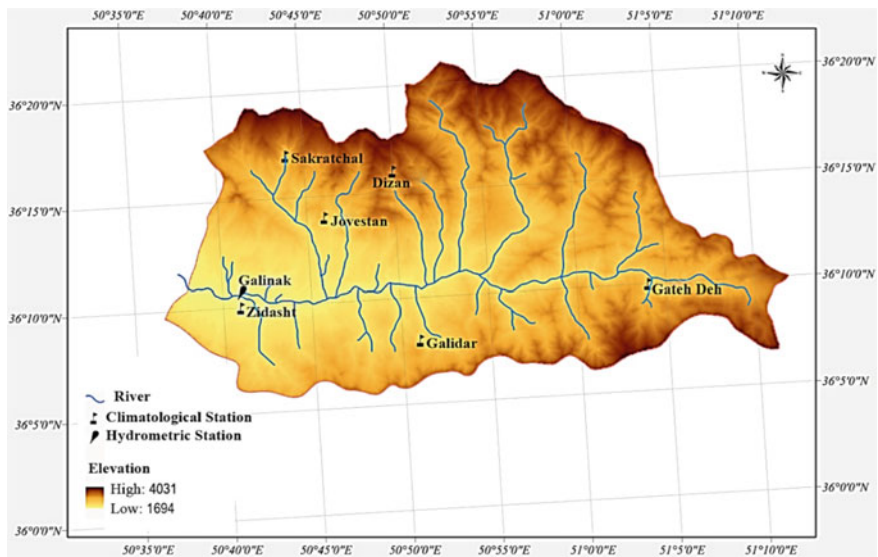
(WRM) and the Committee of Dam and Basin Management (CDBM), IRCOLD. For each station, the period of available records is represented in Table 1. Then, the records were analyzed and evaluated to be used for developing a long-term dataset. In this dataset, monthly average data of precipitation and temperature were gathered from six climatology stations: Gateh Deh, Dizan, Galidar, Jovestan, Sakratchal, and Zidasht. Discharge data were collected from a hydrometry station named Galinak. The locations of these stations are shown in Fig. 5 (Fig. 4).

The distribution of precipitation differs from 250 to 1000 mm/year in different locations of this basin. Regarding the precipitation amount and catchment height, the

**Fig. 4** Long-term average precipitation in each station and whole of the basin [57–59, 85]







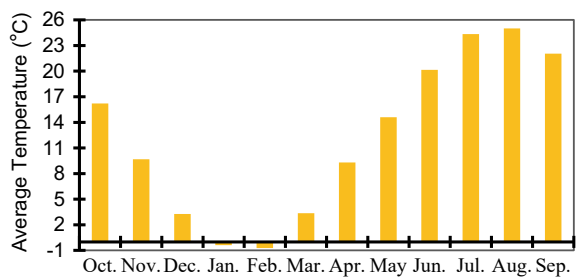
**Fig. 5** Climatological and hydrometric stations in the upstream basin of Taleghan [57–59, 85]

Jovestan, Gatch Deh, and Galidar stations have a high impact on average precipitation and on the basin’s runoff [59].

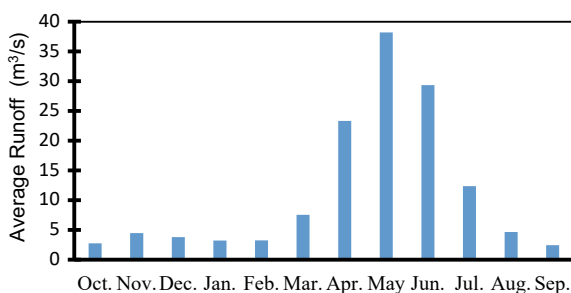
According to Fig. 4, the long-term annual average of precipitation of the Taleghan upper basin in this period is 600 mm. The maximum precipitation occurs in April and May with an average of around 126 mm/year and 119 mm/year, respectively. The minimum precipitation occurs from July to September with an average of below 8 mm/year. The precipitation in winter is more than 45 mm/year in all stations.

The average elevation of the basin is 3722 m. Since the Zidasht station has a similar elevation (2000 m), the monthly average data of this station’s temperature is considered the basis of calculations. The maximum and minimum temperatures are 37 °C in July and 18 °C in March. This basin’s long-term monthly average temperature is 7.8 °C [59]. By using the temperature gradient and height relationship, the long-term average temperature of the basin is calculated and represented in Fig. 6.

**Fig. 6** Long-term average temperature [57–59, 85]



**Fig. 7** Long-term average runoff [57–59, 85]



Since the Galinak station is located at the entrance of the Taleghan Dam's reservoir, it could be considered as the outlet of the basin. The data of this point could be considered as the representative of all basin discharges. Therefore, the runoff of this basin can be calculated by subtracting the baseflow from discharge [59]. According to Fig. 7, the highest runoff occurs from April to June, and from September to February, the runoff is at its lowest. The long-term average runoff of this basin shows that the maximum and minimum runoff of this basin is about  $38 \text{ m}^3/\text{s}$  and  $2.4 \text{ m}^3/\text{s}$ , which occur in May and September, respectively.

### 3 Model Structure

In order to develop an ANN-based runoff estimation model, the temperature and precipitation were considered as inputs and runoff as the output, and the dataset was randomly disturbed. Therefore, the 40 years of monthly observed data were randomized and then divided into three datasets: 60% for training, 20% for cross-validation (C.V.), and 20% for test. These datasets were saved separately to ensure the training, C.V., and test data were the same for all developed models.

A feedforward MLP neural network was used as the base model. Two sub-models were developed: one using EKF for training and one using GD for training. All experiments were completed with a single-layer MLP with 1–10 neurons and 1,000 training epochs. For evaluation of the training and model's performance, three evaluation criteria were used: correlation ( $R$ ), mean absolute error (MAE), and mean square error (MSE).

The weights were updated by either the EKF or GD algorithms during each training process. After each training epoch, cross-validation was completed to check the model performance. New weights were saved based on an increase of  $R^2$  or decrease in error (MAE). The corresponding epoch number and evaluation criteria values were saved and considered the 'best epoch.' A unique 'minimum epoch of the best' (MEB) was recorded for each evaluation criteria, and the related weights were used for testing. The following three approaches were considered for the training process.

### ***3.1 Approach 1: Single Dataset***

*In this approach, all training was completed with a randomized but single dataset. Therefore, the order of inputs and desired outputs would be the same in the training queue.*

### ***3.2 Approach 2: Single, Randomized Dataset in Case of No Improvement***

In this approach, initial training was completed with a training dataset as per the first approach. After each cross-validation, in case of no improvement in the model results, the next training was completed using a new randomized version of the training dataset. The randomized datasets were the same for all models. Therefore, the results of this approach could be compared with similar models.

### ***3.3 Approach 3: Randomized Dataset for Each Training Epoch***

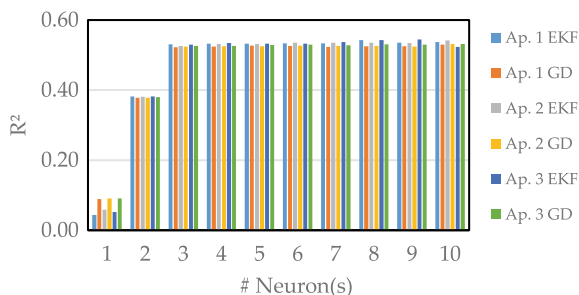
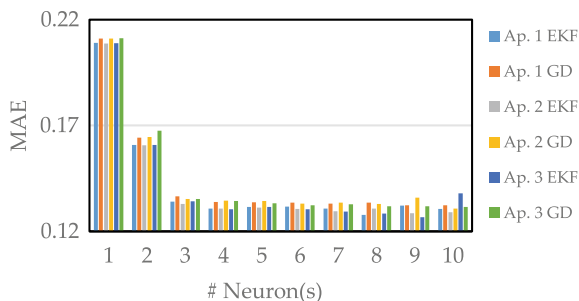
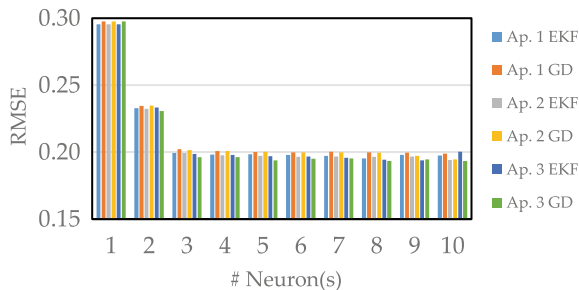
In this approach, each training process was completed using a new randomized version of the training dataset. To ensure all models were trained in a comparable manner, the number of randomized training datasets was equal to the maximum number of training epochs.

The combination of the three approaches and evaluation using up to ten neurons in the hidden layer resulted in a total of 60 models.

## **4 Results and Analysis**

An ideal model has a maximum correlation and minimum error between the predicted and observed data. The higher values of  $R^2$  and lower values of RMSE and MAE were considered to compare and rank the developed models. Also, the number of neurons and computational complexity is considered. The comparison of  $R^2$ , MAE, and RMSE of the best models is represented in Figs. 8, 9 and 10, respectively.

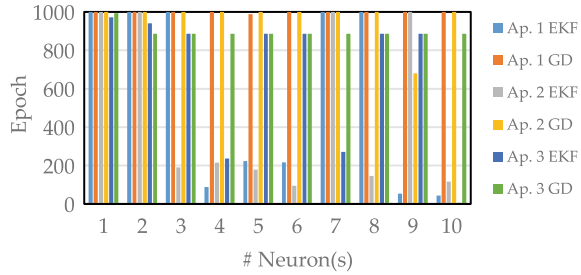
For Approach 1, an EKF-based ANN with eight neurons and an epoch of the best (EB) equal to 1000 was considered the best model. For Approach 2, the best model was EKF-based with ten neurons and an EB equal to 116. This was 8.62 times (862%) less than the same EB for the GD-based model. Finally, for Approach 3, an EKF-based model with nine neurons and an EB equal to 886 was considered the best

**Fig. 8**  $R^2$  of the best models**Fig. 9** MAE of the best models**Fig. 10** RMSE of the best models

model. In addition, this model was considered the best among all of the 60 different developed models used to estimate runoff.

As shown in Fig. 8, the comparison of the  $R^2$  of the best models demonstrates that for one neuron, GD performs better than the EKF method. However, for one neuron, the value of  $R^2$  is quite low that no model could be considered effective. For models with three neurons or more, the results are nearly the same. However, the EKF demonstrates minor improvements over the GD method. Interestingly, for this case, it appears that using more than three neurons does not yield substantial improvements in model accuracy. In order to lower computational demand, ANNs could be designed with only three neurons as there appears to be a diminishing return (less than 4% improvement when using ten neurons).

**Fig. 11** EB of the best models



According to Fig. 9, although the differences are small, the MAE of nearly all EKF-based ANNs were less than GD-based ANNs. According to the MAE evaluation criteria, this case’s optimum model could be considered an EKF-based ANN with four neurons.

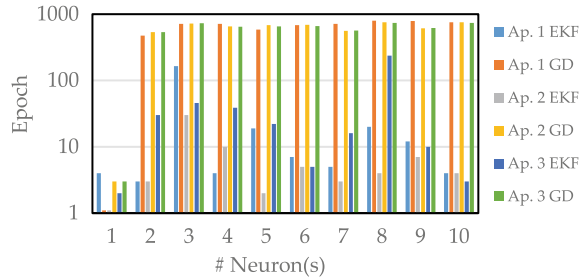
As shown in Fig. 10, the RMSE of EKF-based ANNs were less than GD-based ones for 1–10 neurons for Approaches 1 and 2. The opposite is true for Approach 3. For more than three neurons, the difference between the highest and lowest RMSE is less than 5%, and the RMSE of all EKF-based ANNs is less than 0.2. Therefore, according to the RMSE evaluation criteria, the optimum model was an EKF-based ANN with three neurons.

According to Fig. 11, for Approach 1, EKF-based ANNs required 4.48–22.72 times (448–2272%) fewer training epochs than GD-based ANNs for 4–6, 9, and 10 neurons. For Approach 2, EKF-based ANNs required 4.65–10.52 times (465–1052%) fewer training epochs than GD-based ANNs for 3–6, 8, and 10 neurons. For Approach 3, EKF-based ANNs required 3.74–177.2 times (374–17,720%) fewer training epochs than GD-based ANNs for 4, 7, and 10 neurons. Therefore, Approach 2 was the most effective in reducing the number of training epochs for the EKF-based ANNs training process.

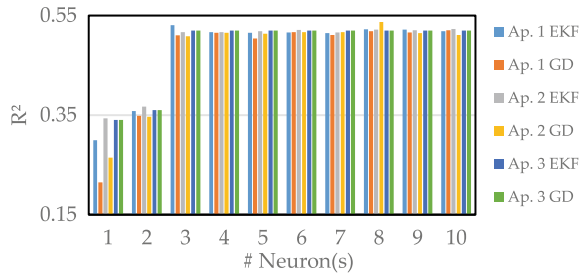
According to Fig. 12, the minimum epoch of the best (MEB) for the EKF-based ANN models were so small that a logarithmic graph is needed. The MEB of EKF-based ANNs was smaller than the GD-based ANNs. On average, the MEB of the EKF was 29.37 or 24.44 times (2444%) fewer than GD for Approach 1. For Approach 2, the MEB of the EKF was 8.12 or 83.45 times (8345%) fewer than GD. Finally, for Approach 3, the MEB of the EKF was 47.37 or 12.43 times (1243%) fewer than GD. Therefore, Approach 2 was considered the best approach for training the EKF-based ANNs in the fewest number of epochs. In addition, these results indicate that the EKF yielded better results in all of the approaches.

According to Fig. 13, for the  $R^2$  evaluation criteria, the results for one and two neurons are substantially poor and can be dismissed. For three neurons or more, the values are very similar and vary between 0.50 and 0.52. However, on average, the EKF has a better performance than GD by 5.06% and 3.98% in Approaches 1 and 2, respectively. For Approach 3, the results of the EKF and GD were nearly identical. In this case, Approach 1 is considered the best approach for training EKF-based ANNs when considering  $R^2$  as a validation criterion.

**Fig. 12** MEB among all models



**Fig. 13**  $R^2$  of the models with MEB

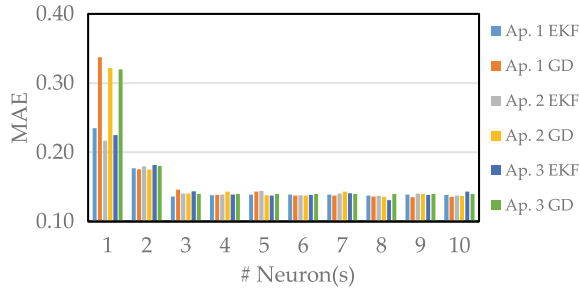


The results shown in Fig. 14 for MAE indicate that the use of 3 or more neurons yields similar errors (less than 0.15 MAE). Although the overall differences between EKF and GD are minimal; on average, the EKF method works better by 3.14% in Approach 1, 2.84% in Approach 2, and 3.53% in Approach 3. According to the MAE evaluation criteria, the performance of EKF-based Approach 3 may be a slightly better choice to train ANNs.

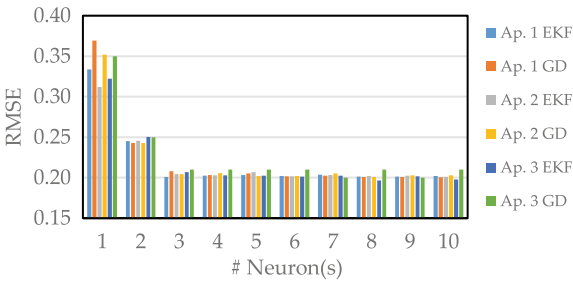
According to Fig. 15, for more than three neurons, RMSE of both methods was about 0.2. The differences between the EKF and GD methods were nearly insignificant for Approaches 1 and 2 (about 1% better for EKF). In Approach 3, the EKF was slightly better than the GD method by about 3.1%.

As shown in Figs. 13, 14 and 15, the results of ANNs with three or more neurons are so similar that the RMSE, MAE, and  $R^2$  differences are nearly negligible. In this case, an EKF-based ANN with 3 or 4 neurons is considered as the MEB optimum

**Fig. 14** MAE of the models with MEB



**Fig. 15** RMSE of the models with MEB



model. Table 3 summarizes the evaluation criteria for the best, optimum, and MEB optimum models. These results indicate that the EKF-based ANN model required a considerably smaller neural network size and fewer training epochs, with a reduction of performance by 2–4% (Table 2).

Table 2 represents the minimum epoch of the best among all the models. The required number of training epochs to achieve the best result is significantly different between GD and EKF. This difference is further highlighted when looking at the difference in terms of percentage.

**Table 2** Minimum epoch of the best among all models

#Neurons	Approach 1		Approach 2		Approach 3	
	GD-EKF	GD/EKF [%]	GD-EKF	GD/EKF [%]	GD-EKF	GD/EKF [%]
1	3	25	2	300	1	150
2	474	15,900	533	17,867	505	1783
3	551	436	695	2417	683	1585
4	707	17,775	641	6510	609	1662
5	569	3095	681	34,150	632	2973
6	674	9729	683	13,760	658	13,260
7	706	14,220	556	18,633	553	3556
8	778	3990	749	18,825	498	309
9	776	6567	607	8771	607	6170
10	749	18,825	747	18,775	733	24,533

**Table 3** Best, optimum, and MEB optimum models specifications

Model	Approach	Training algorithm	#Neurons	EB	$R^2$	MAE	RMSE
BEST	3	EKF	9	886	0.544	0.127	0.194
OPTIMUM	3	EKF	4	237	0.534	0.130	0.198
MEB OPTIMUM	1	EKF	4	4	0.516	0.138	0.203

According to Table 3, the best, optimum, and MEB optimum ANN-based models are using the EKF training algorithm. The validation criteria indicate a slight difference between the accuracy of all models in this list. The accuracy of the best and optimum models is very close. However, comparing the required training epochs to achieve such an accuracy (epoch of the best, EB) indicates that a reasonable accuracy was achieved in the optimum model using 3.74 times (373.78%) fewer training epochs. Moreover, the optimum model could achieve such accuracy using 2.25 times (225%) fewer neurons, which makes this network smaller than the best model. Based on these results, the optimum model offers the most reasonable balance between the training epochs, network size, and accuracy.

## 5 Conclusions

This study provides an easy to understand and implement method to develop ANN-based estimation models using MLP and EKF or GD as the training algorithm. To study these models and the three approaches for feeding training data in optimizing the developed feedforward neural networks in reducing the training epochs and size of the network, a runoff estimation problem was considered, which is an important and challenging issue in water resources management. Two groups of one-layered feedforward ANNs were developed with 1–10 neurons, and three approaches for handling the training data were applied, which resulted in 60 models. Model estimations were evaluated by  $R^2$ , RMSE and MAE. For this study, results indicate that the GD method required 11–67% more neurons than the EKF for a relatively similar performance. In terms of training time, the ANNs applied EKF as the training algorithm required 300–34,200% fewer training epochs than GD. Based on the results, in comparison with GD, using EKF as a learning algorithm improved estimation models by reducing the number of training epochs and size of ANNs. Further studies are recommended.

**Author Contributions** Conceptualization, A.Y. and S.A.G.; Investigation, A.Y. and S.M.T.; Methodology, A.Y.; Project administration, S.A.G.; Software, S.M.T.; Supervision, S.A.G. and W.D.L.; Validation, S.M.T.; Visualization, A.Y.; Writing—original draft, A.Y. and S.M.T.; Writing—review and editing, A.Y., S.A.G., W.D.L. and M.K. All authors have read and agreed to the published version of the manuscript.

## References

1. Moore RJ (1985) The probability-distributed principle and runoff production at point and basin scales. *Hydrol Sci J* 30(2):273–297. <https://doi.org/10.1080/02626668509490989>
2. Quan Z, Teng J, Sun W, Cheng T, Zhang J (2015) Evaluation of the HYMOD model for rainfall-runoff simulation using the GLUE method. *IAHS-AISH Proc* 368(51579007):180–185. <https://doi.org/10.5194/piahs-368-180-2015>



3. Yoosefdoost I, Khashei-Siuki A, Tabari H, Mohammadrezapour O (2022) Runoff simulation under future climate change conditions: performance comparison of data-mining algorithms and conceptual models. *Water Resour Manag* 36(4):1191–1215. <https://doi.org/10.1007/s11269-022-03068-6>
4. Russell SJ, Norvig P (2009) Artificial intelligence: a modern approach. Prentice Hall. Retrieved from [https://www.researchgate.net/publication/235890207\\_Artificial\\_Intelligence\\_A\\_Modern\\_Approach\\_Prentice\\_Hall](https://www.researchgate.net/publication/235890207_Artificial_Intelligence_A_Modern_Approach_Prentice_Hall). Accessed on 27 Feb 2019
5. Apter MJ, McCorduck P (2006) Machines who think: a personal inquiry into the history and prospects of artificial intelligence. *Leonardo* 15(3):242. <https://doi.org/10.2307/1574702>
6. Nilsson NJ (1998) Artificial intelligence : a new synthesis. Morgan Kaufmann Publishers. Retrieved from <https://dl.acm.org/citation.cfm?id=280491>. Accessed on 27 Feb 2019
7. Poole DL, Mackworth AK, Goebel R (1998) Computational intelligence : a logical approach. Oxford University Press
8. Ligeza A (1995) Artificial intelligence: a modern approach. *Neurocomputing* 9(2):215–218. [https://doi.org/10.1016/0925-2312\(95\)90020-9](https://doi.org/10.1016/0925-2312(95)90020-9)
9. Kohavi R, Provost F (1998) Glossary of terms. *Mach Learn* 30(2/3):271–274. <https://doi.org/10.1023/A:1017181826899>
10. Sarve A, Sonawane SS, Varma MN (2015) Ultrasound assisted biodiesel production from sesame (*Sesamum indicum* L.) oil using barium hydroxide as a heterogeneous catalyst: comparative assessment of prediction abilities between response surface methodology (RSM) and artificial neural network (ANN). *Ultrason Sonochem* 26:218–228. <https://doi.org/10.1016/J.ULTSONCH.2015.01.013>
11. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5(4):115–133. <https://doi.org/10.1007/BF02478259>
12. Rojas R (1996) Neural networks: a systematic introduction. Springer, Berlin, p 29. <https://doi.org/10.7312/zuri90466-007>
13. Hebb DO (2009) The organization of behavior: a neuropsychological theory. Taylor & Francis, Oxfordshire
14. Farley B, Clark W (1954) Simulation of self-organizing systems by digital computer. *Trans IRE Prof Gr Inf Theory* 4(4):76–84. <https://doi.org/10.1109/TIT.1954.1057468>
15. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65–386. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.588.3775>. Accessed on 22 Feb 2019
16. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117. <https://doi.org/10.1016/J.NEUNET.2014.09.003>
17. Ivakhnenko A (1973) Cybernetic predicting devices. CCM Information Corp., New York. Retrieved from <https://www.worldcat.org/title/cybernetic-predicting-devices/oclc/219866001?referer=di&ht=edition>. Accessed on 23 Feb 2019
18. Ivakhnenko A (1967) Cybernetics and forecasting techniques. American Elsevier Pub. Co., New York. Retrieved from <https://www.worldcat.org/title/cybernetics-and-forecasting-techniques/oclc/537162>. Accessed on 23 Feb 2019
19. Minsky M, Papert S (1969) Perceptrons: an introduction to computational geometry. MIT Press, Cambridge
20. Werbos PJ (1975) Beyond regression: new tools for prediction and analysis in the behavioral sciences. Harvard University, Cambridge
21. Riedmiller M, Braun H (1993) Direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: 1993 IEEE international conference on neural networks, pp 586–591. <https://doi.org/10.1109/icnn.1993.298623>
22. Gadsden A, Habibi S, Dunne D, Kirubarajan T (2012) Nonlinear estimation techniques applied on target tracking problems. *J Dyn Syst Meas Control* 134(5):054501. <https://doi.org/10.1115/1.4006374>
23. Afshari HH, Gadsden SA, Habibi S (2017) Gaussian filters for parameter and state estimation: a general review of theory and recent trends. *Signal Process* 135:218–238. <https://doi.org/10.1016/J.SIGPRO.2017.01.001>

24. Haykin SS (1994) *Neural networks: a comprehensive foundation*. Macmillan, New York City
25. Astrom KJ, Wittenmark B (1994) *Adaptive control*, 2nd edn. Addison-Wesley, Boston
26. Ljung L, Söderström T (1983) *Theory and practice of recursive identification*. MIT Press
27. Anderson BDO, Moore JB (1979) Optimal filtering. In: *Random processes for image and signal processing*. Englewood Cliffs, New Jersey, pp 307–482. <https://doi.org/10.1117/3.268105.ch4>
28. Söderström T (2002) *Discrete-time stochastic systems*. Springer London, London. <https://doi.org/10.1007/978-1-4471-0101-7>
29. Matthews MB (1990) Neural network nonlinear adaptive filtering using the extended Kalman filter algorithm. In: *Proceedings of the international neural networks conference*, vol 1, pp 115–119. Retrieved from <https://ci.nii.ac.jp/naid/10004332129/>. Accessed on 09 Mar 2019
30. Shah S, Palmieri F, Datum M (1992) Optimal filtering algorithms for fast learning in feedforward neural networks. *Neural Netw* 5(5):779–787. [https://doi.org/10.1016/S0893-6080\(05\)80139-X](https://doi.org/10.1016/S0893-6080(05)80139-X)
31. Singhal S, Wu L (1989) Training multilayer perceptrons with the extended Kalman algorithm. Retrieved from <http://papers.nips.cc/paper/101-training-multilayer-perceptrons-with-the-extended-kalman-algorithm.pdf>. Accessed on 09 Mar 2019
32. Iiguni Y, Sakai H, Tokumaru H (1992) A real-time learning algorithm for a multilayered neural network based on the extended Kalman filter. *IEEE Trans Signal Process* 40(4):959–966. <https://doi.org/10.1109/78.127966>
33. Ruck DW, Rogers SK, Kabrisky M, Maybeck PS, Oxley ME (1992) Comparative analysis of backpropagation and the extended Kalman filter for training multilayer perceptrons. *IEEE Trans Pattern Anal Mach Intell* 14(6):686–691. <https://doi.org/10.1109/34.141559>
34. Puskorius GV, Feldkamp LA (1994) Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks. *IEEE Trans Neural Netw* 5(2):279–297. <https://doi.org/10.1109/72.279191>
35. Plumer ES (1995) Training neural networks using sequential extended Kalman filtering. In: *1995 world Congress on neural networks*, Washington, DC (United States). Retrieved from <https://digital.library.unt.edu/ark:/67531/metadc678272/>. Accessed 01 Mar 2019
36. Sum JPF (1998) Extended Kalman filter based pruning algorithms and several aspects of neural network learning. The Chinese University of Hong Kong
37. Sum J, Leung C-S, Young GH, Kan W-K (1999) On the Kalman filtering method in neural network training and pruning. *IEEE Trans Neural Netw* 10(1):161–166. <https://doi.org/10.1109/72.737502>
38. Williams RJ (1992) Training recurrent networks using the extended Kalman filter. In: *Proceedings of 1992 IJCNN international joint conference on neural networks*, vol 4, pp 241–246. <https://doi.org/10.1109/ijcnn.1992.227335>
39. Suykens JAK, De Moor BLR, Vandewalle J (1995) Nonlinear system identification using neural state space models, applicable to robust control design. *Int J Control* 62(1):129–152. <https://doi.org/10.1080/00207179508921536>
40. Puskorius GV, Feldkamp LA (2002) Decoupled extended Kalman filter training of feedforward layered networks. In: *IJCNN-91-Seattle international joint conference on neural networks*, vol 1, pp 771–777. <https://doi.org/10.1109/ijcnn.1991.155276>
41. Caliskan F, Aykan R, Hajiyev C (2008) Aircraft icing detection, identification, and reconfigurable control based on Kalman filtering and neural networks. <https://doi.org/10.1061/ASCE0893-1321200821251>
42. Leung CS, Chan LW (2003) Dual extended Kalman filtering in recurrent neural networks. *Neural Netw* 16(2):223–239. Retrieved from [www.elsevier.com/locate/neunet](http://www.elsevier.com/locate/neunet). Accessed on 01 Mar 2019
43. Jondhale SR, Deshpande RS (2019) Kalman filtering framework-based real time target tracking in wireless sensor networks using generalized regression neural networks. *IEEE Sens J* 19(1):224–233. <https://doi.org/10.1109/JSEN.2018.2873357>
44. Pietruszkiewicz W (2010) A comparison of nonlinear Kalman filtering applied to feedforward neural networks as learning algorithms. In *2010 IEEE 9th international conference on cybernetic intelligent systems*, pp 1–6. <https://doi.org/10.1109/UKRICIS.2010.5898137>

45. Kurylyak Y, Barbe K, Lamonaca F, Grimaldi D, Van Moer W (2013) Photoplethysmogram-based blood pressure evaluation using Kalman filtering and neural networks. In: 2013 IEEE international symposium on medical measurements and applications (MeMeA), pp 170–174. <https://doi.org/10.1109/MeMeA.2013.6549729>
46. Krok A (2013) The development of Kalman filter learning technique for artificial neural networks. *J Telecommun Inf Technol* 2013(4):16–21
47. Ahmat Ruslan F, Adnan R, Manan Samad A, Md Zain Z (2013) Flood prediction modeling using hybrid BPN-EKF and hybrid ENN-EKF: a comparative study. *Zainazlan Md Zain/Int J Eng Res Appl* 3(4):290–297. [www.ijera.com](http://www.ijera.com)
48. de Vos Kamerlingh NJ (2013) Hydrology and earth system sciences echo state networks as an alternative to traditional artificial neural networks in rainfall-runoff modelling. *Hydrol Earth Syst Sci* 17:253–267. <https://doi.org/10.5194/hess-17-253-2013>
49. Ahmat Ruslan F, Samad AM, Adnan R (2017) Modelling of flood prediction system using hybrid NNARX and extended Kalman filter. In: Proceedings—2017 IEEE 13th international colloquium on signal processing and its applications, CSPA 2017, no March, pp 149–152. <https://doi.org/10.1109/CSPA.2017.8064941>
50. Santhusitha D, Karunasingha K, Liong S-Y (2018) Enhancement of chaotic hydrological time series prediction with real-time noise reduction using extended Kalman filter. <https://doi.org/10.1016/j.jhydrol.2018.08.044>
51. Karunasinghe DSK, Liong SY (2006) Chaotic time series prediction with a global model: artificial neural network. *J Hydrol* 323(1–4):92–105. <https://doi.org/10.1016/J.JHYDROL.2005.07.048>
52. Hosseini A, Golabi MR, Marofi S, Khaledian N, Solatani M (2020) Evaluation of extended Kalman filter-based neural network (EKFNN) model and gene expression planning in rainfall-runoff modelin. *Watershed Eng Manag* 12(3):771–784. <https://doi.org/10.22092/IJWMSE.2019.121031.1457>
53. Wang JJ, Wang J, Sinclair D, Watts L (2006) A neural network and Kalman filter hybrid approach for GPS/INS integration. In: 12th IAIN congress on 2006 international symposium, vol 3, p 3
54. Moody JE (1991) Note on generalization, regularization, and architecture selection in nonlinear learning systems. In: *Neural networks for signal processing*, pp 1–10. <https://doi.org/10.1109/nnspp.1991.239541>
55. Reed R (1993) Pruning algorithms-a survey. *IEEE Trans Neural Netw* 4(5):740–747. <https://doi.org/10.1109/72.248452>
56. Chen N et al (2016) Automatic detection of pearlite spheroidization grade of steel using optical metallography. *Microsc Microanal* 22(01):208–218. <https://doi.org/10.1017/S1431927615015706>
57. YoosefDoost A, Sadeghian MS, Bazargan Lari MR (2014) Analysis and evaluation of using artificial parameters generated by data mining in runoff estimation by neural networks considering to the climate change. Retrieved from <https://civilica.com/doc/319125/>
58. YoosefDoost A, Sadeghian MS, Bazargan Lari MR (2014) Analysis and evaluation the inputs which provided from data mining and RPROP learning algorithm in optimization FTDNN and FGam artificial neural networks. Retrieved from <https://civilica.com/doc/319126/>
59. YoosefDoost A, Sadeghian MS, Bazargan Lari MR (2014) Data mining and optimization of runoff estimation by artificial neural networks. Retrieved from <https://en.civilica.com/doc/319126/>
60. YoosefDoost A, Lubitz WD (2021) Design guideline for hydropower plants using one or multiple archimedes screws. *Processes* 9(12):2128. <https://doi.org/10.3390/pr9122128>
61. YoosefDoost A, Lubitz WD (2021) Archimedes screw design: an analytical model for rapid estimation of Archimedes screw geometry. *Energies* 14(22):7812. <https://doi.org/10.3390/en14227812>
62. Cash S, Yuste R (1999) Linear summation of excitatory inputs by CA1 pyramidal neurons. *Neuron* 22(2):383–394. [https://doi.org/10.1016/S0896-6273\(00\)81098-3](https://doi.org/10.1016/S0896-6273(00)81098-3)

63. Morel D, Singh C, Levy WB (2018) Linearization of excitatory synaptic integration at no extra cost. *J Comput Neurosci* 44(2):173–188. <https://doi.org/10.1007/s10827-017-0673-5>
64. Liou D-R, Liou J-W, Liou C-Y (2013) Learning behaviors of perceptron. iConcept Press, Hong Kong
65. Yousofi MH (2014) Utilizing Automatic recognition and classification of images for pattern recognition. *Int J Intell Inf Syst* 3(6):80. <https://doi.org/10.11648/j.ijis.s.2014030601.25>
66. Novikoff AJ (1963) On convergence proofs for perceptrons. Washington, D.C. Retrieved from <http://classes.engr.oregonstate.edu/eecs/fall2017/cs534/extra/novikoff-1963.pdf>. Accessed on 05 May 2021
67. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 2(4):303–314. <https://doi.org/10.1007/BF02551274>
68. Rosenblatt F (1961) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan Books (1962)
69. Rumelhart DE, McClelland JL (1986) Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition, vol 1. MIT Press, pp 318–362. SDPRG University of California. Retrieved from <https://dl.acm.org/citation.cfm?id=104279.104293>. Accessed on 24 Apr 2019
70. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. Retrieved from <https://www.deeplearningbook.org/contents/mlp.html#pf25>. Accessed on 06 May 2021
71. McClelland JL, Rumelhart DE (1986) Parallel distributed processing. MIT Press, Cambridge. PDPR Group, and others
72. Judd K (2003) Nonlinear state estimation, indistinguishable states, and the extended Kalman filter. *Phys D Nonlinear Phenom* 183(3–4):273–281. [https://doi.org/10.1016/S0167-2789\(03\)00180-5](https://doi.org/10.1016/S0167-2789(03)00180-5)
73. Russell I (2012) The delta rule. University of Hartford. <https://web.archive.org/web/20160304032228/http://uhavax.hartford.edu/compsci/neural-networks-delta-rule.html>. Accessed on 05 May 2021
74. Dabbura I (2017) Gradient descent algorithm and its variants. Towards Data Science. <https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3>. Accessed on 21 Dec 2017
75. Ruder S (2016) An overview of gradient descent optimization algorithms. Insight Centre for Data Analytics at NUI Galway
76. Nise NS (2011) Control systems engineering, 6th edn. Wiley, Hoboken
77. Kalman RE (2011) A new approach to linear filtering and prediction problems. *J Basic Eng.* <https://doi.org/10.1115/1.3662552>
78. Grewal MS, Andrews AP (2008) Kalman filtering: theory and practice using MATLAB®, 3rd edn. Wiley, New York. <https://doi.org/10.1002/9780470377819>
79. Bar-Shalom Y, Li X-R, Kirubarajan T (2003) Estimation with applications to tracking and navigation. <https://doi.org/10.1002/0471221279>
80. Anderson BDO, Moore JB (2005) Optimal filtering. Dover Publications
81. Lee Rodgers J, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. *Am Stat* 42(1):59–66. <https://doi.org/10.1080/00031305.1988.10475524>
82. Willmott C, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* 30:79–82. <https://doi.org/10.3354/cr030079>
83. Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York
84. YoosefDoost A, Sadegh Sadeghian M, Ali Node Farahani M, Rasekhi A (2017) Comparison between performance of statistical and low cost ARIMA model with GFDL, CM2.1 and CGM 3 atmosphere-ocean general circulation models in assessment of the effects of climate change on temperature and precipitation in Taleghan Basin. *Am J Water Resour* 5(4):92–99. <https://doi.org/10.12691/ajwr-5-4-1>
85. Yoosefdoost A, Yoosefdoost I, Asghari H, Sadeghian MS (2018) Comparison of HadCM3, CSIRO Mk3 and GFDL CM2. 1 in prediction the climate change in Taleghan River Basin. *Am J Civ Eng Archit* 6(3):93–100. <https://doi.org/10.12691/ajcea-6-3-1>

86. YoosefDoost A, Asghari H, Abunuri R, Sadegh Sadeghian M (2018) Comparison of CGCM3, CSIRO MK3 and HADCM3 models in estimating the effects of climate change on temperature and precipitation in Taleghan Basin. *Am J Environ Prot* 6(1):28–34. <https://doi.org/10.12691/env-6-1-5>
87. Regional Water Company of Tehran (2015) Taleghan dam's structure. THRW. [https://www.thrw.ir/SC.php?type=component\\_sections&id=229&sid=7](https://www.thrw.ir/SC.php?type=component_sections&id=229&sid=7). Accessed 16 Apr 2019
88. MehrNews (2017) Taleghan dam's overflow has not finished yet. <https://www.mehrnews.com/news/4001375/>. Accessed 16 Apr 2019

# On Instantaneous Behaviour of Microplastic Contaminants in Turbulent Flow



Arefeh Shamskhany and Shooka Karimpour

**Abstract** The global plastic waste generation increased drastically during the past decade, and therefore, microplastic (MP) input to the aquatic environment is growing exponentially. MP distribution in the aquatic environment is linked to particles' physical characteristics. However, due to different origins and exposure to various weathering processes, MPs possess an extensive range of densities, shapes, and sizes. Turbulent structure induced by different mechanisms, such as temperature gradient, wind, and sudden topography changes, is another dominant factor affecting MP transport and distribution. This study aims to investigate the combined effect of size and density on the entrainment and mixing behaviour of MPs in a turbulent flow induced over the backward facing step. We used well-known sediment analogy criteria to study MP mixing and entrainment. Settling parameter, Stokes number, and trapping radius are used in this study to understand the entrainment and distribution of MPs in turbulent structures with different turbulent intensities. The first parameter describes the particle entrainment in the ambient fluid, while the other two parameters describe the distribution and mixing behaviour of particles. Our analysis demonstrates universal trends for MPs entrainment and distribution with the ambient flow dominated by MP size and density, combined with the turbulent flow hydrodynamics.

**Keywords** Microplastic Contaminants in Turbulent Flow • Instantaneous Behaviour

---

A. Shamskhany (✉) · S. Karimpour  
Department of Civil Engineering, Lassonde School of Engineering, York University, Toronto, Canada  
e-mail: [shams019@yorku.ca](mailto:shams019@yorku.ca)

S. Karimpour  
e-mail: [shooka.karimpour@lassonde.yorku.ca](mailto:shooka.karimpour@lassonde.yorku.ca)

## 1 Introduction

Plastic generation has increased drastically over the last decade, leading to a proportional increase in plastic waste generation [14]. Mismanaged plastic wastes can get into aquatic environments through different pathways, such as rivers, wastewater treatment systems, and stormwater runoff. Plastic debris remains in the aquatic environment for countless years, where weathering processes break particles into smaller parts [3]. Therefore, the presence of microplastics (MPs), plastic particles sized smaller than 5 mm, will grow exponentially over the years. Despite their benefits, the exponential growth of plastics and the corresponding increase in MP abundance in aquatic environments resulted in critical environmental risks due to the detrimental impact on aquatics' health. The fragmentation process and diversity of origins result in a broad range of shapes and sizes for aquatic MPs, from infinitesimal spheres and fragments, which is difficult to detect with naked eye, to sheets and fibres in the size of small gravels. Also, MPs possess a miscellaneous range of polymers due to their origin and utilities [25, 27]. The physical characteristics of a particle are linked to its mobility and fate in the ambient flow. Therefore, the vast range of aquatic MPs' physical characteristics results in the unique behaviour of these particles. Recent studies demonstrate MPs' selective presence in different aquatic compartments, which is affected by particles' physical properties [6, 24]. Similar to sediment particles, MPs' motion is derived from the forces acting on them. The force balance between gravity, buoyancy, and drag of MP particles dictates the trajectory and fate of these particles. Therefore, many studies focus on the density and buoyancy of plastic polymers, which undeniably play key roles in MP distribution [9, 25]. However, MPs shape and size also affect the active force terms, especially drag, drag to buoyancy ratio, and therefore are undeniably important in particles' mobility and distribution [8, 18, 30]. Khatmullina and Isachenko [18] have conducted laboratory experiments to assess the settling patterns of regularly shaped sinking MPs in analogy with sediments. Wang et al. [31] also carried out laboratory settling experiments for irregularly shaped heavy MPs.

The absence of MPs from the surface layer can be associated with many factors, including biofouling, settling, and vertical mixing with the turbulent flow [19, 22, 17]. Turbulent diffusion plays a significant role in the vertical distribution and transport of MP particles [15, 13, 23]. Yet, hydrodynamics of MPs and the role of turbulent mixing on their distribution and transport has not been thoroughly studied. On the other hand, the wide range of physical properties makes the investigation of MPs' distribution and fate very complicated in the aquatic environment. However, the existing literature on sediments' mixing and entrainment can help understand the behaviour of MPs in turbulent aquatic systems [13, 18, 31]. The growing sediment analogy knowledge demonstrates the significant effect of size on particle mixing and entrainment with the turbulent flow [28, 11, 12]. Thus, MPs of varying sizes from 5 mm to 10  $\mu\text{m}$  can have different trajectories and fates in the same flow. The aim of this paper is to study the entrainment and mixing behaviour of MP particles of various sizes and densities at different turbulent intensities. Here we conducted

numerical experiments in an open channel system using OpenFoam. Our numerical experiments include spherical MP particles ranging in size from 100  $\mu\text{m}$  to 2 mm and densities associated with abundant polymers from buoyant polypropylene polyethylene to heavy polyethylene terephthalate particles. We injected these particles to fully developed two-dimensional turbulent flows of various intensities and observed their mixing and entrainment with the flow. Results of this study are quantified using dimensionless hydrodynamic parameters in sedimentology to explain the combined effect of size and density on MPs' entrainment with the turbulent flow.

## 2 Methodology

In the present study, simulations are conducted using OpenFOAM, an open-source computational fluidic tool. In using OpenFOAM, we modified one of the pre-existing hybrid Eulerian–Lagrangian solvers to adjust it to MPs' characteristics. MPs are injected into turbulent flows to investigate the effect of turbulent circulations on the vertical mixing and transport of particles. In order to induce a coherent turbulent structure, flow over a backward facing step (BFS) is modelled. Once the flow behind the step reaches a fully developed state, MP particles are injected into the turbulent ambient flow. This section provides an overview of governing equation and model setup for both Eulerian and Lagrangian particle-tracking sub-models.

### 2.1 Eulerian Sub-model

In the first step, the turbulent flow structure is simulated using the Eulerian sub-model. For this purpose, a set of conservation of mass and momentum equations are solved using a finite volume scheme:

$$\frac{\partial(\bar{u}_i)}{\partial x_i} = 0 \quad (1)$$

$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial(\bar{u}_i \bar{u}_j)}{\partial x_j} = -\frac{1}{\rho_f} \frac{\partial \bar{p}}{\partial x_i} + \nu \frac{\partial^2 \bar{u}_i}{\partial x_j \partial x_j} - \frac{1}{\rho_f} \frac{\partial \tau_{ij}^{\text{SGS}}}{\partial x_i} \quad (2)$$

where  $\rho_f$  is the density of the ambient flow, which in our simulations, it is assumed as 1020  $\text{kg/m}^3$  (the density of seawater),  $\bar{p}$  is the pressure,  $\nu$  is the kinematic viscosity of the fluid,  $\bar{u}_i$  is the velocity of the flow, and  $\tau_{ij}^{\text{SGS}}$  represents turbulent stresses. The turbulent flow over the BFS is modelled using the large eddy simulation (LES) turbulent modelling approach. Grid size selection is critical in the LES model, as turbulent scales larger than the grid size are resolved, while the effect of smaller than scale, i.e. grid size, is modelled. In the present study, Smagorinsky sub-grid-scale viscosity is used [26]:



$$\tau_{ij}^{\text{SGS}} - \frac{1}{3}\tau_{kk}\delta_{ij} = 2\nu_t\bar{S}_{ij} \quad (3)$$

$$\nu_t = C_s\Delta^2\sqrt{2\bar{S}_{ij}\bar{S}_{ij}} \quad (4)$$

$$\bar{S}_{ij} = \frac{1}{2}\left(\frac{\partial\bar{u}_i}{\partial x_j} + \frac{\partial\bar{u}_j}{\partial x_i}\right) \quad (5)$$

where  $\nu_t$  is the turbulent viscosity,  $S_{ij}$  is the strain rate tensor,  $\delta_{ij}$  is the Kronecker delta tensor,  $\Delta = \sqrt{\text{d}x\text{d}y} = \text{d}x$  is the grid size, and  $C = 0.094$  is the Smagorinsky constant used in current simulations.

## 2.2 Lagrangian Sub-model

Ambient flow hydrodynamics affect the trajectories and fate of particles. Therefore, the injection instant alters the particle mobility with the ambient flow. In order to have a comprehensive understanding of the relation between particle behaviour and flow hydrodynamics, we used multiple injections at different times. First, the Eulerian model starts running for a specific period, until the ambient flow reaches into a quasi-steady fully developed situation. Next, MP particles with the same physical characteristics are injected in a time-sequential behaviour every 10 s. The concentration of MPs in the ambient flow is quite dilute, and therefore, MP particles do not affect the ambient flow or each other as a cloud of particles [23]. In other words, it is assumed that microplastics are discrete independent passive particles with no effective collision, which is an equivalent to the one-way coupling system [10]. Therefore, each microplastic particle is considered as a free body diagram, and based on Newton's second law, the instantaneous velocity of the particle is calculated (Eq. (6)). Here we assumed drag, buoyancy, and gravity as active force components on a MP particle:

$$m_p \frac{\text{d}v_p}{\text{d}t_p} = F_D + F_G + F_B \quad (6)$$

where  $m_p$  is the particle's mass,  $v_p$  is the particle's instantaneous velocity in vertical direction,  $\text{d}t_p$  is the Lagrangian time step, and  $F_D$ ,  $F_G$ , and  $F_B$  are drag, gravity, and buoyancy force components.

In the present study, we assumed all MP particles have spherical shapes, based on two supporting facts. First, drag calculation is very critical in dynamic particle-tracking studies. The drag force and terminal velocity calculation of spherical particles is well-understood based on the existing literature. However, for irregular shapes, the calculation of the drag coefficient is still a challenging open-ended question and requires further research [4, 18]. For instance, fibres are among the most abundant

reported MP shapes, especially among bed sediments [5, 20, 29, 32]. However, the settling pattern of fibres is completely random, and therefore, the same particle might have different settling velocities as it sinks [1, 18]. The second reason for using spherical MPs is the consistent abundance of these particle shapes among reported MPs at different aquatic compartments [2, 27]. Therefore, in this study, the active vertical force components are calculated based on Eqs. (7)–(9) for a spherical MP particle:

$$F_D = \frac{1}{2} \rho_f C_D A_p |v_f - v_p| (v_f - v_p) \quad (7)$$

$$F_G = \rho_p g V_p \quad (8)$$

$$F_B = -\rho_f g V_p \quad (9)$$

where  $A_p$  is the projected area,  $v_f$  is the ambient flow instantaneous velocity in the vertical direction,  $\rho_p$  is the density of the particle,  $V_p$  is the volume of the particle, and  $C_D$  is the drag coefficient calculated using the [23] method for spherules:

$$C_D = \begin{cases} \frac{24}{\text{Re}_p} \left( 1 + \frac{1}{6} \text{Re}_p^{\frac{2}{3}} \right), & \text{Re}_p \leq 1000 \\ 0.424, & \text{Re}_p > 1000 \end{cases} \quad (10)$$

where  $\text{Re}_p$  is the particle Reynolds number and is calculated based on Eq. (11):

$$\text{Re}_p = \frac{d_p |v_f - v_p|}{\nu} \quad (11)$$

where  $d_p$  is the particle diameter. In each iteration, after finding the instantaneous particle velocity, MP displacement in two directions is calculated using Eq. (12):

$$\frac{dx_p}{dt_p} = v_p \quad (12)$$

where  $x_p$  is the instantaneous particle location.

### 2.3 Numerical Simulation

OpenFOAM has multiple predefined solvers for both the Eulerian and Lagrangian simulations. Current simulations are based on the PimpleFoam solver, which employs the PIMPLE algorithm for pressure calculation, and is suitable for transient flow simulation. The temporal and spatial discretization schemes are set to the Upwind

and Gauss Linear, respectively. These numerical simulations are based on the explicit scheme. Therefore, the Courant number needs to be considered as the stability criterion.

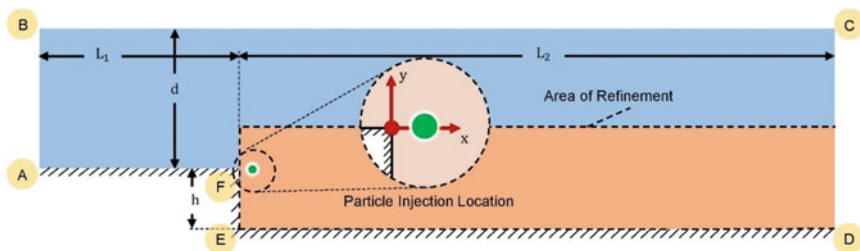
$$C_o = \frac{\bar{u}_i dt_E}{dx} \quad (13)$$

where  $dt_E$  is the Eulerian time step. In these simulations, two different time steps are considered for the Eulerian and Lagrangian sub-models. The grid size in the current simulations is always greater than the particle size. Therefore, in each time step, the particle is located inside one of the grids, and the ambient flow characteristics need to be interpolated at the location of the particle. Furthermore, Courant number is also used for the regulation of the Lagrangian time step,  $dt_p$ , in the particle-tracking sub-model. The Lagrangian time step,  $dt_p$ , is then defined as the time particle requires to exit the containing cell, based on the maximum particle Courant number. The Lagrangian time steps can be smaller than or equal to the Eulerian time step.

Figure 1 depicts the general geometry of the simulations. As demonstrated,  $d$  is the inlet water depth,  $h$  is the step height, and  $L_1$  and  $L_2$  are lengths before and after the step, respectively. Here we used a uniform flow boundary condition with an inlet velocity of  $U_0$  at the upstream inlet boundary (A-B), zero-gradient at the outlet and the free surface boundaries (C-D and B-C, respectively), and no-slip at the lower wall boundary (A-F-E-D). In order to minimize the effect of numerical instabilities, MP particles are injected into the ambient flow when the turbulent flow is fully developed and reaches the quasi-steady state, at a location which is adequately far from the boundaries. Also, boundaries are located far enough from the recirculation zone, to prevent probable numerical instabilities and errors.

The grids are structured, orthogonal, square mesh in the computational domain. In the present simulations, we considered a refinement zone, with a grid size of  $dx = dy = 1.25$  cm (demonstrated in Fig. 1), and in other parts of the geometry, a coarser grid size of  $dx = dy = 5$  cm is used.

The hydrodynamic and physical properties of all executed simulations are summarized in Table 1. In this study, based on the most abundant polymer types reported in different compartments of the aquatic environment, we considered three different



**Fig. 1** Geometry of the BFS;  $L_1 = 20$ ,  $L_2 = 30$ ,  $d = 7$ , and  $h = 3$  m; injection point = (0.5, 0.0). The coordinate system is located at the corner of the BFS

**Table 1** Executed simulations properties and IDs. These cases are simulated at three inlet velocities of 0.025, 0.100, and 0.400 m/s

Density (kg/m <sup>3</sup> )	Size (mm)			
	0.1	0.2	0.5	2.0
PE:940	PE-S1	PE-S2	PE-S3	PE-S4
PS:1100	PS-S1	PS-S2	PS-S3	PS-S4
PET:1410	PET-S1	PET-S2	PET-S3	PET-S4

densities both for positively buoyant and negatively buoyant plastics [24]. The most abundant polymer is polyethylene (PE) with a density range of 880–970 kg/m<sup>3</sup>, followed by polypropylene with densities ranging from 900 to 920 kg/m<sup>3</sup> [24]. Here, PE stands for both polyethylene and polypropylene. Also, polystyrene (PS) with densities ranging from 1040 to 1100 kg/m<sup>3</sup> is one of the most abundant polymer types, especially in the bed sediment. The last polymer considered in this study is polyethylene terephthalate (PET), which represents heavy particles' behaviour. Furthermore, here PET also presents the average density for polyvinyl chloride.

As discussed before, MPs have a wide range of sizes due to different weathering processes. Thus, we considered four size groups to investigate the effect of size on MPs mixing behaviour in turbulent flows. The size selection in this study is based on MPs size distribution in different compartments of the aquatic environments [24]. The size range in this study includes 2 mm, MP particles visible to the naked eye, to a smaller MP range of 200  $\mu$ m. This size range is also based on the size range of sand and silt particles, which are spheroid sediment particles and a comparison target in this study. In order to investigate the role of turbulent flow in MPs vertical diffusion, we conducted multiple numerical experiments with three different inlet velocities of  $U_o = 0.025, 0.1$ , and  $0.4$  m/s. For each inlet velocity, twelve cases with different sizes and densities are simulated.

### 3 Results and Discussion

In turbulent flow, velocities can be decomposed into a mean value,  $\overline{U}$ , and a fluctuating term,  $u'$ :

$$u = \overline{U} + u' \quad (14)$$

$$v = \overline{V} + v' \quad (15)$$

where  $u$  and  $v$  are the instantaneous velocity components in  $x$ - and  $y$ - directions, respectively. To determine the mean and the fluctuating components, instantaneous velocities are ensemble averaged over a period of detachment/reattachment after the quasi-steady state is achieved at  $t = t_q$ . Turbulent kinetic energy,  $k$ , is the parameter

**Table 2** Range of mean velocity in  $x$ - and  $y$ - direction,  $\bar{U}$  and  $\bar{V}$ , respectively, and the square root of maximum turbulent kinetic energy,  $k_{\max}$ , for different inlet velocities

$U_0$ (m/s)	$\bar{U}$ (m/s)	$\bar{V}$ (m/s)	$\sqrt{k_{\max}}$ (m/s)
0.025	[− 0.013, 0.027]	[− 0.008, 0.0104]	0.016
0.100	[− 0.048, 0.105]	[− 0.038, 0.032]	0.062
0.400	[− 0.193, 0.428]	[− 0.120, 0.158]	0.234

frequently used to quantify turbulent intensity and in a 2D flow is defined as:

$$k = \frac{1}{2} (\overline{u'^2} + \overline{v'^2}) \quad (16)$$

Table 2 summarizes the range of mean velocities,  $\bar{U}$  and  $\bar{V}$ , and turbulent kinetic energy,  $k$ , for different ambient flow velocities of  $U_o = 0.025, 0.100$ , and  $0.400$  m/s.

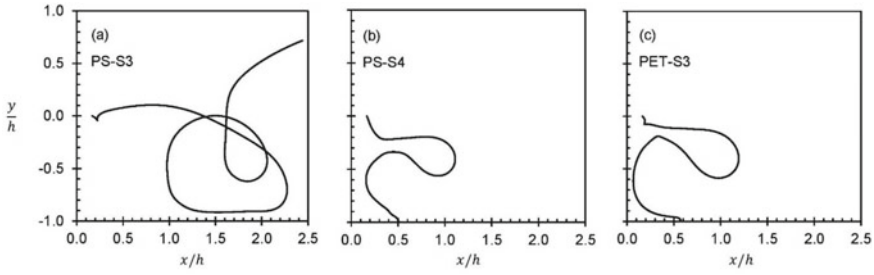
The maximum and minimum velocity range of mean velocities and turbulent kinetic energy is proportional to the inlet velocity magnitude. In other words, the instantaneous characteristics of a coherent turbulent structure behind the BFS is dominated by the inlet velocity. Therefore, here we used the inlet velocity,  $U_o$ , as a representative for turbulent structure intensity. In the following sections, based on the combined characteristics of the ambient flow and MP particles, results and findings are presented in this section.

### 3.1 Particle Relaxation Time and the Settling Parameter

The particle relaxation time,  $\tau_p$ , is defined as the time a particle needs to respond to the ambient flow changes:

$$\tau_p = \frac{|\rho_p - \rho_f| d_p^2}{18\mu} \quad (17)$$

Based on Eq. (14), particle relaxation time depends on both the particle size and the difference between the particle and ambient flow densities. This means when the particle size or its marginal density with respect to the ambient flow is relatively large compared to the ambient flow viscosity, the particle resistance to the ambient flow movements is considerable, and therefore, the particle movement needs more time to update based on flow changes. However, the particle is moving based on its natural gravitational sinking/rising behaviour in the meantime. Therefore, if the terminal velocity of the particle is considerable, the particle can skip the turbulent circulations rather than getting entrained. Figure 2 shows the effect of size and marginal density on particles entrainment with the ambient flow. As demonstrated in the picture, ambient



**Fig. 2** Effect of particle relaxation time on MPs entrainment. The ambient flow inlet velocity,  $U_o$ , is 0.1 m/s, and all particles have the same injection time. **a** PS-S3 with diameter of 0.5 mm, **b** PS-S4 with diameter of 2.0 mm, and **c** PET-S3 with diameter of 0.5 mm. **b** and **c** demonstrate the effect of size and marginal density, respectively

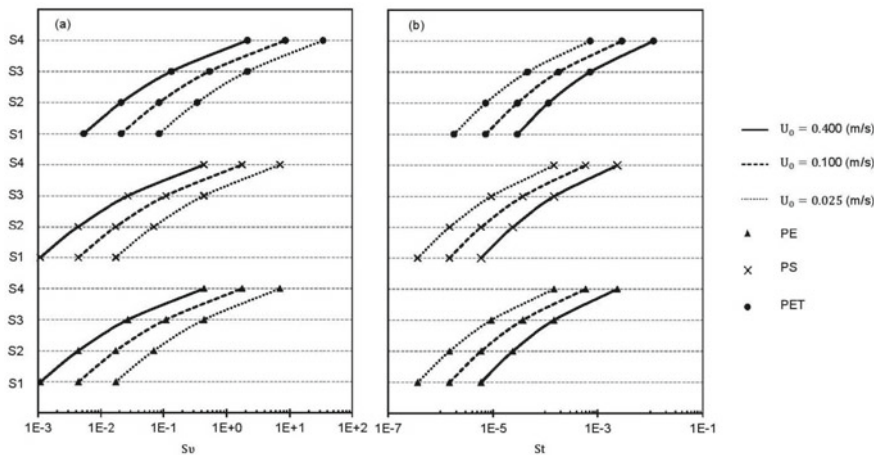
flow characteristics and the injection times are the same. However, the difference in particle relaxation time due to particles characteristics results in different particle trajectories and fates. Large particles' trajectories are minimally affected by the ambient flow and are mainly governed by the particle's gravitational sinking/rising behaviour. In contrast, small relaxation time indicates the prompt particle response to ambient flow changes and therefore higher entrainment with the ambient turbulent flow.

Particle relaxation time only depends on physical properties of the particle and flow hydrodynamics. The effect of ambient flow on particle entrainment is signified by settling parameter, defined as the ratio of Stokes settling velocity,  $\tau_p g$ , to the ambient flow velocity scale [12, 28]. As demonstrated in Table 2, the range of mean velocity components and the turbulent kinetic energy are directly proportional to the inlet velocity,  $U_o$ . Therefore, here we used  $U_o$  as the velocity scale representative:

$$Sv = \frac{\tau_p g}{\tau_f} = \frac{|\rho_p - \rho_f| g d_p^2}{18 \mu U_o} \quad (18)$$

The settling parameter determines how the particle's natural settling or rising behaviour is affected by the ambient flow movements. When the settling parameter is small, the particle tendency to its natural sinking/rising behaviour is low, and thus, the particle gets entrained with the ambient flow. On the other hand, when the settling parameter is high, the particle mainly navigates in the flow while minimally affected by the turbulent flow. Therefore, particles with high settling parameter tend to move based on their natural gravitational behaviour, which can be sinking or rising due to particle's density. Figure 3a illustrates the range of settling parameter,  $Sv$ , for particle properties enlisted in Table 1.

In consideration for Kolmogorov microscales, dissipative time scales should be incorporated in the definition of settling parameter [23]. However, the definition of settling parameter based on Eq. (15) only incorporates the role of large-scale turbulent motions and eliminates the effect of SGS turbulent fluctuations.



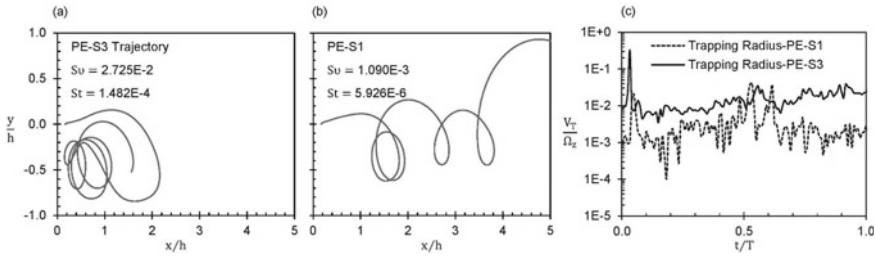
**Fig. 3** **a** Range of settling parameter,  $Sv$ , and **b** the range of stokes number,  $St$ , in the present simulations as listed in Table 1

### 3.2 Stokes Number and the Radius of Circulation Trapping

Another important dimensionless parameter which is commonly used in sediment analogy for particles mixing behaviour with the ambient flow is the Stokes number. Stokes number is defined as the ratio of the particle relaxation time to the flow time-scale,  $\tau_l$ .

$$St = \frac{\tau_p}{\tau_l} = \frac{|\rho_p - \rho_f| d_p^2 U_0}{18\mu h} \quad (19)$$

where  $\tau_l$  is considered as the large eddy turnover time, assumed as the ratio of the step height,  $h$ , to the inlet velocity,  $U_0$ . Stokes number defines particles deviation from ambient flow parcels' movement. Therefore, when the particle is entrained with the ambient flow, Stokes number can explain whether the particle trajectory is close to the ambient flow path-lines, or it deviates from the flow. Similar to the settling parameter, Stokes number has a direct relationship with the particle size. As the size of the particle increases, the stokes number increases as well. However, in contrast with the settling parameter, Stokes number and ambient flow inlet velocity have a direct relationship, which means the Stokes number increases at higher inlet velocities. Figure 3b demonstrates the range of Stokes numbers for all executed cases. When the Stokes number is small, particle's mixing behaviour is similar to a passive tracer, and the particle follows the ambient flow path-lines closely. However, when the Stokes number is large, particle's trajectory deviates from the flow parcels' trajectories.



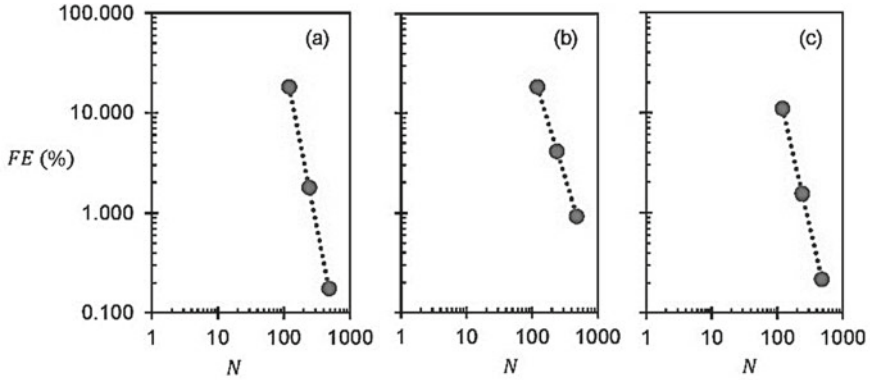
**Fig. 4** Particles' trajectories, **a** and **b**, and corresponding trapping radius, **c**, at the inlet velocity of  $U_o = 0.4$  m/s. As the settling parameter increase, the trapping radius increases as well. As the stokes number decreases, particle trajectory gets closer to the ambient flow path-lines

The radius of trapping explains how the particle motion deviates from the ambient flow movements. Here we used the ratio of terminal velocity to the vorticity magnitude,  $V_T / \Omega_z$ , as the trapping radius [7]. When the trapping radius is small, the particle follows the flow path-lines very closely. However, when the trapping radius is high, even in a strong turbulent structure, the particle deviates from the ambient flow path-line. The trapping radius determines the circular deviation from fluid parcels. Figure 4 shows the effect of size on the radius of trapping. As the settling parameter increase, particles' resistance to turbulent diffusion intensifies. Although both trajectories in Fig. 5a and b are demonstrating fully entrained case, but only large circulations with high vorticity magnitude are capable of transporting particles with high settling parameters. As a result, as the settling parameter increases, the trapping radius increases as well. Stokes number also represents the distribution of the entrained MP particles in a turbulent flow. As demonstrated in Fig. 4, the particle with a lower stokes number is transported with the ambient flow to farther downstream locations, while the particle with a larger stokes number keeps circulating behind the step. Therefore, the Stokes number can explain the deviation of particle trajectory from flow path-lines and therefore its transport with the ambient flow.

### 3.3 Numerical Verification and Convergence Studies

In order to verify the results of this study, we conducted convergence experiments in computational hydrodynamic simulation to estimate the error and accuracy [16]. For this purpose, one of the fully entrained simulations, here the PE-S1 case at the inlet velocity of 0.1, is selected as the pilot simulation for convergence test. The pilot simulation is executed at three grid sizes with a refinement ratio of  $r = 2$ . Next, the instantaneous position of the first thirty injected particles during a full period of detachment after each particle's injection is recorded per second. In the next step, the normalized vertical position of particles, measured from the bottom of the channel at  $y/h = -1$ , is used to evaluate the convergence of errors and instabilities. Then





**Fig. 5** Fractional computational error, FE, convergence for three grid sizes; **a**, **b**, and **c** correspond to 85, 90, and 95 percentiles, respectively

85, 90, and 95 percentiles (PR) are calculated for dataset of vertical locations in each of the simulations. The order of convergence,  $P_k$ , is then defined as

$$P_k = \frac{1}{\ln r} \ln \left( \frac{PR_k - PR_{k-1}}{PR_{k+1} - PR_k} \right) \quad (20)$$

where PR is the percentile rank, and the subscript demonstrates the level of refinement, as  $k$  for the original grid size, and  $k-1$  and  $k+1$  for the coarsened and refined cases, respectively. The anticipated real solution,  $PR_{\Delta x \rightarrow 0}$ , is then extrapolated, and the fractional computational error, FE, for each of the three simulations is calculated.

$$PR_{\Delta x \rightarrow 0} = \frac{r^{P_k} PR_{k+1} - PR_k}{r^{P_k} - 1} \quad (21)$$

$$FE = \frac{PR_n - PR_{\Delta x \rightarrow 0}}{PR_{\Delta x \rightarrow 0}} \times 100 \quad (22)$$

where  $PR_n$  is the corresponding percentile. Table 3 reports the properties as well as the convergence parameters for each of the three cases.

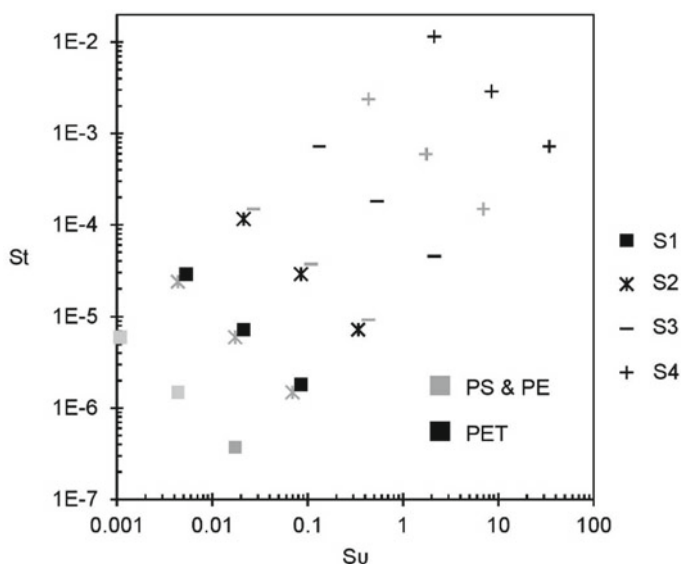
The fractional computational error versus the grid refinement is demonstrated in Fig. 5. The horizontal axis,  $N$ , represents the number of grids along the step height,  $h$ . The slope of the errors trendline is aligned with our spatial interpolation scheme (Gauss linear interpolation) order of error, which is two.

**Table 3** Percentiles, fractional error, and order of convergence after two levels of refinement

Percentile (%)	Corresponding y- position (m)	$\Delta x$ (m)	$P_k$	$PR_{\Delta x \rightarrow 0}$	FE (%)	$N$
85	1.286	2.500	3.345	1.088	18.188	120
	1.107	1.250			1.790	240
	1.090	0.625			0.176	480
90	1.384	2.500	2.154	1.169	18.382	120
	1.217	1.250			4.131	240
	1.180	0.625			0.928	480
95	1.510	2.500	2.829	1.360	11.019	120
	1.381	1.250			1.551	240
	1.363	0.625			0.218	480

4 Conclusion

The distribution and fate of aquatic MPs depend on the physical characteristics of the particle and the fluid as well as ambient flow hydrodynamics. Turbulent coherent structure plays a significant role in MPs distribution and entrainment with the flow. In this study, we conducted thirty-six numerical experiments to study the combined effect of particle properties, size and density, and turbulent intensity. We used three well-established parameters in sediment analogy to investigate MPs entrainment and distribution in a turbulent flow. The first criterion is the settling parameter which determines the tendency of a particle to its potential natural vertical settling or rising behaviour. Our results demonstrate an inverse relationship between the settling parameter and particle entrainment with the ambient flow. When the settling parameter is high, the particle tends to move based on its natural gravitational behaviour. As the settling parameter increases, particles get more entrained with the ambient flow. The second criterion is the Stokes number, which explains the distribution of particles in the turbulent flow. When the Stokes number is low, the particle’s behaviour is similar to a passive tracer. However, when the Stokes number is large, the particle deviates from the ambient flow. The third criterion is the trapping radius, which similar to the Stokes number, demonstrates the circular particle diffusion around the potential position dictated by the flow. When the trapping radius is small, the particle deviation from the ambient flow is negligible, and particles closely follow the same trajectories as the fluid parcels. Therefore, in a constant flow structure, when the trapping radius is high, the particle’s vertical diffusion is more considerable. However, the role of Stokes number and the trapping radius only emerge in fully entrained cases. Figure 6 shows the settling parameter versus the Stokes number for executed cases in the present simulation. When the settling parameter is high, particles are not entrained, and therefore, the Stokes number and trapping radius are not applicable. When both the settling parameter and the Stokes number are low, particles move close to flow parcels. Thus, particles are distributed to farther downstream locations



**Fig. 6** Stokes number versus settling parameter for executed cases. PS and PE particles of similar size possess similar relaxation times and therefore have similar stokes numbers and settling parameters. Settling parameter determines the entrainment of the particles, while the stokes number explains the distribution of particles in the ambient flow

behind the step. In contrast, when the Stokes number is high, particles are stuck in circulations close to the step. Therefore, although particles are still fully entrained with the turbulent flow, they are concentrated in regions close to the source of injection. Further investigation is required to explain MP behaviours using combined Stokes number and settling parameter.

## References

1. Bagaev A, Mizyuk A, Khatmullina L, Isachenko I, Chubarenko I (2017) Anthropogenic fibres in the Baltic Sea water column: field data, laboratory and numerical testing of their motion. *Sci Total Environ* 599:560–571
2. Ballent A, Corcoran PL, Madden O, Helm PA, Longstaffe FJ (2016) Sources and sinks of microplastics in Canadian Lake Ontario nearshore, tributary and beach sediments. *Mar Pollut Bull* 110(1):383–395
3. Barnes DK, Galgani F, Thompson RC, Barlaz M (2009) Accumulation and fragmentation of plastic debris in global environments. *Philos Trans Royal Soc B: Biol Sci* 364(1526):1985–1998
4. Chubarenko I, Bagaev A, Zobkov M, Esiukova E (2016) On some physical and dynamical properties of microplastic particles in marine environment. *Mar Pollut Bull* 108(1–2):105–112
5. Courteney-Jones W, Quinn B, Ewins C, Gary SF, Narayanaswamy BE (2020) Microplastic accumulation in deep-sea sediments from the rockall trough. *Mar Pollut Bull* 154:111092

6. Cózar A, Echevarria F, González-Gordillo JJ, Irigoien X, Úbeda B, Hernández-León S, Palma AT, Navarro S, García-de-Lomas J, Ruiz A, Fernández-de-Puelles ML (2014) Plastic debris in the open ocean. *Proc Natl Acad Sci* 111(28):10239–10244
7. Dey S, Ali SZ, Padhi E (2019) Terminal fall velocity: the legacy of Stokes from the perspective of fluvial hydraulics. *Proc Royal Soc A* 475(2228):20190277
8. Dietrich WE (1982) Settling velocity of natural particles. *Water Resour Res* 18(6):1615–1626
9. Erni-Cassola G, Zadjelovic V, Gibson MI, Christie-Oleza JA (2019) Distribution of plastic polymer types in the marine environment; a meta-analysis. *J Hazard Mater* 369:691–698
10. Elghobashi S (1994) On predicting particle-laden turbulent flows. *Appl Sci Res* 52(4):309–329
11. Fornari W, Picano F, Sardina G, Brandt L (2016) Reduced particle settling speed in turbulence. *J Fluid Mech* 808:153–167
12. Good GH, Ireland PJ, Bewley GP, Bodenschatz E, Collins LR, Warhaft Z (2014) Settling regimes of inertial particles in isotropic turbulence. *J Fluid Mech* 759
13. Jalón-Rojas I, Wang XH, Fredj E (2019) A 3D numerical model to track marine plastic debris (TrackMPD): Sensitivity of microplastic trajectories and fates to particle dynamical properties and physical processes. *Mar Pollut Bull* 141:256–272
14. Jambeck JR, Geyer R, Wilcox C, Siegler TR, Perryman M, Andrady A, Narayan R, Law KL (2015) Plastic waste inputs from land into the ocean. *Science* 347(6223):768–771
15. Kane IA, Clare MA (2019) Dispersion, accumulation, and the ultimate fate of microplastics in deep-marine environments: a review and future directions. *Front Earth Sci* 7. <https://doi.org/10.3389/feart.2019.00080>
16. Ghannadi SK, Chu VH (2015) High-order interpolation schemes for shear instability simulations. *Int J Num Methods Heat Fluid Flow*
17. Kane IA, Clare MA, Miramontes E, Wogelius R, Rothwell JJ, Garreau P, Pohl F (2020) Seafloor microplastic hotspots controlled by deep-sea circulation. *Science* 368(6495):1140–1145
18. Khatmullina L, Isachenko I (2017) Settling velocity of microplastic particles of regular shapes. *Mar Pollut Bull* 114(2):871–880
19. Kooi M, Nes EHV, Scheffer M, Koelmans AA (2017) Ups and downs in the ocean: effects of biofouling on vertical transport of microplastics. *Environ Sci Technol* 51(14):7963–7971
20. Mu J, Qu L, Jin F, Zhang S, Fang C, Ma X, Zhang W, Huo C, Cong Y, Wang J (2019) Abundance and distribution of microplastics in the surface sediments from the northern Bering and Chukchi Seas. *Environ Poll* 245:122–130
21. Putnam A (1961) Integratable form of droplet drag coefficient. *Ars J* 31(10):1467–1468
22. Reisser J, Slat B, Noble K, Du Plessis K, Epp M, Proietti M, Pattiaratchi C (2015) The vertical distribution of buoyant plastics at sea: an observational study in the North Atlantic Gyre. *Biogeosciences* 12(4):1249–1256
23. Shamskhany A, Karimpour S (2021) The role of microplastic characteristics on vertical transport and mixing. In: Canadian society of civil engineering annual conference
24. Shamskhany A, Li Z, Patel P, Karimpour S (2021) Evidence of microplastic size impact on mobility and transport in the marine environment: a review and synthesis of recent research. *Front Marine Sci* 8
25. Schwarz AE, Lighthart TN, Boukris E, van Harmelen T (2019) Sources, transport, and accumulation of different types of plastic litter in aquatic environments: a review study. *Mar Pollut Bull* 143(March):92–100. <https://doi.org/10.1016/j.marpolbul.2019.04.029>
26. Smagorinsky J (1963) General circulation experiments with the primitive equations: I. Basic. *Exper Monthly Weather Rev* 91(3):99–164
27. Song YK, Hong SH, Jang M, Kang JH, Kwon OY, Han GM, Shim WJ (2014) Large accumulation of micro-sized synthetic polymer particles in the sea surface microlayer. *Environ Sci Technol* 48(16):9014–9021
28. Stout JE, Arya SP, Genikhovich EL (1995) The effect of nonlinear drag on the motion and settling velocity of heavy particles. *J Atmos Sci* 52(22):3836–3848
29. Tekman MB, Wekerle C, Lorenz C, Primpke S, Hasemann C, Gerdts G, Bergmann M (2020) Tying up loose ends of microplastic pollution in the Arctic: distribution from the sea surface through the water column to deep-sea sediments at the HAUSGARTEN observatory. *Environ Sci Technol* 54(7):4079–4090

30. Terfous A, Hazzab A, Ghenaim A (2013) Predicting the drag coefficient and settling velocity of spherical particles. *Powder Technol* 239:12–20
31. Wang Z, Dou M, Ren P, Sun B, Jia R, Zhou Y (2021) Settling velocity of irregularly shaped microplastics under steady and dynamic flow conditions. *Environ Sci Pollut Res* 28(44):62116–62132
32. Zhang D, Liu X, Huang W, Li J, Wang C, Zhang D, Zhang C (2020) Microplastic pollution in deep-sea sediments and organisms of the Western Pacific Ocean. *Environ Pollut* 259:113948

# Influence of Biochar Amendment on Runoff Retention and Vegetation Cover for Extensive Green Roofs



Jad Saade, Samantha Pelayo Cazares, Wenxi Liao, Giuliana Frizzi, Virinder Sidhu, Liat Margolis, Sean Thomas, and Jennifer Drake

**Abstract** Canadian cities broadly promote green roofs as a sustainable solution that mitigates urban flooding and combined sewage overflows. This study assessed the performance of eight extensive green roof testbeds for rainwater retention, discharge intensity, and vegetation growth in a highly urbanized area of Toronto, Canada, over a complete growing season (May–October 2021). The 3.24 m<sup>2</sup> testbeds were constructed with high-organic substrate and two planting types: (1) a mix of native forbs and grasses planted from seed and (2) commercial *Sedum spp.* mats. The two vegetation treatments have very different establishment times and, thus, illustrated vegetation cover's role in green roof hydrology. The substrate in four testbeds was amended with 5.4% (v/v) biochar made from pyrolysis of sugar-maple sawdust. Discharge from each testbed was measured with a 0.2 mm tipping bucket, and vegetation cover was assessed monthly using a three-dimensional pin-frame. Subject to 90 natural rain events, the green roof testbeds completely retained 87% and 40% of small (0.2–4.8 mm) and medium (5–19.8 mm) rain events, respectively. Testbeds with *Sedum spp.* had higher retention than those planted with native plant *spp.* likely due to higher vegetated coverage. Biochar had positive effects on water retention and peak discharge for testbeds with native plants for small events and the vegetation growth of these plant species.

**Keywords** Extensive green roofs • Biochar amendment • Runoff retention

---

J. Saade (✉) · S. P. Cazares · G. Frizzi · V. Sidhu · J. Drake  
Department of Civil and Mineral Engineering, University of Toronto, 35 St. George Street,  
Toronto, ON M5S 1A4, Canada  
e-mail: [jad.saade@mail.utoronto.ca](mailto:jad.saade@mail.utoronto.ca)

W. Liao · S. Thomas  
Institute of Forestry and Conservation, John H. Daniels Faculty of Architecture Landscape and Design, University of Toronto, 33 Willcocks St, Toronto, ON M5S 3B3, Canada

L. Margolis  
John H. Daniels Faculty of Architecture, Landscape and Design, University of Toronto, 1 Spadina Crescent, Toronto, ON M5S 2J5, Canada

## 1 Introduction

According to the Canadian Disaster Database, flooding as a form of Meteorological-Hydrological natural disaster is a recurring disaster that Canadians suffer from [29]. Since 2010, flooding has accounted for 6.44 billion CAD, ranking it as top-ranked disaster occurring in Canada. With the increase of urbanization, flooding is exacerbated since the vegetated land has been converted to impermeable surfaces, decreasing infiltration and increasing runoff. The stormwater runoff from the paved surfaces, such as parking lots, sidewalks, and roofs, has imposed pressure on traditional stormwater systems and thus increases the risk of urban flooding.

An effective approach to mitigate stormwater runoff is to develop green infrastructure that mimics the pre-development water balance using innovative stormwater management philosophy. Low Impact Development (LID) technologies can serve as a means to this approach. LID technologies were first introduced in Maryland [7, 28]. Design guidance for LID was published in Ontario through both the 2003 MOE Stormwater management guidelines ([27] and then in 2010 through the CVC/TRCA LID manual [10]. Unlike traditional urban stormwater management systems, LID technologies are retrofit designs relying on evapotranspiration and infiltration. One such stormwater-retention-based LID technique is the green roof [11]. Green roofs date back to the fifth century in Babylon, where hanging gardens were built on multi-level stone terraces for aesthetic purposes in cities [2]. Recently, green roofs have been developed in cities to provide urban ecosystem services and socio-economic benefits, including stormwater management, biodiversity enhancement, air and noise pollution reduction, and property price improvement [34].

As one of the leading cities in green roof development in North America, Toronto has implemented the green roof bylaw to standardize and improve roof system designs while meeting the Ontario Building Code requirements [4, 6]. Economic benefit models show that near 270 million CAD in municipal capital cost savings and 30 million CAD in annual savings can be achieved by installing 12,000 acres of green roofs in the City of Toronto through mitigating combined sewer overflows and urban flood risks [38]. These savings are predominantly due to the ability of green roofs to retain more water and delay peak runoff than traditional roofs. Qin et al. [30] used the US EPA SWMM model to analyze LID scenarios and reported more flood reduction from green roofs as the size of rain events increased, but less flood reduction from green roofs as the duration of the events increased. In China, water retention of green roofs with different substrate configurations (e.g., substrate depth) ranges from 34 to 59%. Razzaghmanesh and Beecham [32] studied extensive green roof systems for two years in Adelaide, Australia (the driest city in Australia), and found an average retention of 74% in 226 rain events (968 mm in total) and an average of 61.5% peak runoff attenuation. Studies reported that the highest runoff retention by green roofs in North America was 69% [16, 35].

In Canada, [35] studied the effect of climate on runoff retention of extensive green roofs. Their results showed cumulative stormwater retention of 34% in the humid, maritime climate of Halifax, Nova Scotia (98 rain events), 48% in the humid

continental climate of London, Ontario (160 rain events), and 67% in the semi-arid, continental climate of Calgary, Alberta (86 rain events) [35]. Their average green roof runoff retention rate for the three climates was 93%, 74%, and 46% for small (<3 mm), medium (3–15 mm), and large (>15 mm) rain events, respectively. A multi-year study on green roofs in downtown Toronto found that the roof systems resulted in a significant decrease in runoff volume and rate, with an annual average volume reduction of 57% for the green roofs with 100 mm substrate depth [24]. Jahanfar et al. [19] found similar results for 51 rain events over a green roof integrated with photovoltaic arrays. They obtained cumulative runoff retention of 73% for the green roof with 150 mm substrate depth, with categorized retention of 99%, 92%, and 55% for small (0.6 mm–3 mm), medium (3–10 mm), and large (>10 mm) rain events, respectively [19]. Similarly, [14] found 70% cumulative retention for the 27 green roof plots under 176 rain events.

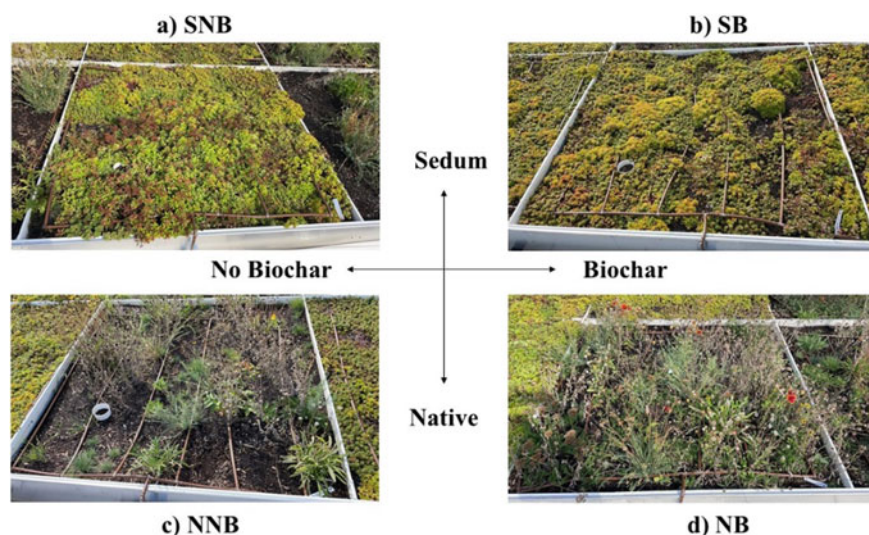
The hydrological performance of green roofs can be improved by adding appropriate substrate additives. Vijayaraghavan and Joshi [39] mixed the substrates with 10% (v/v) brown seaweed and tested the green roof performance under rainfall simulations (5–65 mm). They found that the amendment of brown seaweed supported plant growth, increased moisture retention capacity, and enhanced heavy metal sorption on green roofs [39]. Biochar is produced from the pyrolysis of biomass (biomass combustion at high temperatures in an oxygen-limited environment) and may be used as another soil amendment on green roofs [21, 33]. Although biochar properties vary according to feedstock [18], pyrolysis conditions [18], and post-processing methods [23, 36], biochar generally enhances carbon sequestration and improves water quality and quantity [17]. Studies have tested the effects of biochar amendment on the hydrological performance of green roofs [1, 5, 15, 20, 22] however, research on biochar effects on substrate water retention with different vegetation types remains scarce. This study aims to quantify the impact of biochar amendment on runoff retention capacity and plant growth of extensive green roofs with different vegetation species.

## 2 Materials and Methods

### 2.1 Experimental Setup

The experimental testbeds were installed in October 2019 at the Green Roof Innovation Testing Lab (GRITlab 2) on the rooftop of the John H. Daniels Faculty of Architecture, Landscape, and Design at the University of Toronto (108 m a.s.l.). The two types of vegetation utilized were nursery-grown *Sedum* spp. mats and native herbaceous species mix (hereafter native plant species). Testbeds with the native plant *spp* were seeded systematically three times throughout 2019–2020 (October–November 2019, June 2020, and October 2020) due to wind erosion and consequent poor germination rates. The experimental design included one non-vegetated





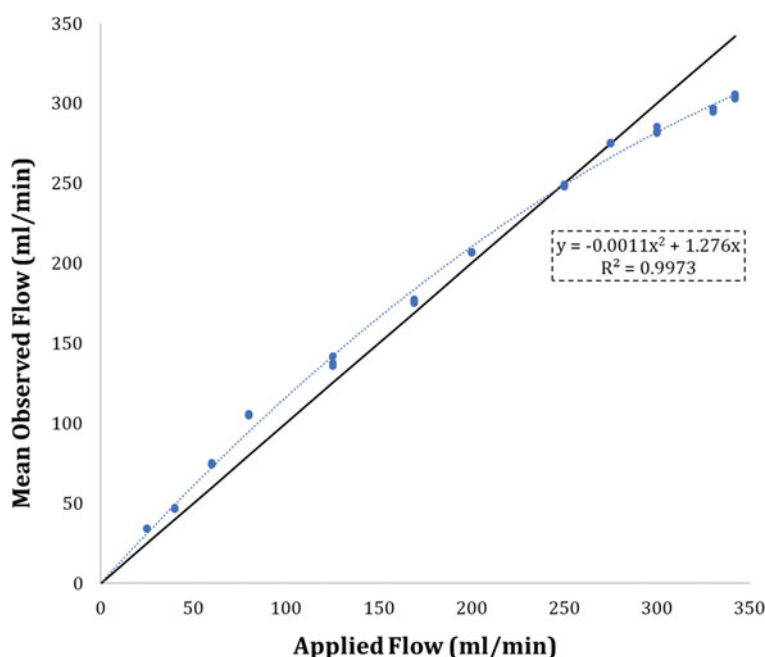
**Fig. 1** Green roof testbed types according to biochar amendment and vegetation type, **a** *Sedum* vegetation without biochar amendment (SNB), **b** *sedum* vegetation with Biochar amendment (SB), **c** native species vegetation without Biochar amendment (NNB), and **d** native species vegetation with biochar amendment (NB)

testbed coated with white roof membrane as a control, and four green roofs, two for each combination of vegetation type (*Sedum* spp. Mat; native spp. mix) and biochar amendment (biochar; no biochar) (Fig. 1).

The extensive green roof testbeds had a dimension of 1.8 m  $\times$  1.8 m and substrate depth of 150 mm, with a 2% slope. The growing media was comprised of two products Bio-mix Eco-blend and VR HydroMix Intensive [3], and sourced and mixed by Gro-Bark in Ontario. The growing media has maximum water retention of 78% and organic matter content of 59%. Biochar (produced from sugar-maple sawdust) was provided by the Haliburton Forest and Wildlife Reserve, Ltd. in Ontario. The biochar was manually mixed with the substrate in each testbed at a rate of 5.4% (v/v) (approximately 6.5 kg per testbed). Supplementary irrigation was provided to the testbeds only under extreme drought conditions during summer 2021 to support plant survival.

## 2.2 Data Collection and Measurements

The discharge from the green roofs was measured using a 0.2 mm single spoon Davis 7345.147 Pro 2 tipping bucket (Scaled Instruments Newberry, FL). The tipping bucket was calibrated for low (25–60 mL/min), medium (60–250 mL/min), and high (250–342 mL/min) steady flows (Fig. 2). The tipping bucket's maximum flow



**Fig. 2** Calibration curve and equation for green roof discharge sensor

capacity was 342 mL/min (60 tips/minute), and the max calibrated tipping capacity for the discharge setup was 380 mL/min. Precipitation was measured using a TB3 Tipping Bucket Rain Gauge (Hoskin Scientific, Burlington, ON) installed in March 2021. Data logging from these rain and discharge sensors was logged at 1-min intervals using Onset Contact Closure Pulse Input Adaptors and HOBO Micro Station H21-USB Data Loggers (Hoskin Scientific, Burlington, ON). Green roof discharge data were collected between April 9th–October 20th, 2021.

Rain events (and their durations) were defined based on the following criteria: (1) with a minimum rain event size of one tip (0.2 mm), a rain event started when one tip on the rain gauge was recorded and ended when no more tips were recorded, (2) rain events were separated based on a 1 h time difference, (3) rain events occurring consecutively were combined into one event if the discharge from the first event continues during the time of the second event, (4) a rain event causing no discharge in both control testbed and green roof testbed was not considered an event. The rainwater retention for each event (i.e., the percentage of rainfall is stored in the testbed and ultimately released as evapotranspiration) was calculated as follows:

$$\text{Retention (\%)} = \frac{\text{Rainfall (mm)} - \text{Discharge (mm)}}{\text{Rainfall (mm)}} \quad (1)$$

Quality assurance and quality control (QA/QC) techniques were implemented during the monitoring process to ensure that both rainfall and discharge data were valid by identifying possible problems with the equipment. The techniques used in this study followed practices described in the USEPA's Stormwater Best Management Practices Monitoring Manual [37]. Rainfall data was validated with measurements from two other gauges, including another UofT gauge located 260 m away). Rain events not occurring in all three sites were flagged and not included in subsequent analysis. Differences between recorded rainfall between these additional gauges and the study site's gauge helped qualify, adjust, or invalidate the data as appropriate.

In addition to discharge measurements, vegetation cover and density were evaluated for June to October 2021 using the pin-frame method [25, 26]. The wooden pin-frame ( $0.6\text{ m} \times 0.6\text{ m} \times 0.6\text{ m}$ ) is comprised of 16 steel rods, 6 mm in diameter, each marked at three different heights (15, 30, and 45 cm) from the rod bottom end. Each testbed was measured for vegetation density at three different locations (relative to the discharge drainage pipe) on the test bed: (1) top left, (2) center, and (3) bottom right, to cover the discharge path to the drainage pipe (influenced by the slope direction). Only green or living plant parts touching a pin were recorded. The number of pin touches (NT) and the number of pins touched (NP) for each height level and each location calculated plant cover and plant density. The plant cover (%) at each location was calculated as the number of pins touched (NP) recorded at each level divided by the total number of pins (16), and the overall testbed cover was the average cover for the three locations. The plant density (dimensionless) at each location was calculated as the number of touches (NT) divided by the number of pins (NP), and the overall testbed density was the average of the three locations.

### 2.3 Statistical Analysis and Data Analysis

Paired t-test was performed to test whether biochar amendment showed a significant effect on retention, peak discharge, and vegetation growth and whether vegetation type affected retention, peak discharge, and vegetation growth. Additionally, the paired *t*-test compares performances of testbeds (retention, peak discharge, and vegetation growth). Variables showing a skewed distribution were transformed, using the natural logarithms, before t-tests were conducted to satisfy the prerequisite assumptions of normality. A *p*-value of 0.05 was adopted for statistical significance. All the data analyses were performed in R version 4.0.2 [31]. After creating events from the logged time series data, calculations were performed in R Software using dplyr [8] and ggplot2 libraries [9].

**Table 1** Rain event occurrences

Event	Occurrence	Cumulative rain (mm)	Average event size (mm)
Small (0.2–4.8 mm)	61	91.0	1.5
Medium (5–19.8 mm)	21	209.0	10.0
Large (>20 mm)	8	280.4	35.0
All events	90	580.4	6.4

### 3 Results and Discussion

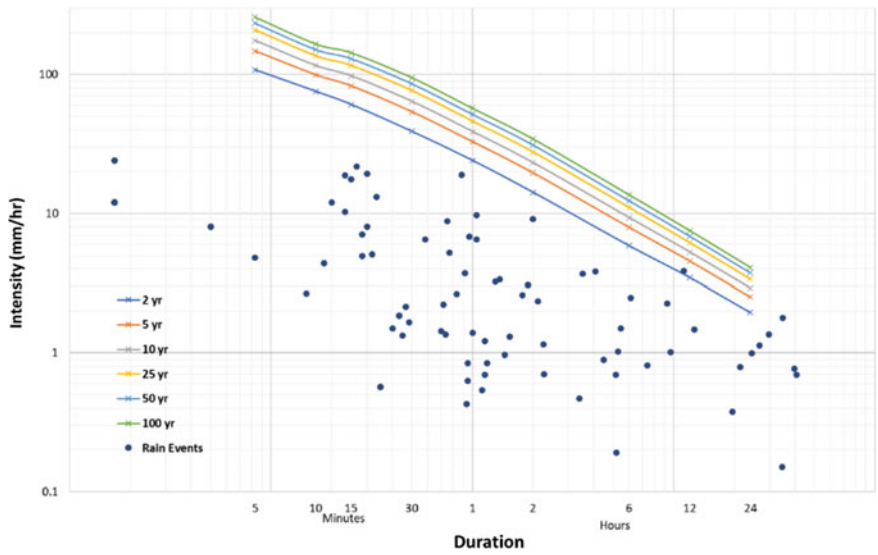
#### 3.1 Recorded Rain Events and Sample Hydrograph

Ninety rain events, totaling 580.4 mm rainfall, were identified from April to August 2021 (Table 1). The Intensity–Duration–Frequency (IDF) curves for downtown Toronto were plotted using 69 years (1940–2017) of historical rain data from ECCC weather station #6,158,355 [12]. Only one rain event (7 September 2021) intersected with the IDF curve, equivalent to a 2-year return period event. The storm produced 43.2 mm of rain and was 11.2 h long. The event had an average rainfall intensity of 3.86 mm/hr. Over 80% of the observed rain events were less than 10 mm, while most large rain events occurred during the fall (September and October, Fig. 4). All rain events generated runoff from the control roof, but only half (45 events) generated discharge from the green roofs (Fig. 3).

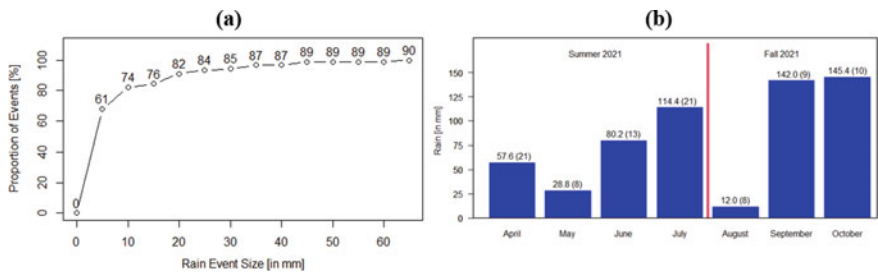
A sample hydrograph and hyetograph are shown in Fig. 5. This large October storm produced 28.4 mm over four days. The antecedent dry period prior to the event was 3.5 days. The control had higher discharge during most of the event duration as compared to all the testbeds, especially during the times when rainfall intensity was frequently high (0.4 mm/min). Testbeds with native species and biochar (NB) had the least discharge overall, with testbeds with *Sedum spp* and biochar (SB) had the highest discharge overall, especially at the beginning of the rain event.

#### 3.2 Vegetation Measurements

Vegetation cover and density were calculated for the testbeds with *Sedum spp.* and native plant spp (Fig. 6). The *sedum* mats were pre-vegetated one year prior to installation and thus plant cover and density were significantly higher for these green roofs compared to the native planted beds. Plant cover on the *sedum* testbed ranged from 93 to 100%, and plant density ranged from 2 to 4. The native grass and forbs were manually seeded and grew for over two years. Seeds were prone to erosion

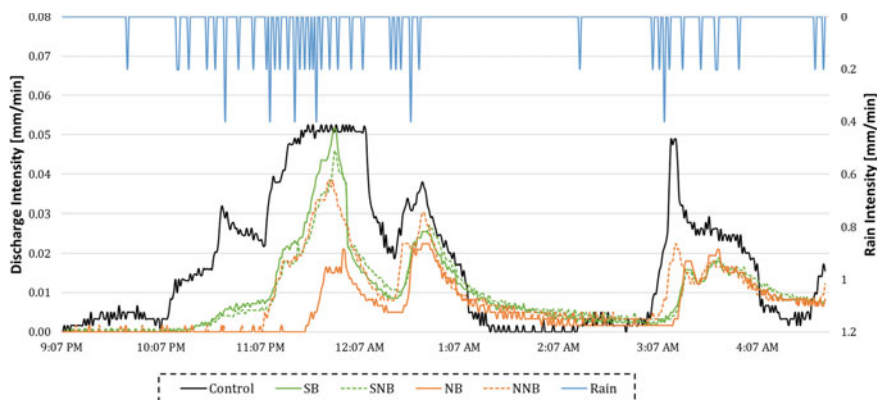


**Fig. 3** GRITlab 2 rainfall events measured over a 7-month period from April 2021 to August 2021. Downtown Toronto region IDF curves are also shown

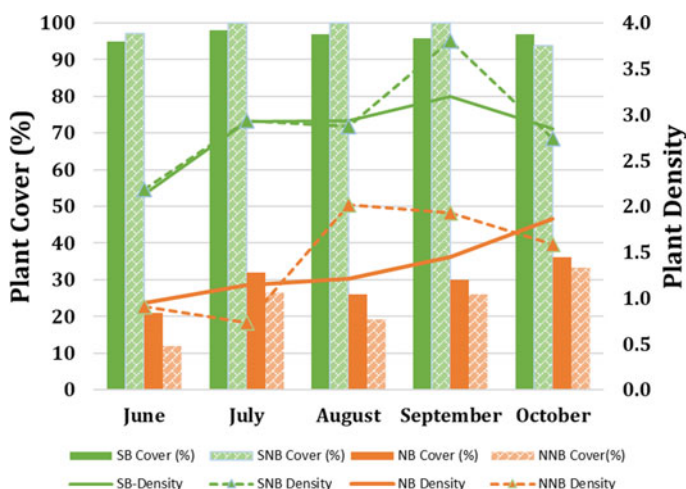


**Fig. 4** Rainfall distribution by size and month **a** the relative frequency of event occurrence by event size, labeled by cumulative number of events with this size, and **b** monthly cumulative rain, labeled  $\alpha(\beta)$  where  $\alpha$  is the cumulative rain (in mm) and  $\beta$  is the number of events per months

from wind and rain. This led to poor germination rates and required several seeding activities to make up for the poor vegetation growth. Lower vegetation coverage was found in testbeds with native plants, having plant cover ranging from 10 to 36% and plant density ranging from 1 to 2. Vegetation cover declined on the native planted testbeds in August, due to low rainfall during that month (12 mm). For the majority of the growing season, the *sedum* testbeds without biochar had slightly high vegetation cover and density.



**Fig. 5** Observed hyetograph and hydrographs for the control testbed and the green roof testbeds SB (*sedum spp.* with biochar), SNB (*sedum spp.* without biochar), NB (native species with biochar), and NNB (native species without biochar) for event 89, occurring on the evening of October 3 2021. Note that the current hydrograph represents part of the whole rain event that lasted till October 6 2021



**Fig. 6** Vegetation cover and density for testbeds SB (*sedum spp.* with biochar), SNB (*sedum spp.* without biochar), NB (native species with biochar), and NNB (native species without biochar) for the period June–October 2021

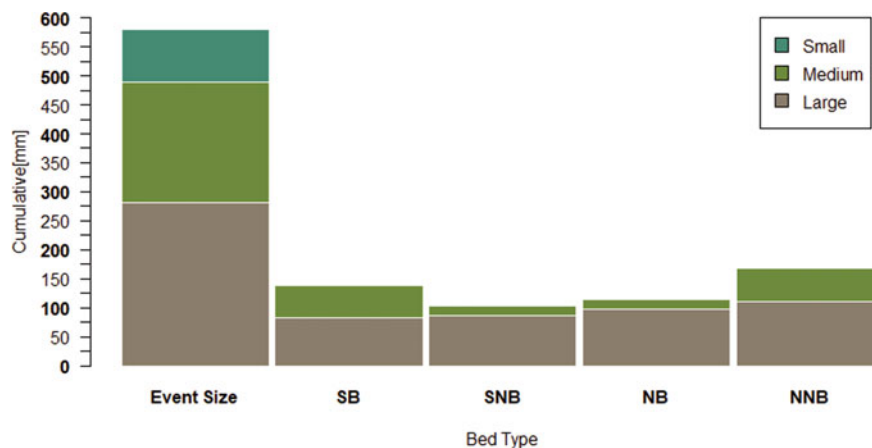
Further monitoring is required as the native planting continues to establish. Green roofs amended with biochar had higher native plant cover throughout the test period compared with those without biochar, which is consistent with other findings on the effects of biochar on plant growth in green roofs [22]. However, no differences were statistically significant.

### 3.3 Green Roof Retention

Rainfall and normalized discharge depths (mm) for the four green roof systems are plotted in Fig. 7. Overall the four green roof testbeds retained 77% of the received rainfall. This is consistent with prior studies in the Toronto area, which have reported cumulative rainfall retention ranging from 70 to 77% over a growing season [14, 19]. The average event-based rainfall retention was 90.7%, similar to other Canadian green roof studies. For example, [35] observed average event retention of 80.6% for a green roof over 86 rain events in Calgary, Alberta.

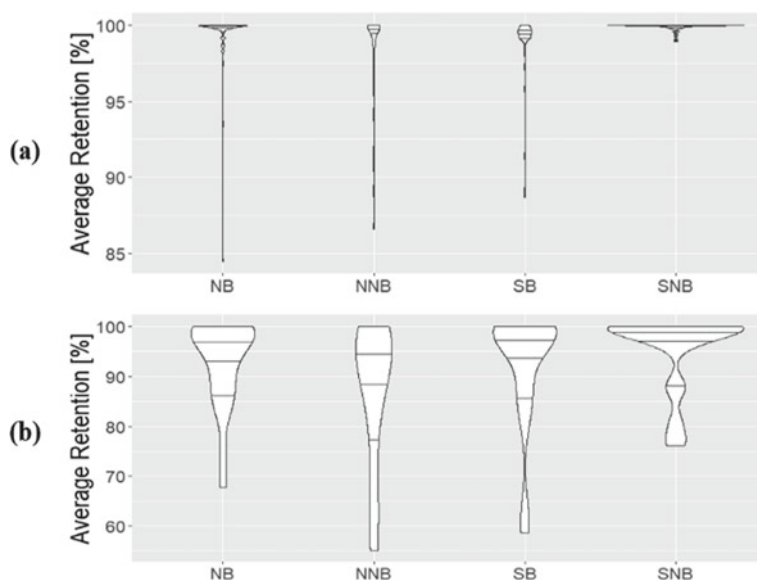
Violin plots of the four green roof types (Fig. 8) illustrate that the rainwater retention varied based on vegetation and amendment. Green roofs without biochar retained significantly ( $p = 0.007$ ) different amounts of rainwater for small storms based on the vegetated cover. The *sedum* beds, which had full coverage in 2021, retained more water than the native species beds, which were still under establishment. The amount of water retained for medium or large rain events was not affected by the presence or absence of vegetation. No significant differences in mean water retention were observed for medium or large rainfall events.

Biochar did not affect the testbeds with *Sedum spp* in small events ( $p$ -value 0.045) where SNB testbeds had higher retention than SB testbeds. On a monthly scale (Fig. 9), the SNB testbed proved dominant in the least cumulative discharge (104 mm) and the NNB testbed proved least efficient (168 mm) during the entire monitoring period. Additionally, the differences in discharge between the testbeds became more evident during September and October, which had the highest cumulative rainfall (Fig. 4), with SB and NB showing a cumulative discharge of 140 mm 114 mm,



**Fig. 7** Cumulative rain and cumulative discharge for all small, medium, and large rain event types for the SB (*sedum spp* with biochar amendment), SNB (*Sedum spp* with no biochar amendment), NB (Native species with biochar amendment), and NNB (native species without biochar amendment) testbeds. Note that small events only produced 0.94, 0.13, 0.38, and 0.78 mm for SB, SNB, NB, and NNB respectively, but discharge is not visible in this graph because of scale





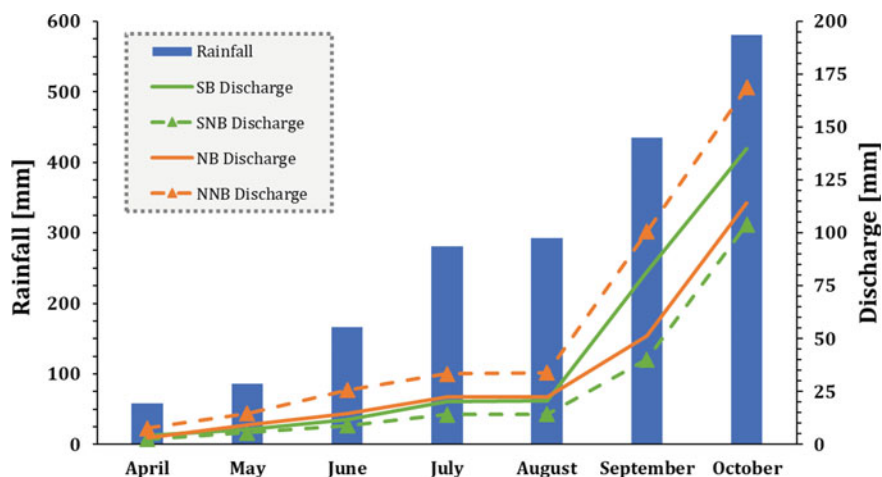
**Fig. 8** Average retention per testbed type and event size **a** small events, **b** medium events for the SB (*sedum* spp with biochar amendment), SNB (*sedum* spp with no biochar amendment), NB (native species with biochar amendment), and NNB (native species without biochar amendment) testbeds. Horizontal lines signify the 25th, 50th, and 75th quartiles

respectively. Though Fig. 9 generally shows that *sedum* testbeds, type SNB showed more pronounced effects than native plants testbeds, type NNB, with no biochar amendment, it is suspected that this was not due to vegetation type, but due to the difference in vegetation cover. With low vegetative cover, evapotranspiration during dry periods between rain events is lower, thus increasing storage until the rain event occurs and eventually decreasing rainwater retention [41].

### 3.4 Peak Discharge

Peak discharge rates from small, medium, and large storms are plotted in Fig. 10. All green roof testbeds had lower values of peak discharge as compared to the control for all three event sizes and values were significant for small and medium events ( $p$ -values < 0.008). This is consistent with the results from [40] who observed peak discharge reductions of 40–58%. Overall, the testbeds reduced peak discharge rates by 18%, 71%, and over 100% for large, medium, and small rain events, respectively, relative to peak rainfall intensity that was 0.83 mm/min, 0.70 mm/min, and 0.33 mm/min, for small, medium, and large events, respectively. SNB had better retention than SB, with lower values of peak discharge for small rain events. The difference in their peak discharge values was found to be significant. Moreover, small events showed





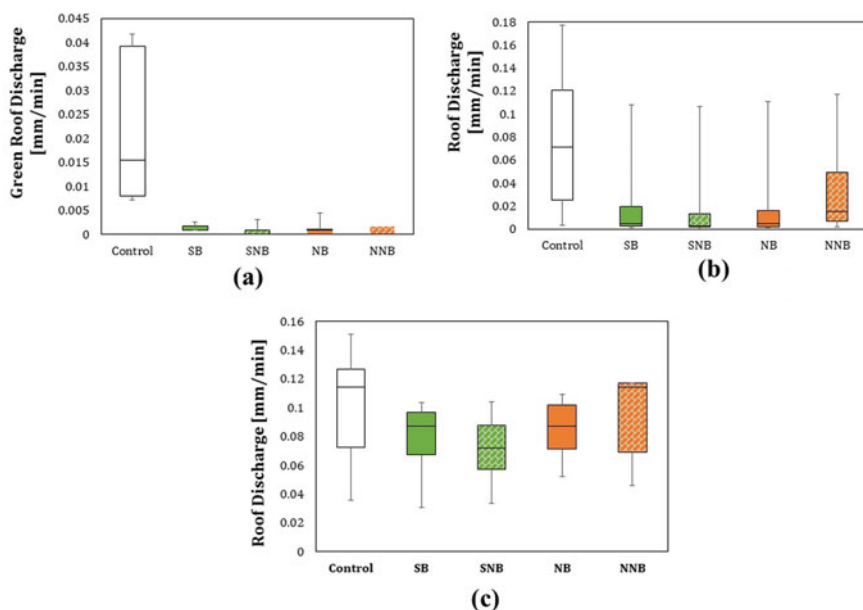
**Fig. 9** Cumulative rain and discharge for green roof testbed types, SB (sedum spp With biochar), SNB (sedum spp without biochar), NB (native species with biochar), NNB (Native species without biochar) across the growing season April–October 2021

peak rain intensity (0.33 mm/min) 1 order of magnitude higher than peak control discharge (0.03 mm/min), and 3 orders of magnitude higher than peak green roof discharge.

Additionally, testbeds with native species showed in general higher values of peak discharge. Again, this was likely due to lack of plant cover and not due to vegetation type itself. While no significant difference was found in medium rain events, SNB testbeds showed the least peak discharge values, while NNB showed the highest peak discharge values for large rain events.

## 4 Ongoing Work

The work in this study is ongoing, and data will continue to be collected throughout the 2022 growing season. Analysis of the discharge time series is also ongoing to better understand the testbed type behavior in terms of other hydrological parameters. These parameters include time to peak discharge, discharge duration, and attenuation (detention lag time and decrease in total discharge volume), which can be used to help identify where a green roof functions best. Data from the control roof often overwhelmed flow sensors in 2021. The tipping bucket configuration has been modified for 2022 to hopefully allow for more direct comparisons between discharge between green roofs and conventional roofs. Finally, in late 2021, soil moisture and temperature sensors have been installed on the green roofs to further expand the data generated from the lab.



**Fig. 10** Peak discharge for the control testbed and the 4 green roof testbed types, SB (*sedum* spp with biochar), SNB (*sedum* spp without biochar), NB (native species with biochar), NNB (native species without biochar) under the 3 rain events sizes **a** small events (<4.8 mm), **b** medium events (>4.8 mm and <19.8 mm), and **c** large events (>19.8 mm)

## 5 Conclusion

This study tests the importance of vegetation and biochar amendment on rainfall retention, peak runoff decrease, and plant growth. Average green roof retention for all four green roof types was found to be 77%. The observed results mainly reveal the importance of vegetative cover for the hydrological performance of the green roof. Testbeds with *sedum* plants had higher retention than testbeds with a herbaceous native plant species mix, especially during September, when rainfall was frequent. Biochar amendment decreased retention for *sedum* testbeds for small rain events. In addition, *sedum* beds had a higher frequency of high retention for all event sizes except for medium-sized rain events, where biochar-amended testbeds planted with native species had better performance than biochar-amended testbeds planted with *Sedum* spp. Peak discharge was lowest for *sedum* testbeds with no biochar amendment and highest for native testbeds with no biochar amendment. In addition, average detention values for all green roof types were found to be 18%, 71%, and over 100% for small, medium, and large rain events.

The *sedum* beds had higher vegetative cover and density than native species because they were installed as pre-grown mats. Biochar-amended testbeds with native plants had higher cover than native plants without biochar amendment. By

contrast, non-amended *sedum* testbeds had higher cover than biochar-amended *sedum* testbeds. This shows the need for further testing of the green roof testbeds, especially after the testbeds with native species are further established.

**Acknowledgements** This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Collaboration Research and Training Experience Program (CREATE) to J. Drake, L. Margolis, and S. C. Thomas. J. Saade received funding from Ontario Graduate Scholarship (OGS), S. Pelayo Cazares received funding from Mitacs, and J. Drake received funding from an Ontario Early Researcher Award (ERA). In-kind services and green roof material were provided by Gro-Bark Inc. and Bioroof Systems Inc.

## References

1. Beck DA, Johnson GR, Spolek GA (2011) Amending green roof soil with biochar to affect runoff water quantity and quality. *Environ Pollut* 159(8–9):2111–2118. <https://doi.org/10.1016/j.envpol.2011.01.022>
2. Berardi U, GhaffarianHoseini AH, GhaffarianHoseini A (2014) State-of-the-art analysis of the environmental benefits of green roofs. *Appl Energy* 115:411–428. <https://doi.org/10.1016/j.apenergy.2013.10.047>
3. Bioroof Systems (2020) Materials Test Report - Turf & Soil Diagnostics
4. Borooah A (2011) Toronto green roof construction standard and the supplementary guidelines 21. <http://www.toronto.ca/greenroofs/pdf/GreenRoof-supGuidelines.pdf>
5. Cao CTN, Farrell C, Kristiansen PE, Rayner JP (2014) Biochar makes green roof substrates lighter and improves water supply to plants. *Ecol Eng* 71:368–374. <https://doi.org/10.1016/j.ecoleng.2014.06.017>
6. City of Toronto. (2022). *Green Roof Bylaw*. <https://www.toronto.ca/city-government/planning-development/official-plan-guidelines/green-roofs/>
7. Coffman L (2000) Low-impact development design: a new paradigm for stormwater management mimicking and restoring the natural hydrologic regime an alternative stormwater. 158–167. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.199.2586>
8. Cran (2021a) dplyr package. <https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html>
9. Cran (2021b) ggplot2 package. <https://cran.r-project.org/web/packages/ggplot2/index.html>
10. Credit Valley Conservation, & Toronto Region Conservation Authority (2010) Low Impact Development Stormwater Management Planning and Design Guide. Version 1.0. Toronto and Region Conservation Authority and Credit Valley Conservation Authority. 4-92-4-94. [https://cvc.ca/wp-content/uploads/2014/04/LID-SWM-Guide-v1.0\\_2010\\_1\\_no-appendices.pdf](https://cvc.ca/wp-content/uploads/2014/04/LID-SWM-Guide-v1.0_2010_1_no-appendices.pdf)
11. Eckart K, McPhee Z, Bolisetti T (2017) Performance and implementation of low impact development – a review. *Sci Total Environ* 607–608:413–432. <https://doi.org/10.1016/j.scitotenv.2017.06.254>
12. Environment and Climate Change Canada (2022) Short duration rainfall intensity-duration-frequency data. [https://climatedata.ca/site/assets/themes/climate-data-ca/resources/app/idf/idf\\_v-3.20\\_2021\\_03\\_26\\_615\\_ON\\_6158355\\_TORONTO\\_CITY.txt](https://climatedata.ca/site/assets/themes/climate-data-ca/resources/app/idf/idf_v-3.20_2021_03_26_615_ON_6158355_TORONTO_CITY.txt)
13. Hill J, Drake J, Sleep B (2016) Comparisons of extensive green roof media in Southern Ontario. *Ecol Eng* 94:418–426. <https://doi.org/10.1016/j.ecoleng.2016.05.045>
14. Hill J, Drake J, Sleep B, Margolis L (2017) Influences of four extensive green roof design variables on stormwater hydrology. *J Hydrol Eng* 22(8):04017019. [https://doi.org/10.1061/\(ASCE\)he.1943-5584.0001534](https://doi.org/10.1061/(ASCE)he.1943-5584.0001534)

15. Huang S, Garg A, Mei G, Huang D, Chandra RB, Sadasiv SG (2020) Experimental study on the hydrological performance of green roofs in the application of novel biochar. *Hydrol Process* 34(23):4512–4525. <https://doi.org/10.1002/hyp.13881>
16. Hutchinson D, Abrams P, Retzlaff R, Liptan T (2003) Stormwater monitoring two ecoroofs in Portland, Oregon, USA. *Greening Rooftops Sustain Commun* 1–18. <http://www.portlandoregon.gov/bes/article/63098%5Cn>; <http://www.portlandonline.com/shared/cfm/image.cfm?id=63098>
17. International Biochar Initiative (2022) Biochar. <https://biochar-international.org/biochar/>
18. Ippolito JA, Cui L, Kammann C, Wrage-Mönnig N, Estavillo JM, Fuertes-Mendizabal T, Cayuela ML, Sigua G, Novak J, Spokas K, Borchard N (2020) Feedstock choice, pyrolysis temperature, and type influence biochar characteristics: a comprehensive meta-data analysis review. *Biochar* 2(4):421–438. <https://doi.org/10.1007/s42773-020-00067-x>
19. Jahanfar A, Drake J, Sleep B, Margolis L (2019) Evaluating the shading effect of photovoltaic panels on green roof discharge reduction and plant growth. *J Hydrol* 568(10):919–928. <https://doi.org/10.1016/j.jhydrol.2018.11.019>
20. Kuoppamäki K, Hagner M, Lehvävirta S, Setälä H (2016) Biochar amendment in the green roof substrate affects runoff quality and quantity. *Ecol Eng* 88:1–9. <https://doi.org/10.1016/j.ecoleng.2015.12.010>
21. Lehmann J, Joseph S (2015) *Biochar for environmental management: science, technology and implementation*. Routledge. <https://books.google.ca/books?hl=en&lr=&id=gWDABgAAQBAJ&oi=fnd&pg=PP1&dq=Biochar+for+environmental+management:+Science,+technology+and+implementation.+Routledge.++&ots=tZWqEQHYpV&sig=UnrGNL9qB39AXyonm4Wd1w1vBF0#v=onepage&q=Biochar+for+environmental+manage>
22. Liao W, Drake J, Thomas SC (2022) Biochar granulation enhances plant performance on a green roof substrate. *Sci Total Environ* 813:152638. <https://doi.org/10.1016/j.scitotenv.2021.152638>
23. Liao W, Thomas S (2019) Biochar particle size and post-pyrolysis mechanical processing affect soil pH, water retention capacity, and plant performance. *Soil Syst* 3(1):14. <https://doi.org/10.3390/soilsystems3010014>
24. Liu KKY, Minor J (2005) Performance evaluation of an extensive green roof. *Greening Rooftops Sustain Commun* 1–11. [https://d1wqtxts1xzle7.cloudfront.net/7265468/Performance\\_evaluation\\_of\\_an\\_extensive\\_green\\_roof-with-cover-page-v2.pdf?Expires=1644878501&Signature=BoeJCGFRK0D7h2kcflOOL96gX~jBrGdJxZaDjiDv4uzr1gD9syuGG6uZQKw1waMVS1Qe1wzEmf3L7~qxZq6HzRGvBsJCOBGyEP3UltrIsz8](https://d1wqtxts1xzle7.cloudfront.net/7265468/Performance_evaluation_of_an_extensive_green_roof-with-cover-page-v2.pdf?Expires=1644878501&Signature=BoeJCGFRK0D7h2kcflOOL96gX~jBrGdJxZaDjiDv4uzr1gD9syuGG6uZQKw1waMVS1Qe1wzEmf3L7~qxZq6HzRGvBsJCOBGyEP3UltrIsz8)
25. MacIvor JS, Lundholm J (2011) Performance evaluation of native plants suited to extensive green roof conditions in a maritime climate. *Ecol Eng* 37(3):407–417. <https://doi.org/10.1016/j.ecoleng.2010.10.004>
26. MacIvor JS, Margolis L, Perotto M, Drake JAP (2016) Air temperature cooling by extensive green roofs in Toronto Canada. *Ecol Eng* 95:36–42. <https://doi.org/10.1016/j.ecoleng.2016.06.050>
27. Ministry Of Environment Ontario (2003) *Stormwater management planning and design manual* March 2003. Water Resour (3)
28. Prince George's County (1999) *Low-impact development hydrologic analysis*. Urban Drainage Modeling (Issue July). [https://doi.org/10.1061/40583\(275\)63](https://doi.org/10.1061/40583(275)63)
29. Public Safety Canada (2022) Canadian disaster database. <https://cdd.publicsafety.gc.ca/rs/ltse-eng.aspx?cultureCode=en-Ca&boundingBox=&provinces=1,2,3,4,5,6,7,8,9,10,11,12,13&eventTypes=%2527EP%2527,%2527IN%2527,%2527PA%2527,%2527AV%2527,%2527CE%2527,%2527DR%2527,%2527FL%2527,%2527GS%2527,%2527HE%2527,%2527HU%2527,%2527SO%2527,%2527SS%2527,%2527ST%2527,%2527TO%2527,%2527>
30. Qin H, Li Z, Fu G (2013) The effects of low impact development on urban flooding under different rainfall characteristics. *J Environ Manage* 129:577–585. <https://doi.org/10.1016/j.jenvman.2013.08.026>
31. Foundation R (2022) R Software. <https://www.r-project.org/>

32. Razzaghmanesh M, Beecham S (2014) The hydrological behavior of extensive and intensive green roofs in a dry climate. *Sci Total Environ* 499(1):284–296. <https://doi.org/10.1016/j.scitotenv.2014.08.046>
33. Santos FM, Gonçalves AL, Pires JCM (2019) Negative emission technologies. *Bioenergy Carbon Capture Storage: Using Nat Resour Sustain Develop* 1–13. <https://doi.org/10.1016/B978-0-12-816229-3.00001-6>
34. Shafique M, Kim R, Rafiq M (2018) Green roof benefits, opportunities, and challenges – a review. *Renew Sustain Energy Rev* 90(March):757–773. <https://doi.org/10.1016/j.rser.2018.04.006>
35. Sims AW, Robinson CE, Smart CC, Voogt JA, Hay GJ, Lundholm JT, Powers B, O’Carroll DM (2016) Retention performance of green roofs in three different climate regions. *J Hydrol* 542:115–124. <https://doi.org/10.1016/j.jhydrol.2016.08.055>
36. Thomas SC (2021) Post-processing of biochars to enhance plant growth responses: a review and meta-analysis. *Biochar* 3(4):437–455. <https://doi.org/10.1007/s42773-021-00115-0>
37. US EPA (2002) Urban stormwater BMP performance monitoring. <https://www3.epa.gov/npdes/pubs/montcomplete.pdf>
38. US EPA (2007) Reducing stormwater costs through low impact development (LID) strategies and practices. [https://www.epa.gov/sites/default/files/2015-10/documents/2008\\_01\\_02\\_nps\\_lid\\_costs07uments\\_reducingstormwatercosts-2.pdf](https://www.epa.gov/sites/default/files/2015-10/documents/2008_01_02_nps_lid_costs07uments_reducingstormwatercosts-2.pdf)
39. Vijayaraghavan K, Joshi UM (2015) Application of seaweed as substrate additive in green roofs: enhancement of water retention and sorption capacity. *Landsc Urban Plan* 143:25–32. <https://doi.org/10.1016/j.landurbplan.2015.06.006>
40. Wong GKL, Jim CY (2014) Quantitative hydrologic performance of extensive green roof under humid-tropical rainfall regime. *Ecol Eng* 70:366–378. <https://doi.org/10.1016/j.ecoleng.2014.06.025>
41. Zaremba GJ, Traver RG, Wadzuk BM (2016) Impact of drainage on green roof evapotranspiration. *J Irrig Drain Eng* 142(7):04016022. [https://doi.org/10.1061/\(ASCE\)ir.1943-4774.0001022](https://doi.org/10.1061/(ASCE)ir.1943-4774.0001022)

# A Non-stationary Stochastic Model of Extreme Rain Events in the Changing Climate



Rituraj Bhadra and Mahesh Pandey

**Abstract** The Intergovernmental Panel on Climate Change has concluded that the frequency and intensity of extremes of heat waves, precipitation, droughts, and cyclones will continue to increase with every additional increment of global warming. Time-dependent changes in intensity and frequency of weather extremes require the non-stationary stochastic modelling of various environmental loads in the structural reliability assessment of infrastructure systems. Since most of the existing design codes and standards assume ‘stationary climate’ conditions, the effects of the transition to non-stationary conditions on infrastructure risk and reliability assessment must be carefully investigated. Whether the current design philosophies will ensure proper safety for our structures under these changing conditions, is a matter of concern from the engineering perspective. Therefore, this paper develops a non-stationary Poisson process model of extreme rain events to support the assessment of the increased risk of flooding in the future. The proposed model has a baseline model of precipitation events that are based on historical data available from existing weather stations. Then, time-dependent amplification functions, based on the findings of climate models, are assigned to the frequency and intensity of rain events. To illustrate the proposed approach, the paper analyses precipitation data from Toronto International Airport Station. The results indicate that the return periods of the extreme one-day rainfall events in Toronto corresponding to higher threshold magnitudes decrease significantly over time. The distribution of the extreme values of one-day rainfall events increases considerably with increase in design life of the structures.

**Keywords** Extreme rain events • Non-stationary stochastic model

---

R. Bhadra (✉) • M. Pandey

Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada  
e-mail: [r3bhadra@uwaterloo.ca](mailto:r3bhadra@uwaterloo.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_72](https://doi.org/10.1007/978-3-031-34593-7_72)

1133

## 1 Introduction

Global climate is changing over the years due to greenhouse gas emissions caused by human activities. Based on the simulations performed on global models, it is found that in Canada the rate of increase in the annual precipitation is most likely to increase following the rates close to the Clausius-Clapeyron rate of about 7% per degree Celsius rise in the local temperature [2]. This means that extreme events are likely to occur at higher frequency and the intensity of such events is also going to increase. The changes in several meteorological variables under a number of emission scenarios are also presented in the report 'Canada's changing climate, 2019' [1]. The sophisticated simulations on the global climate has presented the projections of climate change by the end of the current century. For example, the report predicts that the recurrence times of the 50-year extreme precipitation events corresponding to 24-h duration, will drop to as low as around 10 years under the RCP8.5 emission scenario by the year 2100.

Some of the recent events of environmental calamities like the unprecedented heatwaves in British Columbia, Canada in June 2021 and crippling heavy rainfall events in the same region in November 2021 are indicative of this climate change and this reinforces the requirement of studying climate change from the perspective of the safety of infrastructure systems and their design.

The current design philosophies and codes are developed on the fundamental assumption of stationarity of the loading. The design loads are usually obtained as some suitable high percentile of the distribution of the extreme loadings. In such models, the effects of climate change cannot be incorporated suitably. This poses a challenge to the safety and reliability of the infrastructure systems designed based on these codes. Therefore, it is of utmost importance that non-stationary models of loading that can account for the effects of climate change are developed.

This paper is intended to present a non-stationary stochastic process model which is capable of modelling time-varying characteristics of the rainfall events. An analytical example for the demonstration of the proposed methodology is also presented.

## 2 Background Literature

For any kind of design of an infrastructure system, it is crucial to have information regarding the maximum load that the system might incur in its lifetime. Information regarding extreme rainfall events is crucial for the prediction of flooding of infrastructure systems, the design of highway drainage systems, and other water-retaining hydrological structures. The extreme values of rainfall for any given duration have been tackled with the stationary models of extreme value distributions like the generalized extreme value distribution and the General Pareto distributions [3, 5, 6] until now. However, the primary assumption of these models is that the processes are

stationary Poisson processes and the intensities are exponentially distributed. This kind of stationary model fails to deal with Climate change. This is because climate change leads to the necessity of change in frequencies and intensities over time which cannot be accounted for in the stationary models.

To model the annual maximum values of any meteorological variable like rainfall, one of the most widely used distribution is the Gumbel distribution which is based on the 'stationary' assumptions. To incorporate climate change in these extreme value distributions, [5], have proposed time-dependent parameters for the Gumbel distribution or the Generalized extreme value (GEV) family of distributions. With a given sample, these parameters can be obtained using Method of moments (MoM) or Maximum likelihood estimation (MLE). Such time-dependent parameters for the GEV distributions can also be estimated using Bayesian inference conditioned on the historical data on the extreme values. One such analysis is presented by [4] for the construction of time-dependent Intensity–Duration–Frequency (IDF) curves for several locations in the United States. Such methodologies are used widely for the non-stationary modelling of extreme values under non-stationary conditions. For example, [9] used similar models for ice accretion in Canada. The problem with GEV is that since it only uses the block maxima, i.e. the annual maxima, the high-intensity events which occur within the same year are not given due importance. To combat this, General Pareto distributions (GPD) are used to model extreme values from selected events that cross some threshold values, referred to as Peak over thresholds (POT) in the literature. For non-stationary analysis with the GPD model, the time-varying scale parameters were used by Sugahara et al. [14] for frequency analysis of extreme rainfall in Sao Paulo, Brazil.

However, the problem with such analysis is that the trends can be evaluated based only on historical data. In Canada, some of the trend analysis studies on the historical rainfall data of the past few decades have shown inconclusive results in an overall sense. In some regions, the trends are decreasing and some regions have increasing trends in the intensities. While in some other regions the trends are statistically insignificant based on Mann–Kendall test analysis. Shephard et al. [12] has reported that among all the climate stations and past 20 years of data, for the rainfall duration of 24 h, 53.8% of the stations have increasing trends and 45.9% of stations have decreasing trends. It is noteworthy that, among all the stations that have shown some trends, only in a total of 4.4% stations, the trends are statistically significant. Similar inconclusive trends in extreme rainfall were reported by [15] for the province of Ontario. However, as discussed earlier, global climate simulations have shown that these rates will increase in the recent future under the current emission scenario. Bush and Lemmen [1] based on the simulations have reported that the intensity of the 24-h annual maximum precipitation for 50-year return periods values is expected to increase up to an overall 29.7% by 2100 in Canada. Therefore, it is not possible to get convincing extreme value models just by using the historical data and using the same model for future predictions.

This necessitates the stochastic process models where the recurrence intervals are more precisely incorporated in the formulations. A more general form of the stationary Poisson processes, i.e. the non-homogeneous Poisson process (NHPP)



can prove to be very effective in the formulation of time-dependent changes in the rates and intensities of the rainfall. Researchers have attempted to model the seasonal variations and trends in the rainfall occurrence for different regions using the NHPP [13, 8], however, the critical concepts like return periods and distribution of the extreme rainfall events are not discussed comprehensively in this framework.

In this paper using the base model from the historical data and using the projections based on the reports (ECCC 2019), a non-stationary model based on NHPP is developed. The maximum value distribution corresponding to any segment of time in the future can be obtained. The return periods which are no longer expected to be constant can be analytically calculated. In the following sections, a brief discussion of the stochastic process models used in the modelling is given, followed by the methodology adopted, is presented. Then some exemplar analysis on a station of Toronto International Airport is performed for demonstration purposes.

### 3 Stochastic Process Models

Of all the stochastic process models the Poisson process models stands to be one of the most widely used processes to model shock processes. This is because of its memoryless property and property of closure under superposition which makes the model simple and convenient to use. A Poisson process is a counting process  $N(t)$  where the number of events ( $k$ ) that have occurred in any interval of time is Poisson distributed (Eq. (1)) [11]. The number of events that have occurred in any disjoint set of time intervals are independent. The Poisson process is completely defined by the mean value function  $E[N(t_1, t_2)] = \Lambda(t_1, t_2)$  which is the mean number of events in the interval  $(t_1, t_2]$ . Here,  $E[.]$  is the expectation operator.

$$P[N(t_1, t_2) = k] = \frac{e^{-\Lambda(t)} [\Lambda(t_1, t_2)]^k}{k!}, \quad k = 0, 1, 2, \dots, \infty \quad (1)$$

The Poisson process for which the rate of the process, i.e.  $\lambda(t)$  is a constant, i.e.  $\lambda$  as in Eq. (2), is called a homogeneous Poisson process or a stationary Poisson process. This model is extensively used for hazard modelling when the occurrences of the events are expected to be stationary over time.

$$\frac{d\Lambda(t)}{dt} = \lambda \quad (2)$$

The Poisson process models for which the rate function is not constant and is a function of time, the process is called Non-homogeneous Poisson process or non-stationary Poisson process. The rate function can be linear or exponential depending on the process being modelled. The mean value function, i.e. the function indicating the mean number of events in any finite domain of time can be obtained by integrating the rate function over that time domain (Eq. (3)).

$$\Lambda(t_1, t_2) = \int_{t_1}^{t_2} \lambda(t) dt = \Lambda(t_2) - \Lambda(t_1) \quad (3)$$

For any kind of hazard analysis, there are two components to the process; first is the occurrence process that involves the time of occurrences and second, the magnitude of the hazard each time the hazard occurs. Both these quantities are uncertain in most of the cases and hence are considered random variables. These magnitudes or marks denoted by  $X_i$ , associated with each occurrence of the event at time ( $S_i$ ) forms the combined process ( $S_i, X_i$ ) can be referred to as a Marked Poisson process. In the current study, the arrival of the non-zero rainfall events is modelled as the NHPP and the magnitude of the one-day rainfall depths will be considered as the marks for the marked Poisson process.

## 4 Return Period

Unlike a stationary process, the return periods for a non-stationary process no longer remains constant over time. There are several ways in which return periods are defined and used in the literature of extreme rainfall analysis. In the following subsections two of the most prominent definitions of return periods in this context are presented.

### 4.1 Mean Inter-arrival Times

For a non-homogeneous Poisson process, with the mean value function (MVF) given as  $\Lambda(t)$ , the events occurring at times ( $S_1, S_2, \dots, S_n$ ), the inter-arrival times ( $T_i$ ) defined as the time difference between the occurrence of the  $(i - 1)$ th and the  $i$ th event as in Eq. (4).

$$T_i = S_i - S_{i-1} \quad (4)$$

The complementary cumulative distribution function (CDF) denoted as  $\overline{F}_{T_1}(t)$  of the first event occurrence time ( $T_1$ ) is defined as follows:

$$\overline{F}_{T_1}(t) = P[T_1 > t] = e^{-\Lambda(t)} \quad (5)$$

Which implies that the probability density function (PDF) denoted as  $f_{T_1}(t)$  of the first occurrence time ( $T_1$ ) is as given by Eq. (6), where  $\lambda(t)$  is the rate function for the NHPP given by the time derivative of  $\Lambda(t)$ .

$$f_{T_1}(t) = \lambda(t)e^{-\Lambda(t)} \quad (6)$$

These inter-arrival times  $T_i$  are not independent or identically distributed. The complementary CDF of the inter-arrival times can be given as in Eq. (7) [10].

$$\overline{F_{T_n}}(t) = \int_0^\infty f_{T_1}(t+u) \frac{[\Lambda(u)]^{n-1}}{(n-1)!} du \quad (7)$$

One formal way of defining the return period can be taken as the mean inter-arrival times  $E[T_i]$ . It can be understood from the foregoing discussion that the return period is no longer constant. This expected value of the inter-arrival times can be given as Eq. (8):

$$E[T_n] = \int_0^\infty \overline{F_{T_1}}(s) \frac{[\Lambda(s)]^{n-1}}{(n-1)!} ds \quad (8)$$

## 4.2 Mean Waiting Time to Subsequent Event

For any given time  $t$ , the time  $W(t)$  till the occurrence of next event is a quantity of interest and the mean value of this quantity can be considered as another representative definition of the return period. The complementary CDF of the waiting time is given (using the independence of number of events in disjoint intervals of time) as in Eq. (9):

$$\overline{F_{W(t)}}(s) = 1 - P[W(t) \leq s] = P[W(t) > s] = P[N(t, t+s) = 0] = e^{-\Lambda(t, t+s)} \quad (9)$$

where  $N(t, t+s)$  is the number of events in the time interval  $(t, t+s]$ . Using the property of expectations in probability theory, the mean of this waiting time can be given as:

$$E[W(t)] = \int_0^\infty e^{-\Lambda(t, t+s)} ds \quad (10)$$

## 5 Extreme Value Distribution

The distribution of the maximum intensity of the rainfall is of critical interest from the design perspective of most infrastructure systems. Consider a Poisson process with a mean value function  $\Lambda(t)$  and rate  $\lambda(t)$ . For a marked Poisson process, the cumulative distribution  $F_{\text{Max}}(x, t_1, t_2)$ , of the maximum magnitude of the events in any given interval of time  $(t_1, t_2]$  can be derived as [7]:

$$F_{\text{Max}}(x, t_1, t_2) = e^{-\Lambda(t_1, t_2)(1 - F_X(x))} \quad (11)$$

where  $F_X(x)$  is the distribution of the marks ( $X_i$ ) associated to the marked NHPP. In this context these represent the magnitude or intensity of the rainfall events. The above expression can be used when the rate of the NHPP is a function of time but the intensities of the events are independent and identically distributed. However, if the intensities of the events are considered to be time dependent and are a function of the occurrence times, then the formulation changes slightly.

Let us consider the following definition of the time-dependent mark  $X$ :

$$X = \psi(S, X_0) \quad (12)$$

where  $S$  is the occurrence time the event and  $X_0$  is the random variable denoting the intensity of the events based on the historical model following a distribution function given by  $F_{X_0}$ . Then, the distribution of the maximum rainfall magnitude can be given by Eq. (13):

$$F_{\text{Max}}(x, t_1, t_2) = \exp \left[ - \int_{t_1}^{t_2} [1 - F_{X_0}(\psi^{-1}(s, x))] \lambda(s) ds \right] = e^{-\Lambda(x, t_1, t_2)} \quad (13)$$

where  $\Lambda(x, t_1, t_2) = \int_{t_1}^{t_2} [1 - F_{X_0}(\psi^{-1}(s, x))] \lambda(s) ds$ , is the mean value function of the thinned NHPP for a threshold value of ' $x$ '.

## 6 Methodology

Initially, the Climate station Toronto Lester B. Pearson International Airport, Ontario (ID: 6,158,733) is considered and rainfall data is collected for the 24-h duration for the years 1983–2012. First a base Marked-HPP model is fitted to the historical data. Then the model rate and intensities are projected as a linear function of time in the future. The parameters of these time-dependent functions are calibrated against the ECCC studies [2].

For the base marked HPP model preparation, all the non-zero data points are extracted then fitted to a Weibull distribution and the parameters are estimated using the maximum likelihood estimation (MLE). The Weibull distribution is selected after several trials with other distributions and comparing the models based on the Akaike's information criteria (AIC), Bayesian information criteria (BIC), and log likelihood values for each distribution. This was accomplished with the *fitdistrplus*-package in R. The shape and scale parameters of the fitted Weibull distribution are estimated as 0.83 and 6.08 mm, respectively. The rate ( $\lambda_0$ ) of the base HPP was calculated as 67.9/year using MLE. This rate gives the mean number of non-zero one-day rainfall events per year based on the historical data.

It is intended to consider the span of 80 years, i.e. 2020–2100 for the analysis and projections. Therefore, the reference time period  $t_{\text{end}}$  in the subsequent expressions is taken as 80 years. For the purpose of demonstration, three threshold values of one-day rainfall are chosen as 60, 80, and 100 mm/day.

## 6.1 Calibration of Parameters

A linear rate function  $\lambda(t)$  for the gradual change in the frequencies of the rainfall events is taken as follows in Eq. (14):

$$\lambda(t) = \lambda_0 \left( 1 + \frac{k_1 t}{t_{\text{end}}} \right) \quad (14)$$

In the above expression of rate function, the parameter  $k_1$  can be considered to be a rate amplification factor or a frequency amplification factor accounting for climate change. Here,  $t_{\text{end}}$  is considered 80 years (2020–2100). Now consider the function defining the time-dependent marks  $X(S)$  as in Eq. (15), which is also considered linear.

$$X(S) = X_0 \left( 1 + \frac{k_2 S}{t_{\text{end}}} \right), \text{ in distribution} \quad (15)$$

The parameter  $k_2$  denotes the intensity amplification factor. This parameters are calibrated based on the projections of percentage change in mean annual rainfall and the 24-h duration (50-year return period) extreme rainfall events, given in the reports [1, 2] for the chosen location. The report suggests that the annual rainfall will increase approximately linearly with the global warming, and that the one-day (50-year Return period) rainfall will follow the Clausius-Clapeyron (CC) condition, i.e. approximately 7% increase per degree of local temperature increment. The return periods provided in the report are based on the annual probability of exceedance, wherein a Gumbel distribution is fitted to the annual maximum data and desired percentiles are obtained. Hence, for the purpose of calibration, the annual maximum value distribution  $F_{\text{Max}}(x, t - 1, t)$  is considered which is equivalent the

annual maximum distribution of the report. Using this model, the probability of exceedance ( $p$ ) is calculated and the return period as per this convention is  $T = \frac{1}{p}$  is obtained.

$$T = \frac{1}{p} = \frac{1}{1 - F_{\text{Max}}(x_i, t_i, k_1, k_2)}$$

(16)

The time between 2020 and 2100 is divided into 10-year segments and for each segment, the one-day (1/50) is interpolated using the global warming timing data and by linear interpolation, then the squared-error is added to form an error function ( $\xi$ ) as follows, which is a function of the parameters  $k_1$  and  $k_2$ , as in Eq. (17) (Table 1).

$$\xi = \frac{1}{n} \sum_{i=1}^n [50 - T]^2$$

(17)

Constraint : 
$$\left[1 + \frac{k_2}{t_{\text{end}}}t\right]\left[1 + \frac{k_1}{t_{\text{end}}}t\right] - 1 = \frac{\Delta M(t)}{100}$$

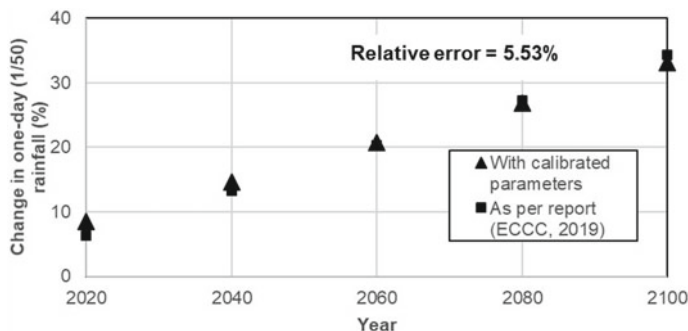
(18)

This error function is minimized numerically in MATLAB using the percentage of annual rainfall increment ( $\Delta M$ ), as a constraint. Equation (18) represents the constraint on the total annual rainfall changes as suggested in the studies (ECCC 2019). The total annual rainfall for a given year is approximate calculated as  $\left[1 + \frac{k_2}{t_{\text{end}}}t\right]\left[1 + \frac{k_1}{t_{\text{end}}}t\right]\lambda_0\mu_{X_0}$  where  $\lambda_0$  and  $\mu_{X_0}$  are the mean base rates and mean base intensity obtained from the historical data. Then the relative change will be given by the left-hand side of Eq. (18). It may be noted that this constraint tolerance is increased, so that the calibration against the extreme values are more pronounced, since the extreme values are a more important concern from the design perspective.

On solving, the parameter values are obtained as  $k_1 = 0.156$ , and  $k_2 = 0.202$ . With the parameters obtained the percentage changes in the one-day (1/50) rainfall is calculated and plotted with the changes recommended by the reports in Fig. 1. The plot shows a good fit of the model with a 5.53% relative error in the percentage changes in one-day extreme rainfall as suggested by ECCC 2019 and that predicted the NHPP model.

**Table 1** Increase in annual and extreme rainfall for Toronto station (ECCC 2019)

Global temperature rise (with respect to 1986–2016 period) (°C)	Local temperature rise (annual average) (°C)	Mean annual rainfall increase (%)	One day rain (50-year RP) increase (%)
+ 0.5	0.9	3.1	6.3
+ 3.5	4.8	22.1	33.6



**Fig. 1** Calibration of the parameters

## 7 Results

For these intensities chosen in Sect. 6, the mean inter-arrival times and mean waiting times, which are the representative definitions of return period, are calculated with the parameters obtained. Then the maximum value distribution is calculated for different periods of time starting from 2020.

### 7.1 Return Periods

In the non-stationary analysis, the return periods unlike the stationary process changes over every subsequent occurrence of the events. In this case the return periods are considered to be the mean inter-arrival times of the subsequent events. As per this definition, the first 10 return periods are calculated and plotted along with the 95% prediction intervals in Fig. 2, for the three threshold values chosen. It can be seen the return periods of the higher threshold values are more affected by the non-stationary process. The return period of one-day rainfall events crossing 100 mm/day drops from 94.0 to 43.2 years which is much higher than the drop for the events crossing 60 mm/day which is 10.1 to 8.9 years. This indicates that higher intensity events are going to occur more frequently, as supported by the studies on climate change discussed in Sect. 1 [1].

### 7.2 Mean Waiting Times

Another way of representing the return periods of rainfall events is the mean waiting time of the next occurrence of the event as per Eq. (10). In the Fig. 3, the mean waiting time and the 95% prediction interval of the waiting time for the occurrence of the next event is calculated and plotted for the duration between 2020 and 2100

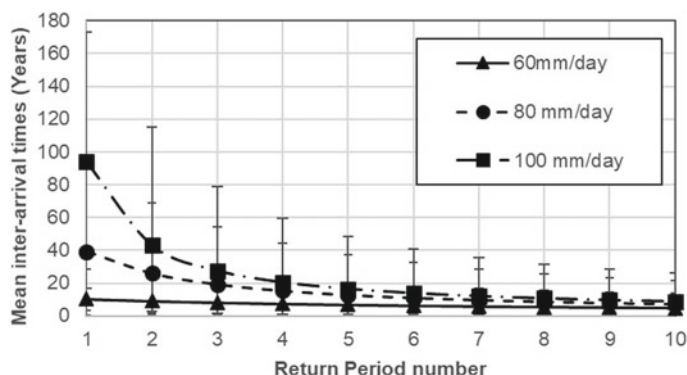


Fig. 2 Non-stationary return periods (mean and 95% prediction interval for inter-arrival times)

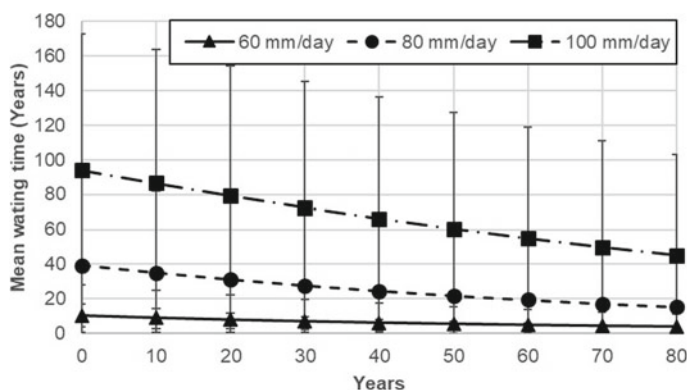


Fig. 3 Mean and 95% prediction interval for waiting time to next event in the period 2020–2100

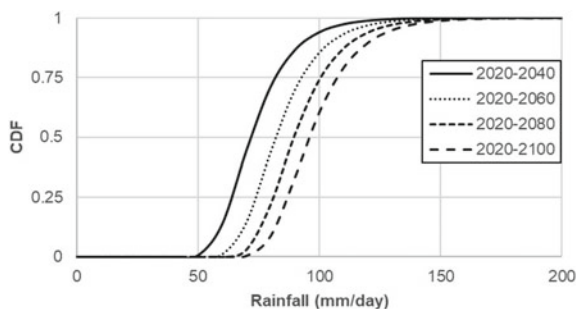
for three threshold values 60, 80, and 100 mm/day. It can be observed from the plot that the mean waiting time for the rainfall events exceeding 100 mm/day decreases from 94.0–45.2 years at the end of the century.

### 7.3 Extreme Value Distribution

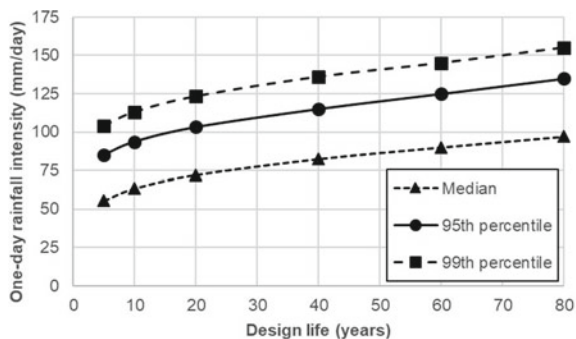
For the design of the hydrological systems and infrastructure, it is necessary to have information on the extreme values over the design life of the structure. Using the formulation of the extreme value in Eq. (13), the cumulative distribution function (CDF) is computed for several periods of times starting from 2020. In the Fig. 4, it can be seen that the CDF shifts towards the right with increase in the duration, i.e. from the period 2020–2040 to the period 2020–2100. This indicates the increase in the percentile values of the maximum one-day rainfall events with the increase



**Fig. 4** Cumulative distribution function (CDF) of the maximum one-day rainfall magnitude over different periods of time



**Fig. 5** Percentiles of the extreme one-day rainfall for different design lives starting from 2020



in the design life from 20 to 80 years, starting from 2020. This is also illustrated in the Fig. 5, where the median, 95th and the 99th percentile of the extreme value distribution is plotted for several design lives starting from 2020.

## 8 Conclusions

In this paper, the requirement of the non-stationary stochastic process models for the calculation of the rainfall extremes under changing climatic conditions, over the traditional extreme value analysis with underlying stationary assumptions is explained in support of sufficient literature and recent reports on impacts of climate change on Canada. To aid the purpose, a stochastic process model based on the marked non-homogeneous Poisson process (NHPP) is proposed. In this model, the occurrence rate, as well as the magnitudes of the rainfall events, are considered time-dependent. These model parameters can incorporate any change in the intensities and frequencies of rainfall events and hence the model is robust while dealing with any time-varying changes due to climate change. Definition of return period changes significantly in a time-dependent frame and hence relevant definitions of return period are proposed in this context. One of them being the mean inter-arrival times between consecutive events and the second being the mean waiting time till the next event. A simple

calibration scheme is proposed to calculate the parameters of this model based on the recommendations of the climate change reports.

As a demonstration, the climate station of Toronto International Airport is chosen and a base HPP model is constructed based on the historical data. With this base model, and using the data of the projections of the one-day (50-year Return period) extreme rainfall events and the mean increase in annual rainfall concerning global climate change, from the ECCC reports an NHPP model is constructed. Then with the model in hand, important quantities like the return periods and the extreme value distributions are obtained and presented for different design lives.

## References

1. Bush E, Lemmen DS (2019) Canada's changing climate
2. Cannon AJ, Jeong DI, Zhang X, Zwiers FW (2020) Resilient buildings and core public infrastructure: an assessment of the impact of climate change on climatic design data in Canada. Government of Canada. <http://publications.gc.ca/pub?id=9.893021&sl=0>
3. Charras-Garrido M, Lezaud P (2013) Extreme value analysis: an introduction. *J de La Soc Française de Statistique & Revue de Statistique Appliquée* 154(2):66–97. <http://www.sfds.asso.fr/journal>
4. Cheng L, Aghakouchak A (2014) Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. *Sci Rep* 4:1–6. <https://doi.org/10.1038/srep07093>
5. Coles S, Bawa J, Trenner L, Dorazio P (2001) An introduction to statistical modeling of extreme values, vol 208, Springer. <https://doi.org/10.1198/tech.2002.s73>
6. Coles SG, Powell EA (1996) Bayesian methods in extreme value modelling: a review and new developments. *Int Stat Rev/Revue Internationale de Statistique* 64(1):119. <https://doi.org/10.2307/1403426>
7. Manzana N, Pandey MD, van der Weide JAM (2019) Probability distribution of maximum load generated by stochastic hazards modeled as shock, pulse, and alternating renewal processes. *ASCE-ASME J Risk Uncertain Eng Syst Part A: Civil Eng* 5(1):04018045. <https://doi.org/10.1061/ajrua6.0000994>
8. Ngailo T, Shaban N, Reuder J, Rutalebwa E, Mugume I (2016) Non homogeneous poisson process modelling of seasonal extreme rainfall events in Tanzania. *Int J Sci Res (IJSR)* 5(10):1858–68. <https://doi.org/10.21275/ART20162322>
9. Pandey MD, Manzana N (2019) An investigation of non-stationary nature of ice accretion data. 1–6
10. Parzen E (1962) *Stochastic processes holden-day*. San Francisco
11. Serfozo R (2009) *Basics of applied stochastic processes*. J Chem Inf Model. Springer Science & Business Media
12. Shephard MW, Mekis E, Morris RJ, Feng Y, Zhang X, Kilcup K, Fleetwood R (2014) Trends in Canadian short-duration extreme rainfall: including an intensity-duration-frequency perspective. *Atmos—Ocean* 52(5):398–417. <https://doi.org/10.1080/07055900.2014.969677>
13. Sirangelo B, Ferrari E, De Luca DL (2011) Occurrence analysis of daily rainfalls through non-homogeneous poissonian processes. *Nat Hazards Earth Syst Sci* 11(6):1657–1668. <https://doi.org/10.5194/nhess-11-1657-2011>

14. Sugahara S, da Rocha RP, Silveira R (2009) Non-stationary frequency analysis of extreme daily rainfall in Sao Paulo, Brazil. *Int J Climatol* 29(9):1339–1349. <https://doi.org/10.1002/joc.1760>
15. Wang Y, McBean EA, Jarrett P (2015) Identification of changes in heavy rainfall events in Ontario, Canada. *Stoch Env Res Risk Assess* 29(8):1949–1962. <https://doi.org/10.1007/s00477-015-1085-6>

# **Environmental Specialty: Environmental Sustainability**

# The Elephant in the Room: Engaging with Communities About Climate Change Uncertainty



J. A. Daraio

**Abstract** The government of Newfoundland and Labrador (NL) provides climate projections for a range of important infrastructure design parameters. A partnership between Memorial University and the Government of NL, funded by Natural Resources Canada, aimed to train engineers and planners on how to use these data for infrastructure planning and design. After several workshops, it became clear that a significant obstacle to implementing some of these measures existed due to differing perceptions of uncertainty between stakeholders at the local level and technical experts. Effective communication between technical personnel and stakeholders must make explicit the meaning of “uncertainty” with which they are working for decisions to be made under uncertainty. Uncertainty in the sense of surprise suggests that it is important not to be faced with an event that was not known to be possible. This represents a risk perspective where uncertainty refers to known outcomes with unknown probabilities that is used by technical experts. There are robust methods to deal with risk from this scientific perspective, but decision makers at the local level feel unable to deal with this kind of uncertainty. A vulnerability perspective of uncertainty, with the sense of ignorance, suggests an aversion to being mistaken, particularly by an error that leads to a negative outcome, which is the meaning with which many stakeholders tend to work. This includes risk and probability, and also allows for ambiguity and vagueness that are qualitative sources of uncertainty. Clarifying these two senses allows for aleatory and epistemic sources to be more clearly defined and dealing with natural and model uncertainty can be put in perspective. The conversation can be shifted from the technical/engineering/scientific aspects of uncertainty to issues under the control of local stakeholders and decision makers. This shift facilitates the inclusion of local knowledge and greatly increases the chance of sustainable actions through a robust process of decision making.

**Keywords** Engaging with communities · Climate change uncertainty

---

J. A. Daraio (✉)

Department of Civil Engineering, Memorial University of Newfoundland, St. John's, NL, Canada  
e-mail: [jadaraio@mun.ca](mailto:jadaraio@mun.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_73](https://doi.org/10.1007/978-3-031-34593-7_73)

1149

## 1 Introduction

The civil engineer has an important role to play in the process of planning and design of civil infrastructure, which has always been done under conditions of uncertainty. Uncertainty due to climate change has added another layer of uncertainty and highlighted many of the difficulties of incorporating it into the design process. It is widely recognized that planning and design of public infrastructure must be done within the context of sustainability, with an emphasis on the development of resilient infrastructure that requires building adaptive capacity into infrastructure systems. In this context, design is *for* uncertainty, not just *despite* or *with* uncertainty. This distinction is important. Not recognizing the distinction has precluded discussion of uncertainty within the engineering community and between engineers (and other technical experts) and community stakeholders and decision makers. A key question is: what do we actually mean when by the term “uncertainty?”.

Focusing on storm water and coastal infrastructure, the complexity of processes that determine the loading and performance of public infrastructure systems makes it difficult to quantify potential impacts of climate change without a high level of uncertainty [13]. However, conditions of uncertainty cannot preclude action. Uncertainty in scientific and engineering knowledge should not be a reason to delay action, or not to act, to prevent or mitigate a potential harm. It is essential to act now to mitigate and adapt to the potential impacts of climate change under conditions of uncertainty. We must act despite the presence of uncertainty, including deep uncertainty when all possible futures cannot be clearly identified [48]. More precisely, we must act *because* of the presence of uncertainty [19], especially deep uncertainty. This creates a significant challenge for engineers and decision makers who must utilize uncertain information to plan, design, and implement resilient sustainable infrastructure systems in a changing climate [64, 68].

The objective of this paper is to describe how an explicit discussion of the meanings of the term “*uncertainty*” can potentially facilitate discussion during planning and design of resilient public infrastructure. Clarity on the meaning of *uncertainty* can improve communication to address the challenge of bridging the gap between knowledge and action in implementing sustainable design of infrastructure. I first present the experience in Newfoundland and Labrador, Canada to incorporate climate change into the design and planning of publicly funded infrastructure. I will argue that a key component missing in attempts to build capacity of professional engineers to consider climate change was a common understanding of uncertainty. Communication between engineers, municipalities, and communities has been hampered due to individuals using the term with different meanings. Clarity on the meaning of the term can lead to a more open dialogue and more sharply define the types of uncertainty to be dealt with by engineers and to be dealt with by stakeholders, community members, and decision makers. I will then discuss key aspects of sustainability from an engineering perspective, its connections with conceptions and design of resilient infrastructure, and locate where engineers in Canada and the United States (US) stand in relation to using community-based approaches to planning and design of

public infrastructure for uncertainty. I conclude the paper by outlining a framework to develop community-based resilience strategies in several small communities in Newfoundland and Labrador.

## 2 Public Infrastructure and Climate Change in NL

Newfoundland and Labrador (NL) is a sparsely populated province of just over 500,000 with half of this population concentrated on the Avalon Peninsula. There are over 260 communities outside the St. John's metro area, and over 90% of NL's population lives along the coast, which means that many communities are subject to the impacts of rising sea level [5]. It is not clear how the projected increases in frequency and magnitude of extreme events, increasing temperatures, and higher sea levels will impact storm water and coastal infrastructure. Higher base-levels and storm surge during an event will lead to more inland flooding and significant economic loss. A significant need exists to assess impacts on local infrastructure throughout the NL and Canada [12]. Given the demographics, climate, and geography of NL, it is necessary to develop and apply climate adaptation tools and resources tailored to the region's needs.

The provincial Government of Newfoundland and Labrador (GNL) planned to invest \$3 billion (CAD) over the 5 year period 2017–2022 through its provincial infrastructure plan [16]. Design of infrastructure utilizes historical data on the occurrence of extreme events to determine and reduce risk of failure without greatly increasing the cost of construction, observed climate trends and climate change projections indicate that historical data can no longer be used to estimate potential future conditions [40]. The absence of stationarity of meteorological and climate data creates a reliance on model projections that indicate warmer, wetter, and more extreme weather conditions across Newfoundland and Labrador (NL) [24]. The GNL has been proactive in its approach to incorporating climate change into infrastructure design and has developed a set of data and tools for this purpose. These include recently updated regionally downscaled climate data for the province, and updated Intensity–Duration–Frequency (IDF) curves for climate change projections through 2070 and 2100 [24]. Climate projections through 2100 are also available for a range of other important infrastructure design parameters.

Despite the availability of these data, engineers were not using this information, except where required by government policy. A 2016 GNL commissioned independent study concluded that key decision makers and professionals were either unaware or lacked the expertise to use these resources and cited several issues related to problems with communication. First, “small municipalities either did not know that including climate change should be made explicit in RFPs or assumed that climate change would be considered by engineering consultants even if it was not explicitly referenced” [16]. Second, under the impression that incorporating climate change into design would increase project costs, engineering consultants did not include climate change adaptation if it was not explicitly requested. Third, consultants also

indicated that regulations and oversight were inconsistent within and between municipalities and the provincial government. Additionally, while not explicitly highlighted in the 2016 GNL study, a lack of understanding and clear communication of uncertainty has been a problem. Or rather, it has not been directly addressed as an issue. For these reasons engineering consultants did not incorporate climate considerations into design if not directly requested, and it was recommended that part of the solution to the lack of uptake was to develop technical training and awareness workshops.

## ***2.1 Building Resilient Infrastructure***

Memorial University of Newfoundland and the GNL received funding from Natural Resources Canada (NRCan) for a project (Building and sustaining infrastructure resilience through targeted climate adaptation training for professionals in Newfoundland and Labrador) to raise awareness and to train professional engineers and planners on how to use the set of available climate data and tools for public infrastructure design in the province. The primary focus of the training has been through a series of workshops and conferences, the first of which was titled “Building Climate Resilience: Incorporating Climate Change into Public Infrastructure Planning and Design” held in St. John’s NL on March 8 and 9, 2018, and the last series of workshops to be held in fall 2022.

The target audience for the first workshop included provincial and municipal engineers and planners, Chief Administrative Officers, engineering and planning consultants, and policy and administrative staff involved in policy, planning, procurement, design, construction, operation, maintenance, and management of public infrastructure in NL. Several plenary sessions were held with the following learning outcomes.

- To understand what is involved in integrating climate change considerations in their work and how it might impact their practice.
- To understand the basics of climate change science and the genesis of lingering climate skepticism.
- To be able to describe how future climate information is produced and how to apply it in planning and design.
- To recognize how to factor adaptation measures into engineering work.
- To be able to describe the availability and application of NL provincial data sets to support planning and design of climate-resilient infrastructure.
- To be able to apply principles of climate risk and vulnerability assessment to public infrastructure.
- To be more confident about planning for the full design and operating life of public infrastructure given climate considerations.



- To be able to explain the benefits of using multi-disciplinary and multi-stakeholder teams.

Pre- and post-workshop surveys were done to assess some aspects of learning and potential changes in attitudes, though these were not done scientifically. An interesting result of the surveys was that participants found the outcomes relating to science aspects of climate change to be the least useful. This result was unexpected since those involved in the workshop are engaged in science-based and evidence-based decision making. How can individuals engaged in this process find the actual science to be of little use? It is not possible to answer this question without further study, however it seemed to be related to the uncertainty, or perceptions of uncertainty, in the climate system and in climate models due to both natural climate variability, emission scenarios, and model uncertainty. *That is, if these results are uncertain, then they are of little use.*

The second workshop funded through NRCan was held on April 9, 2019 in St. John's. This workshop focused on climate law and liability that stressed the need for due diligence by engineers to incorporate climate change, and on asset management with climate change. The workshop was useful in making more clear the need to incorporate climate change into infrastructure planning and design, however there was no discussion about uncertainty.

## **2.2 Highlighting Uncertainty**

After these workshops, the importance of perspectives on uncertainty become even more clear after the author wrote a special feature as an attempt to further the dialogue on design of infrastructure under uncertainty. This was to be published in a monthly newsletter developed under the NRCan project and hosted by the GNL. The purpose was also, to some extent, to show the value in having a basic understanding of the science behind climate change, impact assessment, and projected changes in design parameters. My partners with the GNL chose not to publish the feature because they feared it would lead readers to believe that there was too much uncertainty and/or it was too difficult to account for it in design and planning of public infrastructure. It was suggested that it could set back the ultimate goals of the project.

### **2.2.1 Uncertainty in Infrastructure Planning and Design**

In the special feature it was pointed out that updated IDF curves made available by the GNL represented a great first step to incorporate climate change into design for various hydrotechnical applications. However there are no existing methods or tools for NL that include the full range of processes that will be impacted by climate change. The updated IDF curves and other climate projections available on the GNL website must be combined with other watershed properties that are important to flood

processes in a basin. This includes snowfall, timing of snow melt, soil moisture, and permafrost which are impacted by both precipitation and temperature. Therefore it is vital to include these processes individually and in combination in order to fully account for the impacts of climate change. More advanced hydrologic modeling is required that entails the use of long-term simulations of rainfall-runoff processes using downscaled climate projections for precipitation and temperature. While it may not seem feasible for every potential infrastructure project to carry out this level of research to set design parameters and quantify uncertainties, some relatively straightforward protocols can be followed to guide this work. There are two general approaches to more fully include climate change in infrastructure design.

One approach is to update and develop a plausible set of design standards based on high-level modeling studies that can be more generally applied to incorporate climate change considerations into design. This approach uses model ensembles that allows for an explicit quantification of uncertainty. Probabilistic methods have been used to quantify and evaluate resilience metrics based on model ensembles for a variety of infrastructure systems [23, 50, 57, 60, 67]. From a risk management perspective, adapting to climate change requires an accurate description of climate over the time period of consideration [46]. Current approaches for decision making in integrated water management systems attempt to identify a most likely future for planning and design [51]. However, a reliance on accurate and precise climate prediction limits adaptation decision making in the presence of uncertainty [19]. Increased uncertainty can be accompanied by increased accuracy while reduced uncertainty, i.e., higher precision, can be less accurate. Decisions made using model results with less uncertainty, but less accuracy, can lead to maladaptation [4, 19].

Considerable challenges remain regarding how to use knowledge from such studies to build resilience and adaptive capacity [36] into storm water infrastructure. This gap in our knowledge includes both “know-what” and “know-how” in applying concepts and results of risk and resilience assessments for climate change adaptation to real world systems. This suggests that the development of general design standards incorporating climate change is limited.

A second approach incorporates of a wider range of climate change impacts into infrastructure design recognizing that resilience is a property of socio-economic-ecological systems [28, 62]. Integrative methods for assessments [20] stress the need for mixed, or combined qualitative and quantitative, methods of analysis of infrastructure resilience [45] and may be better suited for civil engineering applications. The use of model ensembles, as in the first approach, represents a probabilistic framework that allows for a focus on engineering judgment. The added dimensions in this approach allows for stress testing, such as sensitivity analysis, to show the extent to which a set of “defensible” judgments would lead to the same choice of action [55]. However, engineering judgment is difficult to quantify, it is not constrained by formal logic, it is difficult to calibrate, can incorporate conservatism, and is often heuristic and holistic [2]. Additionally, there may be cases where it is not possible to enumerate all members of the set of judgments (deep uncertainty).

High levels of uncertainty and/or the inability to quantify uncertainty cannot preclude action. There are often modes of knowledge from risk-relevant perspectives, such as local knowledge and how communities may prefer to manage uncertainties [15]. In cases where it is not possible to quantify such judgments and there is deep uncertainty, it is necessary to use approaches and methods that incorporate more qualitative knowledge from researchers, practitioners, and local communities. Additionally, inclusive and comprehensive methods should be used to incorporate such knowledge into the planning and design of infrastructure systems.

Ultimately, both approaches are closely linked. There will be cases where research-based design standards that incorporate climate change can be used, and cases where these design standards are best used within a community-based planning approach. The choice of approach is highly dependent on the level of uncertainty of risk due to climate change. It is clear that incorporating climate change considerations is not a simple task. Recognizing this fact points to the need for engineering judgment in the design of infrastructure using any approach and requires some knowledge of the underlying science and uncertainty.

The proposed feature article was intended to encourage dialogue and collaboration among engineers and within technical and non-technical teams. From the perspective of the GNL, this could cause confusion and possibly suggest that dealing with uncertainty is too complex and overwhelming to integrate into engineering design. Given this response from the GNL and based on responses from participants at the first workshop, it became clear that a key part of the problem was the perception of uncertainty in climate change projections, risk, and the decision making process at all levels, including the technical and municipal level, in particular for those that ultimately determine the scope of local infrastructure projects. It seemed that the gap between knowledge available and application of that knowledge for decision making was one of uncertainty that could be navigated in different ways depending on one's perspective. In order to explore this conjecture, as the focus of the third conference and workshop held under the NRCan project (held on November 18–20, 2019, also in St. John's), a day was devoted to discussing methods that deal with uncertainty in design and planning of sustainable infrastructure. The following section describes some of the content of the introductory session I gave at the conference, and the basis of my thinking toward the need to change the conversation on uncertainty.

### ***2.3 Perceptions of Uncertainty***

Prior to the sessions held on November 19, 2020, a survey was sent out to registrants in order to gauge perspectives on a number of issues to be discussed over the course of the day. The survey included a set of questions about uncertainty (Table 1). While the results shown cannot be used to draw a scientific conclusion, it is interesting to note that only 52% of respondents indicated that uncertainty does not preclude the ability to make decisions, and that almost 40% of respondents believed that uncertainty can be eliminated by technical knowledge. There was a post-session survey with the

**Table 1** Pre- and post-conference surveys

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total
<i>Uncertainty is a result of our ignorance</i>						
	0	14	10	18	11	53
Pre-	0.00%	26.40%	18.90%	34.00%	20.80%	
	1	4	0	5	2	12
Post-	8.30%	33.30%	0.00%	41.70%	16.70%	
<i>Uncertainty precludes the ability to make decisions</i>						
	1	10	14	15	12	52
Pre-	1.90%	19.20%	26.90%	28.80%	23.10%	
	0	1	1	6	4	12
Post-	0.00%	8.30%	8.30%	50.00%	33.30%	
<i>Uncertainty limits the ability to make decisions</i>						
	1	24	5	18	5	53
Pre-	1.90%	45.30%	9.40%	34.00%	9.40%	
	1	3	3	2	3	12
Post-	8.30%	25.00%	25.00%	16.70%	25.00%	
<i>Uncertain information is unreliable</i>						
	1	11	14	23	4	53
Pre-	1.90%	20.80%	26.40%	43.40%	7.50%	
	0	0	5	3	4	12
Post-	0.00%	0.00%	41.70%	25.00%	33.30%	
<i>Uncertainty can be eliminated by scientific/technical knowledge</i>						
	3	18	14	12	6	53
Pre-	5.70%	34.00%	26.40%	22.60%	11.30%	
	0	4	3	3	2	12
Post-	0.00%	33.30%	25.00%	25.00%	16.70%	

same set of questions, however there were only 12 respondents, so it is difficult to draw concrete conclusions about the outcomes. The post-session survey indicated some shift in perceptions, and provided support to my conjecture. Therefore, it is worth pursuing further.

### 3 Defining “Uncertainty”

It is useful to think of uncertainty in terms of the following.

1. Known knowns: what we know we know

2. Known unknowns: what we know we don't know
3. Unknown unknowns: what we don't know we don't know
4. Unknown knowns: what we don't know we know.

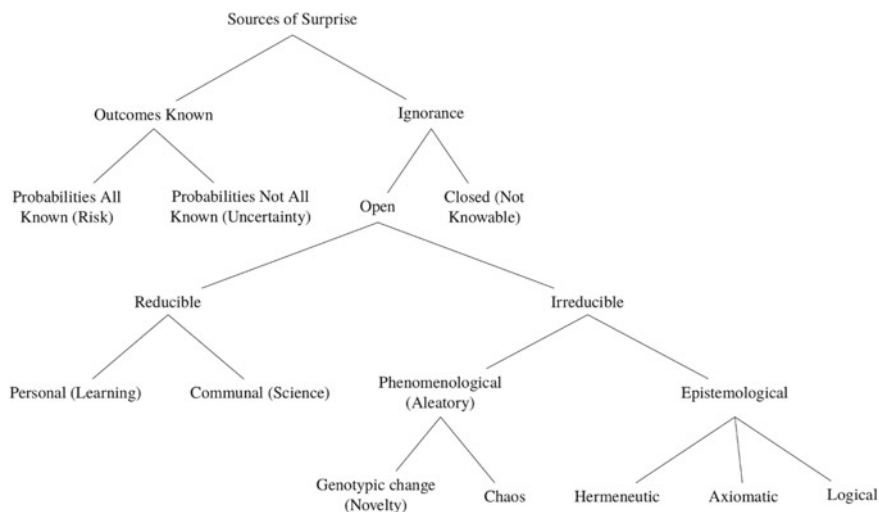
While (4) seems counter-intuitive, it is an important aspect of uncertainty [18], in the context of decision making at the community level. A single team member may not know what they even need to know, but it is likely that someone has the relevant knowledge. Looking at more precise definitions of “uncertain” provides a little more insight into the biases and preconceptions attached to the concept.

Merriam-Webster, American Heritage, and Oxford English dictionaries each define “uncertainty” as the “state” or “quality” of being uncertain, which is of no help. These dictionaries define “uncertain” as: not known beyond doubt; not known or established; questionable; not having certain knowledge; not having sure knowledge; not clearly identified or defined; not determined; undecided; not constant; subject to change; variable; not certain to occur; **not reliable**. There is a normative element in these definitions where something that is uncertain becomes untrustworthy and problematic. Indeed, many of the conference participants held that uncertain information is unreliable. A clear negative connotation is evident.

The point of looking at these definitions is not to decide on the best definition or meaning but highlight what may be viewed as the common or shared uses of the term. When dealing with uncertainty in the context of sustainability and decision making, we need to be more precise about our use of the term. When it comes to a term that often has negative connotations, and can be ill-defined, there is a great chance for misunderstanding and disinformation [14]. Most notably to be aware that not everyone has the same *sense* or *referent* for the term. Sense, or “mode of presentation,” and referent, or the thing or object itself, are the two key elements of how words get meaning [27]. It is very important that stakeholders, community members, engineers, and decision makers are aware of the meaning, both sense and referent, of uncertainty they have in mind when at the table.

### 3.1 *Senses and Referents of “Uncertainty”*

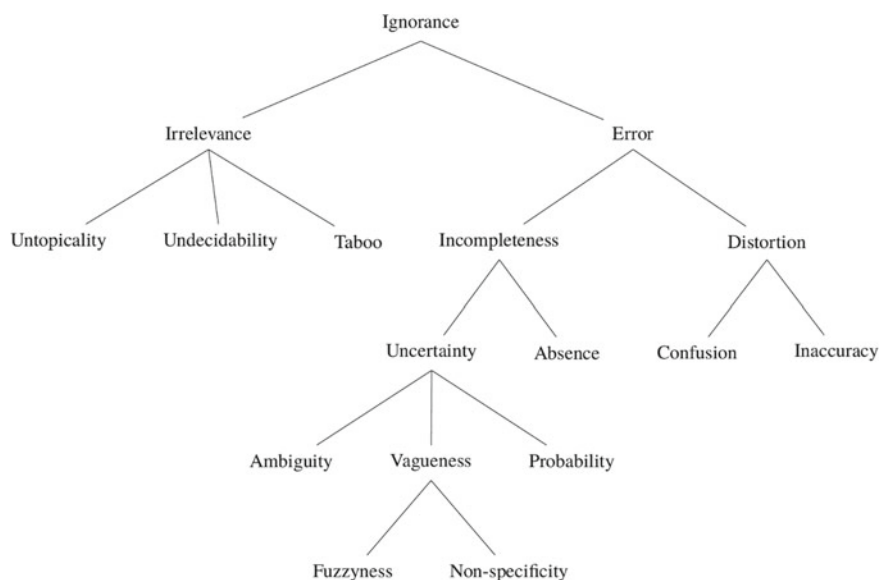
While uncertainty has a range of senses and referents, I focus on two perspectives that I believe represent relevant viewpoints for this discussion, risk, and vulnerability. The risk perspective, or *sense of risk*, stems from the idea of avoiding surprise where it is important not to be faced with some kind of event that was not known to be possible. The *sense of vulnerability* reflects a view from ignorance where we don't want to be unaware of our situation or make an error that leads to a negative outcome, possibly also by surprise. These senses of uncertainty are not mutually exclusive (Fig. 1). Both views can be held by the same person at different times, and not recognizing the intended sense of a speaker can lead to miscommunication. In particular, it is vital to see that each sense of the term can refer to quite different things.



**Fig. 1** Sources of surprise. Adapted from [22]

The risk perspective sees uncertainty as referring to known outcomes with unknown probabilities (Fig. 1). This seems to be the dominant perspective used in the design and planning of public infrastructure and fits squarely within the engineer's toolbox. With respect to climate change, projections provide a set of plausible future conditions (known outcomes), and where the physics are well-known, high-resolution climate models will well describe the climate system. Climate model uncertainty is due to uncertain boundary and initial conditions, incomplete theoretical understanding, parameter uncertainty, and imperfect models [35]. The probabilities of projections based on climate models are known, so climate risk can be determined in a relatively straightforward way. Uncertainty is quantified based on whether a model is credible, plausible, or consistent with observations [35]. This is the uncertainty in risk assessment under climate change, and is the perspective held by most technical experts. One caveat regards the emission scenarios. In the shorter term,  $\approx 20$  years, there is little difference in the outcomes and probabilities between the full ranges of emission scenarios [33]. However over longer time scales the uncertainty in future socio-economic conditions becomes more difficult to quantify, and the probabilities may not be known.

Ignorance under the risk perspective represents a lack of knowledge and could become invisible to the decision making process, except perhaps for aleatory uncertainty (see Fig. 1), which includes natural or inherent variability [26]. While aleatory uncertainty is partially due to ignorance (i.e., more research will reduce or eliminate this uncertainty), it is recognized as an important source of uncertainty in addition to model uncertainty under the risk perspective, and it is quantified (estimated with



**Fig. 2** Sources of ignorance. Adapted from [59]

uncertainty?) as a known probability. The key point is that in the sense of risk, uncertainty has a clear referent as *a known outcome with unknown probabilities*. I don't believe that this is intuitive, and it is not what most (non-technical) people mean by the term "uncertainty".

Compared to uncertainty in the sense of risk, both sense and referent differ in the sense of vulnerability based on the schema for uncertainty developed by Faber et al. [22], Fig. 2. The desire to reduce or remove ignorance suggests an aversion to being mistaken, particularly by an error that leads to a negative outcome. This sense leads to a view of uncertainty that includes risk and probability, but also allows for ambiguity and vagueness that have sources other than scientific uncertainty.

Looking at the schema in Fig. 2, there are things that we choose to be ignorant of for a wide variety of reasons. This includes a lack of interest, the fact that it's not possible to know everything even if we wanted to, and that there are things we shouldn't know (taboo). On the other hand, ignorance that leads to error, or is caused by error, can have several forms, one of which is uncertainty. From this perspective, uncertainty refers to not just probabilities, but to ambiguity and vagueness, which is a much wider range of things than known outcomes with unknown probabilities. I prefer this sense of the term "uncertainty" as I believe it allows for a more explicit recognition of the full range of referents. Further, I agree with [18] that "[t]o be uncertain is to be ignorant, but it is also to know something". Uncertainty is not ignorance, but ignorance is a component of it. Importantly, uncertainty implies that

one has some knowledge about something, though it may be incomplete, ambiguous or conflicting, and it can (should) be incorporated into both perspectives. The referent for uncertainty in the sense of vulnerability *is fuzzy; it can be quantifiable in many ways and can be qualitative.*

### 3.2 *Talk About It!*

When engaging community stakeholders and decision makers, it is important to recognize the range of meanings of “uncertainty,” clearly identify sources of uncertainty as identified from a risk perspective along with the robust methods available for dealing with natural and model uncertainty. This will allow a non-technical audience to put in perspective what engineers are talking about and addressing when *designing with* climate change. For example, cities that have been successful in engaging climate science were not immobilized by uncertainty. They tend to take advantage of available data, are cautious but take incremental action in order to limit expenditures and the potential for large mistakes [10]. They also recognize the need to update projections and analyses. Carmin and Dodman [10] concluded that (1) scientific data (from assessments) were used as a basis to gain insight into projected impacts and establish the need for action, (2) available evidence allows for setting of priorities and appropriate measures needed to build adaptive capacity; (3) science is used as a means to generate support. Actions with uncertainty should be cautious and flexible.

Though it is important for stakeholders to understand the limits of scientific knowledge, it is clearly the job of technical experts to provide the scientific data in its proper context. Dealing with uncertainty from the risk perspective can be overly technical and overwhelming at the community level, and this is not where this should be handled. Recognition of this fact has the potential to shift the focus of decision making for sustainable action at the local level. The conversation can shift from the technical/engineering/scientific aspects of uncertainty to focus on the uncertainty from the perspective of vulnerability, or about how individuals/communities/governments will choose to deal with the salient issues. The latter are well within the control of local stakeholders and decision makers. This shift facilitates the inclusion of different modes of knowledge from risk-relevant perspectives, such as local knowledge and how communities may prefer to manage and *design for* uncertainties. Sustainable actions can be enhanced through a robust process of decision making, and by recognizing that uncertainty provides an opportunity to do things right. It is necessary to use methods that incorporate more qualitative knowledge from researchers, practitioners, and local communities. Inclusive and comprehensive quantitative methods should be used to incorporate such knowledge into decision models (see Sect. 5). Civil engineers have started to move in this direction.

Using this understanding of uncertainty as a starting point, to better engage and communicate with community stakeholders as engineers, it is important to clarify a larger context for uncertainty to reduce the fuzziness of the vulnerability perspective.



Or at least develop a clear enough picture while recognizing there are other viewpoints on what uncertainty means and how we should deal with it. In the next sections I outline a position from which civil engineers can work with community stakeholders to find common ground and facilitate a discussion about how to plan, design, and implement resilient infrastructure. I briefly discuss the concept of sustainability, the importance of incorporating climate change into infrastructure planning and design and the current position of civil engineering in relation to sustainability in Canada and the US. In Sect. 5 I outline a way forward that brings together risk and vulnerability perspectives on uncertainty and could allow for better planning and design for resilient infrastructure by engaging at the community level.

## 4 Sustainability, Civil Engineering, and Uncertainty

The notion of adaptation to the impacts of extreme events or resilience infrastructure with or without the occurrence of climate change starts from the concept of sustainability. Sustainability and sustainable development came to prominence with the Bruntland report in [9]. The definition for sustainable development from the report states that it “meets the needs of the present without compromising the ability of future generations to meet their own needs”. While this part of the definition is quoted frequently in the literature, the Bruntland Commission report goes on to state.

...sustainable development is not a fixed state of harmony, but rather a process of change in which the exploitation of resources, the direction of investments, the orientation of technological development, and institutional change are made consistent with future as well as present needs [emphasis added].

This latter aspect of sustainable development described by Bruntland [9], while often overlooked, is implicit in many ways in more recent ideas of sustainability. For instance, Sachs [56] maintains that sustainability is a way of looking at the world, and a way of describing “shared aspirations for a decent life”. It is both an analytic theory, and a normative or ethical framework. This means that it is “a way to understand the world as a complex interaction of economic, social, environmental, and political systems”. It is also a world view (*Weltanschauung*), or a normative view of the world. A detailed review of the concept of sustainability is beyond the scope of this paper, and concepts of sustainability have been widely discussed and debated.

In relation to resilient infrastructure and climate change adaptation, the extended definition of sustainable development as a process allows for the direct incorporation of uncertainty into the discussion. That is, it is very difficult to not compromise “the ability of future generations to meet their own needs” if decision makers today are highly uncertain about the outcomes of their decisions. Setting aside a discussion on the uncertainty of the needs of future generations, sustainable development under uncertainty requires a robust process that can adjust to changing knowledge conditions.

## 4.1 *Call to Action*

The American Society of Civil Engineers (ASCE) and the Canadian Society for Civil Engineering (CSCE) have made strong efforts to incorporate sustainability into the civil engineering. The ASCE Call to Action aligns professional codes of ethics with sustainable development, and while not explicit, includes a call and responsibility to act on climate change. The ASCE, in partnership with the CSCE, has recently made a Call to Action for members regarding sustainability. As stated by Wright and Perks [66]:

Whereas

- I civil engineers provide essential infrastructure;
- II that infrastructure is inadequate and dangerously deteriorating;
- III availability of resources and future conditions are highly uncertain;
- IV current approaches, practices, and standards do not address the full range of societal needs.
- V Therefore,
  - civil engineering practice must be transformed, and
  - ASCE is responsible to lead this charge.
  - Goals:
    - 1.1. Transform how infrastructure is conceived, delivered, and operated to enable a sustainable future.
    - 1.2. Establish ASCE as the trusted leader and preferred resource for sustainable civil engineering practices.
    - 1.3. Make the Institute for Sustainable Infrastructure's Envision the broadly adopted framework for sustainable infrastructure.
    - 1.4. Expand the capacity of civil engineers to create relationships of trust and respect.

The ASCE defines of sustainability in general as a set of environmental, economic, and social conditions—the “Triple Bottom Line”—in which all of society has the capacity and opportunity to maintain and improve its quality of life indefinitely, without degrading the quantity, quality or the availability of natural, economic, and social resources. The foundational principles of the ASCE's definition of sustainable development are, “doing the right project” and “doing the project right” [65]. All this is a clear indication that the engineering profession, civil engineering in particular, is ready to move forward, and there are a growing number cases where such practices have been put into place.

However there is still no standard for how best to deal with uncertainty. From an engineering perspective Ayyub [3] summarizes types and sources of uncertainty as follows (compare with above):

1. Recognized and well-characterized (e.g., properties of materials)
2. Recognized and moderately characterized (future climate variables)

3. Recognized and poorly characterized (future energy use)
4. Recognized but cannot be characterized (governance and cooperation)
5. Unknown, not yet discovered.

The first two categories fit well into the risk perspective of uncertainty (Sect. 3.1), (3) and (4) represent uncertainty that is either much more difficult to quantify or not quantifiable, and (5) is an admission of ignorance. There are generally two ways to approach addressing these issues, top-down and bottom-up. According to [52], “[i]n order to recognize the potential impacts of climate change, top-down, scenario-driven approaches are most effective. In order to catalyze the choice to act, bottom-up, vulnerability-driven approaches are needed for personal relevancy”. The former represents the risk perspective and the latter the vulnerability perspective of uncertainty. Projects that have attempted to integrate these aspects have mostly been large scale in North America, and there is not a lot of guidance for smaller projects involving more rural communities [8]. Engineers often do not know how to proceed to properly address sustainability and move in the wrong direction by following traditional (risk) methods.

## 5 A Path Forward

Gibbs et al. [28] identify three scales for resilience in civil engineering: community or urban, organizational, and individual assets. Traditional approaches to engineering design are often focused on optimization of reliability and robustness of infrastructure assets, however resilience to climate change does not apply to a single piece of infrastructure, resilience is a property of a system [62]. Conventional approaches to planning and design also tend to be reactive, not proactive, and do not tend to include contingencies that may arise over the lifetime of a project [41]. Additionally, most approaches to planning and design for resilient infrastructure have used a top-down risk assessment and management framework to identify and assess risks prior to implementation of a response [49].

I have argued that the risk assessment perspective on uncertainty is limited, and I believe there are some open questions regarding how uncertainty is quantified through the use of probability distributions. For instance, while a probability distribution is supposed to be representative of a complete description of uncertainty, a confidence interval is not a representation of uncertainty, it represents the variability of an estimate of a given variable [54]. While a detailed review these issues is beyond the scope of this paper, there are a variety of assessment tools that use probabilities for analysis of resilience, in addition to or instead of risk, over a wide range of applications [58]. For instance, there has been a move toward assessing resilience in urban water systems [38], in particular with consideration of the changing climate [53, 68]. Furthermore, recent studies have focused on bottom-up assessments of vulnerability

and adaptive capacity of existing storm water systems to climate change [34, 42], including communities that have combined sewer systems [25], and the impacts of projected changes of rainfall on drainage systems [47, 63].

Building adaptive capacity to climate change requires integration across multiple scales, which includes institutional, technical, economic, financial, environmental, and socio-cultural aspects [1]. This is a key part of resilient infrastructure, and successful adaptation requires iteration, monitoring, evaluation, and social learning [43]. A participatory approach is required that increases the likelihood of cooperation and coordination of actions toward sustainable design strategies [41]. Communication, in general, of uncertainty, in particular, is critical to reach sustainable solutions [11]. A process that allows input and discussion from stakeholders, and those who bear the consequences, is more likely to lead to less conflict and better decisions [61].

Incorporation of all these considerations leads to a state of deep uncertainty. Fortunately there are methods available for decision making under deep uncertainty [39]. A full review of decision tools and methods is beyond the scope of this paper, but I will briefly mention several that are used in ongoing research by the author. One such method is Info-Gap decision theory (IG) that provides a non-probabilistic method of quantifying deep or severe uncertainty. It focuses on “the disparity between what is known and what could be known,” without a statement about the structure of the uncertainty as with a probability distribution [6] [emphasis original]. The IG model has been used in application to water resources systems, and it has performed as well as other probabilistic approaches [51]. Roach et al. [51] suggested that application of IG methods would be improved if combined with relevant measures of uncertainty, and hybrid and combined IG models exist to allow for integration of probability distributions into the IG model [7]. Info-Gap methods have been used in practice combined with frequentist and Bayesian methods [44], and have been used in a climate change context with regard to policy decisions [29]. Info-Gap methods fit well and can be used with a range of other methods that can deal with deep uncertainty. This includes Robust Decision Making [37], Engineering Options Analysis [17], Dynamic Adaptive Policy Pathways [30, 31], and other pathways or scenario approaches [21].

Methods that can incorporate uncertainty that cannot be quantified by probability distributions allows and requires social science-based inputs and community-based knowledge to be brought directly into decision-support tools. This creates the potential to create novel decision-support methodologies relevant to civil engineering and design and planning of public infrastructure systems. As of this writing, I am working within this framework on a project funded by the Leslie Harris Center of Regional Policy and Development at Memorial University of Newfoundland. This project seeks to involve the communities in the Baccalieu Trail region of Newfoundland in developing sustainable strategies for upgrading storm water and road infrastructure using a watershed-based systems approach that incorporates climate change. In conjunction with this, a project funded by the Ocean Frontier Institute (led by Dalhousie University, Memorial University, and the University of Prince Edward Island) on future ocean and coastal infrastructure that began in 2021. A multi-disciplinary team is working on this project, including an engineer, geographer,

social psychologist, and sociologist. A key part of the project entails the design of a survey for stakeholders, exploring knowledge, and concerns around climate adaptation (needs, options, local context). These ongoing and future projects will allow some of the ideas expressed in this paper to be tested in a more rigorous manner with the purpose of contributing to the development of an accessible guide to adaptation planning.

## 6 Conclusions

Investments in public infrastructure by the GNL recognized the need to incorporate climate change considerations into planning a design. The GNL funded work to develop data and tools for this purpose. Engineers in the province were not using this information due to a lack of awareness and inability to utilize available data and tools to incorporate climate change into design. An NRCan funded project has facilitated increased awareness and attempted to build the capacity of engineers in order to increase utilization of these data and tools. Even with increased awareness and capacity, engineers are not compelled to bring sustainability and climate adaptation until adaptation capacity, resilience, and sustainability become mainstream in the engineering design and planning process. At present, there is movement within the engineering community to follow the principles of sustainability and incorporate climate change into planning and design. However this process takes time, and engineers have been required to do so in Newfoundland and Labrador by provincial regulations, such as including updated IDF curves and climate change into flood plain mapping. Engineers in Canada will also be compelled to incorporate climate change in federally funded projects that require risk assessment be done through the climate lens [32]. I have argued that misunderstanding and miscommunication of uncertainty plays a large part in the reluctance of engineers to include uncertainty into the infrastructure planning and design process. The profession must encourage the inclusion of new approaches and methods to meet current needs while engaging stakeholders at the community level.

Explicit and open discussion about perceptions, attitudes, and the meaning of uncertainty will help locate the type and sources of uncertainty that need to be considered in the planning and design of public infrastructure. Technical aspects of uncertainty can be handled by those with the proper technical expertise. This has the potential to shift the focus of decision making for sustainable action at the local level to focus on other sources and types of uncertainty within the control (and understanding) of local stakeholders and decision makers. Discussions can shift from the technical aspects of uncertainty to focus on how communities will choose to deal with climate change and uncertainty. Local knowledge and how communities may prefer to manage uncertainties can be used as input into the design process by engineers in conjunction with parameter and model uncertainties. Decision-support tools and models can fully incorporate the quantitative and qualitative uncertainty to analyze and identify strategies and pathways to meet sustainable development goals.

Sustainable actions can be enhanced through a robust process of decision making, and by recognize that uncertainty provides an opportunity to do things right.

## References

1. Amadei B (2014) Engineering for sustainable human development. ASCE, Reston
2. Aspinall WP, Cooke RM (2013) Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier J, Sparks S, Hill LJ (eds) Risk and uncertainty assessment for natural hazards. Cambridge University Press
3. Ayyub BM (2018) Ayyub BM (ed) Climate-resilient infrastructure: adaptive design and risk management. American Society of Civil Engineers, Reston
4. Barnett J et al (2013) Reducing the risk of maladaptation in response to sea level rise and urban water security. In: Moser SC, Boykoff MT (eds) Successful adaptation to climate change: linking science and policy in a rapidly changing world. Routledge, New York
5. Batterson M, Liverman D (2010) Past and future sea-level change in newfoundland and labrador: guidelines for policy and planning. Newfoundland and Labrador Department of Natural Resources Geological Survey, Report 10-1, pp 129–141
6. Ben-Haim Y (2006) Info-gap decision theory: decisions under severe uncertainty, 2nd edn. Academic Press, Amsterdam
7. Ben-Haim Y (2019) Info-gap decision theory (IG). In: Marchau VAWJ et al (eds) Decision making under deep uncertainty: from theory to practice. Springer International Publishing, pp 93–115
8. Birkmann J, Wenzel F, Greiving S, Garschagen M, Vallée D, Nowak W, Welle T et al (2017) Extreme events, critical infrastructures, human vulnerability and strategic planning: emerging research issues. *J Extreme Events* 03(04):1650017
9. Bruntland GH, WCED (1987) Our common future: report of the world commission on environment and development. United Nations; Oxford University, Oxford
10. Carmin J, Dodman D (2013) Engaging science and managing scientific uncertainty in urban climate adaptation planning. In: Successful adaptation to climate change: linking science and policy in a rapidly changing world. Routledge, New York
11. Cash DW et al (2003) Knowledge systems for sustainable development. *Proc Natl Acad Sci* 100(14):8086–8091
12. Catto N (2010) A review of academic literature related to climate change impacts and adaptation in Newfoundland and Labrador. Memorial University
13. Chen J, Brissette FP, Leconte R (2011) Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. *J Hydrol* 401(3):190–202
14. Coates T, Tapsell S (2019) Planning for an uncertain future: the challenges of a locally based collaborative approach to coastal development decisions. *Environ Sci Policy* 101:24–31
15. Cornell SE, Jackson MS (2013) Social science perspectives on natural hazards risk and uncertainty. In: Rougier J et al (eds) Risk and uncertainty assessment for natural hazards. Cambridge University Press
16. Daraio JA, Khan AA, Finnis J (2019) Incorporating climate change considerations into flood mapping and infrastructure design in newfoundland and labrador. In: CSCE general conference
17. de Neufville R, Smet K (2019) Engineering options analysis (EOA). In: Marchau VAWJ et al (eds) Decision making under deep uncertainty: from theory to practice. Springer Publishing, pp 117–132
18. DeNicola DR (2017) Understanding ignorance: the surprising impact of what we don't know. The MIT Press, Cambridge
19. Dessai S, Hulme M, Lempert R, Pielke R Jr (2009) Climate prediction: a limit to adaptation? In: Adger WN et al (eds) Adapting to climate change: thresholds, values, governance. Cambridge University Press, Cambridge, UK, pp 64–78

20. Di Baldassarre G, Nohrstedt D, Mård J et al (2018) An integrative research framework to unravel the interplay of natural hazards and vulnerabilities. *Earth's Future* 6(3):305–310
21. Eisenhauer DC (2016) Pathways to climate change adaptation: making climate change action political. *Geogr Compass* 10(5):207–221. <https://doi.org/10.1111/gec3.12263>
22. Faber M, Manstetten R, Proops JLR (1992) Humankind and the environment: an anatomy of surprise and ignorance. *Environ Values* 1(3):217–241
23. Feofilovs M, Romagnoli F (2017) Resilience of critical infrastructures: probabilistic case study of a district heating pipeline network in municipality of Latvia. *Energy Proc* 128:17–23
24. Finnis J, Daraio J (2018) Projected impacts of climate change for the province of newfoundland and labrador: 2018 update. Memorial University of Newfoundland
25. Fortier C, Mailhot A (2015) Climate change impact on combined sewer overflows. *J Water Resour Plan Manag* 141(5):04014073–04014077
26. Freer J, Beven KJ, Neal J, Schumann G, Hall J, Bates P (2013) Flood risk and uncertainty. In: Rougier J, Sparks S, Hill LJ (eds) *Risk and uncertainty assessment for natural hazards*. Cambridge University Press, pp 190–233
27. Frege G (1996) On sense and nominatum. In: Martinich AP (ed) *The philosophy of language*, 3rd edn. Oxford University Press, Oxford, pp 186–198
28. Gibbs M, Lemay L, Vinson T (2017) Resilience. In: Kelly WE, Luke B, Wright RN (eds) *Engineering for sustainable communities: principles and practices*. ASCE, Reston, pp 269–281
29. Groves DG et al (2019) Robust decision making (RDM): application to water planning and climate policy. In: Vincent AW et al (eds) *Decision making under deep uncertainty: from theory to practice*, pp 135–163. Springer International Publishing, Cham
30. Haasnoot M, Kwakkel JH, Walker WE, Maat J (2013) Dynamic adaptive policy pathways: a method for crafting robust decisions for a deeply uncertain world. *Glob Environ Chang* 23(2):485–498. <https://doi.org/10.1016/j.gloenvcha.2012.12.006>
31. Haasnoot M, Warren A, Kwakkel JH (2019) Decision making under deep uncertainty, from theory to practice, pp 71–92. [https://doi.org/10.1007/978-3-030-05252-2\\_4](https://doi.org/10.1007/978-3-030-05252-2_4)
32. Infrastructure Canada (2019) Climate lens: general guidance version 1.2.
33. IPCC (2022) Climate change 2022 impacts, adaptation, and vulnerability summary for policy makers. [https://report.ipcc.ch/ar6wg2/pdf/IPCC\\_AR6\\_WGII\\_SummaryForPolicymakers.pdf](https://report.ipcc.ch/ar6wg2/pdf/IPCC_AR6_WGII_SummaryForPolicymakers.pdf). Accessed 5 Mar 22
34. Kang N, Kim S, Kim Y, Noh H, Hong S, Kim H (2016) Urban drainage system improvement for climate change adaptation. *Water* 8(7):268
35. Knutti R (2008) Should we believe model predictions of future climate change? *Philos Trans R Soc A Math Phys Eng Sci* 366:4647–4664
36. Knutti R (2019) Closing the knowledge-action gap in climate change. *One Earth* 1(1):21–23
37. Lempert RJ (2019) Robust decision making (RDM). In: Marchau VAWJ, Walker WE, Bloemen PJTM, Popper SW (eds) *Decision making under deep uncertainty: from theory to practice*. Springer International Publishing
38. Makropoulos C, Nikolopoulos D, Palmen L, Kools S, Segrave A, Vries D, Koop S et al (2018) A resilience assessment method for urban water systems. *Urban Water J* 15(4):316–328
39. Marchau VAWJ, Walker WE, Bloemen PJTM, Popper SW (eds) (2019) *Decision making under deep uncertainty, from theory to practice*. Springer, Cham
40. Milly PCD, Betancourt J, Falkenmark M, Hirsch RM, Kundzewicz ZW, Lettenmaier DP, Stouffer RJ (2008) Stationarity is dead: whither water management? *Science* 319(5863):573–574
41. Moallemi EA, Malekpour S (2017) A participatory exploratory modelling approach for long-term planning in energy transitions. *Energy Res Soc Sci* 35:205–216 (*Energy Res Soc Sci* 35, 2018). <https://doi.org/10.1016/j.erss.2017.10.022>
42. Moore TL, Gulliver JS, Stack L, Simpson MH (2016) Stormwater management and climate change: vulnerability and capacity for adaptation in urban and suburban contexts. *Clim Change* 138(3–4):491–504. <https://doi.org/10.1007/s10584-016-1766-2>
43. Moser SC, Boykoff MT (2013) Climate change and adaptation success: the scope of the challenge. In: Moser SC, Boykoff MT (eds) *Successful adaptation to climate change: linking science and policy in a rapidly changing world*. Routledge, New York, NY, pp 1–34

44. O'Malley D, Vesselinov VV (2015) Bayesian-information-gap decision theory with an application to CO<sub>2</sub> sequestration. *Water Resour Res* 51(9):7080–7089
45. Opdyke A, Javernick-Will A, Koschmann M (2017) Infrastructure hazard resilience trends: an analysis of 25 years of research. *Nat Hazards* 87(2):773–789
46. Patt A (2013) Climate risk management: laying the groundwork for successful adaptation. In: *Successful adaptation to climate change: linking science and policy in a rapidly changing world*. Routledge, New York, NY
47. Pereira MJ et al (2014) Climate change impacts in the design of drainage systems: case study of Portugal. *J Irrig Drain Eng* 05014009(2):1–11
48. Pérez-Blanco CD (2022) Navigating deep uncertainty in complex human–water systems. In: Kondrup C et al (eds) *Climate adaptation modelling*. Springer International Publishing, Cham, pp 169–178
49. Ploberger C, Filho WL (2016) Towards long-term resilience: the challenge of integrating climate change related risks into a risk analysis framework. In Filho WL et al (eds) *Climate change adaptation, resilience and hazards*. Springer International Publishing Switzerland, pp 369–379
50. Rehak D, Senovsky P, Hromada M, Lovecek T, Novotny P (2018) Cascading impact assessment in a critical infrastructure system. *Int J Crit Infrastruct Prot* 000:1–14
51. Roach T, Kapelan Z, Ledbetter R (2015) Comparison of info-gap and robust optimisation methods for integrated water resource management under severe uncertainty. *Proc Eng* 119:874–883. <https://doi.org/10.1016/j.proeng.2015.08.955>
52. Roberts C (2013) Introduction to adaptation and resilience. In: McGregor A, Roberts C, Cousins F (eds) *Two degrees: the built environment and our changing climate*. Routledge, New York, NY, pp 177–183
53. Roca M et al (2016) Methodology to assess coastal infrastructure resilience to climate change. In: *FLOODrisk 2016—3rd European conference on flood risk management*, vol 02004, pp 1–9
54. Rougier JC (2013) Quantifying hazard losses. In: Rougier J, Sparks S, Hill LJ (eds) *Risk and uncertainty assessment for natural hazards*. Cambridge University Press, pp 19–39
55. Rougier JC, Beven KJ (2013) Model and data limitations: the sources and implications of epistemic uncertainty. In: Rougier J, Sparks S, Hill LJ (eds) *Risk and uncertainty assessment for natural hazards*. Cambridge University Press, pp 40–63
56. Sachs JD (2015) *The age of sustainable development*. Columbia University Press, New York
57. Shadabfar M et al (2022) Resilience-based design of infrastructure: review of models, methodologies, and computational tools. *ASCE-ASME J Risk Uncert Eng Syst Part A Civ Eng* 8(1):03121004. <https://doi.org/10.1061/AJRUA6.0001184>
58. Sharifi A (2016) A critical review of selected tools for assessing community resilience. *Ecol Ind* 69:629–647. <https://doi.org/10.1016/j.ecolind.2016.05.023>
59. Smithson M (1989) *Ignorance and uncertainty: emerging paradigms*. Springer. <https://doi.org/10.1007/978-1-4612-3628-3>
60. Tsavdaroglou M, Al-Jibouri SHS, Bles T, Halman JIM (2018) Proposed methodology for risk analysis of interdependent critical infrastructures to extreme weather events. *Int J Crit Infrastruct Protect* 21:57–71. <https://doi.org/10.1016/j.ijcip.2018.04.002>
61. Voinov A, Bousquet F (2010) Modelling with stakeholders. *Environ Model Softw* 25(11):1268–1281. <https://doi.org/10.1016/j.envsoft.2010.03.007>
62. Walker B, Salt D (2006) *Resilience thinking: sustaining ecosystems and people in a changing world*. Island Press, Washington, DC
63. Weesakul U, Chaowiwat W, Rehan MM, Weesakul S (2017) Modification of a design storm pattern for urban drainage systems considering the impact of climate change. *Eng Appl Sci Res* 44(3):161–169. <https://doi.org/10.14456/easr.2017.24>
64. Willems P, Arnbjerg-Nielsen K, Olsson J, Nguyen VTV (2012) Climate change impact assessment on urban rainfall extremes and urban drainage: methods and shortcomings. *Atmos Res* 103(C):106–118
65. Wright R, Lake B, Perks A (2017) Introduction. In: *Engineering for sustainable communities: principles and practices*. ASCE Press, Reston



66. Wright R, Perks A (2017) Preface. In: Engineering for sustainable communities: principles and practices. ASCE Press, Reston
67. Zhao S, Liu X, Zhuo Y (2017) Hybrid hidden Markov models for resilience metrics in a dynamic infrastructure system. Reliab Eng Syst Saf 164:84–97. <https://doi.org/10.1016/j.ress.2017.02.009>
68. Zhou Q (2014) A review of sustainable urban drainage systems considering the climate change and urbanization impacts. Water 6(4):976–992. <https://doi.org/10.3390/w6040976>

# Prioritization of Barriers for Photovoltaic Solar Waste Management in Saskatchewan



Monasib Romel, Golam Kabir, and Kelvin Tsun Wai Ng

**Abstract** The worldwide exponential upsurge of photovoltaic panel installations and the subsequent heights of photovoltaic waste is a matter of intense concern. There is an estimation that within next 2050 year, the worldwide generation of photovoltaic waste may rise upto 60–78 million tons. The objective of this study is to identify the crucial barriers to photovoltaic solar waste management in Canada and prioritize them. At first, the barriers to photovoltaic solar waste management were identified through literature review and expert feedback. Face-to-face interviews were conducted with the selected seven experts who have comprehensive knowledge and expertise on solar management in Saskatchewan. In this study, Analytic Hierarchy Process (AHP) is used to analyze and find prioritization among these barriers. Some crucial barriers from each category are lack of legislative framework, lack of monitoring and supervision, generation of the low volume of solar waste, low profitability in recycling, lack of consumer awareness, and lack of knowledge about business opportunities. Among the rest, lack of restriction on landfill disposal, undefined role of stakeholders, lack of subsidy and tax rebate, additional cost for consumers, lack of knowledge about business opportunities, insufficient campaigns are also worth mentionable barriers. This study is expected to contribute to the concerned government agencies to assess, evaluate, and utilize the priority of barriers to establish a sustainable and resilient solar waste management plan in Saskatchewan, Canada.

**Keywords** Photovoltaic panel · Solar waste · Analytic hierarchy process · Waste management · Barriers

---

M. Romel (✉) · G. Kabir · K. T. W. Ng  
University of Regina, Regina, Canada  
e-mail: [mar070@uregina.ca](mailto:mar070@uregina.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_74](https://doi.org/10.1007/978-3-031-34593-7_74)

1171

# 1 Introduction

To ensure the maximum benefit from renewable energy, the management of its waste generated in the process has to be identified and managed in a sustainable way to minimize the potential adverse impact on the environment [18]. Solar, hydro, wind, tidal, biomass, geothermal are some of the mentionable renewable energy technologies in the world. With the transition from fossil to renewable sustainable sources of energy, the waste generation is also matter of concern. For solar, there is an estimation that by 2050 the generation of photovoltaic (PV) solar waste will reach between 60 and 78 million tones [8]. Considering this growth of photovoltaic panels, it can be estimated that this can reach the share of 10% among all kinds of electrical and electronic waste by the year 2050 [8]. Although photovoltaic panels are known as reliable, efficient, silent, environment friendly, and zero carbon emission devices, the impact on humans and the environment as waste has been realized gradually [25]. Photovoltaic panels, batteries, inverters, and other accessories contain hazardous materials such as cadmium, lead, tin, and lithium, which can be harmful to both humans and the environment if these elements are not properly managed when they become waste [9]. On the other hand, some materials are scarce and worthy enough for recovering elements like gallium, tellurium, indium, ruthenium, etc. [7]. This means that solar energy is not free of impacts and it can become harmful at the end of its life when it becomes waste [1, 19]. To ensure environmental safety and to avoid a shortage of crucial materials in the future, these expired or damaged or out-of-order panels needed to be treated appropriately, i.e., reuse, recycle, and refurbish [16].

In the year 2020, the volume of photovoltaic waste in Canada was 700 MT whereas this volume is forecasted at 13,000 MT in the year 2030 as mentioned in the report by International Renewable Energy Agency (IRENA) and the International Energy Agency (IEA) [12]. In a recent announcement, the Canadian Government declared “A healthy environment and a healthy economy” at the UN climate change conference 2020 where the concept of “Net Zero” was discussed by which Canada will reach net zero emission by 2050 (Ambiente Cambio Climatico Canada, 2020). Installation of solar panels is one of the effective solutions of producing electricity for Canada to reach their target of 2050 (Ambiente Cambio Climatico Canada, 2020). The growth of solar waste is now inevitable. These expired solar panels will generate hazardous waste because of the leaching of heavy metals like tin or lead or contaminated soil or ground [17, 22].

This study identifies the gap for working with the barriers of management of solar wastes management. One of the applications of the Multi-Criteria Decision Analysis (MCDA) called Analytic Hierarchy Process (AHP) method is used here for analysis of barriers of solar waste management. The main research objective of this project is to identify the barriers of solar waste management and then rank them to prioritize for implementation of sustainable solar waste management system. The findings can be used by concerned government authority, policy makers, or personnel who are related to the dialogue, collaboration, stewardship, or solar firms’ operation and maintenance in Saskatchewan, Canada.

## **2 Literature Review**

### ***2.1 Economic Barriers***

Economic barriers are considered to be one of the crucial barriers [10]. Some of them are chemical prices of the recycling process, recyclers transportation cost along with the capital investment cost of the collection center, recycling plants, and necessary recycling machinery of the plant. In this study, D'Adamo et al. [6] conducted economic feasibility and found that PV wastes are not profitable for recycling. On average, the recycling cost per panel is \$15–\$20, whereas for landfilling a panel the cost is only \$5–\$7. All the associated costs like dismantle cost, transportation cost, treatment cost, labor cost, disposal cost, etc. are taken by the customer. Solar panels are not included in any provincial Extended Producers Responsibility (EPR) as mentioned in the report published by Canadian Council of Ministers of the Environment [3].

### ***2.2 Social and Technological Barriers***

Consumers choose curbside disposal rather than taking Solar all the way to the specific collection center facility. This was found in a report done by Song et al. [23]. They have also concluded that customer does not believe in paying for any of the waste management process which strengthens the need for a suitable business model or endorsement for evolving effective comprehensive PV waste management system. Another barrier is shared by Kim and Jeong [13] that in some cases manufacturers of PV panels do not share the patented material information. That is why recyclers are finding it difficult to find the best technique for recycling these PV modules.

### ***2.3 Policy and Regulation Barriers***

Proper regulations will ensure the consciousness of manufacturer, traders, importer, exporter, and retailer to maintain the orders imposed by the government [15]. So far there is no an incentive package from the government to endorse the collection and recycling program. Law enforcement for identification and monitoring of such PV panels are required for registration of the solar system for the household consumers unless they are seeking a rebate [10, 17] (Table 1).

**Table 1** Summary of barriers against solar waste management

General aspect	Barriers		Key points	References
Economic (EC) barriers	Profitability in recycling	EC1	There might be a material loss in the process of collecting the panels which might hamper the profitability for the recyclers	[4, 5, 8, 10, 11]
	Cost for consumer	EC2	Cost like dismantle cost, transportation cost, treatment cost, labor cost, disposal cost is taken by the customer	[10, 13, 14]
	Responsibility of producer	EC3	Producers are not taking responsibility of the solar waste because of lack of government incentives and irregular	[3, 8]
	No subsidy and tax rebate	EC4	There is still no subsidy or tax rebate for formal recyclers	Expert opinion
	Unwillingness to pay	EC5	Experts says it might not be possible to encourage consumers to pay to remove expired panels	Expert opinion
	Low volume of solar waste	EC6	Uncertainty about the volume of PV waste because they are removed for a wide range of economic reasons	[10], Expert opinion
Policy and regulation (PR) barriers	Legislative framework	PR1	Authorities are still seating back thinking that they have time to think about the waste in future	[8, 15]
	Monitoring and supervision	PR2	The already installed solar panels are not traceable, the waste will not be traceable as well	[10, 17]

(continued)

**Table 1** (continued)

General aspect	Barriers		Key points	References
	Undefined role of stakeholders	PR3	The role is not defined in current solar waste management practices	[27]
	No restriction on landfill disposal	PR4	There is no such restriction on disposal of solar panels in Regina landfill in Regina and Saskatoon	[2, 4]
Social and knowledge (SK) barriers	Lack of consumer awareness	SK1	Consumers know and appreciate the use of solar energy for the benefit zero carbon emission but they do not know about the fate of expired solar panels	[10, 17, 24]
	Insufficient campaign	SK2	To provide importance to the solar waste management, an integrated campaign can be launched to grab attention of the users	Expert opinion
	Less knowledge about business opportunities	SK3	Entrepreneurs have not found the business opportunity in recycling of expired panels	[26]

### 3 Methodology

In this study the AHP method has been used to rank the barriers of photovoltaic solar waste management. Here AHP has been used in basic form as described by Satty [20]. A converted scale has been used from the fundamental scale as defined by Satty [21] as described in Table 2.

The steps of the AHP method are described as follows:

Step-1: Selection of barriers for solar waste management through literature review, open data source and finally getting feedback from decision-makers of the industry.

Step-2: All these selected barriers are then distributed into a set of Main Barriers to be ranked. For each main barrier, a set of sub-barriers are also finalized for ranking among themselves under the umbrella of the main barrier.

**Table 2** Conversion from fundamental scale by Satty [21]

Intensity importance	Description	Explanation	Conversion	Description
1	Equal importance	Two activities contribute equally to the objective	1	Equal importance
2	Weak or slight			
3	Moderate importance	Experience and judgment slightly favor one activity over another	2	Moderate importance
4	Moderate plus			
5	Strong importance	Experience and judgment slightly favor one activity over another	3	Strong importance
6	Strong plus			
7	Very strong	An activity is favored very strongly over another; its dominance demonstrated in practice	4	Very strong importance
8	Very very strong			
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation	5	Extreme importance

Step-3: Determination of a pair-wise comparison of relative importance among the  $m$  barriers.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \cdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \tag{1}$$

In the above equation,  $a_{ij} > 0$ ,  $a_{ij} = \frac{1}{a_{ji}}$ ,  $a_{ii} = 1$ ,  $a_{ij}$  is the rating of importance of barrier  $i$  in respect barrier  $j$ . For example, if the barriers are of equal importance, then  $a_{ij} = a_{ji} = 1$ .

Step-4: Numerical weight and ranking: In this step, the numerical weights are to be calculated which will be assigned against each barrier  $B_1, B_2, B_3 \dots B_m$  on the basis of pair-wise comparison matrix  $A$ . The obtained weight vector is shown as follows:

$$W = [W_1 \ W_2 \ \cdots \ W_m] \tag{2}$$

This is known as normalized eigenvector of matrix  $A$ . Element of each column is divided by the summation of that column for normalization of the value. Then the elements of each row are summed up and divided by the total number of barriers

**Table 3** Randomly generated consistency index (RI)

$m$	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

of that category. The formula is shows as below:

$$W_i = \frac{1}{m} \sum_j \frac{a_{ij}}{\sum_i a_{ij}}. \quad (3)$$

Ranking is calculated of each barrier based on the value of its Weight obtained in by Eq. (3).

Step-5: Consistency Check: In this step, the consistency of the respondents is check by calculating the Consistency Index, CI with the following formula.

$$CI = \frac{\lambda_{\max} - m}{m - 1} \quad (4)$$

Here,  $\lambda_{\max}$  is the maximum Eigen value which is calculated with the following formula.

$$\lambda_{\max} = \frac{1}{m} \sum_{i=1}^m \frac{(Aw)_i}{w_i} \quad (5)$$

After comparing with the Randomly Generated Consistency Index (RI) through the consistency Ratio,  $CR = CI/RI$ , if we find the value of  $CI < 0.1$ , the degree of consistency is satisfactory (Table 3).

## 4 Case Study

In this study, nine decision-makers were selected out of twenty initial contacts based on their experience of more than four years for either in the practical installation or related to the dialogue, collaboration, stewardship, or solar firms' operation and maintenance. Primarily seventeen barriers were selected from literature review and open data source. These barriers were emailed for the final selection of analysis of prioritization among themselves. Finally, thirteen barriers were selected after eliminating or merging the barriers, since some of them were similar or overlapping. The decision-makers then participated in the set of questionnaires for AHP. The selection of decision-makers is shown in Table 4.



**Table 4** Summary of the decision-makers

Institution/company/forum	Designation	Number of responses	Service length
Government organizations	Project Engineer, Electrical Engineer, Recycling Executive, Waste Manager	3	More than 5 years
Academic expert	Assistant Professor	1	More than 4 years
Enlisted Solar Panel Installer in SaskEnergy	Site Engineer, Foreman, and Site Manager	3	More than 5 years
Autonomous Institution/forum for dialogue and collaboration	Policy Manager, Events Manager, Business Development Manager	2	More than 8 years

Now, with the help of Eq. (3), normalized numerical weights are calculated and rankings have been established. The weights and rankings are shown below from Tables 5, 6, 7 and 8.

**Table 5** Normalized weights and ranking of main barriers

Sub-barriers	Normalized weight	Rank
PR	0.54	1
EC	0.32	2
SK	0.13	3

**Table 6** Normalized weights and ranking for sub-barriers of policy and regulation barriers

Sub-barriers	Normalized weight	Rank
PR1	0.44	1
PR2	0.27	2
PR3	0.12	4
PR4	0.17	3

**Table 7** Normalized weights and ranking for sub-barriers of economic barriers

Sub-barriers	Normalized weight	Rank
EC1	0.17	2
EC2	0.10	5
EC3	0.15	3
EC4	0.10	4
EC5	0.05	6
EC6	0.26	1

**Table 8** Normalized weights and ranking for sub-barriers of social and knowledge barriers

Sub-barriers	Normalized weight	Rank
SK1	0.54	1
SK2	0.18	3
SK3	0.28	2

**Table 9** Consistency check eigen value method

Parameters symbols	Main barriers	Policy and regulation barriers	Economic barriers	Social and knowledge barriers
$\lambda$	3.11	4.18	6.61	3.08
$m$	3	4	6	3
RI	0.58	0.9	1.24	0.58
CI	0.0571	0.061	0.122	0.042
CR	9.85%	6.82%	9.88%	7.30%
	< 10% (accepted)	< 10% (accepted)	< 10% (accepted)	< 10% (accepted)

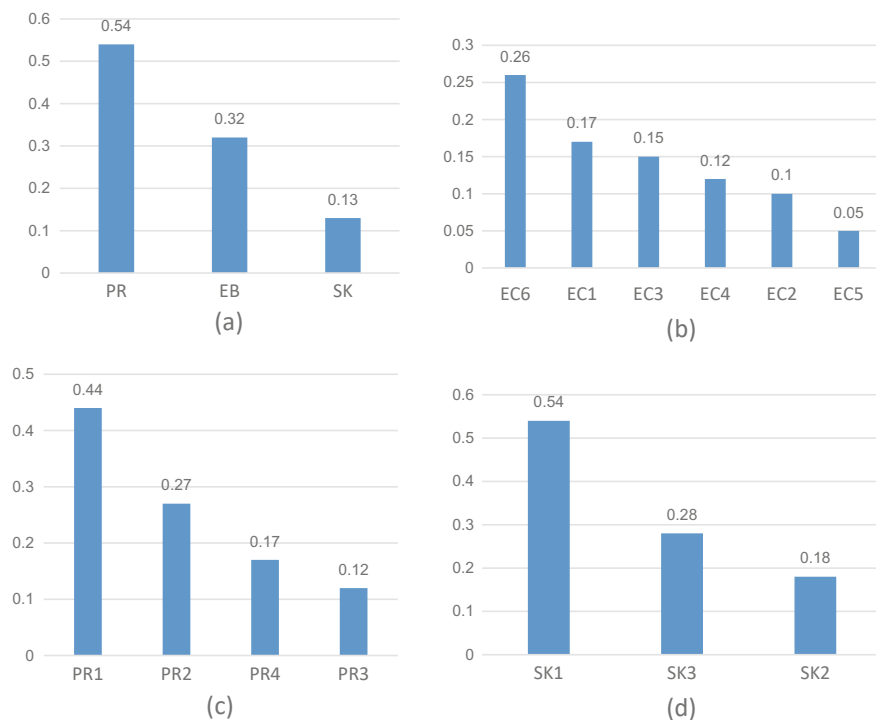
Finally, consistency check was performed for both the main barriers and sub-barriers as per Eqs. (4) and (5) and Table 3. The consistency is shown in Table 9.

Figure 1 shows the weight and ranking of main barriers and sub-barriers.

## 5 Model Comparison and Discussion

The study reveals that the Policy and regulation (PR) barrier is the most crucial among all the three main barriers for solar waste management process shown in Fig. 1a. Economic (EC) barrier is the second and Social and Knowledge (SK) barrier is the third crucial in the analysis. Within the concept of Policy and Regulations (PR), the ranking among sub-barriers is found as legislative framework (PR1), monitoring and Supervision (PR2), Lack of restriction on landfill disposal (PR4) and undefined role of stakeholders (PR3) respectively shown in Fig. 1c. This ranking indicates that formulation of a distinct law for solar waste management is the crucial for sustainable solar waste management in Canada. As mentioned by Malandrino et al. [15], one sustainable policy can ensure awareness of all different stakeholders, manufacturer, traders, importer, exporter, and retailer to follow the law imposed by the government.

The second crucial dimension of barrier is the Economic Barrier. Among all its' sub-barriers, Low Volume of Solar Waste (EC6) has achieved the highest priority and Profitability in Recycling (EC1) and Responsibility of Producer (EC3) being the second and third priority as per respondents' opinion shown in Fig. 1b. For



**Fig. 1** Weight and ranking of main barriers and sub-barriers

example, Canadian Council of Ministers of the Environment [3] has mentioned in their published report that Solar panels are not included in any provincial Extended Producers Responsibility (ERP) in Canada. On the other hand, among the rest of the sub-barriers, Subsidy and Tax Rebate (EC4), Cost for Consumer (EC2), and Unwillingness to Pay (EC5) are worth mentionable as per chronological order of priority in the study.

Among all the sub-barriers in Social and Knowledge aspect, Lack of Consumer Awareness (SK1) should be given the highest priority among all others, whereas lack of knowledge about business opportunities (SK3) and Insufficient campaign (SK2) are ranked as second and third priority respectively shown in Fig. 1d. We find the similar comment in the research report of Song et al. [23] where consumers choose curbside disposal rather than taking solar all the way to the specific collection center facility. Again, awareness program related to the solar waste management needs to be started for awareness of users and commercial companies for their business windows. Door to door collection system can be created through integration of different stakeholders of solar industry.

## 6 Conclusion

The generation of solar panel waste will be increasing quickly. In this study, we tried to identify and investigate barriers associated with the solar panel waste management process in the context of Saskatchewan. In this study, AHP has been used for the investigation and prioritization of these barriers. There is a number of sub-barriers under each of these main barriers. Out of thirteen barriers of all categories, the findings exhibit that the top two barriers from each category are measured to be the most crucial barriers in the implementation of solar waste management such as Legislative Framework (PR1), Monitoring and Supervision (PR2), Low Volume of Solar Waste (EC6), Profitability in Recycling (EC1), Lack of Consumer Awareness (SK1), and Lack of knowledge about business opportunities (SK3). The role of the manufacturers and distributors of solar panels is crucial and their role of taking back the solar waste panels should also be clearly defined in the policy or regulatory framework. In terms of social and knowledge-based categories, the government needs to provide different platforms for human resource training in centers like Regina trade and skill center, Regina's open-door society on solar waste management technology. Integrated solar waste management with the participation of all stakeholders appears to be the right solution for the management.

For analysis, experts' opinion has been obtained, justified opinions were converted into numerical input carefully for analysis. The number of experts could have been increased more for different disciplines to get the impact from a diversified perspective. The framework is applicable for the other provinces of Canada for identification and prioritization for analysis of waste management of various other industries like leather, chemical, hospital construction, and others. Future research related to this work is based on the application of some of the multicriteria methods, for example, the TOPSIS for barrier evaluation and their ranking.

**Acknowledgements** Authors wish to acknowledge the support of Faculty of Graduate Studies and Research (FGSR) and all the respondents of the survey from Saskpower, City of Regina, City of Saskatoon, Canadian Solar Industries Association (CanSIA), Canadian Renewable Energy Association (CanREA), CanCORE Canadian Council on Renewable Electricity, and Canadian Solar.

## References

1. Behi Z, Ng KTW, Richter A, Karimi N, Ghosh A, Zhang L (2022) Exploring the untapped potential of solar photovoltaic energy at a smart campus: Shadow and cloud analyses. *Energy Environ* 33(3):511–526. <https://doi.org/10.1177/0958305X211008998>
2. Canadian Council of Ministers of the Environment (2006) National guidelines for hazardous waste landfills
3. Canadian Council of Ministers of the Environment (2014) Progress report on the Canada-wide action plan for extended producer responsibility, pp 1–17
4. City of Regina (2020) City of Regina acceptable landfill materials. <https://www.regina.ca/home-property/recycling-garbage/landfill/>

5. Collins MK, Anctil A (2017) Implications for current regulatory waste toxicity characterisation methods from analysing metal and metalloid leaching from photovoltaic modules. *Int J Sustain Energy* 36(6):531–544. <https://doi.org/10.1080/14786451.2015.1053392>
6. D'Adamo I, Miliacca M, Rosa P (2017) Economic feasibility for recycling of waste crystalline silicon photovoltaic modules. *Int J Photoenergy* 2017:1–7. <https://doi.org/10.1155/2017/4184676>
7. Faircloth CC, Wagner KH, Woodward KE, Rakkwamsuk P, Gheewala SH (2019) The environmental and economic impacts of photovoltaic waste management in Thailand. *Resour Conserv Recycl* 143:260–272. <https://doi.org/10.1016/j.resconrec.2019.01.008>
8. Farrell CC, Osman AI, Doherty R, Saad M, Zhang X, Murphy A, Harrison J, Vennard ASM, Kumaravel V, Al-Muhtaseb AH, Rooney DW (2020) Technical challenges and opportunities in realising a circular economy for waste photovoltaic modules. *Renew Sustain Energy Rev* 128:109911. <https://doi.org/10.1016/j.rser.2020.109911>
9. Fthenakis V, Athias C, Blumenthal A, Kulur A, Magliozzo J, Ng D (2020) Sustainability evaluation of CdTe PV: an update. *Renew Sustain Energy Rev* 123:109776. <https://doi.org/10.1016/j.rser.2020.109776>
10. FutureBridge (2021) The future of solar panel recycling: a circular economy insight. Limited, Cheers Interactive (India) Private. <https://www.futurebridge.com/industry/perspectives-energy/the-future-of-solar-panel-recycling-a-circular-economy-insight/>
11. Heelan J, Gratz E, Zheng Z, Wang Q, Chen M, Apelian D, Wang Y (2016) Current and prospective Li-Ion battery recycling and recovery processes. *JOM* 68(10):2632–2638. <https://doi.org/10.1007/s11837-016-1994-y>
12. IRENA (2016) End of life management solar PV panels. International Renewable Energy Agency (IRENA) and the International Energy Agency (IEA)
13. Kim S, Jeong B (2016) Closed-loop supply chain planning model for a photovoltaic system manufacturer with internal and external recycling. *Sustainability (Switz)* 8(7):1–17. <https://doi.org/10.3390/su8070596>
14. Liu C, Zhang Q, Wang H (2020) Cost-benefit analysis of waste photovoltaic module recycling in China. *Waste Manage* 118:491–500. <https://doi.org/10.1016/j.wasman.2020.08.052>
15. Malandrino O, Sica D, Testa M, Supino S (2017) Policies and measures for sustainable management of solar panel end-of-life in Italy. *Sustainability (Switz)* 9(4):1–15. <https://doi.org/10.3390/su9040481>
16. Markert E, Celik I, Apul D (2020) Private and externality costs and benefits of recycling crystalline silicon (c-Si) photovoltaic panels. *Energies* 13(14). <https://doi.org/10.3390/en13143650>
17. Mathur D, Gregory R, Simmons T (2020) End-of-life management of solar PV panels. [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjyKr9jqnxAhUNU30KHatrANEQFnoECAyQAA&url=https%3A%2F%2Fwww.cdu.edu.au%2Fsites%2Fdefault%2Ffiles%2Fthe-northern-institute%2Fsolmanagemnetsolarpv\\_final\\_e-version.pdf&usq=AOvVaw2qjdkLu](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjyKr9jqnxAhUNU30KHatrANEQFnoECAyQAA&url=https%3A%2F%2Fwww.cdu.edu.au%2Fsites%2Fdefault%2Ffiles%2Fthe-northern-institute%2Fsolmanagemnetsolarpv_final_e-version.pdf&usq=AOvVaw2qjdkLu)
18. Pradhan P, Costa L, Rybski D, Lucht W, Kropp JP (2020) A systematic study of sustainable development goal (SDG) interactions. *Earth's Future* 5(11):1169–1179. <https://doi.org/10.1002/2017EF000632>
19. Salim HK, Stewart RA, Sahin O, Dudley M (2019) Drivers, barriers and enablers to end-of-life management of solar photovoltaic and battery energy storage systems: a systematic literature review. *J Clean Prod* 211:537–554. <https://doi.org/10.1016/j.jclepro.2018.11.229>
20. Satty TL (1988) The analytic hierarchy process. McGraw-Hill, New York, USA; RWS Publications, Pittsburgh, PA
21. Satty TL (1990) The analytic hierarchy process. RWS Publications, Pittsburgh, PA
22. Song BP, Zhang MY, Fan Y, Jiang L, Kang J, Gou TT, Zhang CL, Yang N, Zhang GJ, Zhou X (2020) Recycling experimental investigation on end of life photovoltaic panels by application of high voltage fragmentation. *Waste Manage* 101:180–187. <https://doi.org/10.1016/j.wasman.2019.10.015>

23. Song Q, Wang Z, Li J (2012) Residents' behaviors, attitudes, and willingness to pay for recycling e-waste in Macau. *J Environ Manage* 106:8–16. <https://doi.org/10.1016/j.jenvman.2012.03.036>
24. Vandeligt K, Sophie P, Yves P (2012) Assessment of the environmental performance of solar photovoltaic technologies. In: Environment Canada. [https://www.ec.gc.ca/scitech/B53B14DE-034C-457B-8B2B-39AFCFED04E6/ForContractor\\_721\\_Solar\\_Photovoltaic\\_Technology\\_e\\_09FINAL-update2-s.pdf](https://www.ec.gc.ca/scitech/B53B14DE-034C-457B-8B2B-39AFCFED04E6/ForContractor_721_Solar_Photovoltaic_Technology_e_09FINAL-update2-s.pdf)
25. Venkatachary SK, Samikannu R, Murugesan S, Dasari NR, Subramaniam RU (2020) Economics and impact of recycling solar waste materials on the environment and health care. *Environ Technol Innov* 20:101130. <https://doi.org/10.1016/j.eti.2020.101130>
26. Wang R, Xu Z (2014) Recycling of non-metallic fractions from waste electrical and electronic equipment (WEEE): a review. *Waste Manage* 34(8):1455–1469. <https://doi.org/10.1016/j.wasman.2014.03.004>
27. Yoon JH, Sim K (2015) Why is South Korea's renewable energy policy failing? A qualitative evaluation. *Energy Policy* 86:369–379. <https://doi.org/10.1016/j.enpol.2015.07.020>

# Forecasting of Solar Installation Capacity in Canada



Monasib Romel, Golam Kabir, and Kelvin Tsun Wai Ng

**Abstract** With the development of the renewable energy sector, solar energy resources are playing a crucial role. Worldwide the existing solar electricity generation is over 450 GW and it is predicted to increase to 5500 GW by 2050. In this study, the forecasting of electricity by photovoltaic solar in Canada has been performed using historical data and regression analysis. Three regression analysis methods (linear regression, polynomial regression, and power regression) are used in this study. After that, a comparative analysis was performed with the estimation of volumes made by the Canada Energy Regulator with an estimation of volumes obtained by all three regression models. The findings of the study show that the polynomial regression model is the best model having the highest  $R^2$  value (0.9359). In contrast, the forecasting trend of volumes indicates that it is matching more with power regression than polynomial regression after the year 2030, whereas it followed the linear regression in the first decade. Following the functional pattern of historical data, the volume is forecasted to be 8646 MW by linear method, 13,486 MW by polynomial method, 28,864 MW by power method in comparison with 23,439 MW by Canada Energy Regulator in the year 2045. The result of the study can be used by government policymakers, other research authorities, or scholars. On the other hand, in practical life, the analysis of comparison of forecasting of electricity can be used for forecasting of electricity from different alternative sources like biomass, wind, hydro, and others.

**Keywords** Solar installation capacity · Forecasting

---

M. Romel (✉) · G. Kabir · K. T. W. Ng  
Faculty of Engineering and Applied Science, University of Regina, Regina, Canada  
e-mail: [mar070@uregina.ca](mailto:mar070@uregina.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_75](https://doi.org/10.1007/978-3-031-34593-7_75)

1185

# 1 Introduction

In this era of searching for an alternative, non-emitting sources of generating electricity solar can play a vital role. The International Energy Agency [9] estimated that solar and wind will grow from 9% of global electricity in 2020 to 68% of global electricity in 2050 in their concept of “Net-Zero Carbon”. They also included that in the next 30 years, solar installation capacity will increase 20-fold which will be the highest growth among all renewable sources [9]. This evolution of speed and scale of growth in the energy sector will provide an opportunity for Canada to take immediate steps to be part of this sustainable growth [4]. Canada can be considered as one of the most prospective countries to install solar-generated electricity as an effective contribution to this global turnaround in the energy sector in the world.

The Government of Canada is eyeing for the generation of 90% of electricity from renewable sources from the current scenario of 81%, including 61% by hydro, 15% by uranium, and 5% by wind, solar, biomass, and others [2]. The Government of Canada has an estimation to replace the 45 TWH of coal-source electricity within the year 2030 [9]. To support this agenda, the government made an announcement in their budget of 2021 to enhance their incentives in the apparatus that are used in the storage of hydroelectric energy, in the equipment for solar heating, in technologies that facilitate geothermal energy, and also on the equipment which produces hydrogen by water-electrolysis [9]. Considering the growth in new electricity generation capacity through renewable sources, wind and solar accounted for about 52% of all in Canada between the time period of 2010–2018. Canada is also looking to increase the renewable-based generation of electricity (hydro, solar, wind biofuel) from 69% in the year 2020 to 80% in the year 2050 [9]. The IEA’s report on world energy outlook, the solar installation has been mentioned as the new king of electricity supply and one of the cheapest among all [4]. The use of solar photovoltaic energy has been widely studied in various academic institutions, including university campuses [1].

With a view to Canada’s decarbonization, the government is aiming for a phase-out of coal-fired generation and solar has an ample opportunity to take place and achieve the target of Net-Zero emission by year 2050 [9]. This study identifies the gap of Canadian forecasting electricity generation by using photovoltaic panels. The main research objective of this project is to forecast the volume of solar installation capacity in Canada. Three of the regression-based methods, linear, 2nd order polynomial, and power models are used for the forecasting and the performance of these models are compared with the estimates provided by the Canada Energy Regulator [3] for understanding the growth pattern. The findings can be used by concerned government authorities, policymakers, or personnel who are related to the dialogue, collaboration, stewardship, or solar firms’ installation project in Canada.



## 2 Literature Review

### 2.1 Forecasting

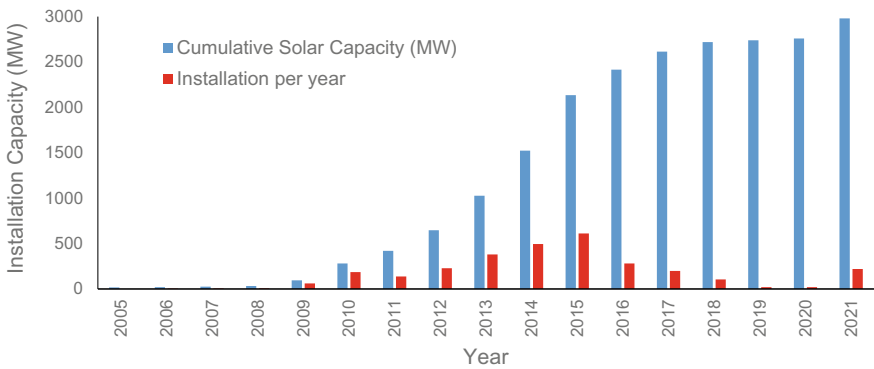
Forecasting refers to a formal method of statistics that may involve either a qualitative or quantitative approach for the prediction process. For quantitative forecasting, past data is used to define the function and finally, it predicts future data. Regression is one of the popular forecasting methods in statistics and is one of the most commonly adopted methods in data forecasting [6]. Depending on the past data trend, the regression function can be defined. The regression models can be linear, non-linear, polynomial, logarithmic, exponential, power regression, and others [7].

### 2.2 Solar Installation Growth in Canada

It is worth mentionable that Canada has just started exploring the enormous and unexploited resources of solar energy. In the last 15 years solar installation has grown from 17 MW in 2005 to 2981 MW in 2021. In the year 2021, 288 MW of solar capacity has been installed among which 250 MW in Alberta, 21 MW in Saskatchewan, 9.5 MW in Quebec, 4.8 MW in Nova Scotia, 1.5 MW in Yukon Territory, 0.3 MW in Ontario, and 0.1 MW in Prince Edward island [4]. This growth of solar installation is because of the search of alternative sources of fossil fuel (Coal and Oil) to reduce the emission of CO<sub>2</sub> and other gases to the atmosphere in Canada.

The year wise installation and cumulative capacity of solar have been shown in Fig. 1.

Canada is ranked 22nd in the world solar energy capacity in 2020 [4]. Because of its large area, Canada has potential solar energy resources. There are more than 140 solar farms in Canada with a capacity of 1 MW mostly situated in the provinces



**Fig. 1** Growth of solar PV capacity [3, 12]

of Ontario and Alberta. Ontario, Alberta, and Saskatchewan are there three potential provinces for the highest solar potentials in Canada. Primarily the solar installation was started in the province of Ontario but gradually the projects are also being approved in Alberta and adjacent provinces. The data is collected from Wikipedia and different newspaper articles [5, 15]. Following is the list of some of the large solar projects in Canada in Table 1.

In Canada, there are 292 off-grid communities and there is insufficient supply for reliable, affordable, and safe electricity. Many of these communities have to produce their own energy by using expensive and polluting diesel generator [13]. So there is opportunity to replace of this diesel with clean energy by installing solar photovoltaic, which could mitigate their local demand [11]. Again, Canada is also well equipped for the supply of raw materials for solar industry like lead, copper zinc for panels and cobalt, nickel graphite for battery production. The price of the solar power for consumers also plunged almost 82% in the last 15 years. This made it acceptable for mainstream people and the installations has increased many folds in the last decade [10].

### ***2.3 Shift from Coal and Oil***

With a view to reduce Green House Gas (GHG), there is a negative growth in coal and oil sources electricity generation plants for last decade. This also opens the door for solar to increase the share of electricity generation. The sector wise growth for electricity installation has been shown in Fig. 2.

### ***2.4 Canadian Energy Regulator***

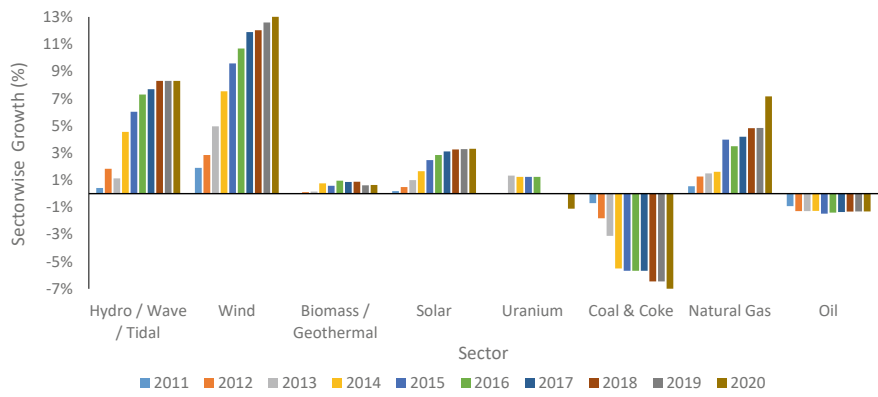
Canada Energy Regulator (CER) is a government owned organization which keeps the record of energy related issues and also review energy related projects in Canada. They also make sure that safety and environment standards are maintained as per Canadian guidelines. CER also keeps record of data of different important energy sector like solar, biomass, nuclear, wind, coal, etc.

## **3 Methodology**

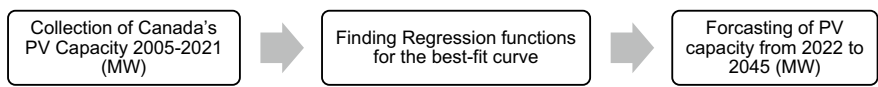
Some of the mentionable uses of regression in forecasting are prediction of continuous values, modeling of time series, and finding relationships between variables [8]. Time series regression analysis is an effective statistical tool for finding the best-fit curve on the responsive history and to do the forecasting of the future amount of that variable following the trend [14]. This study is an attempt to quantify the PV

**Table 1** List of solar farms having installation capacity more than 50 MW in Canada

		Total capacity (MW)	Land size (acres)	Year on installation on grid	Province
1	Travers solar project	465	3300	2022	Alberta
2	Brooks solar farm	400	3900	2022	Alberta
3	Dunmore solar project	216	623	2023	Alberta
4	Vauxhall solar project	150	1300	2022	Alberta
5	Airport city solar project	120	627	2022	Alberta
6	Lathom solar project	120	1000	2022	Alberta
7	Saddlebrook solar and storage project	102	300	2024	Alberta
8	Grand renewable solar project	100	800	2015	Ontario
9	Kingston solar project	100	810	2015	Ontario
10	Enterprise solar project	94	750	2022	Alberta
11	Loyalist solar project	85	400	2019	Ontario
12	Sarnia PV power plant	80	1100	2009	Ontario
13	Vulcan solar project	77	960	2023	Alberta
14	Claresholm 2 solar	75	1280	2021	Alberta
15	Enchant solar project	75	550	2022	Alberta
16	Prairie sunlight I solar project	74	552	2021	Alberta
17	Wheat crest solar project	60	320	2022	Alberta
18	Sault Ste. Marie project	58	500	2011	Ontario
19	Claresholm 1 solar project	58	1280	2021	Alberta
20	Southgate solar project	50	235	2016	Ontario
21	Windsor solar project	50	300	2016	Ontario



**Fig. 2** Sector wise growth of electricity installation in Canada [3]



**Fig. 3** Approach to estimate PV panel waste in Canada

installation capacity in Canada until 2045. A three-step method has been used to quantify PV waste amount in Canada as outlined in Fig. 3.

The first step comprises of collecting secondary data PV installation capacity form reliable source. In the second step, the three regression methods were selected for formulating the function using the collected data. The regression analysis used here are linear regression, polynomial regression, and power regression. In the final step, forecasting has been made using the function until year 2045 and finally a comparative analysis was performed with the government estimation for the year 2025, 2030, 2035, and 2045 with the volumes obtained in all three regression methodologies. Regression and trend analyses are more applicable to short-to-mid-term forecasts and a 25-year forecast period is forecast period is selected in this study. The regression methods were conducted using Microsoft Excel 2019.

### 3.1 Linear Regression

Linear regression is considered to be one of the fundamental types of regression. In this regression, the linear relationship between the dependent and independent variables is analyzed.

$$Y = \lambda_0 + \lambda_1 X + C \tag{1}$$

In the second step, the sum of errors  $C$  are minimized for calculating  $\lambda = [\lambda_0, \lambda_1]$ :

$$P(\lambda_0, \lambda_1) = \min C^K C = \min (y_i - \lambda_0 - \lambda_1 x_i)^K (y_i - \lambda_0 - \lambda_1 x_i) \quad (2)$$

Here,  $P(\lambda_0, \lambda_1) = \min \sum (y_i - \lambda_0 - \lambda_1 x_i)^2$ , and  $i = [1, \dots, n]$

Formulation of  $[x_1, x_2, x_3, \dots, x_n]$  into  $n \times 2$  which can be named as design matrix.

$$U = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad (3)$$

According to the invertibility of  $U^K U$ . The  $\lambda$  is also calculated as follows:

$$\lambda = (U^K U)^{-1} U^K Y \quad (4)$$

In the third step,  $C$  and  $\lambda$  are taken to get the finalized model as follows:

$$Y = \lambda U + C \quad (5)$$

### 3.2 Polynomial Regression

Polynomial is one of the non-linear techniques of regression analysis. This method performs evaluation of values minimize the sum of the squares of the distances of the data points to the curve [16].

Following model of equation will be used for analysis of regression:

$$Y = \lambda_0 + \lambda_1 X + \lambda_2 X^2 + C \quad (6)$$

In the first step, the responses of variable and measurable variables are gathers which are defined as follows:

In the second step, the sum of errors  $C$  are minimized for calculating  $\lambda = [\lambda_0, \lambda_1, \lambda_2]$ :

$$P(\lambda_0, \lambda_1, \lambda_2) = \min C^K C = \min (Y - \lambda_0 - \lambda_1 X_i - \lambda_2 X_i^2)^K (Y - \lambda_0 - \lambda_1 X_i - \lambda_2 X_i^2) \quad (7)$$

Here,  $P(\lambda_0, \lambda_1, \lambda_2) = \min \sum (Y - \lambda_0 - \lambda_1 X_i - \lambda_2 X_i^2)^2$ , and  $i = [1, \dots, n]$ .

Formulating  $[X_1, X_2, X_3, \dots, X_n]$  into  $n \times 3$  matrix termed the design matrix.

$$Q = \begin{pmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{pmatrix} \quad (8)$$

According to the invertibility of  $Q^K Q$ . The  $\lambda$  is also calculated as follows:

$$\lambda = (Q^K Q)^{-1} Q^K Y \quad (9)$$

In the third step,  $C$  and  $\lambda$  are taken to get the finalized model as follows:

$$Y = Q^K \lambda + C \quad (10)$$

### 3.3 Power Regression

Another type of non-linear regression is power regression.

In the first step, the responses of variable and measurable variables are gathered which are defined as follows:

$$Y = \lambda_0 X^K \quad (11)$$

In the second step, the sum of errors  $C$  are minimized for calculating  $\lambda = [\lambda_0]$ :

$$P(\lambda_0) = \min C^K C = \min (Y - \lambda_0 X_i^K)^K (Y - \lambda_0 X_i^K) \quad (12)$$

Here,  $P(\lambda_0) = \min \sum (Y - \lambda_0 X_i^K)^2$ , and  $i = [1, \dots, n]$ .

Formulating  $[X_1, X_2, X_3, \dots, X_n]$  into  $n \times 1$  matrix termed the design matrix.

$$S = \begin{pmatrix} X_1^K \\ X_2^K \\ \vdots \\ X_n^K \end{pmatrix} \quad (13)$$

According to the invertibility of  $S^K S$ , the  $\lambda_0$  is also calculated as follows:

$$\lambda = (S^K S)^{-1} S^K Y \quad (14)$$

**Table 2** Linear, polynomial, power regression functions

S. No.	Function	Equation	Values of $R^2$
1	Linear	$Y = 228.92X - 739.36$	0.9280
2	Polynomial	$Y = 4.8399X^2 - 19257X + 2E+07$	0.9359
3	Power function	$Y = 4.4717X^{2.3623}$	0.8877

In the third step,  $C$  and  $\lambda$  are taken to get the finalized model as follows:

$$Y = X^K \lambda \quad (15)$$

## 4 Result and Discussion

The methods mentioned in the previous section were used to forecast the volume of solar installation capacity until 2045 in Canada. To perform the analysis, relevant linear and polynomial and power regression-based functions were developed.

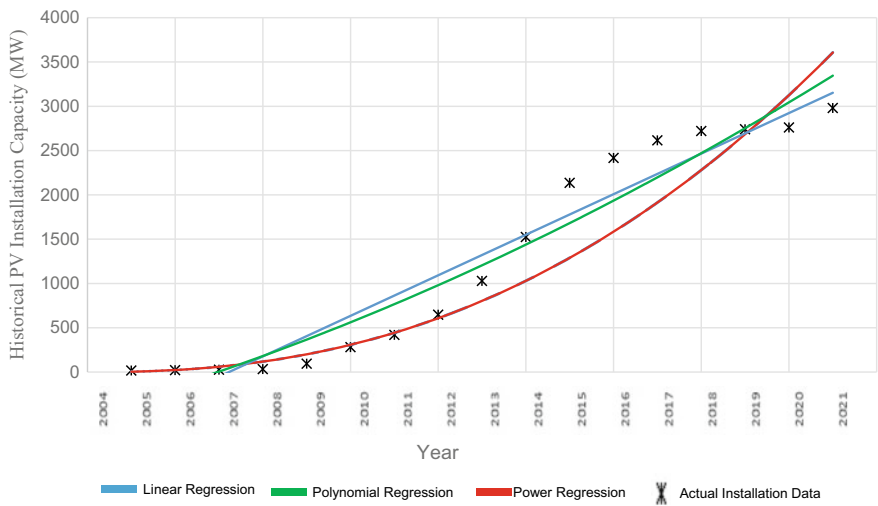
### 4.1 Collection and Analysis of Data

The existing solar installation capacity from 2005 to 2021 was collected from the report of Canada Energy Regulator [3]. As per the 2021 report, Canadian solar capacity is 2981 MW in Canada which has grown from 17 MW in 2005. In this study, the relevant linear and polynomial and power regression-based functions are developed. These functions are mentioned in the following Table 2.

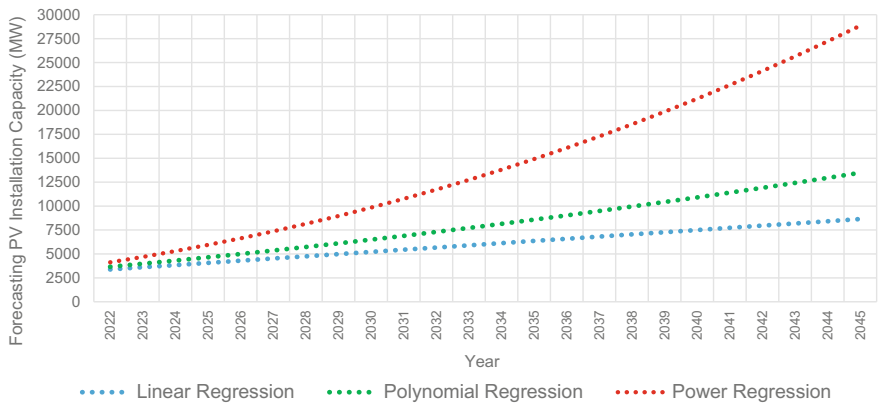
### 4.2 Comparison Between Curves

To estimate the time series model and to find the best-fit curve for the data, linear and non-linear regressions analysis are one of the key analyses for forecasting the future. In this study, comparative analysis has been performed between linear and second order polynomial regression and power regression models. It is noticeable from Table 2 that in comparison of  $R^2$ , the polynomial regression function has higher Value (0.9359) than both the linear regression function (0.9280) and the power regression function (0.8877). All the  $R^2$  values appear acceptable.

The linear and polynomial and power regression curves for historical data and for forecasting are shown in Figs. 4 and 5 respectively below. As shown in Fig. 4, it appears the regression models captures the overall data trend but not the characteristics of the set. This is normal using simple regression models.

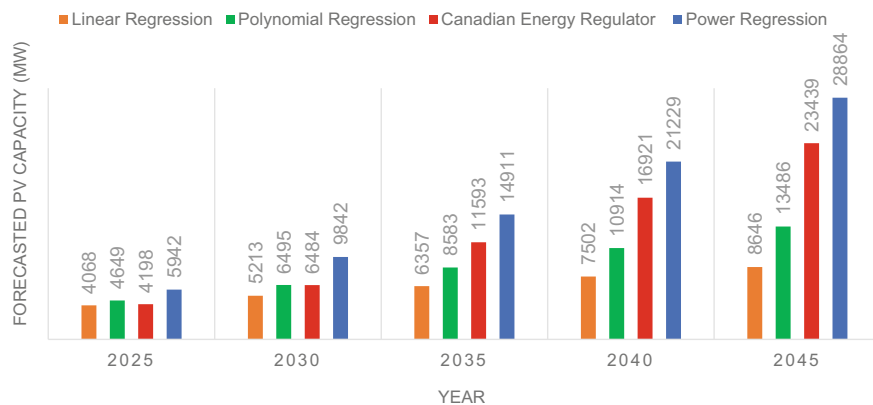


**Fig. 4** Historical installation capacity (MW) and linear, 2nd order polynomial and power regression trends



**Fig. 5** Forecasting of solar installation capacity by linear, regression, and power regression trends





**Fig. 6** Comparison of PV installation forecasting

### 4.3 Forecasting Comparison with Canadian Energy Regulator 2021

In this section, comparison of forecasting has been performed between volume of PV installation (MW), obtained by different regression methods and volume of PV installation (MW) obtained from Canada Energy Regulator [3] as a secondary source. Forecasting by the power regression method shows that 28,864 MW electricity needs to be included in the grid by PV panels within 2045, whereas the estimation is 13,486 MW by polynomial method and estimation is 8,646 MW by linear method. In this study, Canada Energy Regulator [3] estimations has been used a benchmark. For example, linear regression model consistently underpredicts PV capacity than CER estimates during the 25-year forecasting period. The 2nd order polynomial model overpredicts in short term (5–10 years) but overpredicts in long term (10–25 years). The power model consistently overpredicts but appears to capture the long-term increasing trend. The comparison is shown in Fig. 6.

### 4.4 Implication of Findings

Implications of these findings can be stated as that this forecasting and comparison will assist the administrators to appraise their decision-making plans and take effective measures for the management of the waste generated from these end-of-life PV panels. Moreover, the analysis of the historical data has opened the scenario of whether there exists any trend of solar installation in the last decade for concerned stakeholders. Again, the implications of different linear, polynomial, and power regression models can be considered in terms of theoretical point of view.

## 5 Conclusion

In this study, a systematic literature review on the use of solar panels in Canada has been conducted. Data from various sources were consolidated, verified, and examined to forecast the solar installation capacity until 2045. Three different regression techniques were tested and then compared with the obtained volume of forecasting as provided by the Canada Energy Regulator [3]. The regression methods used in analysis are linear regression, polynomial regression, and power regression. Modeling results showed that the solar installation capacity forecast functions have been formulated by giving input of historical data obtained from Canada Energy Regulator [3] into these regression functions.

In summary, linear model performs best in the short term (5 year), the 2nd polynomial model performs best in mid-range (10 years) and power model performs best in long-range (25 years). This prepared framework for PV capacity estimation can be used for installation capacity estimation of some other alternative sources of electricity generation in Canada including biomass, wind, hydro, etc. For future research work, some other models of regression, i.e., higher orders of polynomial regression, exponential regression, logarithmic regression can be used to compare volume of PV capacity in Canada.

**Acknowledgements** Authors wish to acknowledge the support of Faculty of Graduate Studies and Research (FGSR) and Canada Energy Regulator, Natural Resources Canada.

## References

1. Behi Z, Ng KTW, Richter A, Karimi N, Ghosh A, Zhang L (2022) Exploring the untapped potential of solar photovoltaic energy at a smart campus: shadow and cloud analyses. *Energy Environ* 33(3):511–526. <https://doi.org/10.1177/0958305X211008998>
2. Canada Energy Regulator (2019) Canada energy regulator 2019. <https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-canada.html#:~:text=MorethanhalfoftheelectricityinCanada,electricitysectoreccursprimarilyattheprovinciallevel>
3. Canada Energy Regulator (2021) Canada's energy future 2021 (macro indicators, electricity capacity). Canada Energy Regulator. <https://apps.cer-rec.gc.ca/ftprpndc/dflt.aspx?GoCTemplateCulture=en-CA>
4. CanREA (2021) Powering Canada's journey to net-zero. Canadian Renewable Energy Association
5. CBC News (n.d.) How Canada's largest solar farm is changing Alberta's landscape. CBC News. <https://www.cbc.ca/news/canada/calgary/travers-solar-project-vulcan-1.6233629>
6. Chowdhury A, Vu HL, Ng KTW, Richter A, Bruce N (2017) An investigation on Ontario's non-hazardous municipal solid waste diversion using trend analysis. *Can J Civ Eng* 44(11):861–870. <https://doi.org/10.1139/cjce-2017-0168>
7. Guerard JB (2013) Introduction to financial forecasting in investment analysis, 2007, pp 1–236. <https://doi.org/10.1007/978-1-4614-5239-3>
8. Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. *J Stat Softw* 27(3):22. <http://www.jstatsoft.org/v27/i03/paper>

9. International Energy Agency (2021) Canada 2022—energy policy review. International Energy Agency. [www.iea.org/t&c/](http://www.iea.org/t&c/)
10. IRENA (2016) End of life management solar PV panels. International Renewable Energy Agency (IRENA) and the International Energy Agency (IEA)
11. Lovekin D, Moorhouse J, Morales V, Salek B (2020) Diesel reduction progress in remote communities modelling approach and methodology about the Pembina Institute. Pembina Institute
12. Natural Resources Canada (2020) Energy factbook 2019–2020. Natural Resources Canada
13. NRCan (2020) Natural Resources Canada
14. Sharma S, Kleinbaum DG, Kupper LL (1978) Applied regression analysis and other multivariate methods. *J Mark Res* 15(3):498. <https://doi.org/10.2307/3150614>
15. Wikipedia (2020) Photovoltaic power stations in Canada. Wikipedia. [https://en.wikipedia.org/wiki/List\\_of\\_photovoltaic\\_power\\_stations\\_in\\_Canada](https://en.wikipedia.org/wiki/List_of_photovoltaic_power_stations_in_Canada)
16. Yang Y (2015) Development of the regional freight transportation demand prediction models based on the regression analysis methods. *Neurocomputing* 158:42–47

# The Meadoway: Urban Ecosystem Restoration at a City-Level Scale Providing Enhanced Regulating and Supporting Services



Ke Qin, Marney Isaac, and Jennifer Drake

**Abstract** The Meadoway project, led by the Toronto and Region Conservation Authority (TRCA), aims to restore 200 ha of meadow habitats and complete a 16-km linear multi-use trail along the Gatineau Hydro Corridor across Scarborough, Ontario. This ambitious project is demonstrating that urban ecosystem restoration can be successfully implemented at a large city-level scale. The hydro corridor's transition from turf grass to deep-rooted native meadow plants is hypothesized to enhance the regulating (e.g., erosion and flood control) and supporting (e.g., soil quality regulation and nutrient cycling) services. This study aims to evaluate this hypothesis. In-situ infiltration, penetrometer tests, and soil sampling were conducted on two pre-restored turf lands and two restored meadows in 2020. Soil samples were analyzed for bulk density, porosity, total carbon, total nitrogen, and available phosphorus. Water balance analysis was conducted by simulating artificial rainfall events with different return periods upon undisturbed vegetated soil samples with accompanying saturated hydraulic conductivity measurements using a mini-disk infiltrometer. The restored meadows had a similar in-situ saturated hydraulic conductivity with turf lands, but had a lower cone index, and a higher bulk density and lower porosity than the turf lands. The soil nutrient content showed large variation among different sites. However, a notable trend in soil available phosphorus was measured where it was consistently higher in surrounding pre-restored turf lands than in restored meadows. The rainfall simulation tests showed that the turf lands generated much more surface runoff than the restored meadows during rainfall events with return periods from 2 to 100 years, while the saturated hydraulic conductivities of soil samples from the restored meadows and the turf lands were not considerably different. This indicated that the saturated hydraulic conductivity measurement is not a good indicator to quantify the hydrological regulating functions of green infrastructure like The Meadoway.

---

K. Qin (✉) · J. Drake

Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON, Canada

e-mail: [ke.qin@mail.utoronto.ca](mailto:ke.qin@mail.utoronto.ca)

M. Isaac

Department of Physical and Environmental Sciences, University of Toronto Scarborough, Scarborough, ON, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_76](https://doi.org/10.1007/978-3-031-34593-7_76)

**Keywords** Urban ecosystem restoration · City-level scale providing

## 1 Background

The Greater Toronto Area (GTA) is currently experiencing rapid urbanization with a sharply increasing population and expanding urbanized lands, shrinking green areas and degrading urban ecosystems. The GTA population is projected to increase by 36.7%, from 7.0 million in 2019 to over 9.5 million by 2046 [23]. Adverse effects of ecosystem degradation caused by urbanization include water and air deterioration, urban heat island effects, increased flood risks, decreased biodiversity, losses and fragmentation of habitats, and aesthetic degradation [9, 40]. Intensified urbanization with changing land covers also contributes to climate change, which causes intensifying rainfall events and higher temperatures, posing risks on municipal infrastructure and ecosystems [1, 6].

One strategy to build a future living city is to rethink and maximize the value of urban greenspace by creating a green infrastructure (GI) network. This infrastructure network will provide residents with a sustainable landscape to live within, promote access to nature, and protect local ecosystems. As a type of GI, urban greenspace plays a critical role in cities' sustainable development and management, providing multiple urban ecosystem services (ESSs) to protect the urban environment and support residents' wellbeing. Revitalization projects are recognized as an effective approach to maximize urban greenspace's potential by creating high-quality green areas and green infrastructure [34]. In built-up communities like Toronto, linear infrastructure corridors such as transportation, power, and pipeline corridors are the few remaining unexploited open spaces that are well-positioned for greenspace restoration. The revitalized corridors can provide better ecosystem functions in the city and enhance the connectivity for people across the city by developing an alternative low-impact transportation.

The Gatineau Hydro Corridor stretching across the City of Scarborough was constructed in the 1920s to connect downtown Toronto to the hydroelectric power plants in Quebec's Gatineau region [34]. The success of the Scarborough Center Butterfly Trail (SCBT) within the Gatineau Hydro Corridor, which was revitalized in 2015 with restored meadow habitats, supports the Gatineau Hydro Corridor Revitalization Project. This project aims to restore 200 ha of meadow habitats named The Meadoway and complete a linear multi-use trail over 16-km connecting downtown Toronto and the Rouge National Urban Park [36].

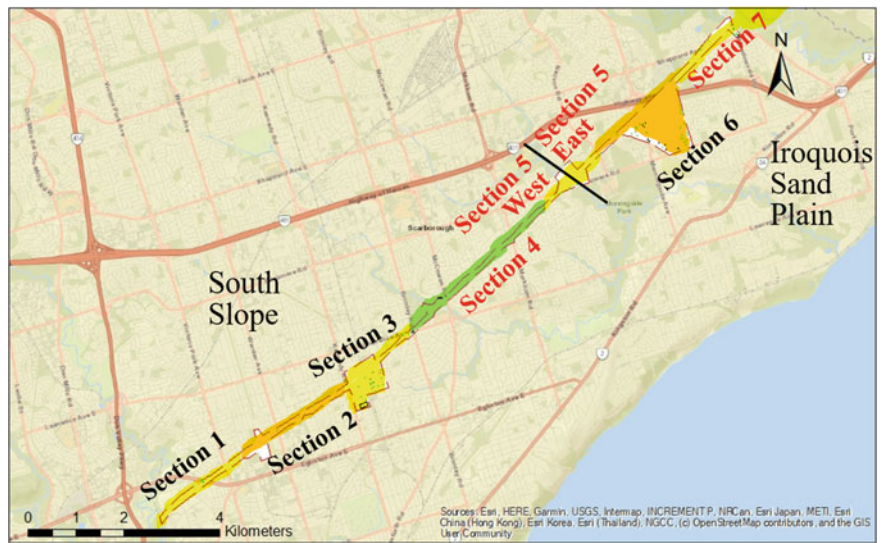
The Meadoway is expected to provide and enhance various types of ecosystem services by restoring urban ecosystems. The transition from turf grasses to native tall meadow plants in The Meadoway is expected to alter soil structure and soil hydraulic properties, which can potentially enhance the delivered hydrological regulating, erosion control, and nutrient cycling services. By characterizing changes in soil physical, hydraulic, and hydrological properties, the impact of restoration efforts on potential ecosystem functions may be validated and measured. These data

support regulators and planners to comprehensively understand the delivered benefits of ecosystem restoration and help establish practices and opportunities to maximize public benefits. This study aims to: (1) characterize the soil properties within restored and unrestored sections of The Meadoway, (2) evaluate if the restoration works are resulting in enhanced regulating (e.g., flood and erosion control) and supporting (e.g., soil quality regulation and nutrient cycling) services, and (3) provide recommendations for future research on evaluating ecosystem services enhanced by The Meadoway.

## 2 Study Site

The restoration of meadows in The Meadoway is being completed in seven phases (Sections 1–7, Fig. 1). The meadow restoration in Sections 4 and 7 was completed in 2015 and 2017, respectively. By the end of 2020, the site preparation in Sections 1 and 2 for restoration had finished, while Sections 3, 5, and 6 remained undisturbed. The Meadoway lies within two physiographic regions separated roughly by Highland Creek. Sections 1–4 and Section 5 West are located within the South Slope comprised of sandy silt to sand glacial till and river deposits of mainly sand and gravel, while Sections 5 East, 6 and 7 are located within the Iroquois Sand Plain comprised of sandy loam soil [36]. In total, six sites were selected, encompassing three land cover types (i.e., meadows, buffer areas, and turf lands), and two physiographic regions (i.e., South Slope and Iroquois Sand Plain) into considerations which are summarized in Table 1. Study sites included the restored meadows and adjacent buffer turf grass strips in Sections 4 and 7 and the pre-restored turf lands in Section 5. Tillage was conducted during the establishment of native plants in the restored meadows in both Sections 4 and 7. Section 5 is divided into a west and an east part to account for the effects of different soil types in the two physiographic regions.

Restored meadows were planted with various native species including tall sunflower (*Helianthus giganteus*), ox-eye (*Heliopsis helianthoides*), tall coreopsis (*Coreopsis tripteris*), cup-plant (*Silphium perfoliatum*), big bluestem grass (*Andropogon gerardii*), Indian grass (*Sorghastrum nutans*), and switch grass (*Panicum virgatum*) [35]. The buffer areas and turf lands at study sites were covered by non-native turf grass including meadow fescue (*Schedonorus pratensis*), red fescue (*Festuca rubra* ssp. *rubra*), and Kentucky blue grass (*Poa pratensis* ssp. *pratensis*).



**Fig. 1** The locations of seven sections in the Meadowway with selected study sites highlighted in red

**Table 1** The physiographic regions, sections, vegetation covers, and IDs of the six study sites

Physiographic region	Section	Vegetation cover	ID
South Slope	4	Meadow	S4M
		Mowed Buffer Grass	S4B
	5 West	Mowed Turf Grass	S5W
Iroquois Sand Plain	5 East	Mowed Turf Grass	S5E
		Mowed Buffer Grass	S7B
	7	Meadow	S7M

### 3 Methodology

#### 3.1 Soil Characterization

##### 3.1.1 Double-Ring Infiltration Tests, Penetrometer Tests, Density, Porosity, and Fine Particle Size Distribution

A 1-ha representative area was selected in each of Section 4 meadow, Section 5 West turf, Section 5 East turf, and Section 7 meadow for field tests and soil sampling. The selected areas were 100 m × 100 m in Section 4 meadow and Section 5 West turf and were 80 m × 120 m in Section 5 East turf and Section 7 meadow due to the linear shape of the two study sites. The turf areas surrounding Section 4 meadow and Section 7 meadow were also selected for the same field tests and soil sampling

procedures. Table 2 gives a summary of methods, sampling numbers, and depths of analyzed parameters in each study site.

Nine infiltration tests were conducted in the selected meadow, buffer, and turf areas in Sections 4, 5 West, 5 East, and 7 using double-ring infiltrometers [10]. Each 1-ha plot was divided by a  $3 \times 3$  grid, and one infiltration test was conducted at the center of each subplot. The detailed dimensions of plot division and the locations of infiltration tests are shown in Fig. 2. Disturbed areas such as informal trails were avoided for infiltration tests to reduce the effects of human disturbance on infiltration capacity. In order to avoid the effects of soil compaction on the results of infiltration tests, the cone index of the soil at each infiltration test location was measured using a proving ring penetrometer. Five penetrometer measurements were taken evenly around each infiltration test location. The average of the five cone index measurements at each infiltration test location should not deviate significantly from the nine cone index measurements taken over the whole 1-ha treatment site.

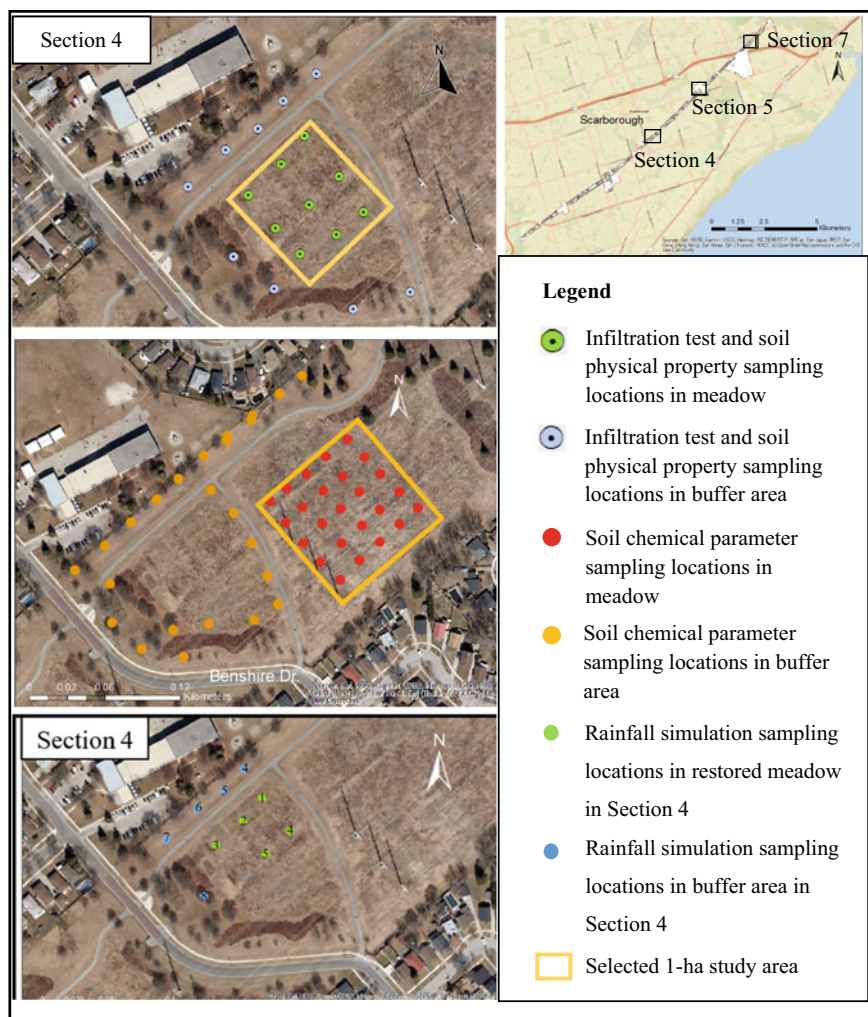
Nine soil samples were collected to analyze for bulk density and porosity at the same locations of infiltration tests in each selected 1-ha plot and buffer mow area (Fig. 2). The Bulk Density Soil Sampling Kit by AMS was used to collect soil bulk density and porosity samples [2]. Bulk density samples were processed and analyzed in the laboratory soon after sample collection. The samples were put in ceramic bowls and were dried in an oven at 110 °C for 48 h. The weight of oven-dried samples and the volume of sample rings were used to calculate the bulk density of each soil sample. The porosity of each sample was then calculated using this bulk density and the specific density analyzed through pycnometer tests.

Soil samples were taken for fine particle size distribution analysis at the ground level to a depth of 15 cm using a shovel at the same bulk density sample locations. Each sample had a fresh weight of at least 500 g and was collected and preserved in a sealable plastic bag for laboratory analysis. Hydrometer tests were then conducted on collected samples to analyze the particle size distribution of materials finer than the No. 200 (75- $\mu$ m) sieve and larger than about 0.2  $\mu$ m. ASTM D7928-17 method

**Table 2** Methods, sampling depths, and sampling numbers of analyzed parameters

Parameter	Method	Sampling depth (cm)	Sampling number
In-situ infiltration capacity	Double-ring infiltration tests	At ground level	9 per site
Cone index	Penetrometer tests	At ground level	
Bulk density and porosity	Analysis of intact soil cores	30 cm	
Soil chemical parameters (total C and total N and available P)	Chemical analysis of bulk soil samples	0–15 cm	25 per site in S4M, S4B, S5W, and S5E 24 per site in S7M and S7B





**Fig. 2** Example sampling locations in Section 4. Additional figures and sampling location information are available in Qin's Master's thesis [28]

[3] was adopted for hydrometer tests, which were performed on materials passing the No. 10 (2.0-mm) or a finer sieve.

### 3.1.2 Soil Chemical Property Analysis

Soil samples were collected from Section 4 meadow and buffer, Section 5 West and East turf, and Section 7 meadow and buffer to analyze for chemical parameters, including total carbon (C), total nitrogen (N), and available phosphorus (P). The data

of total C, total N, and available P content in the soil provides information on the soil's ability to support plant productivity and fertility, which helps to analyze the influence of soil types on the successfulness of native plant restoration. In addition, the soil chemical parameter data collected at the early stage can establish a baseline, upon which future data can be compared to investigate the influence of meadow restoration on soil nutrient cycling process (i.e., one of supporting services).

Soil samples for chemical analysis were collected at 25 locations in each of Section 4 meadow and buffer area and Section 5 West and East turf lands, and at 24 locations in each of Section 7 meadow and buffer area (Fig. 2). The selected sampling locations were undisturbed and representative. Disturbed lands such as gravel piles and recently fertilized areas were avoided for soil sampling. At each site, 9 of the 25 or 24 sampling locations for nutrient analysis were near the nine sampling locations for physical parameter analysis to provide paired data in all sites except for Section 4 meadow. The sampling rationale was to collect one soil sample in each  $20\text{ m} \times 20\text{ m}$  subplot in the selected 1-ha meadow and turf plots and collect one sample at an interval of 20 cm along the linear buffer mow areas. The selected 1-ha sampling areas were  $100\text{ m} \times 100\text{ m}$  squares in Section 4 and Section 5 West, which were divided by a  $5 \times 5$  grid, and were  $80\text{ m} \times 120\text{ m}$  rectangles in Section 5 East and Section 7, which were divided by a  $4 \times 6$  grid. Soil samples were collected at the center of each subplot. At each location, one chemical analysis sample was taken at a depth of 0–15 cm using a shovel. Each sample had a fresh weight of at least 500 g and was collected and preserved in a sealable plastic bag for laboratory analysis.

Total nitrogen and total carbon in soil samples were analyzed using the LECO CN628 Elemental Analyzer. The LECO CN628 is a combustion instrument that burns soil samples into combustion gases with pure oxygen in the furnace, which are then swept by a helium carrier gas into infrared cells to detect  $\text{H}_2\text{O}$  and  $\text{CO}_2$  and a thermal conductivity cell to detect nitrogen [18]. Available phosphorus in soil samples was analyzed using QuikChem 8500 Series Flow Injection Analysis (FIA) system [4]. Before FIA, soil samples were extracted using Bray I solution. The extracting solution and reagents were then pumped into and mixed in the manifold (reaction module), where the sample was diluted, dialyzed, extracted, incubated, and derivatized with photometric detection.

### 3.1.3 Statistical Analysis

Several statistical tools in R software [29] were used to do covariance and correlation analysis on the data. Firstly, some tests were conducted to determine whether non-parametric or parametric tests should be used. Shapiro–Wilk tests were conducted on infiltration capacity, cone index, bulk density, porosity, total C, total N, and available P measurements at each site for normality. Although most of the datasets passed the Shapiro–Wilk test indicating normality, most distribution curves did not have a bell shape, indicating a deviation from the normal distribution. In addition, due to the small sample sizes, which tend to increase the probability of misinterpreting a normal

distribution, non-parametric tests were selected to analyze the data. All analyzed parameters were compared between paired sites by non-parametric Wilcoxon tests. A  $p$ -value of 0.05 was used to accept or reject the null hypothesis. Spearman's rank correlation test evaluated correlations between soil parameters and infiltration measurements. A  $p$ -value of 0.05 was used to accept or reject the null hypothesis (i.e., there is no significant correlation between two variables).

### ***3.2 Rainfall Simulation Tests and Water Balance Analysis***

Five pairs of undisturbed vegetated soil samples were each collected from Section 4 of The Meadoway in the restored meadow and turf buffer (Fig. 2). The terms “turf land” and “buffer area” are used interchangeably here. The restored meadow was covered with native plants, while the buffer area was covered with turf grasses. When possible, sampling locations were selected to overlap with the double-ring infiltration tests conducted in 2020. The ID number of each sample was the same as the one used for in-situ infiltration tests and soil sampling in 2020. Two additional vegetated soil samples were collected from a cultural meadow in the Rouge National Urban Park (RNUP), located approximately 12 km northeast of Section 4. The RNUP cultural meadow is located within the South Slope and was planted with native grasses and wildflowers in 2009 [36], making it several years older than the restored meadow of Section 4 which was restored in 2015 [36].

Before collecting soil-vegetation samples, three cone index measurements were taken using a ELE International proving ring penetrometer [11] to check if the soil was similar to the overall conditions observed in 2020. The 2006 Toronto Wet Weather Flow Management Guidelines Intensity–Duration–Frequency (IDF) curves were applied to calculate rainfall depths for 10-min storms of return periods 2, 5, 10, 25, 50, and 100. These return periods were selected because (1) City of Toronto has adopted the 100-year storm as the level of protection for properties against surface flooding [8], and (2) post-development peak flows of a development site located within the Highland Creek watershed, where Section 4 of The Meadoway is located, should be controlled to pre-development levels for all storms up to and including the 100-year storm (i.e., 2, 5, 10, 25, 50, and 100-year storms) [8]. The rainfall intensity was multiplied by the rainfall duration (10 min) and multiplied by the basal area of the rainfall simulation container (23 cm  $\times$  28 cm) to calculate the volume of rain-water used in each test. The information of designed rainfall events with different return periods is summarized in Table 3.

A rainfall simulation container was placed upon the soil cutter, from which synthetic “rainfall” drained out through the small holes (1.73 mm) onto the surface of the vegetated soil. The Fisher Scientific GP1000 peristaltic pump [12] was used to control rainfall intensities by adjusting the revolution rate. During rainfall events, surface runoff flowed out through an outlet into the runoff container, and subsurface discharge percolated from the bottom of the soil sample was collected in the

**Table 3** Simulated rainfall events with 2-, 5-, 10-, 25-, 50-, and 100-year return periods

Return period (year)	Rainfall duration (min)	Rainfall intensity (mm/h)	Rainfall volume (mL)	Pump revolution rate (RPM)
2	10	88.2	947	17
5	10	131.8	1415	25
10	10	162.3	1742	30
25	10	189.5	2034	36
50	10	224.3	2408	42
100	10	250.3	2687	47

exfiltration container. Tap water was supplied to the rainfall simulation container at a constant rate with the pump set to match the desired rainfall intensity.

Before conducting rainfall simulation tests, soil samples were saturated and left drain overnight to reach the field capacity. To saturate the sample, rainfall was applied at an intensity of 84 mm/h (90 mL/min) until surface runoff and exfiltration rates stabilized. This typically took 30 min. The 30-min rainfall intensity was less than the designed 2-year rainfall to avoid any unintended alterations of soil structure. Then, the sample was left to drain overnight for at least 12 h to reach the field capacity. On the next day, the saturated sample was re-weighed. Then, the rainfall simulation kit was assembled, and the rainwater volume of the first event with a 2-year return period was measured using 1000 mL graduated cylinders and transferred in a bucket. The peristaltic pump was used to pump water from the bucket into the leveled rainfall simulation container at the desired rate. At the end of the rainfall simulation, any water remaining in the tubing or the bucket was poured manually into the rainfall simulation container. Once water stopped draining from the rain simulation container, it was flipped over, and any remaining water was poured directly onto the soil surface. When the rainfall ended, the weight of the soil sample plus the weight of the soil cutter was measured immediately to calculate the amount of water retained by the soil. The runoff collected in the runoff container and the exfiltration collected in the exfiltration container was measured for volume, respectively. All data were recorded in a Microsoft Excel sheet for analysis. The soil sample was then left to drain for at least 10 min until percolation water stopped draining from the bottom. The sample was re-weighed, and then the next rainfall event (e.g., 5-year) was applied. The simulated rainfall events were conducted in the same way and in order of increasing rainfall intensity. Rainfall simulation tests on paired buffer and meadow samples were conducted on the same day to avoid possible effects of the changes in soil characteristics due to biodegradation or drought with time. The volume of water collected in the runoff container ( $V_{\text{runoff}}$ ) and the exfiltration container ( $V_{\text{exfiltration}}$ ) was converted into the ratio to the total precipitation ( $P$ ) in each simulated rainfall event as  $R_{\text{runoff}}$  and  $R_{\text{exfiltration}}$ . The volume of water retained by the soil sample ( $V_{\text{retained}}$ ) was calculated based on the increased mass of the soil sample after each test, which was also converted into the ratio to the total precipitation as  $R_{\text{retained}}$ . For rainfall events with a specific return period, the mean of  $R_{\text{runoff}}$ ,  $R_{\text{exfiltration}}$ , and

$R_{\text{retained}}$  of all soil samples from each type of land (e.g., restored meadow, turf land, and cultural meadow) was calculated. In addition, non-parametric Wilcoxon tests were conducted to assess potential differences in the  $R_{\text{runoff}}$ ,  $R_{\text{exfiltration}}$ , and  $R_{\text{retained}}$  of soil samples between the restored meadow, the turf land, and the cultural meadow.

After rainfall simulation tests were completed on each sample, the sample was left until percolation water stopped draining from the base of the sample. Subsequently, three infiltration tests were conducted on each soil sample using the mini-disk infiltrometer [21], which is a very robust tool to measure the saturated hydraulic conductivity of soils at a location. Measurements were spaced evenly over the soil surface and away from the edge of the soil cutter to eliminate the effects of no-flow boundary, which was likely to reduce the infiltration rate.

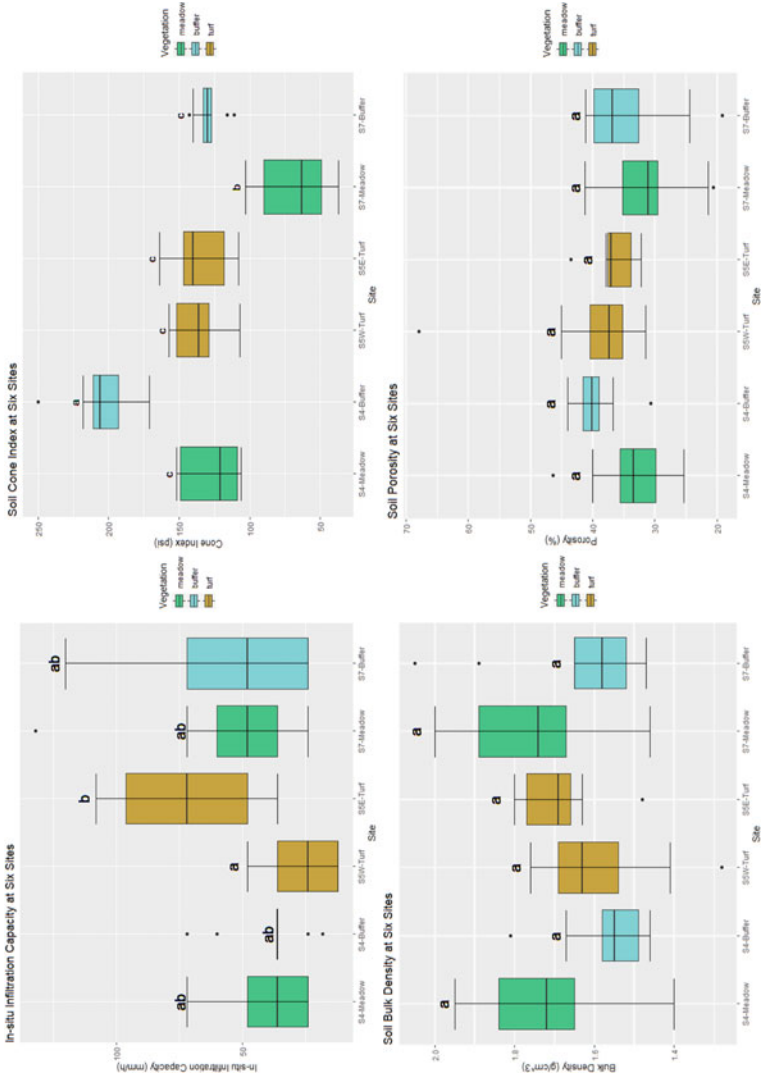
## 4 Results and Discussion

### 4.1 Soil Characterization

The results of soil physical property analysis, including in-situ infiltration capacity, cone index, bulk density, and porosity at different study sites, in the three land types, and in the South Slope and the Iroquois Plain, are plotted in Fig. 3. All sites had moderate infiltration rates. The infiltration capacity measured by the double-ring infiltrometer is collected under near-saturated conditions, and thus the results can be compared to soil saturated hydraulic conductivity. The mean in-situ saturated hydraulic conductivities at the six sites ranged from 26 to 71 mm/h (0.6–1.7 m/d), suggesting that The Meadoway sites are comprised of loamy sand to fine sand [7]. As is common with soils, each site exhibited a wide range in infiltration capacity. Individual measurements ranged from 12 mm/h at Section 5 West to 132 mm/h at Section 7 meadow. In general, significant differences ( $p$ -value = 0.0017) in infiltration capacity were only observed between sites located in the South Slope (Section 4 and 5 West) and the Iroquois Sand Plain (Section 5 East and 7). Average infiltration capacity was approximately two times higher in the Iroquois Sand Plain (60 mm/h) than in the South Slope (36 mm/h). Vegetation treatments (turf, meadow or buffer) were not associated with any differences in infiltration capacity.

The average cone index at the six sites ranged from 67 to 206 psi. Unlike the infiltration capacity, trends in soil compaction were noted between the turf, meadow, and buffer treatments. The two turf sites (Section 5 West and East) had similar average cone indices of 136 and 137 psi, respectively. In contrast, the meadow treatment was associated with a significant decrease in cone index. In Section 4, the average cone index was 206 psi in the buffer but only 127 psi in the meadow. Similarly, in Section 7, the average cone index declined from 129 psi in the buffer to 67 psi in the meadow.

The mean bulk densities at the six sites were very close, ranging from 1.6 to 1.8 g/cm<sup>3</sup> with a very small standard deviation of 0.1 g/cm<sup>3</sup> among all sites. The mean bulk density in meadows was greater than that in buffer areas and turf lands, and the mean



**Fig. 3** Plots of in-situ infiltration capacity, cone index, bulk density, and porosity of soils

bulk density in the South Slope was slightly greater than that in the Iroquois Sand Plain. The mean porosities at the six sites were also very close, ranging from 33 to 41% and having a very small standard deviation of 2.6% among all sites. The mean porosity in meadows was smaller than that in buffer areas and turf lands, and the mean porosity in the South Slope was greater than that in the Iroquois Sand Plain. No significant differences in bulk density and porosity between different sites and between the South Slope and Iroquois Sand Plain were observed. However, it was found that the bulk density of the soil in meadows was significantly greater than that in buffer areas ( $p$ -value = 0.032), and the porosity of the soil in meadows was significantly lower than that in buffer areas and turf lands ( $p$ -values = 0.023). In addition, a significant negative correlation was observed between the cone index of surface soil and the bulk density of subsurface soil ( $p$ -value = 0.002) and a significant positive correlation between the cone index of surface soil and the porosity of subsurface soil ( $p$ -value = 0.0002).

The near-saturated infiltration capacity at the six study sites is all greater than  $10^{-8}$  m/s, indicating good water drainage in these areas [33]. The differences in the in-situ infiltration capacity between different sites are primarily caused by different soil types. The soil in the Iroquois Sand Plain has a significantly greater in-situ infiltration capacity than that in the South Slope. The lower in-situ infiltration capacity in the South Slope can be explained by the results of fine particle size distribution analysis that the soil in the South Slope was finer and had a greater portion of clay than the soil in the Iroquois Sand Plain. This result is supported by findings from past literature that soil physical properties, including soil texture, the amount of organic matter, soil structure, and soil permeability, affected soil hydraulic and hydrological characteristics, including infiltration capacity and moisture retention capacity [13, 16].

Vegetation cover did not significantly affect the in-situ infiltration capacity in the study sites. Many previous studies have observed that plant covers greatly affect a soil's infiltration capacity [30, 31]. Reyes Gomez et al. [30] discussed that rather than preferentially promoting native grass species, species-specific functional traits determined the infiltration capacity. This might explain the insignificant difference in the in-situ infiltration capacity between the study areas covered with turf grasses and native species. Characterizing species-specific functional traits of turf grasses and native plants such as root length density in the study areas may allow us to better understand the dominant pathways of infiltration water and evaluate the influence of native plant restoration on infiltration and percolation. In addition, since the area of each restored meadow is very large, nine double-ring infiltration tests in each selected 1-ha study area may be inadequate to capture the representative hydrological performance of the land. Double-ring infiltration tests are small-scale and may not be sufficient to measure the variability of conditions in test areas [27]. Based on the Stormwater Management Criteria by the TRCA [20], at least two infiltration tests should be conducted per soil test pit, and a minimum of one soil test should be conducted for each  $450\text{ m}^2$  of study area. Therefore, at least 44 double-ring infiltration tests should be conducted at 22 locations at each 1-ha study site in the future.



In general, the plant coverage in restored meadows is less than that in turf lands and buffer areas, which might limit the performance of native plants in enhancing infiltration. Furthermore, the age of restored plants also affects the soil's infiltration capacity, as the roots of newly established plants can block flow channels, and the root decay of mature plants can enhance infiltration rate by creating macropores in the soil [19]. Long-term monitoring and investigation are necessary to trace the influence of native plant restoration on the infiltration capacity of the land as the meadow matures.

The cone index, also referred to as penetration resistance, reflects the compression conditions of the soil. According to TRCA [32], soil with a surface resistance (cone index) greater than 110 psi is considered to have been compacted to the degree that limits plant root growth. Based on this criterion, the cone indexes at the six sites are high enough to restrict the growth of plants. The cone index is significantly affected by soil physical properties (e.g., soil texture), soil moisture, plant covers, and tillage practices [17]. Tillage practices usually decrease the cone index of the soil, especially in the top layer [5]. The study by [15] indicated that introducing plants with deep, strong taproots could mitigate soil compaction. Meadows have a much lower cone index than turf lands and buffer areas, which the tillage might cause during restoration and the replacement of native plants. When collecting soil samples, it was observed that the soil from meadows was much looser than the soil from adjacent buffer areas, indicating the effects of tillage. The native plants in restored meadows also have thicker and deeper roots than turf grasses in pre-restored turf lands, which may also play an important role in decreasing soil compaction.

Statistical analysis showed that the Iroquois Sand Plain had a significantly lower cone index than the South Slope in the study sites, which could be explained by the differences in soil structure. There was no significant correlation between in-situ infiltration capacity and cone index, indicating that cone index is not an effective indicator of in-situ infiltration capacity of the soil. This reduces the noise in in-situ infiltration capacity analysis caused by possible differences in soil compression and facilitates the investigation of the influence of vegetation on the hydrological performance of the land. Statistical tests showed that the cone index was negatively correlated with the subsurface bulk density and positively correlated with the subsurface porosity. Since the cone index measured the penetration resistance of surface soil and the samples for bulk density and porosity were taken from 30 cm below the surface, the revealed correlation might not indicate a causal relationship between the variables.

The bulk density and the porosity did not differ much between different sites and between the South Slope and Iroquois Sand Plain. Meadows soils had a significantly higher soil bulk density than buffer areas and a significantly lower soil porosity than buffer areas and turf lands. Bulk density is an indicator of soil compaction and is affected by inherent soil properties, including soil texture, the densities of soil minerals and organic matter particles, and particle packing arrangement and soil management practices such as cultivation [38]. Since the soil samples for bulk density and porosity analysis were taken from a depth of 30 cm, the results reflect the properties of subsurface soil and not necessarily that of the surface soil. The

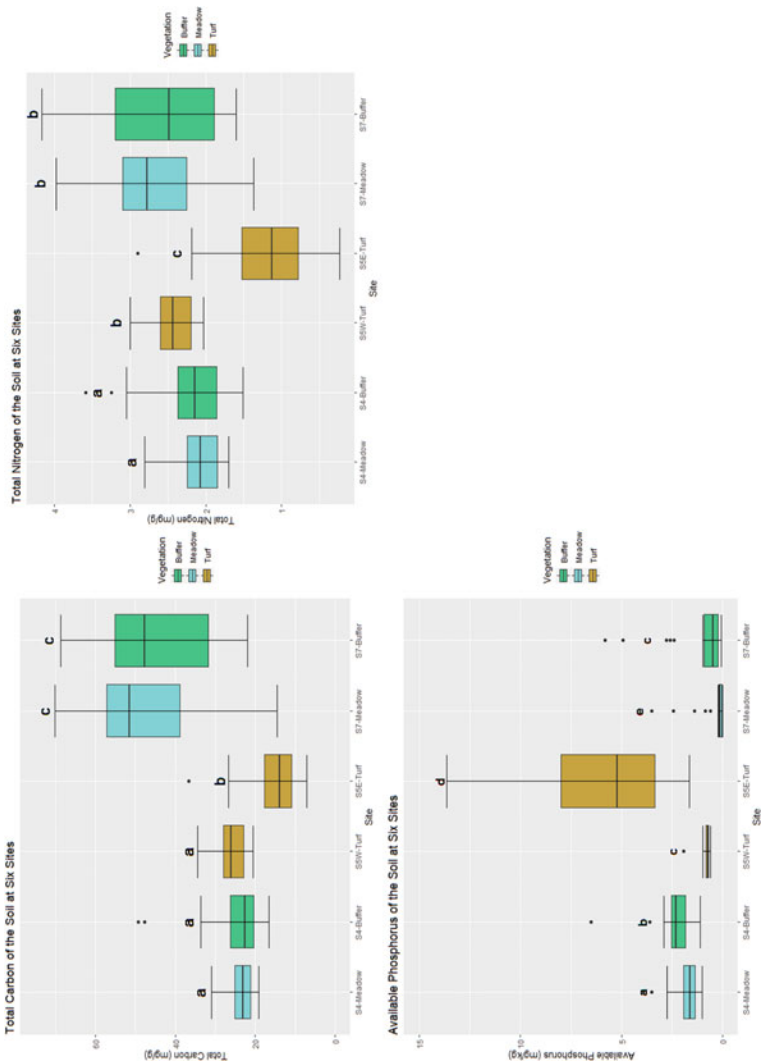


bulk density data indicated that the subsurface soil in meadows is more compacted than the subsurface soil in turf lands and buffer areas. Past studies showed that surface tillage increased subsurface compaction, creating a “hardpan” or “plow pan” due to the weight of tillage equipment [14], which might be the case of the soil in The Meadowway. A more comprehensive analysis of the surface and subsurface soil samples for soil textural and structural parameters (e.g., particle size distribution, pore size distribution, pore structure) is needed to investigate the effects of vegetation covers and soil management practices on soil physical and hydraulic properties.

The results of soil chemical property analysis, including total carbon, total nitrogen, and available phosphorus at different study sites, in the three land types, and in the South Slope and the Iroquois Sand Plain are plotted in Fig. 4. The mean total carbon of the soils at the six sites ranged from 15.33 to 47.30 mg/g. There were no clear trends in total carbon associated with vegetation treatment or geology. Soils at Section 7 had a significantly higher total carbon than that at all other sites ( $p$ -values  $\leq 10^{-5}$ ). Total carbon levels were similar ( $p$ -value = 0.686 and 0.585, respectively) between meadow and buffer samples in Sections 4 and 7. Soil carbon rates differed between each turf site (Section 5 East and West) (values  $\leq 10^{-7}$ ). Soil nitrogen levels followed a similar pattern. The mean total nitrogen of the soils at the six sites ranged from 1.22 to 2.73 mg/g. Total nitrogen of the soil at Section 5 East turf land was significantly lower than that at all other sites ( $p$ -values  $\leq 10^{-6}$ ). Total nitrogen levels were similar ( $p$ -value = 0.878 and 0.721, respectively) between meadow and buffer samples in Sections 4 and 7. Soil nitrogen rates differed between each turf site (Section 5 East and West) ( $p$ -values  $\leq 10^{-6}$ ).

The mean plant-available phosphorus of the soils at the six sites had a wide range from 0.44 to 6.12 mg/g. The mean plant-available phosphorus contents of the soils at Section 7 meadow and Section 5 West turf land were much lower than that at all other sites. In contrast, the mean plant-available phosphorus of the soil at Section 5 East turf land was much higher than that at all other sites. The sites ranked according to the soil plant-available phosphorus content from low to high are: Section 7 meadow, Section 5 West turf, Section 7 buffer, Section 4 meadow, Section 4 buffer, and Section 5 East turf. Meadows had a significantly lower soil plant-available phosphorus content than turf lands and buffer areas ( $p$ -value = 0.0029 and 0.0058, respectively). The plant-available phosphorus of the soil deviated greatly between different sites. A significant difference was found between most paired sites ( $p$ -value  $< 0.001$ ), reflecting a great variance of the soil plant-available phosphorus content along The Meadowway. No significant difference was found between the soil collected from the South Slope and Iroquois Sand Plain. Total carbon and total nitrogen in soils have a significant positive relationship, while available phosphorus has a negative relationship with total carbon and total nitrogen.

The chemical analysis showed that the soil in Section 5 East had a significantly lower total carbon than the soils in other sites, indicating that the soil in Section 5 East might have a smaller amount of organic matter. Studies showed that the organic carbon in the soil was positively related to silt and clay content [25]. During soil sampling, it was observed that the soil in Section 5 East was much more sandy and less clayey in texture than the soils in other sites, which might explain its low



**Fig. 4** Plots of total carbon, total nitrogen, and available phosphorus of soils at six sites

organic matter content. In contrast, the soils in Section 7 meadow and buffer area had a significantly higher total carbon than soils in other sites, indicating that they had a greater amount of organic matter. Organic matter plays an important role in nutrient cycling as the mineralization of organic matter produces available nitrogen and phosphorus for plants [22]. Organic matter also provides food sources for soil microorganisms, which are heavily involved in nutrient cycling.

The soil in Section 5 East had significantly lower total nitrogen and significantly higher available phosphorus than all other sites. The soils in Section 7 meadow and buffer area had greater total nitrogen and lower available phosphorus than soils in other sites. Organic matter holds 90–95% of the nitrogen in soils [22]. The low content of total nitrogen in Section 5 East can be explained by the low content of organic matter, which is indicated by the low total carbon. Similarly, the high content of soil nitrogen in Section 7 can be explained by the high content of organic matter indicated by the high total carbon. This relationship is revealed by the significant positive correlation between total carbon and total nitrogen from Spearman correlation tests. The research by Wibowo and Kasno [39] also found a strong positive correlation between soil organic carbon content and total nitrogen, showing that the nitrogen retention capacity of the soil increases with a greater amount of soil organic matter. The available phosphorus in soils is primarily affected by soil inherent properties (e.g., soil moisture, aeration, salinity, pH, and amount of organic matter) and climate (e.g., rainfall and temperature) [37]. Soils with a greater amount of organic matter have a larger pool of phosphorus for mineralization.

Soils with pH values below 5.5 or between 7.5 and 8.5 have very limited phosphorus availability to plants. Well-aerated soils release phosphorus faster than poorly aerated soils (i.e., saturated soil). Although the soil in Section 5 East has less total carbon, it has significantly more available phosphorus.

In contrast, the soil in Section 7 had a smaller amount of available phosphorus even though it had a high total carbon content. In addition, it was notable that the soil in the buffer area had significantly higher available phosphorus than the soil in the restored meadow in both Sections 4 and 7. However, no significant differences were found in the total carbon between the two land types. To find out the root cause of the variation in soil available phosphorus, soil properties including soil pH, soil organic matter content, salinity, and aeration should be analyzed in different sites.

## 4.2 Rainfall Simulation Tests and Water Balance Analysis

The means of runoff to precipitation ratio ( $R_{\text{runoff}}$ ), exfiltration to precipitation ratio ( $R_{\text{exfiltration}}$ ), and retained water to precipitation ratio ( $R_{\text{retained}}$ ) for soil samples from the restored meadow, the turf land, and the cultural meadow in simulated rainfall events with different return periods are summarized in Table 4. The weight of soil samples increased between the 2-year and 5-year rainfall simulation suggesting that samples may have not been at field capacity. Almost no surface runoff was generated from all soil samples for some tests (i.e., meadow 2 vs. turf land 6, meadow 1 vs.

turf land 5). These samples exhibited great shrinkage when exposed to the simulated rainwater. Consequently, although water was observed pooling on the sample surface during the tests, it was unable to flow into the runoff collection container. Therefore, the data of rainfall simulation tests on these samples was excluded from analysis. In subsequent simulation tests, soil samples were collected with an extra thickness of about 1 cm to mitigate the effect of soil shrinking.

In successful tests, the turf grass samples generated much more surface runoff than the restored meadow and the cultural meadow (Fig. 5). A small amount of surface runoff occasionally occurred in the restored meadow in 5-, 10-, 25-, and 50-year rainfall events. In addition, as rainfall intensity increased with greater return periods, the surface runoff ratio of the turf land increased gradually from 1.1 to 31.6%. This indicates that as rainfall intensity increased, the land had a smaller capacity to infiltrate water and a greater chance to generate runoff. The cultural meadow generated no surface runoff in all rainfall events. Both the restored meadow and the cultural meadow had a greater capacity to infiltrate rainfall and mitigate surface runoff compared with the turf land. The results of Wilcoxon tests showed that the turf land had a significantly greater surface runoff ratio and a significantly smaller exfiltration ratio than the restored meadow and the cultural meadow ( $p$ -values  $\leq 0.0017$ ).

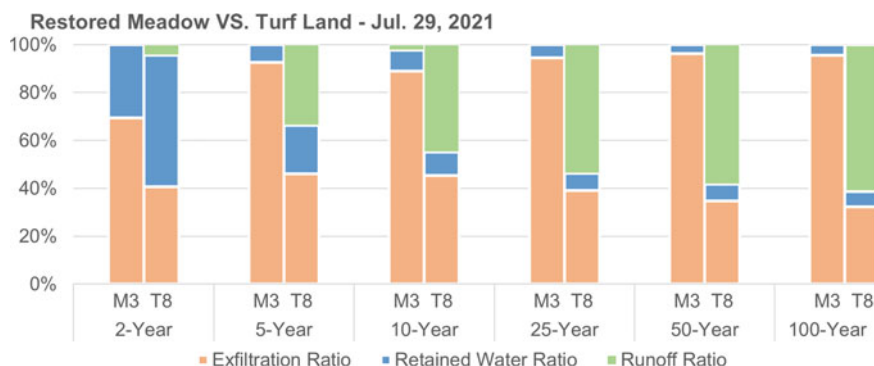
The saturated hydraulic conductivity of each sample from the restored meadow, the turf land, and the cultural meadow ranged from 3.10 to 9.82 cm/h, which is within the range for the textural class of sandy loam [24]. Although rainfall simulation tests showed that the turf land had a greater tendency to generate runoff than the restored meadow and the cultural meadow, this hydrologic behavior could not be predicted based on samples' saturated hydraulic conductivities. The saturated hydraulic conductivity is an important measurement to estimate the infiltration in the water balance analysis of lands [26]. The results of this study suggest that the saturated hydraulic conductivity measurement cannot be interpreted solely to characterize the hydrological response of different land types to storms.

The results of the water balance analysis showed that the turf land had a much greater tendency to generate surface runoff than the restored meadow and the cultural meadow. As rainfall intensity increased, the turf land infiltrated less rainfall and generated more runoff, increasing flood risks. Both the restored and cultural meadows had a great capacity to infiltrate rainfall and mitigate surface runoff even during 50-year and 100-year rainfall events with significant rainfall intensities. In conclusion, the restoration of native meadows can enhance the hydrological regulating functions of urban greenspace by flood control.

Regardless of the significantly different hydrological performances, the restored meadow, the turf land, and the cultural meadow had similar saturated hydraulic conductivities based on mini-disk infiltrometer tests. Therefore, the saturated hydraulic conductivity measurement is not an effective indicator to quantify the enhanced hydrological regulating services of green infrastructures like The Meadoway. In the future, other hydrological experiments, including large-scale rainfall simulation, water balance analysis, and lysimeter tests, should be conducted to evaluate the enhanced hydrological functions of the land.

**Table 4** The runoff ratio, exfiltration ratio, and retained water ratio of soil samples from restored meadow, turf land, and cultural meadow in rainfall events with return periods of 2, 5, 10, 25, 50, and 100 years

Return period (year)	Runoff ratio (%)			Exfiltration ratio (%)			Retained water ratio (%)		
	Restored meadow	Turf land	Cultural meadow	Restored meadow	Turf land	Cultural meadow	Restored meadow	Turf land	Cultural meadow
2	0.00	1.1	0.00	55	61	53	45	38	47
5	1.6	7.9	0.00	86	75	81	12	12	19
10	3.5	11	0.00	87	82	90	10	7.3	10
25	4.6	24	0.00	88	70	94	7.3	6.8	6.2
50	0.30	25	0.00	94	70	92	5.7	5.7	8.2
100	0.00	32	0.00	95	61	95	5.0	7.5	4.6



**Fig. 5** The results of water balance analysis of rainfall simulation tests: runoff ratio, exfiltration ratio, and retained water ratio of restored meadow versus turf land sampled on Jul. 29, 2021

## 5 Conclusion and Recommendations

Through the field investigation and soil sample analysis, it was revealed that the meadow restoration along a hydro corridor in Toronto enhanced regulating (e.g., erosion and flood control) and supporting services (e.g., soil quality regulation and nutrient cycling) of the linear greenspace. A notable finding is that in-situ infiltration tests cannot comprehensively reveal the enhanced hydrological regulating function of the land (e.g., runoff reduction and infiltration capacity), and it should be supplemented by other tests (e.g., rainfall simulation tests). Although the meadow restoration did not significantly increase the infiltration capacity, the water balance analysis of rainfall simulation tests showed the great capacity of restored meadows to facilitate infiltration and reduce surface runoff. This effect might be the mixed result of native plant restoration and the tillage practice in meadows, which caused the alteration of soil physical and hydraulic properties. Meadow restoration decreased surface soil compaction but increased subsurface soil (30 cm) compaction, which might be the consequence of tillage. The decreased surface runoff can potentially reduce water-related soil erosion in restored meadows. The soil chemical parameters (i.e., total C, total N, and available P) varied significantly among different sites, with adjacent pre-restored turf lands having consistently greater available phosphorus in the soil than restored meadows. The mitigated soil erosion in restored meadows can potentially decrease the risk of nutrient losses from the land especially for P and organic C. The different traits between native plants and turf grasses (e.g., root structure, turnover rate, and litter quality) can alter soil structure (e.g., organic content, pore structure, and aggregate stability), which plays an important role in erosion control and nutrient cycling. The study also provides many indications on future research directions and designs for The Meadowway as well as the performance evaluation of similar green infrastructures.

Recommendations are provided here for future research work. To better understand the effects of meadow restoration on soil hydraulic properties and the root

causes, a greater number of in-situ infiltration tests should be conducted in the future, accompanied with more comprehensive surface soil analysis for structural and textural parameters (e.g., complete particle size distribution, pore size distribution, pore structure) and root trait analysis (e.g., root density, root length density, root radius). Long-term monitoring is also recommended to capture The Meadoway's performance as it matures with time. As a very effective method to reveal the hydrological function enhanced by meadow restoration, rainfall simulation tests can be adopted and upgraded in the future with a larger scale. More realistic designed storms integrated with water quality analysis of runoff and exfiltration, which can indicate the improvement of erosion control and water quality regulating services. To further investigate the influence of meadow restoration on soil nutrient cycling, the analysis of soil nutrient content and other properties, including soil pH, organic matter content, texture, salinity, and aeration, should be conducted. Lastly, the nutrient analysis of different plant species (e.g., allocation of nutrients in foliage and roots) and investigating the nutrient interactions between plants, roots, and microorganisms are also recommended.

**Acknowledgements** This research was funded by the Natural Sciences and Engineering Research Council (NSERC) CREATE Grant and a Mitacs-Alliance Grant in partnership with the Toronto and Region Conservation Authority (TRCA). Special thanks to the TRCA, especially Katie and Chris, who provided great support and many suggestions throughout the whole project. Thanks to my colleagues as well as summer students who helped with field and lab work, including but not restricted to Sylvie, Jody, Jad, Giuliana, Samantha, Kelsey, Virinder, Ernie, Nathanael, Jihwan, and Sahildeep.

## References

1. Allen SMJ, Gough WA, Mohsin T (2015) Changes in the frequency of extreme temperature records for Toronto, Ontario, Canada. *Theoret Appl Climatol* 119(3–4):481–491. <https://doi.org/10.1007/s00704-014-1131-1>
2. AMS Inc. (n.d.) AMS bulk density soil sampling kit manual. <https://www.ams-samplers.com/pdfs/Bulk-Density-Soil-Sampling-Kit.pdf>
3. ASTM International (2017) D7928-17 standard test method for particle-size distribution (gradation) of fine-grained soils using the sedimentation (hydrometer) analysis
4. ATS Scientific (n.d.) QuikChem 8500 series 2 FIA system: frequently asked questions. <https://www.google.com/search?q=QuikChem+8500+Series+2+FIA+System+Frequently+Asked+Questions&aq=chrome..69i57.316j0j15&sourceid=chrome&ie=UTF-8>
5. Bueno J, Amiama C, Hernanz JL, Pereira JM (2006) Penetration resistance, soil water content, and workability of grasslands soils under two tillage systems. *Trans ASABE* 49(4):875–882
6. Chase TN, Pielke RA Sr, Kittel TGF, Nemani RR, Running SW (2000) Simulated impacts of historical land cover changes on global climate in northern winter. *Clim Dyn* 16(2–3):93–105. <https://doi.org/10.1007/s003820050007>
7. Chin DA (2013) *Water-resources engineering*, 3rd edn. Pearson Education
8. City of Toronto (2006) Wet weather flow management guidelines. <https://www.toronto.ca/wp-content/uploads/2017/11/9191-wwfm-guidelines-2006-AODA.pdf>

9. Das M, Das A (2019) Dynamics of urbanization and its impact on urban ecosystem services (UESs): a study of a medium size town of West Bengal, Eastern India. *J Urban Manag* 8(3):420–434. <https://doi.org/10.1016/j.jum.2019.03.002>
10. Eijkelkamp Soil and Water (2018) Double ring infiltrometer. [https://www.eijkelkamp.com/download.php?file=M0904e\\_Double\\_ring\\_infiltrometer\\_ac22.pdf](https://www.eijkelkamp.com/download.php?file=M0904e_Double_ring_infiltrometer_ac22.pdf)
11. ELE International. (n.d.) Proving ring penetrometer. Retrieved 26 Dec 2021, from <https://www.ele.com/product/proving-ring-penetrometer>
12. Fisher Scientific (n.d.) Fisherbrand™ GP1000 general-purpose peristaltic pumps. Retrieved 29 Dec 2021, from <https://www.fishersci.com/shop/products/fh100-fh100x-general-purpose-peristaltic-pumps-1/13310656>
13. Franzluebbers AJ (2002) Water infiltration and soil structure related to organic matter and its stratification with depth. *Soil Tillage Res* 66(2):197–205
14. Government of Alberta (2010) Agricultural soil compaction: causes and management. [https://www1.agric.gov.ab.ca/\\$department/deptdocs.nsf/all/agdex13331/\\$file/510-1.pdf?OpenElement#:~:text=Discedorcultivatedsurfacesoils,movingwaterthroughthesoil](https://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/agdex13331/$file/510-1.pdf?OpenElement#:~:text=Discedorcultivatedsurfacesoils,movingwaterthroughthesoil)
15. Hamza MA, Anderson WK (2005) Soil compaction in cropping systems: a review of the nature, causes and possible solutions. *Soil Tillage Res* 82(2):121–145
16. Helalia AM (1993) The relation between soil infiltration and effective porosity in different soils. *Agric Water Manag* 24(1):39–47
17. Kumar A, Chen Y, Sadek MA-A, Rahman S (2012) Soil cone index in relation to soil texture, moisture content, and bulk density for no-tillage and conventional tillage. *Agric Eng Int CIGR J* 14(1):26–37
18. LECO Corporation (2016) 628 series elemental determinators. <http://lecoperu.com/wp-content/uploads/2018/05/CHN628-ESPECIFICACIONES-ENG-29052018.pdf>
19. Meek BD, Rechel ER, Carter LM, DeTar WR, Urie AL (1992) Infiltration rate of a sandy loam soil: effects of traffic, tillage, and plant roots. *Soil Sci Soc Am J* 56(3):908–913. <https://doi.org/10.2136/sssaj1992.03615995005600030038x>
20. Meerow S (2019) A green infrastructure spatial planning model for evaluating ecosystem service tradeoffs and synergies across three coastal megacities. *Environ Res Lett* 14(12). <https://doi.org/10.1088/1748-9326/ab502c>
21. Meter Environment (n.d.) Mini disk infiltrometer. <https://www.metergroup.com/environment/products/mini-disk-infiltrometer/>
22. Murphy B (2014) Soil organic matter and soil function—review of the literature and underlying data. Department of the Environment, Canberra, Australia
23. Ontario Ministry of Finance (2020) Ontario population projections update, 2019–2046. [https://www.fin.gov.on.ca/en/economy/demographics/projections/#:~:text=TheGreaterTorontoArea\(GTA,over9.5millionby2046.&text=Thefiveotherregionsare,populationovertheprojectionperiod](https://www.fin.gov.on.ca/en/economy/demographics/projections/#:~:text=TheGreaterTorontoArea(GTA,over9.5millionby2046.&text=Thefiveotherregionsare,populationovertheprojectionperiod)
24. Pachepsky Y, Park Y (2015) Saturated hydraulic conductivity of US soils grouped according to textural class and bulk density. *Soil Sci Soc Am J* 79(4):1094–1100. <https://doi.org/10.2136/sssaj2015.02.0067>
25. Plante AF, Conant RT, Stewart CE, Paustian K, Six J (2006) Impact of soil texture on the distribution of soil organic matter in physical and chemical fractions. *Soil Sci Soc Am J* 70(1):287–296
26. Post R, Owens D (2020) Overview of water balance practices in the Greenbelt. [https://www.nvca.on.ca/SharedDocuments/GreenbeltWaterBalanceReport\\_FINAL\\_24-Jan-2021.pdf](https://www.nvca.on.ca/SharedDocuments/GreenbeltWaterBalanceReport_FINAL_24-Jan-2021.pdf)
27. Puget Sound Action Team and Washington State University Pierce County Extension. 2005. Low impact development: technical guidance manual for puget sound. [https://www.jtc.sala.ubc.ca/reports/LID\\_manual\\_pugetsound2005.pdf](https://www.jtc.sala.ubc.ca/reports/LID_manual_pugetsound2005.pdf)
28. Qin K (2022) Regulating and supporting ecosystem services provided by urban greenspace and restored meadows along a hydro corridor in Toronto. University of Toronto
29. R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing. <http://www.r-project.org/>



30. Reyes Gomez VM, Viramontes Olivas O, Arredondo Moreno JT, Huber Sannwald E, Rodriguez AR (2015) Functional ecohydrological differences among native and exotic grassland covers in sub-urban landscapes of chihuahua city, Mexico. *Landsc Urban Plan* 139:54–62. <https://doi.org/10.1016/j.landurbplan.2015.03.005>
31. Tang B, Jiao J, Yan F, Li H (2019) Variations in soil infiltration capacity after vegetation restoration in the hilly and gully regions of the Loess Plateau, China. *J Soils Sediments* 19(3):1456–1466
32. TRCA (2012) Preserving and restoring healthy soil: best practices for urban construction. [https://trcaca.s3.ca-central-1.amazonaws.com/app/uploads/2021/10/20103048/preserving\\_and\\_restoring\\_healthy\\_soil\\_trca\\_2012.pdf](https://trcaca.s3.ca-central-1.amazonaws.com/app/uploads/2021/10/20103048/preserving_and_restoring_healthy_soil_trca_2012.pdf)
33. TRCA (2012) Stormwater management criteria. <https://trca.ca/conservation/stormwater-management/understand/swm-criteria-2012/download>
34. TRCA (2019) Restoration opportunities planning primer
35. TRCA (2019) The Meadoway: vegetation, bird and butterfly monitoring update report 2016, 2018, 2019
36. TRCA (2019) The Meadoway multi-use trail municipal class environmental assessment—schedule C: environmental study report
37. United States Department of Agriculture (n.d.) Soil phosphorus: soil quality kit—guides for educators. [https://www.nrcs.usda.gov/Internet/FSE\\_DOCUMENTS/nrcs142p2\\_053254.pdf](https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_053254.pdf)
38. United States Department of Agriculture (2008) Soil quality indicators: bulk density. <https://www.nrcs.usda.gov/sites/default/files/2023-01/Soil%20Quality-Indicators-Bulk%20Density.pdf>
39. Wibowo H, Kasno A (2021) Soil organic carbon and total nitrogen dynamics in paddy soils on the Java Island, Indonesia. *IOP Conf Ser Earth Environ Sci* 648(1):12192. <https://doi.org/10.1088/1755-1315/648/1/012192>
40. Zang S, Wu C, Liu H, Na X (2011) Impact of urbanization on natural ecosystem service values: a comparative study. *Environ Monit Assess* 179(1–4):575–588. <https://doi.org/10.1007/s10661-010-1764-1>

# Investigation of Climate Risks Within the St. Lawrence Marine Corridor Supported by Ultra-High-Resolution Climate Modelling



Bernardo Teufel, Keihan Kouroshnejad, Laxmi Sushama, Enda Murphy, and Julien Cousineau

**Abstract** Climate change adaptation planning and solutions for coastal infrastructure and navigation in the St. Lawrence marine corridor, which plays a key role in Canada's economy and supply chain, are highly dependent on the availability of climate change information at high spatial and temporal resolutions. In this study, ultra-high-resolution regional climate model simulations are implemented using Environment and Climate Change Canada's Global Environmental Multiscale (GEM) model for current and future climates. Advanced and targeted diagnostics are used to identify vulnerability hotspots and opportunities to address specific climate risks within the corridor. First, an ultra-high spatial resolution ( $\sim 4$  km) simulation spanning the 1989–2010 period for a domain covering the St. Lawrence marine corridor is performed using the GEM model driven by the ERA5 reanalysis. Comparisons of modelled climate fields and parameters relevant to infrastructure and navigation with available observations confirmed the ability of the model to simulate important processes, mechanisms, and seasonality. This is followed by future climate simulations, spanning the 2041–2060 and 2081–2100 periods for Representative Concentration Pathway 8.5 scenario, driven by Canadian Earth System Model (CanESM2) outputs. Given the coarse resolution of CanESM2, a grid-telescoping approach is used, i.e. a 10 km spatial resolution GEM simulation driven by CanESM2 is first performed, the outputs of which are used as lateral boundary conditions for high-resolution GEM simulations at 4 km horizontal resolution. Advanced diagnostics focused on extreme weather and climate are used to understand and pinpoint potential climate risks within the St. Lawrence marine corridor, particularly with

---

B. Teufel (✉) · K. Kouroshnejad · L. Sushama

Department of Civil Engineering and Trottier Institute for Sustainability in Engineering and Design, McGill University, Montreal, Canada

e-mail: [bernardo.teufel@mcgill.ca](mailto:bernardo.teufel@mcgill.ca)

E. Murphy · J. Cousineau

Ocean, Coastal and River Engineering Research Centre, National Research Council, Montreal, Canada

respect to navigability, and the potential climate resiliency of key transportation assets in the study region. This paper will present these results, which will form the basis for additional detailed investigations on climate-infrastructure interactions and other climate resiliency studies.

**Keywords** Climate risks · St. Lawrence marine corridor · Ultra-high-resolution climate modelling

## 1 Introduction

The St. Lawrence marine corridor plays a key role in Canada's economy and supply chain. It extends from Lake Ontario to the Gulf of St. Lawrence, passing through Montreal and Quebec City. Upstream of Montreal, navigation is only possible from mid-March to late December, as lock systems cannot operate in the presence of river ice. Downstream of Montreal, navigation is possible year-round, but hazardous in winter, given icy waters and unfavourable weather conditions. Climate change is expected to have significant impacts on coastal infrastructure and navigation along the corridor, and efforts to adapt these systems to the changing climate require high-resolution climate projections.

Previous climate change studies for the Gulf of St. Lawrence have projected significant decreases in ice cover during the twenty-first century [1, 2]. Given that the presence of ice (in current climate) hinders the generation and growth of waves, projected increases to significant wave heights in the Gulf of St. Lawrence and neighbouring coastal waters are expected in future climate [2], which could be hazardous to coastal infrastructure and navigation in the region.

At their typical horizontal resolutions of tens to hundreds of kilometres (e.g., [3, 4]), global and regional climate models (GCMs and RCMs) rely on parameterizations to approximate processes that occur at smaller scales (e.g., convective cloud systems). The errors and uncertainties introduced by these approximations [5–10] lead to low confidence in model projections of local phenomena, hindering local planning, and decision-making processes.

Convection-permitting models (CPMs) are climate models operating at horizontal resolutions near kilometre scale, which allow for the explicit representation of major convective cloud systems and additionally improve the representation of fine-scale orography and surface heterogeneity (e.g., [11–15]). As model resolution increases, shorter time scale extreme events are expected to be represented more realistically, given that they are often associated with smaller spatial scale structures.

CPM climate simulations typically use a grid-telescoping approach where sequentially finer grids are nested within each other until convection-permitting scales are reached over a limited area (e.g., [14]). Due to their fine horizontal resolution, impact-relevant information can be directly derived from CPM climate simulations [16].

Diro and Sushama [17] used the limited area version of the Global Environmental Multiscale model (GEM, [18]) to perform the first five-year simulation at 3 km grid spacing (i.e. convection permitting scales) over an eastern Canadian Arctic domain. They noted substantial improvements with respect to simulations at 12 km grid spacing in, e.g., extreme precipitation events during summer and winter temperatures. Over a slightly larger domain, [19] performed 4 km grid spacing simulations for the 1989–2040 period and demonstrated that GEM adequately simulates base climate variables, as well as climatic hazards, such as extreme rainfall, extreme wind gust and fog.

In this study, both reanalysis-driven and GCM-driven CPM simulations are performed at 4 km grid spacing over the St. Lawrence marine corridor. The analysis focuses on variables that are expected to benefit from high-resolution and that are also crucial to navigation, such as river ice onset and breakup, and freezing spray conditions.

The paper is organized as follows: Sect. 2 describes the model, simulations, and methods. Section 3 deals with model evaluation. Projected changes are presented in Sect. 4. Finally, the discussion and conclusions are presented in Sect. 5.

## 2 Model and Methods

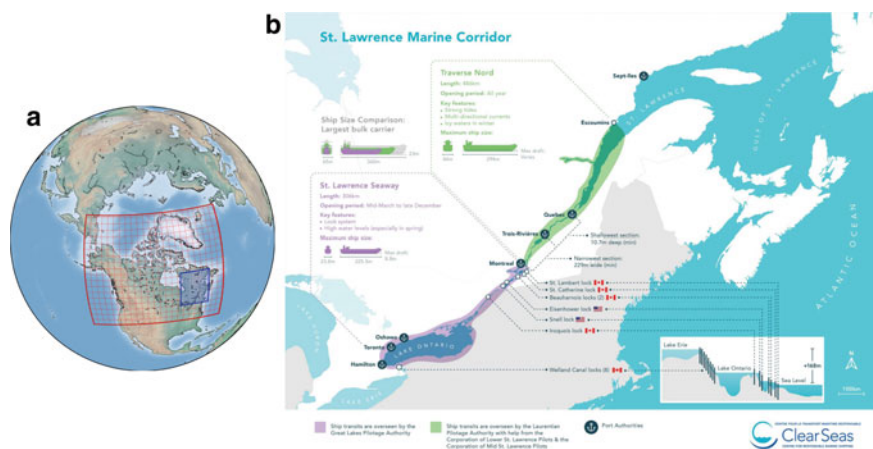
### 2.1 Climate Model

GEM solves non-hydrostatic, deep atmosphere dynamics with an implicit, two-time-level semi-Lagrangian numerical scheme. The model uses a regular latitude–longitude horizontal grid with Arakawa C staggering. To minimize changes in grid spacing across the domain, a rotated pole projection is used such that the domain is approximately centred on the rotated equator. Following [20], Charney–Phillips staggering is used in the vertical coordinate. The planetary boundary layer scheme follows [21] and 22. The radiation scheme is represented by Correlated K solar and terrestrial radiation of [23]. The scheme employed for condensation processes is the double-moment microphysics scheme of [24]. In addition to the large-scale precipitation schemes, the model includes the shallow convection based on [25] and the deep convection scheme of [26]. As convection is neither fully resolved nor can it be assumed to be smaller than the  $\sim 4$  km grid spacing used in this study, the use of convection parameterization is still a topic of debate [27]. Rivers and lakes are represented using FLake [28], a one-dimensional model.

## 2.2 Climate Simulations

Figure 1 shows the model domain, with the inner domain covering the St. Lawrence marine corridor at 4 km grid spacing. A simulation driven by ERA5 reanalysis [29] to minimize boundary forcing errors spans the 1989–2010 period and is hereafter referred to as GEM4\_ERA5. A transient climate change simulation (GEM10\_CanESM2) spanning the 1989–2100 period is performed over the outer pan-Canadian 10 km grid spacing domain (Fig. 1) and is driven by the Canadian Earth System Model (CanESM2; [30]) for Representative Concentration Pathway 8.5 (RCP8.5, [31]). The outputs of GEM10\_CanESM2 are used as lateral boundary conditions for the 4 km horizontal grid spacing GEM4\_CanESM2 simulation over the inner domain. Three 20-year periods are simulated for GEM4\_CanESM2: the reference 1991–2010 period, a mid-century period (2041–2060), and a period towards the end of the century (2081–2100). These simulation periods were chosen to maximize the usefulness of the projections and were constrained by limited computational resources and storage space.

Comparison of GEM4\_ERA5 with available observations will help assess the ability of the model in simulating current climate (1991–2010). Projected changes will be assessed by comparing the future (2041–2060 and 2081–2100) and current periods of the GEM4\_CanESM2 simulation. In this study, model validation and assessment of projected changes are undertaken for a set of climate variables and diagnostics related to navigability.



**Fig. 1** a Grid telescoping for convection permitting climate simulations at 4 km resolution over the St. Lawrence marine corridor (blue grid). The red grid is the pan-Canadian 10 km resolution grid. b Infographic depicting the St. Lawrence Marine Corridor, obtained from the Clear Seas Centre for Responsible Marine Shipping (<https://clearseas.org>)

Validation is achieved by comparing simulations with station observations and with the Daymet dataset [32], which is a 1 km horizontal resolution dataset derived from daily observations of precipitation and near-surface maximum and minimum air temperature at weather stations. At unobserved locations, the Daymet algorithm uses a combination of interpolation and extrapolation, applying weights that reflect the spatial and temporal relationships between a Daymet grid cell and multiple surrounding weather stations. To facilitate comparison, Daymet data is aggregated over each GEM4\_ERA5 grid cell. Other site observations are directly compared to the simulated values for the grid cell in which the observing station is located.

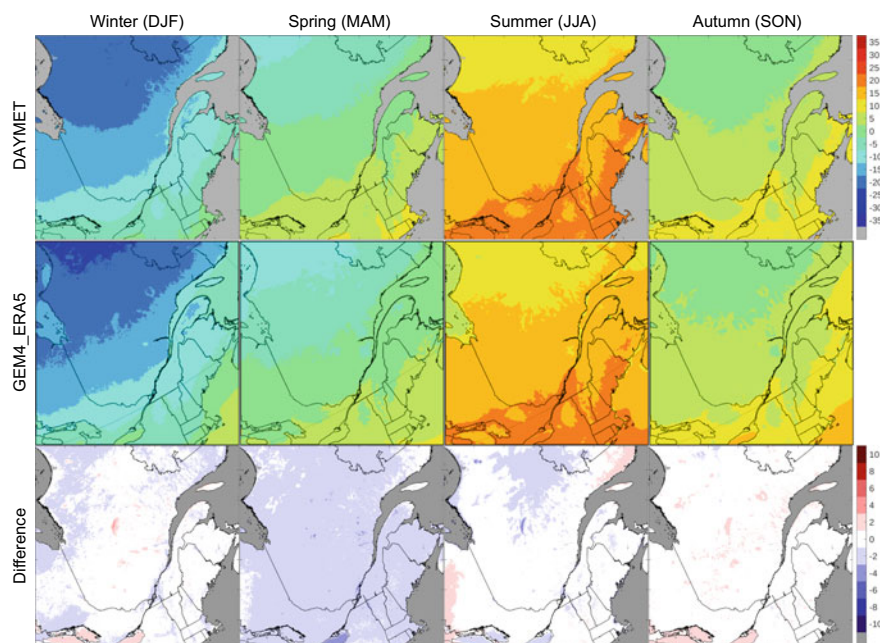
### 3 Validation of Ultra-High-Resolution Climate Simulations

In this study, a two-step process is used for validation. Near-surface air temperature is the most important base variable influencing river ice and winter hazards and is thus validated first. Validation then proceeds to other important variables for which site observations exist, such as the distributions of wind speed and direction, and river ice thickness.

Figure 2 shows that GEM4\_ERA5 captures the spatial and seasonal 2-m air temperature variability in the Daymet dataset quite well. During the spring season, due to slight overestimation of snow cover, GEM4\_ERA5 exhibits a slight cold bias (1–3 °C) over most of the domain, while during the rest of the year, the temperatures in GEM4\_ERA5 and DAYMET are within 1 °C of each other over most locations in the domain. These biases are smaller than in previous climate simulations over the domain (e.g., [33, 34]) and represent a clear example of the added value of high-resolution simulations.

GEM4\_ERA5 is able to reasonably capture the observed distributions of wind speed and direction at most observing stations in the domain (Fig. 3). The differences in wind direction (e.g., stations B, F, G) can be explained by local terrain effects, which operate on scales smaller than 4 km and thus cannot be fully captured by GEM4\_ERA5. Most observing stations also show higher prevalence of strong winds than GEM4\_ERA5. This is expected, as stations are located in open terrain, while GEM calculates wind speed for  $4 \times 4$  km grid cells, which generally include trees and other obstacles to wind. Nonetheless, winds in GEM4\_ERA5 are in much better agreement with observations than winds in coarser resolution simulations (e.g., [35]).

GEM4\_ERA5 captures the annual cycle of ice thickness reasonably well along the St. Lawrence River, as shown in Fig. 4. As in observations, modelled ice thickness is lower near Montreal and higher near Quebec City, also following the differences in climate along the river (Fig. 1, top row). Ice begins forming in December, peaks in early March and persists until late April. GEM4\_ERA5 tends to slightly overestimate ice thickness, which can be explained by the slight cold bias in 2-m air temperature during the winter/spring seasons (Fig. 1, bottom row). It should also be mentioned that the ice scheme in GEM does not account for snow on ice, nor ice dynamics. The



**Fig. 2** Seasonal averages of 2-m air temperature (°C) during the 1991–2010 period, from the DAYMET dataset (top row), from the GEM4\_ERA5 simulation (middle row), and GEM4\_ERA5 biases with respect to DAYMET (bottom row)

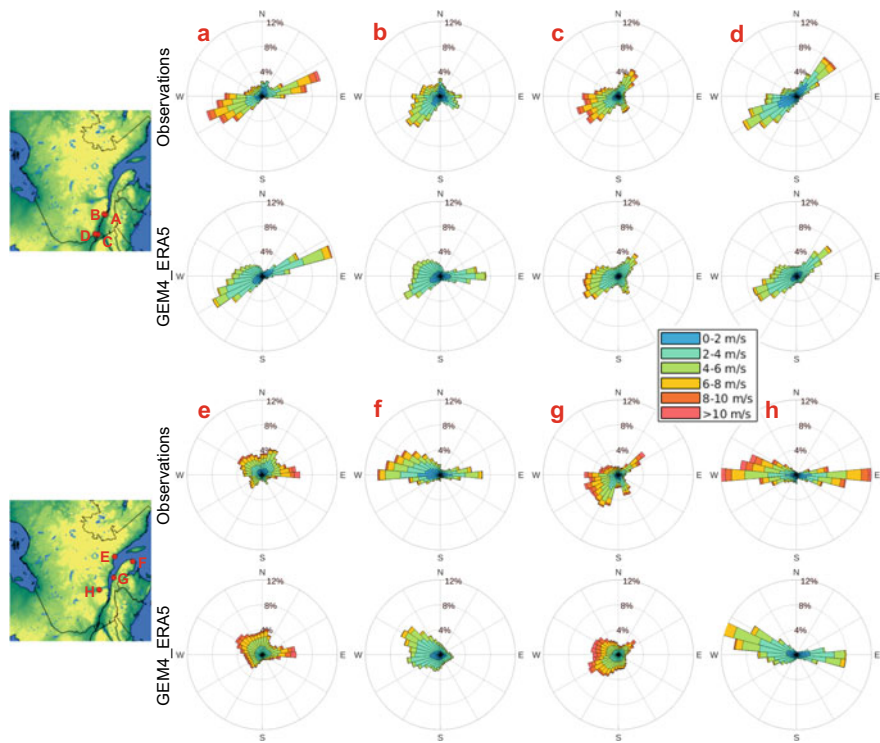
latter can induce large gradients in observed thickness when ice is pushed towards the shore.

Given that GEM4\_ERA5 and GEM4\_CanESM2 differ by the addition of boundary forcing errors, it is important to confirm that the GCM-driven simulation performs reasonably for current climate. GEM4\_CanESM2 exhibits slightly higher air temperatures than GEM4\_ERA5, leading to lower ice thicknesses, while being very similar for wind speed and direction when compared to observations. In general, boundary forcing errors are small and thus the performance of GEM4\_ERA5 is an accurate indicator of the reliability of the projections presented in Sect. 4.

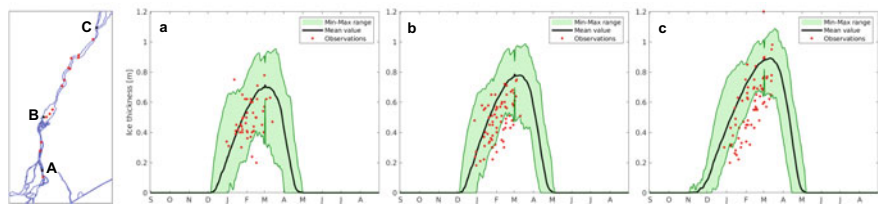
## 4 Navigability of the St. Lawrence in a Changing Climate

As in the validation section, a two-step process is followed. First, changes to near-surface air temperature are presented, given that they drive the changes presented in the second part, which include river ice onset and breakup dates, and freezing spray conditions.





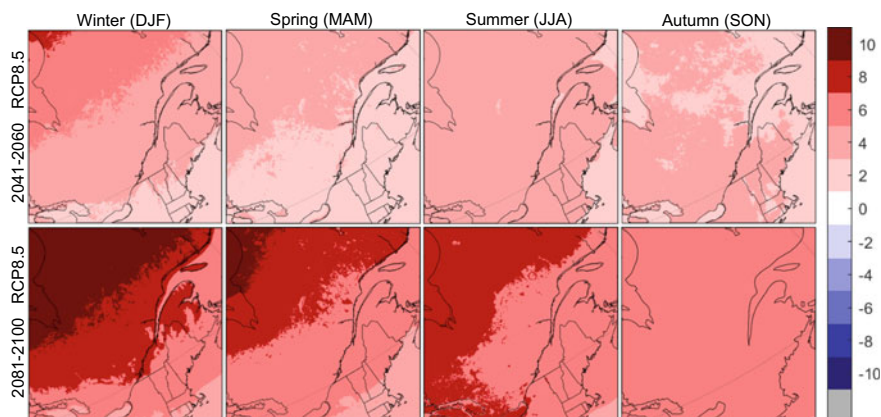
**Fig. 3** Observed and GEM4\_ERA5-modelled distributions of wind speed and direction for the 1991–2010 period at eight airports (A—Quebec City, B—Valcartier, C—Dorval, D—Mirabel, E—Sept-Îles, F—Gaspé, G—Mont-Joli, H—Bagotville)



**Fig. 4** Observed and GEM4\_ERA5-modelled annual cycle of ice thickness (m) for the 1991–2010 period at 3 locations (A—near Montreal, B—Lac St-Pierre, C—near Quebec City)

Following the RCP8.5 high-emissions scenario, near-surface air temperatures are projected to increase significantly (2–4 °C) by the 2041–2060 period (Fig. 5), and even more (4–8 °C) by the end of the century (2081–2100). Consistent with previous research, larger warming is projected for the cold season, due to the albedo-temperature feedback.

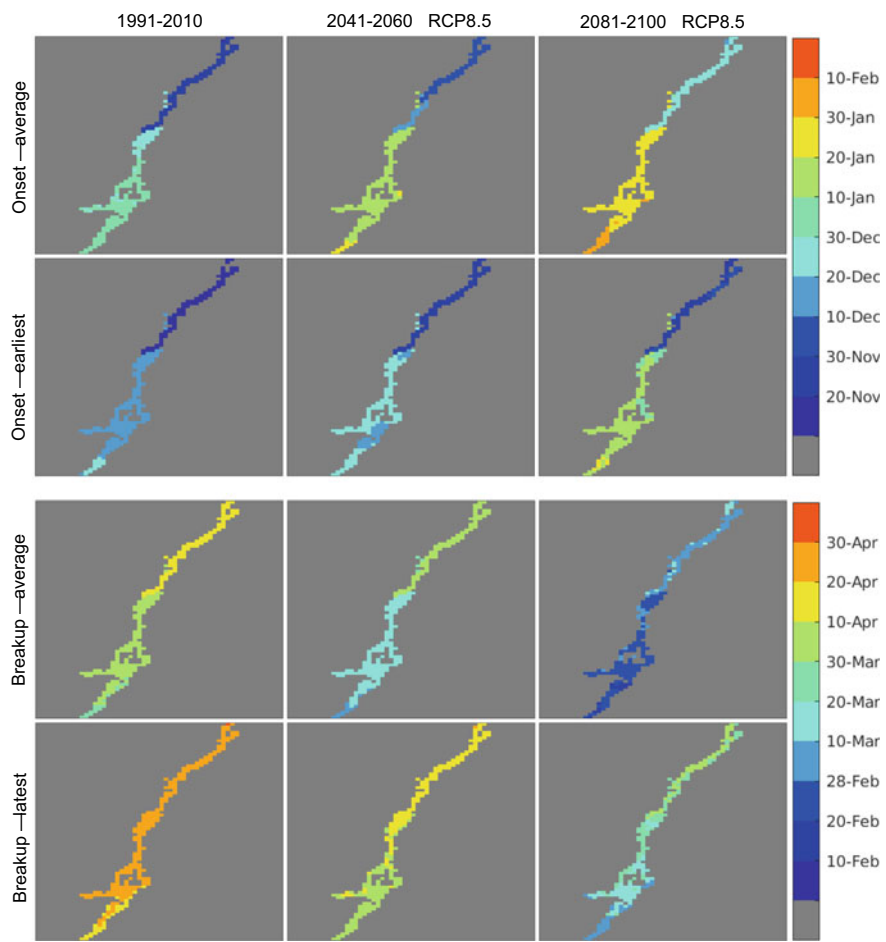




**Fig. 5** Projected changes to seasonal average 2-m air temperature ( $^{\circ}\text{C}$ ) in GEM4\_CanESM2 for the 2041–2060 period (top row) and the 2081–2100 period (bottom row), with respect to the 1991–2010 period

Warmer temperatures (Fig. 5) are projected to increase the duration of the open water period along the St. Lawrence, with ice projected to form later and break up earlier than in current climate. Figure 6 shows that the average ice onset date for most locations is projected to be 10–20 days later by mid-century (2041–2060) and 30–40 days later by the end of the century (2081–2100). Similarly, the average ice breakup date for most locations is projected to be 10–20 days earlier by mid-century and 30–40 days later by the end of the century. Potential benefits of the shortening of the ice season include reduced need for ice breakers downstream of Montreal, and an extension of the shipping season between Lake Ontario and Montreal, which currently is ice-limited during  $\sim 3$  months per year, but this is projected to be reduced to 1–2 months per year by 2081–2100. Potentially damaging impacts of the shortened ice season would include increased exposure of infrastructure and navigation waterways to storm surges [36, 37], including negative surge events, and waves [38].

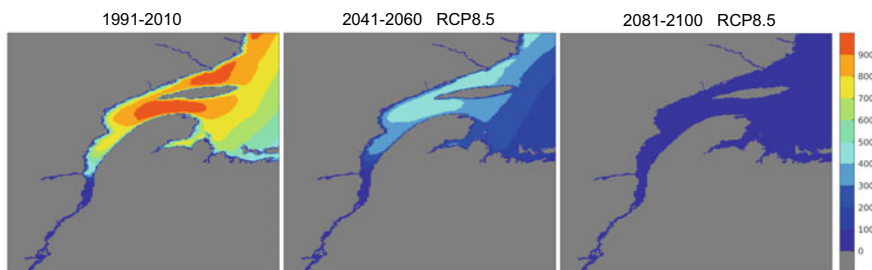
Freezing spray occurs when wind-driven water droplets freeze on impact on exposed surfaces, which can severely impact the stability and safety of vessels, due to rapid ice accumulation. For freezing spray to occur, air temperature needs to be below the freezing point of water ( $0^{\circ}\text{C}$  for fresh water and  $-2^{\circ}\text{C}$  for sea water), water temperature should be below  $6^{\circ}\text{C}$ , wind speed needs to exceed  $8\text{ m/s}$  (Beaufort scale 5) and fractional ice cover should be below 60%. In GEM4\_CanESM2, these conditions occur frequently between Quebec City and the Gulf of St. Lawrence, with up to 900 h of freezing spray conditions per year during the 1991–2010 period (Fig. 7). Warming air and water temperatures are projected to reduce the maximum number of freezing spray hours to below 500 by mid-century (2041–2060) and below 100 by the end of the century (2081–2100). This implies a marked reduction in the projected need for de-icing of vessels, reducing associated risks and expenses, subject to vessel characteristics.



**Fig. 6** Dates of river ice onset and breakup in GEM4\_CanESM2 for the 1991–2010 period (first column), the 2041–2060 period (second column), and the 2081–2100 period (third column). The average date of ice onset (first row), the earliest date of ice onset (second row), the average data of ice breakup (third row), and the latest date of ice breakup (fourth row) are shown

## 5 Summary and Conclusions

The reanalysis-driven ultra-high-resolution (4 km) GEM simulation developed as part of this work is able to well capture the observed behaviour of air temperature, wind speed and direction, and ice thickness over the St. Lawrence marine corridor and adjoining regions. As expected, improvements in the simulation of these variables are noted when compared to previous work at coarser resolutions. The good performance of GEM when compared to observations lends credibility to high-resolution



**Fig. 7** Average number of hours with freezing spray conditions in GEM4\_CanESM2 for the 1991–2010 period (first column), the 2041–2060 period (second column) and the 2081–2100 period (third column)

projections extending to the year 2100. Driven by increasing near-surface air temperatures in a high-emission scenario (RCP8.5), the duration of the river ice season is projected to be significantly reduced—approximately one month by 2041–2060 and up to two months by 2081–2100, when compared to the 1991–2010 reference. Similarly, the frequency of freezing spray conditions is expected to be reduced by up to 50% around mid-century (2041–2060) and up to 90% by the end of the century (2081–2100). While both of these climate impacts would positively impact navigability in the St. Lawrence marine corridor, other hazards (such as fog, storm surges, and high waves) might increase in future climate and merit further study. Additional high-resolution climate projections over the region would be important additions to the work presented here, and of very high value to local planning and decision-making processes.

**Acknowledgements** This research was funded by National Research Council (NRC) of Canada, Natural Sciences and Engineering Research Council of Canada (NSERC), Trottier Institute for Sustainability in Engineering and Design (TISED) and the McGill Sustainability Systems Initiative (MSSI). The GEM simulations in this study were performed on supercomputers managed by Calcul Québec and Compute Canada.

## References

1. Ruest B, Neumeier U, Dumont D, Bismuth E, Senneville S, Caveen J (2016) Recent wave climate and expected future changes in the seasonally ice-infested waters of the Gulf of St. Lawrence, Canada. *Clim Dyn* 46(1):449–466. <https://doi.org/10.1007/s00382-015-2592-3>
2. Wang L, Perrie W, Long Z, Blokhina M, Zhang G, Toulany B, Zhang M (2018) The impact of climate change on the wave climate in the Gulf of St Lawrence. *Ocean Model* 128:87–101. <https://doi.org/10.1016/j.ocemod.2018.06.003>
3. Jacob D, Petersen J, Eggert B, Alias A, Christensen OB, Bouwer LM, Braun A, Colette A, Déqué M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Kröner N, Kotlarski S, Kriegsmann A, Martin E, van Meijgaard E, Moseley C, Pfeifer S, Preuschmann S, Radermacher C, Radtke K, Rechid D, Rounsevell M, Samuelsson P, Somot S, Soussana J-F, Teichmann C, Valentini R, Vautard

- R, Weber B, Yiou P (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg Environ Change* 14(2):563–578. <https://doi.org/10.1007/s10113-013-0499-2>
4. Morrison A, Villarini G, Zhang W, Scoccimarro E (2019) Projected changes in extreme precipitation at sub-daily and daily time scales. *Global Planet Change* 182:103004. <https://doi.org/10.1016/j.gloplacha.2019.103004>
  5. Ban N, Schmidli J, Schär C (2014) Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. *J Geophys Res Atmos* 119(13):7889–7907. <https://doi.org/10.1002/2014JD021478>
  6. Dai A (2006) Precipitation characteristics in eighteen coupled climate models. *J Clim* 19(18):4605–4630
  7. Hohenegger C, Brockhaus P (2008) Towards climate simulations at cloud-resolving scales. *Meteorol Z* 17(4):383–394
  8. Palmer T (2014) Climate forecasting: build high-resolution global climate models. *Nature* 515(7527):338–339. <https://doi.org/10.1038/515338a>
  9. Prein AF, Langhans W, Fosser G, Ferrone A, Ban N, Goergen K, Keller M, Tölle M, Gutjahr O, Feser F, Brisson E, Kollet S, Schmidli J, van Lipzig NPM, Leung R (2015) A review on regional convection-permitting climate modeling: demonstrations, prospects, and challenges. *Rev Geophys* 53(2):323–361. <https://doi.org/10.1002/2014RG000475>
  10. Zhang GJ, Song X (2010) Convection parameterization, tropical Pacific double ITCZ, and upper-ocean biases in the NCAR CCSM3. Part II: Coupled feedback and the role of ocean heat transport. *J Clim* 23(3):800–812
  11. Coppola E, Sobolowski S, Pichelli E, Raffaele F, Ahrens B, Anders I, Ban N, Bastin S, Belda M, Belusic D, Caldas-Alvarez A, Cardoso RM, Davolio S, Dobler A, Fernandez J, Fita L, Fumiere Q, Giorgi F, Goergen K, Güttler I, Halenka T, Heinzeller D, Hodnebrog Ø, Jacob D, Kartsios S, Katragkou E, Kendon E, Khodayar S, Kunstmann H, Knist S, Lavín-Gullón A, Lind P, Lorenz T, Maraun D, Marelle L, van Meijgaard E, Milovac J, Myhre G, Panitz HJ, Piazza M, Raffa M, Raub T, Rockel B, Schär C, Sieck K, Soares PMM, Somot S, Srncic L, Stocchi P, Tölle MH, Truhetz H, Vautard R, de Vries H, Warrach-Sagi K (2020) A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Clim Dyn* 55(1):3–34. <https://doi.org/10.1007/s00382-018-4521-8>
  12. Lauwaet D, van Lipzig NPM, Van Weverberg K, De Ridder K, Goyens C (2012) The precipitation response to the desiccation of Lake Chad. *Q J R Meteorol Soc* 138(664):707–719. <https://doi.org/10.1002/qj.942>
  13. Prein AF, Gobiet A, Suklitsch M, Truhetz H, Awan NK, Keuler K, Georgievski G (2013) Added value of convection permitting seasonal simulations. *Clim Dyn* 41(9):2655–2677. <https://doi.org/10.1007/s00382-013-1744-6>
  14. Prein AF, Holland GJ, Rasmussen RM, Done J, Ikeda K, Clark MP, Liu CH (2013) Importance of regional climate model grid spacing for the simulation of heavy precipitation in the Colorado headwaters. *J Clim* 26(13):4848–4857
  15. Trusilova K, Früh B, Brienens S, Walter A, Masson V, Pigeon G, Becker P (2013) Implementation of an urban parameterization scheme into the regional climate model COSMO-CLM. *J Appl Meteorol Climatol* 52(10):2296–2311
  16. Zhang X, Zwiers FW, Li G, Wan H, Cannon AJ (2017) Complexity in estimating past and future extreme short-duration rainfall. *Nat Geosci* 10(4):255–259. <https://doi.org/10.1038/ngo2911>
  17. Diro GT, Sushama L (2019) Simulating Canadian Arctic climate at convection-permitting resolution. *Atmosphere* 10(8):430
  18. Côté J, Gravel S, Méthot A, Patoine A, Roch M, Staniforth A (1998) The operational CMC–MRB Global environmental multiscale (GEM) model. Part I: design considerations and formulation. *Monthly Weather Review* 126(6):1373–1395. [https://doi.org/10.1175/1520-0493\(1998\)126<1373:tocmge>2.0.co;2](https://doi.org/10.1175/1520-0493(1998)126<1373:tocmge>2.0.co;2)
  19. Teufel B, Sushama L (2022) High-resolution modelling of climatic hazards relevant for Canada's northern transportation sector. *Clim Dyn*. <https://doi.org/10.1007/s00382-022-06265-6>

20. Girard C, Plante A, Desgagné M, McTaggart-Cowan R, Côté J, Charron M, Gravel S, Lee V, Patoine A, Qaddouri A, Roch M, Spacek L, Tanguay M, Vaillancourt PA, Zadra A (2014) Staggered vertical discretization of the canadian environmental multiscale (GEM) Model using a coordinate of the log-hydrostatic-pressure type. *Mon Weather Rev* 142(3):1183–1196. <https://doi.org/10.1175/mwr-d-13-00255.1>
21. Benoit R, Cote J, Mailhot J (1989) Inclusion of a Tke boundary-layer parameterization in the Canadian regional finite-element model. *Mon Weather Rev* 117(8):1726–1750. [https://doi.org/10.1175/1520-0493\(1989\)117%3c1726:loatbl%3e2.0.Co;2](https://doi.org/10.1175/1520-0493(1989)117%3c1726:loatbl%3e2.0.Co;2)
22. Delage Y (1997) Parameterising sub-grid scale vertical transport in atmospheric models under statically stable conditions. *Bound-Layer Meteorol* 82(1):23–48. <https://doi.org/10.1023/A:1000132524077>
23. Li J, Barker HW (2005) A radiation algorithm with correlated-k distribution. Part I: local thermal equilibrium. *J Atmos Sci* 62(2):286–309. <https://doi.org/10.1175/Jas-3396.1>
24. Milbrandt JA, Yau MK (2005) A Multimoment bulk microphysics parameterization. Part I: analysis of the role of the spectral shape parameter. *J Atmos Sci* 62(9):3051–3064. <https://doi.org/10.1175/jas3534.1>
25. Belair S, Mailhot J, Girard C, Vaillancourt P (2005) Boundary layer and shallow cumulus clouds in a medium-range forecast of a large-scale weather system. *Mon Weather Rev* 133(7):1938–1960. <https://doi.org/10.1175/Mwr2958.1>
26. Kain JS, Fritsch JM (1990) A one-dimensional entraining detraining plume model and its application in convective parameterization. *J Atmos Sci* 47(23):2784–2802. [https://doi.org/10.1175/1520-0469\(1990\)047%3c2784:Aodepm%3e2.0.Co;2](https://doi.org/10.1175/1520-0469(1990)047%3c2784:Aodepm%3e2.0.Co;2)
27. Gerard L, Piriou J-M, Brožková R, Geleyn J-F, Banciu D (2009) Cloud and precipitation parameterization in a meso-gamma-scale operational weather prediction model. *Mon Weather Rev* 137(11):3960–3977
28. Mironov DV (2008) Parameterization of lakes in numerical weather prediction. Part 1: description of a lake model
29. Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, De Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, de Rosnay P, Rozum I, Vamborg F, Villaume S, Thépaut J-N (2020) The ERA5 global reanalysis. *Q J R Meteorol Soc* 146(730):1999–2049. <https://doi.org/10.1002/qj.3803>
30. Arora VK, Scinocca JF, Boer GJ, Christian JR, Denman KL, Flato GM, Kharin VV, Lee WG, Merryfield WJ (2011) Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys Res Lett* 38:L05805. <https://doi.org/10.1029/2010gl046270>
31. Riahi K, Rao S, Krey V, Cho C, Chirkov V, Fischer G, Kindermann G, Nakicenovic N, Rafaj P (2011) RCP8.5—a scenario of comparatively high greenhouse gas emissions. *Climatic Change* 109(1):33. <https://doi.org/10.1007/s10584-011-0149-y>
32. Thornton PE, Thornton MM, Mayer BW, Wei Y, Devarakonda R, Vose RS, Cook RB (2016) Daymet: daily surface weather data on a 1-km grid for North America, Version 3: ORNL distributed active archive center
33. Garnaud C, Sushama L, Verseghy D (2014) Impact of interactive vegetation phenology on the Canadian RCM simulated climate over North America. *Clim Dyn* 45(5–6):1471–1492. <https://doi.org/10.1007/s00382-014-2397-9>
34. Teufel B, Sushama L (2019) Abrupt changes across the Arctic permafrost region endanger northern development. *Nat Clim Chang* 9(11):858–862. <https://doi.org/10.1038/s41558-019-0614-6>
35. Dukhan T, Sushama L (2021) Understanding and modelling future wind-driven rain loads on building envelopes for Canada. *Build Environ* 196:107800. <https://doi.org/10.1016/j.buildenv.2021.107800>

36. Kim J, Murphy E, Nistor I, Ferguson S, Provan M (2021) Numerical analysis of storm surges on Canada's Western Arctic coastline. *J Marine Sci Eng* 9(3):326
37. Provan M, Ferguson S, Murphy E (2022) Storm surge contributions to flood hazards on Canada's Atlantic Coast. *J Flood Risk Manag* (in press)
38. Greenan B, James T, Loder J, Pepin P, Azetsu-Scott K, Ianson D, Hamme R, Gilbert D, Tremblay J, Wang X, Perrie W (2019) Changes in oceans surrounding Canada. In: Bush and Lemmen (eds.), *Canada's Changing Climate Report*, pp 343–423. Ottawa, Ontario: Government of Canada

# Development of Performance Index for Small and Medium-Sized Drinking Water Systems



Sarin Raj Pokhrel, Gyan Chhipi-Shrestha, Haroon Mian, Kasun Hewage, and Rehan Sadiq

**Abstract** Drinking water system (DWS) is an important public infrastructure that plays a significant role in attainment of a healthy lifestyle. To ensure that a DWS is functioning efficiently, the first step is to know the existing condition of a water system. A complete understanding of DWS is only possible when its current performance is both measured and compared against the similar water systems. Performance assessment becomes much more critical specifically in small and medium-sized water systems (SMWSs) because these systems constantly grapple with various constraints, including but are not limited to insufficient fund, limited staff, inadequate infrastructure, and absence of advanced water treatment. A questionnaire was distributed to the municipalities across the Okanagan Valley to identify the key performance indicators (KPIs). Thirty-nine KPIs were identified across six performance criteria: environment, finance, infrastructure, staff, and operation and monitoring, and social and institutional. Municipal representatives rated these KPIs based on relevancy, clarity, reliability, and comparability. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), a multicriteria decision-making method was employed to calculate the performance index. Analytical Hierarchical Process was used to calculate weights of performance criteria, whereas KPIs aggregation was based on an Ordered Weighted Average technique. In this study, performance of three small and medium-sized municipal water systems is measured. Necessary data corresponding to each criterion for all the KPIs were also collected. The results show that performance index of water system “X” was found to be the highest (62), followed by Y (33) and Z (21). Water system “X” had the highest aggregated score for operation and monitoring criterion, while “Y” and “Z” had the same score for environment criterion. The findings from the study indicated key areas that require municipal interventions to enhance drinking water system performance and ensure high quality life.

---

S. R. Pokhrel (✉) · G. Chhipi-Shrestha · H. Mian · K. Hewage · R. Sadiq  
School of Engineering, The University of British Columbia, Okanagan Campus, Kelowna,  
BC V1V1V7, Canada  
e-mail: [sarin.pokhrel@ubc.ca](mailto:sarin.pokhrel@ubc.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_78](https://doi.org/10.1007/978-3-031-34593-7_78)

1235

**Keywords** Drinking water · Key performance indicators · Performance index · Small and medium-sized water systems · Municipality

## 1 Introduction

Delivery of safe drinking water is important for sustainable development. Out of 17 United Nations (UN), Sustainable Development Goal (SDG), goal 6 aims to provide clean water and sanitation globally by 2030. According to a recent report released by the UN, progress of six key areas (drinking water, sanitation, water quality and treatment, water use efficiency and stress, integrated water resource management implementation, and freshwater ecosystems) surrounding goal 6 is termed as “off track,” which means that the attainment of sustainable water management is currently positioned in a precarious situation (UN [1]). Population growth, urbanization, industrialization, and climate change are considered as the primary factors that have led to unsustainable management of drinking water systems [2]. To make the matter worst, worldwide two billion people live in the water stress areas, highlighting the need of necessary actions in managing drinking water supplies [3].

Drinking water system (DWS) is an important public infrastructure that has a direct connection to human health and ecosystem. The definition of DWS is based on population size, number of connections, and volume of water flow per day. For example in British Columbia (BC), water systems that serve population less than 500 is small, where as in Quebec, the number ranged between 201 and 1000 to fulfill the same definition [4]. In this study, DWS that provides services to less than 5000 people is regarded as “small,” whereas medium-sized water system is any system providing services to 5000–30,000 individuals [5]. Compared to large water systems, small and medium-sized water systems (SMWSs) face severe challenges due to limited financial resources, inadequate advanced treatment systems, and insufficient staff [6]. Overcoming these challenges requires a comprehensive understanding of the existing DWSs condition such that necessary interventions can be applied to improve water system performance.

Performance is defined as a function of effectiveness (i.e., measures DWS capacity, quality, and delivery of service), reliability (measures DWS effectiveness over a period of time), and cost (measures DWS services are affordable by consumers) [7]. To measure DWS performance, key performance criteria are separated, which aims to assess water system’s performance under a specific objective (e.g., environmental, economical). Performance criteria are composed of performance indicators (PIs), which measures effectiveness and efficiency of water system’s ability and behavior [8]. After obtaining various qualitative and quantitative information through the calculation of PIs, a final score (involves weighting and aggregation process) is achieved, which is called performance index [9].

Different studies have been carried out to compute performance index of large water systems. Nevertheless, only a few authors have attempted to develop and analyze performance index for SMWSs. Parra [10] developed a priority index to



identify pipelines of a water distribution system based on three criteria (hydraulics, water quality, and vulnerability). The authors found that network performance based on quality and hydraulics of pipelines range between 60 and 100% [10]. Scheili [11] developed a drinking water quality index to calculate the spatiotemporal variability of drinking water quality and monitoring the type of contaminants present in the distribution systems based on four scenarios [11]. Bereskie [12] developed a continuous performance improvement framework based on water quality index to identify the underperforming indicators in small communities and recommended the best management strategies to enhance overall drinking water system performance [12].

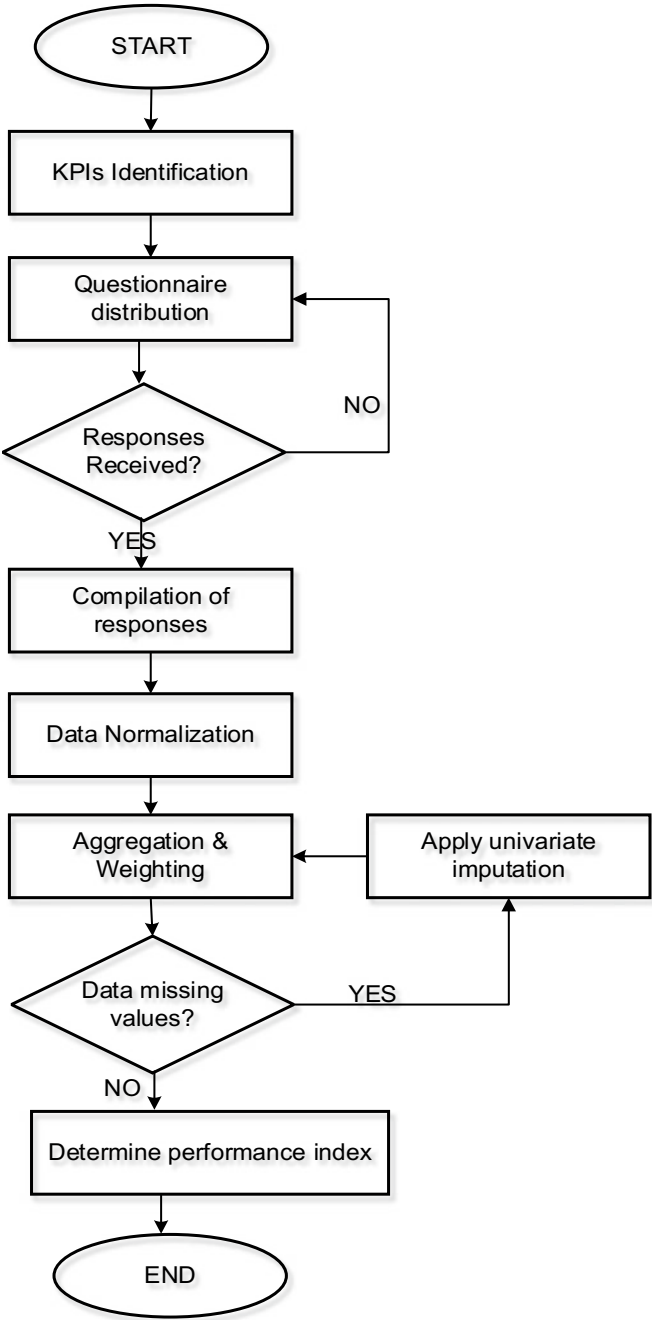
This study aims to develop performance index for three small and medium-sized DWSs. Developing performance index will require collection of relevant data related to each KPIs through a questionnaire. Various performance gap pertaining to different water systems are investigated through which, initial performance of each water system is determined by generating weights and aggregating scores. A performance index is developed, which can be used by the decision-makers and municipal managers to understand the current status of their DWS performance. The clear understanding of the water system's performance will also allow the top-level management to make necessary interventions in enhancing the overall system performance.

## 2 Methodology

In this study, performance of three drinking water systems is assessed. For this purpose, performance index is calculated by employing the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS is one of the most commonly used multicriteria and decision analysis methods. The wide application of this method is primarily attributed to three reasons: (a) It is simple and easy to understand; (b) it helps in robust decision-making as human alternatives are given to priority; and (c) it is based on the distance calculation and identifies positive and negative solutions [13]. The detailed methodology adopted for this study is shown in Fig. 1, which is also elucidated in the five main phases as listed in Sects. 2.1–2.5.

### 2.1 Identifying KPIs and Collecting Data

DWS is composed of various components that defines its effectiveness, efficiency, and functionality. The workability of these components is dependent on the KPIs. Thirty-nine KPIs have been identified across six criteria (environment, finance, infrastructure, staff, social, and operation and monitoring). The municipal representatives, including engineers, managers, technicians, and operators, are participated in the identification process. The detailed identification procedure of these KPIs is explained in another paper.



**Fig. 1** Methodology adopted in this study

Regarding data collection, an excel sheet comprising of all the KPs were distributed to municipalities in the Okanagan Valley as a questionnaire. Respondents were requested to provide the relevant data for the first three years prior to 2021. Data collection process started in April 2021 and completed in December 2021.

## 2.2 *Generating Criteria Weights*

Analytical Hierarchical Process (AHP) was used to develop a weighting scheme to provide weights for the six criteria. For this purpose, the above-mentioned respondents provided the ratings based on the importance of criteria. Each criterion was rated against all the criteria and were prioritized accordingly.

## 2.3 *Aggregating KPIs Values*

Prior to the aggregation, KPI values undergo normalization. These values are normalized between 0 and 1 and referred to as scores. These scores are then expressed in percentages. Ordered Weighted Average (OWA) was employed to aggregate the scores. In this method, another set of weights are assigned to each KPI and are normalized. The obtained scores of each KPI are organized in a descending order. These scores are multiplied with the corresponding weights, which provide the aggregated score for a criterion.

Next positive and negative ideal solutions (NIS) are determined. A positive ideal solution (PIS) is defined when the maximum score is obtained for a criterion, whereas NIS referred to as the minimum score received for a criterion. The role of PIS and NIS is dependent on the desire of outcome to be achieved. For example, if the desired outcome requires a lower value, it becomes PIS for that specific criterion and vice versa. Finally, Euclidean distance is computed, which calculates the distance of criteria from net PIS and NIS as outlined in Eqs 1 and 2.

$$A_i^+ = \sqrt{\sum_{q=1}^p (X_{ij} - X_j^+)^2} \quad (1)$$

$$A_i^- = \sqrt{\sum_{q=1}^p (X_{ij} - X_j^-)^2} \quad (2)$$

## 2.4 Computing Missing KPI Values

In this study, a univariate imputation known as penalty method is employed to fill the missing KPI values. In this method, the lowest KPI score for a specific criterion across the participating systems is inserted as the new value for the specific KPI. Assuming that utility Z do not have a score for the “water price,” and utility X has the lowest score for the corresponding KPI. In such case, water system Z score will be same as of the utility X.

## 2.5 Calculating Performance Index

In order to calculate performance index ( $P_i$ ) of each water system, performance value must be determined. Performance value ( $P_v$ ) is computed when the distance of the NIS is divided by the sum of the distances from PIS and NIS. The performance value multiplied with 100 provides performance index value for each water system. As  $P_i$  is a ratio, it does not have units.  $P_i$  value ranges between 0 and 100, where a higher value represents better performance of water system. Mathematically, the equations of performance value and index are represented as the following:

$$P_v = \frac{A_I^-}{A_I + A_i^-} \quad (3)$$

$$P_i = \frac{A_I^-}{A_I + A_i^-} * 100 \quad (4)$$

## 3 Results and Discussions

Three drinking water systems provided their data for the year 2020. Population of these systems vary between 5000 and 33,000. Although eight months were provided to water systems employee to send the research team a complete data set, there were a few KPI data values which went missing. Responses were provided via email, while for utility Z, the first and second authors visited the municipality to collect data points. Table 1 shows drinking water KPIs.

Performance index for each utility is represented in Fig. 2.

It is clear from the figure that X (62) has the highest performance followed by Y (33) and Z (21). Among all the criteria, environment criterion had the most consistency performance with the aggregated scores (Fig. 3) ranging between 53 and 61% for all three water systems.

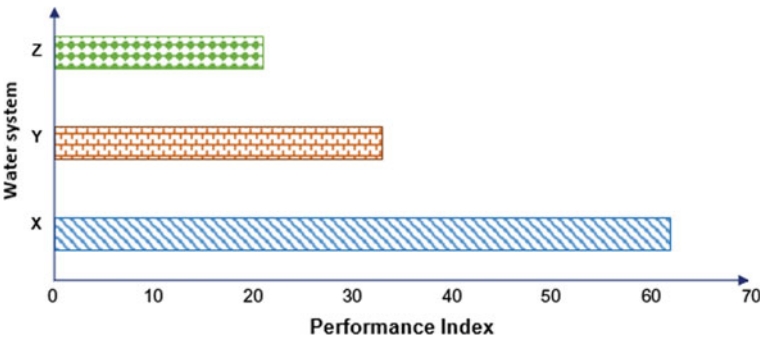
**Table 1** Drinking water KPIs

<b>Environment</b>
1. Residential water use (KPIE <sub>1</sub> )
2. Source water quality (KPIE <sub>2</sub> )
3. Annual water production (KPIE <sub>3</sub> )
4. Agricultural water use (KPIE <sub>4</sub> )
5. Annual agriculture water use per hectare (KPIE <sub>5</sub> )
6. Groundwater distributed (KPIE <sub>6</sub> )
7. Treatment energy intensity (KPIE <sub>7</sub> )
8. Industrial, commercial, and institutional (ICI) water use (KPIE <sub>8</sub> )
9. Treated surface water distributed (KPIE <sub>9</sub> )
10. Non-potable water distributed (KPIE <sub>10</sub> )
11. Maximum Daily Demand: Average daily demand (KPIE <sub>11</sub> )
12. Alternative source water (KPIE <sub>12</sub> )
<b>Finance</b>
1. Water price (KPIF <sub>1</sub> )
2. Operation and maintenance cost for raw water withdrawal from source to the treatment plant (KPIF <sub>2</sub> )
3. Operational and maintenance cost of water treatment (KPIF <sub>3</sub> )
4. Operation and maintenance cost for water distribution (KPIF <sub>4</sub> )
5. Debt service ratio (KPIF <sub>5</sub> )
6. Residential and ICI service revenue generated (agriculture revenue is excluded) (KPIF <sub>6</sub> )
<b>Infrastructure</b>
1. Total main length (KPII <sub>1</sub> )
2. Reported pipe failures (KPII <sub>2</sub> )
3. Water leakage (KPII <sub>3</sub> )
4. Total mains replaced/renewed/rehabilitated (KPII <sub>4</sub> )
5. New mains installed (KPII <sub>5</sub> )
6. Average Annual Life Cycle Investment (AALCI) (KPII <sub>6</sub> )
<b>Operation and monitoring</b>
1. Compliance with microbiological tests (KPIO <sub>1</sub> )
2. Compliance with physical and chemical tests (KPIO <sub>2</sub> )
3. Monitoring (Compliance) of THMs and HAAs (KPIO <sub>3</sub> )
4. Frequency of hydrant inspection (KPIO <sub>4</sub> )
5. Compliance to leak detection program (KPIO <sub>5</sub> )
<b>Staff</b>
1. Full-time operators (FTO) (KPIST <sub>1</sub> )

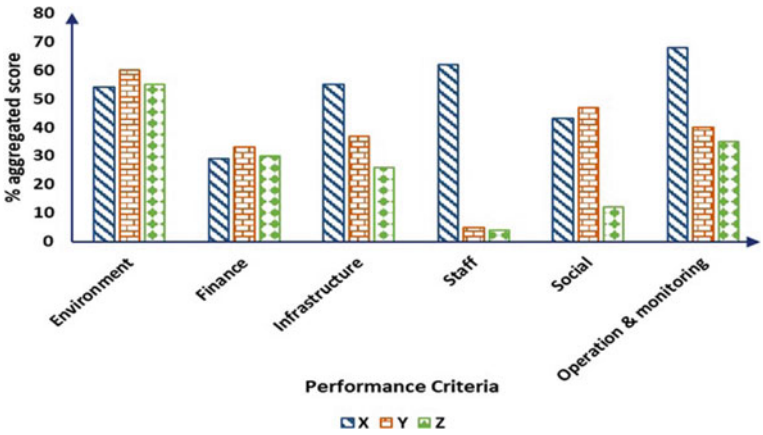
(continued)

**Table 1** (continued)

Environment
2. Distribution of the workforce by age (KPIST <sub>2</sub> )
3. Experience of operators (KPIST <sub>3</sub> )
4. Absentees due to accidents (KPIST <sub>4</sub> )
5. Overtime hours (KPIST <sub>5</sub> )
<b>Social and institutional</b>
1. Water quality complaints (KPIS <sub>1</sub> )
2. Growth within Service Area (KPIS <sub>2</sub> )
3. Boil water notice (BWN) (KPIS <sub>3</sub> )
4. Water quality advisory (WQA) (KPIS <sub>4</sub> )
5. Emergency response readiness (KPIS <sub>5</sub> )



**Fig. 2** Water systems performance index



**Fig. 3** Performance criteria aggregated scores

In terms of individual KPIs, water system Z has the highest (98%) score for the “annual water production,” whereas the score drops to 41 and 1% for the water systems X and Y. It indicates that utility Z has sufficient water availability in their system and are least likely to face water shortage issues in case the local government enforces any water restrictions (e.g., because of flood, heat waves, sudden contamination). Both systems X and Z distributed non-potable water to their consumers. Typically, agriculture demands, landscaping, and recreational activities (e.g., watering wineries, parks) are common non-potable water uses. None of the water systems provided information about their energy consumption for water treatment, which suggests that these systems might not have kept a track of their treatment energy efficiency. Energy consumption provides useful information related to total emissions produced by the treatment plants, baseline energy uses, and intensive energy uses applications (such as pumping and aeration for the water system).

In the USA, close to 2% of energy uses are consumed by drinking water and wastewater systems, contributing to the production of 45 million tons of greenhouse gases every year [14]. Only water system Y (94%) provided groundwater services to their customers, while remaining systems relied on surface water (X = 60 % and Z = 100%) for drinking purposes. Some consumers might have been dependent on private water wells to fulfill their drinking water consumptions.

The following sections describe performance of individual water systems in detail.

### 3.1 *Water System X*

Although water system X had the highest performance index value among three systems, two criteria [Finance (29%) and Social (43%)] low and mediocre aggregated scores impacted their overall performance. Residential and industrial, commercial, and institutional (ICI) revenues generated (KPI<sub>F7</sub>), agriculture water price (KPI<sub>F2</sub>), and operation and maintenance (O&M) cost for water distribution (KPI<sub>F5</sub>) are three KPIs that primarily resulted into a low financial criterion performance. KPI<sub>F2</sub> recorded the most expensive price for agriculture water uses across the systems. However, this result contradicts with its 100% distribution of non-potable water uses, which minimizes treatment and distributions costs. Having said that, the water system might have also supplied potable drinking water to its agriculture customers, while non-potable water consumption might have been limited to other uses. Similarly, lower revenue generated for residential and ICI uses suggest that residential and ICI customers might depend more on their private wells to fulfill their water demands. Excess levels of non-revenue water, high volume of unpaid bills, and unsystematic records of keeping database are other attributing factors that caused lower revenue generation [15].

Water quality complaints and introduction of water quality advisories (WQA) are two main factors responsible to low performance of water system X under social criterion. Nine and zero percent are the scores for these KPIs, respectively. In small water systems, drinking water complaints include but are not limited to taste and

odor of chlorine, poor water quality, aesthetic water quality, water color, and non-compliance with standard and guidelines [16]. It is clear that the water system had been under a WQA every single day in a year, leading them to 0% score. However, no boil water notices (BWNs) were issued during this period. WQA is the lowest degree of advisory that is administered in a water system, which indicates that the public health threat is minimum and can be overcome by mediums other than BWN and do not water use notice. Exposure of contaminants leading into long-term health risks, disturbance in water supplies due to maintenance work, and turbid water conditions are the primary reasons that allow public health officers to impose WQA restrictions in a water system [17].

One of the reasons behind water system X performing better than other systems must be attributed to its 68% aggregated score to O&M criterion. This criterion is related to compliance of various physical (turbidity and temperature), microbiological (*E.coli*, total coliform, residual chlorine), chemical [pH, trihalomethanes (THMs), haloacetic acids (HAAs)] parameters, and leak detection programs. 100% compliance record was achieved for *E.coli* and temperature parameters, while scores for residual chlorine and pH were 99.6 and 92%. Similarly, hydrants were inspected routinely that results to its 100% score. Since a majority of parameters have close to 100% compliance rate, it is fair to state that the overall water quality of the system is good. Timely monitoring schedules, immediate interventions in overcoming water quality issues and failures, and absence of BWNs are accountable factors responsible for the improved O&M performance.

### 3.2 Water System Y

Low aggregated scores for staff (5%), finance (33%), and infrastructure (37%) are key reasons responsible to the overall low performance index (33) for the water system Y. Only 36% score received by the full time operators, FTOs ( $KPI_{ST1}$ ) clearly suggests that the water system is heavily dependent on the part-time operators. In SMWSs, recruitment of FTOs is always a big challenge as these operators are shouldered with additional responsibilities other than their usual work load activities. In addition, operators are not sufficiently paid as the water system struggle with various issues, such as lower revenue generation, inadequate funding, and personal money being used in participating additional courses. This is also one of the reasons why operators in this system work plenty of overtime hours. These hours are related to on call duties, which might occur frequently in small systems due to shortage of operators, conduct multiple added responsibilities (e.g., prepare budgets and master plans for water systems) by a single person, and inconsistent schedule of part-time operators [16, 18].

Regarding the infrastructure criterion, replacement, renewable, and rehabilitation programs are not periodically conducted, which can be linked with their low performance scores that range between 16 and 50%. In addition, the utility also underwent with various reported pipe failure cases and water leakage issues. Pipe



failures are categorized into three types, pipe-intrinsic (corrosion, poor construction and mishandling, defects in joints) operational (pipe pressure fluctuation, blockages, and thermal shocks), and environmental (soil conductivity and moisture, differential settlement, and changing temperature). Leakage is a burning issue for many systems as it accounts to generation of non-revenue water, cross connections, and in extreme cases leading into flooding due to overflow of water. Leaks are usually observed in pipes, toilets, hot water tanks, and irrigation systems. For example, anywhere between 20 and 40 liters of water per hour is lost when toilets continuously run after flushing, contributing around 960 liters of water lost per day. Depending on the water price of a water system, there is a typical addition of around \$200 in the quarterly billing period [19].

### 3.3 Water System Z

Performance index value (21) is the lowest for utility Z. Excluding environment criterion (55%), all the criteria had the aggregated scores less than 50%. Staff (4%) and social (12%) were the two most underperforming criteria for this water system. Distribution of workforce by age ( $KPI_{ST2}$ — $KPI_{ST5}$ ) had the lowest scores, which indicates that the water system struggled to hire operators of varying age groups. In such scenario, newly hired operators will miss an opportunity to learn from their senior operators, which is one of the biggest loopholes for SMWS operators. Besides, 42% score for FTO ( $KPI_{ST1}$ ) also depict that similar to system Y, Z also found themselves in a difficult position to recruit FTOs. Absence of experienced operators ( $KPI_{ST6}$ ) was another cause for a low performance. Because of advanced treatment installations lacking in small systems, the operators' experience become more pivotal as they are responsible in maintaining adequate disinfectants concentration and ensuring proper water quality changes in treatment plants and distribution networks [20].

This water system experienced the highest BWN (15%) among three systems. When BWNs are issued in a community, it means that drinking such waters might put consumers under a serious public health threat. However, this issue can be resolved by boiling the water [17]. BWNs are administered because of drinking water contamination, limited treatment and disinfection, pipe failures and leakage, broken mains, and non-compliance of water quality parameters [21]. Although, the water system had emergency response readiness plans, these places have not yet been fully implemented. This means that in case if the water system undergoes some serious emergency situation, they will not have enough resources and experience to tackle the issues.

Drinking water price ( $KPI_{F1}$ ) had the lowest score (76%) among three systems, which means that it has the highest water price. Increased water price may be because the revenues generated from the water system is not enough to support regular O&M work. This statement is verified from the score of  $KPI_{F7}$  as the water system collects the least revenues among all the participating systems. Similarly, funds might have been excessively used to upgrade the drinking water infrastructure (e.g., fixing leaked

pipes, water mains). Debt service ratio is also maximum, which suggests that the water system experience losses and lower growth. In extreme cases, this situation might take the water system into bankruptcy, which will also prevent the system from welcoming newer water efficient programs [22].

## 4 Conclusion

Performance index is developed for three small and medium-sized water systems across six criteria, including 39 key performance indicators (KPIs). TOPSIS was the preferred multicriteria and decision-making method that was used to analyze the aggregated scores to define performance indices for each water system. Missing data KPI values were inserted by employing a univariate imputation known as penalty method. Data were collected via questionnaire, which also includes structured interviews and in-person meetings.

Overall performance index was found to be maximum for the water system X (62), followed by Y (33) and Z (21). Environment was the most consistent performance criterion in terms of relatively better performance for all the three systems with the aggregated scores 54% (X), 60% (Y), and 55% (Z). Water system X had the best water use both in terms of residential (summer and winter) and industrial, commercial, and institutional. Water quality at source and provision of alternate source water uplifted environmental performance for water system Y. Agriculture water use was minimum for the water system Z. On contrary, finance criterion had the lowest aggregate scores that range between 28 and 32% for all the three systems. Inadequate revenues generated from their customers to support operational and maintenance cost for water withdrawal, treatment, and distribution was the primary reason for such a low performance. Excluding water system X, both systems had the worst aggregated scores (5 and 4%) for staff criteria. This performance is mainly attributed to their availability of insufficient full time operators and appropriate distribution of workforce by age.

Development of the performance index will assist urban water managers to understand their existing performance and identify underperforming areas that requires decision-makers attention in enhancing overall DWS performance. Future studies must be focused in analyzing data for multiple years and comprehend the performance trend over time such that specific and useful interventions can be recommended to the systems.

**Acknowledgements** The authors would like to thank the municipal respondents in taking part in the survey. In addition, the authors would also like to thank the Natural Sciences and Engineering Research Council of Canada for providing the financial support to conduct this research.

## References

1. UN Water (2021) Progress on integrated water resources management
2. Pokhrel SR, Chhipi-Shrestha G, Hewage K, Sadiq R (2021) Sustainable, resilient, and reliable urban water systems: Making a case for “one water approach.” *Environ Rev.* <https://doi.org/10.1139/er-2020-0090>
3. WHO (2019) Almost 2 billion people depend on health care systems without basic water services. Retrieved February 28 2022. <https://www.who.int/news/item/14-12-2020-almost-2-billion-people-depend-on-health-care-systems-without-basic-water-services-who-unicef>
4. Pons W, Young I, Truong J, Jones-Bitton A, McEwen S, Pintar K, Papadopoulos A (2015) A systematic review of waterborne disease outbreaks associated with small non-community drinking water systems in Canada and the united states. *PLoS ONE* 10(10):e0141646–e0141646. <https://doi.org/10.1371/journal.pone.0141646>
5. Canadian Infrastructure Report Card 2019 (2019) Monitoring the State of Canada’s Core Public Infrastructure
6. Moffatt H, Struck S (2011) Water-borne disease outbreaks in Canadian small drinking water systems
7. NRC (1995) Committee on measuring and improving infrastructure performance
8. Alegre H (1999) Performance indicators for water supply systems: current trends and on-going projects. Drought management planning in water supply systems, pp 148–178. Springer Netherlands. [https://doi.org/10.1007/978-94-017-1297-2\\_7](https://doi.org/10.1007/978-94-017-1297-2_7)
9. Haider H, Sadiq R, Tesfamariam S (2016) Inter-utility performance benchmarking model for small-to-medium-sized water systems: aggregated performance indices. *J Water Resour Plan Manage* 142(1):4015039. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000552](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000552)
10. Parra S, Krause S, Angermair G (2018) Recommendations for the adaptation planning of water distribution systems. In: WDSA/CCWI Joint Conference Proceedings vol 1
11. Scheili A, Rodriguez MJ, Sadiq R (2015) Development, application, and sensitivity analysis of a water quality index for drinking water management in small systems. *Environ Monit Assess* 187(11):1–15. <https://doi.org/10.1007/s10661-015-4908-5>
12. Bereskie T, Haider H, Rodriguez MJ, Sadiq R (2017) Framework for continuous performance improvement in small drinking water systems. *Sci Total Environ* 574:1405–1414. <https://doi.org/10.1016/j.scitotenv.2016.08.067>
13. Vega A, Aguarón J, García-Alcaraz J, Moreno-Jiménez JM (2014) Notes on dependent attributes in TOPSIS. *Procedia Comput Sci* 31:308–317
14. USEPA (2021) Energy efficiency for water systems. Retrieved Feb 26 2022. <https://www.epa.gov/sustainable-water-infrastructure/energy-efficiency-water-systems>
15. Libey A, Adank M, Thomas E (2020) Who pays for water? comparing life cycle costs of water services among several low, medium and high-income systems. *World Dev* 136:105155. <https://doi.org/10.1016/j.worlddev.2020.105155>
16. Pokhrel SR, Chhipi-Shrestha G, Rodriguez MJ, Hewage K, Sadiq R (2020) Unfolding “big” problems of small water system performance: a qualitative study in British Columbia. *Canadian Water Resour J* 45(4):269–286. <https://doi.org/10.1080/07011784.2020.1800517>
17. Provincial Government of British Columbia (2021) Water Quality Notifications. Retrieved February 26 2022. <https://www2.gov.bc.ca/gov/content/environment/air-land-water/water/water-quality/drinking-water-quality/notices-boil-water-advisories>
18. Kot M, Castleden H, Gagnon GA (2011) Unintended consequences of regulating drinking water in rural Canadian communities: examples from Atlantic Canada. *Health Place* 17(5):1030–1037. <https://doi.org/10.1016/j.healthplace.2011.06.012>
19. City of Surrey (2022) Detecting Water Leaks. Retrieved February 26 2022. <https://www.surrey.ca/services-payments/property-payment-services/utility-billing-services/detecting-water-leaks>
20. Scheili A, Rodriguez MJ, Sadiq R (2016) Impact of human operational factors on drinking water quality in small systems: an exploratory analysis. *J Clean Prod* 133:681–690. <https://doi.org/10.1016/j.jclepro.2016.05.179>

21. Lui E (2015) On notice for a drinking water crisis in Canada.
22. Santos E, Lisboa I, Eugénio T (2021) Economic sustainability in wastewater treatment companies: a regional analysis for the Iberian Peninsula. *Appl Sci* 11(21):9876. <https://doi.org/10.3390/app11219876>

# Improved NASM Framework for Food Processing Wash-Water and Solid Residuals



Richard G. Zytner, Connor Dunlop, and Bassim Abbassi

**Abstract** Research was conducted on non-agriculture source material (NASM) data and legislation to identify knowledge gaps and limitations associated with the regulatory framework within Ontario. The framework for NASM land application on farm fields is set out in Ontario Regulation 267/03 under the (Nutrient Management Act, 2002, S. O. 2002, Chap. 4. <https://www.ontario.ca/laws/statute/02n04>). Currently, Category 2 NASM (including waste and wash-water from processed fruits and vegetables) and Category 3 NASM (including dairies, abattoirs, and sewage biosolid facilities) have stringent requirements that agri-food processors must follow when looking into land application disposal. Agri-food processors have suggested that the regulations are overly burdensome for smaller facilities due to the cost for ongoing sampling, documentation, and reporting requirements, which hinders their economic expansion. Following the requirements set out in Ontario Regulation 267/03, agri-food processors have identified that their food processing wash-water samples rarely exceed the regulatory limits for heavy metals under Ontario Regulation 267/03. Thus, processors are questioning why they need to continue analyzing for these constituents as this is a repetitive cost to their business, and if any improvements can be made to the regulatory framework to make it more economically feasible without harming the environment. Overall, this project looked at identifying possible improvements to the Ontario Regulation 267/03 framework to assist agri-food processors with cost-effective land application, with a focus on heavy metal sampling requirements. In the first phase of the project, data was collected and summarized from wash-water facilities to characterize the NASM they produced. The NASM data from various agri-food facilities showed that most of the samples were routinely below the lowest threshold for concentration (CM1 level) and always below CM2 levels, a higher concentration but still within the allowable regulatory limits. Next, a thorough jurisdiction scan of North American standards gave insight into possible improvements to Ontario's land application framework. Findings show that Ontario has established one of the best systems for land application to protect the public and the environment. However, it is evident that this regulatory framework is overly burdensome for non-variable food

---

R. G. Zytner (✉) · C. Dunlop · B. Abbassi  
School of Engineering, University of Guelph, Guelph, ON, Canada  
e-mail: [ryztner@uoguelph.ca](mailto:ryztner@uoguelph.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_79](https://doi.org/10.1007/978-3-031-34593-7_79)

1249

processing sources and hinders the feasibility of diverting this beneficial organic matter in the form of NASM from landfills. Recommended improvements include decreased sampling frequencies for heavy metals if there is an evidence of low metal occurrence and increased application timeframes if the weather permits.

**Keywords** Food processing wash-water • Solid residuals • Framework

## 1 Background

Wastewaters and solid residuals that are produced from agricultural operations can contain high levels of organics, nutrients, pathogens, and many other constituents. These parameters can cause harm to the natural environment if improperly managed. To help the agriculture sector better account for their generated wastes, the Ontario government enacted the Nutrient Management Act (NMA) in 2002. The overall purpose of the NMA is stated as “providing for the management of materials containing nutrients in ways that will enhance the protection of the natural environment and provide a sustainable future for agriculture operations and rural development” [1]. The regulatory framework outlines various requirements that must be followed based on the specific agriculture operation to protect Ontario’s environment. The NMA covers many areas including inspections, management of materials and enforcement. This project focuses on the land application of wastewaters and solid residuals from agri-food operations for final disposal within Ontario.

The land application of wastewaters and solid residuals from agri-food operations has shown to provide many benefits to the soil and the environment if managed properly [2]. The high levels of nutrients, such as nitrogen and phosphorus in agri-food wastewater, are essential for crop growth. Additionally, the proper application of agri-food wastewaters and solid residuals has been shown to improve the health of agriculture soils by allowing a build-up of soil organic matter (SOM), which is essential for maintaining soil productivity [2]. Furthermore, the land application of the generated wastes for final disposal will divert this usable material in the form of SOM away from unnecessary landfilling. However, proper practices and standards need to be followed to ensure safe application of the generated wastes, as it can also contain high levels of pathogens and heavy metals.

Under the NMA, Ontario Regulation 267/03 (O. Reg. 267/03) outlines all of the requirements that the various sources of materials have to meet prior, during, and post land application. These regulatory constraints help mitigate any negative environmental and human health risks from the land application of these wastes. Under O. Reg. 267/03, wastes such as treated and recycled leaf and yard wastes, fruit and vegetable processing wash-waters, or sewage biosolids that are being land applied to provide a beneficial use are defined as non-agriculture source materials (NASM) [2]. O. Reg. 267/03 classifies the different sources of materials into three NASM categories as given in Table 1, and a full list of materials from each category can be found in Schedule 4 of O. Reg. 267/03.

**Table 1** Example materials in NASM categories obtained from O. Reg. 267/03

NASM category 1	NASM category 2	NASM category 3
Leaf and yard waste that has not been composted	Leaf and yard waste that has been composted, but does not meet requirements for category AA or A in part II of the compost standards	Wash-water, including materials containing food-grade cleaners, from meat, egg, or dairy processing facilities
Organic waste matter derived from the drying, cleaning and processing of field and nut crops	Organic waste matter that contains no meat or fish and is derived from food processing at sources such as a bakery or a brewery	Organic waste matter from the processing of fish
Manure from non-farm herbivorous animals, including associated bedding materials	Fruit and vegetable processing water that contains no chemicals other than food-grade chemicals	Sewage biosolids, or any other material, other than untreated septage, that contains human body waste

The constraints for land application increase from Category 1 to Category 3 NASM, as these latter categories can potentially contain higher levels of regulated constituents. Currently, agri-food processors falling into Category 2 and Category 3 NASM have stringent requirements they must follow to ensure safe application of their generated material within Ontario. Under O. Reg. 267/03, processors need to sample for a variety of parameters including nutrients, heavy metals, and pathogens to ensure they are below regulatory limits. Based on the completed analysis, processors then have different constraints to follow for their specific NASM material during land application to ensure protection of the environment such as greater setback distances to residential areas or surface waters.

For heavy metal sampling, the regulatory framework within Ontario requires the analysis of 11 individual heavy metals for any material falling into Category 2 or Category 3 NASM. After analysis, the NASM will then be categorized as either “CM1” or “CM2” based on the concentrations of regulated metals. When used in reference to NASM, “CM1” means that the “content of a regulated metal does not exceed the concentrations set out in Column 2 or 3 of Table 1 of Schedule 5 within O. Reg. 267/03. If a content of a regulated metal does exceed the CM1 NASM but not the concentration set out in Column 2 or 3 of Table 2 of Schedule 5 of O. Reg. 267/03, then it will be defined as “CM2.” The CM1 and CM2 limits for aqueous NASM have been summarized in Table 2.

As given in Table 2, there are two different standards prior to land application for heavy metals within the generated NASM, with CM1 being the lower limits. These limits are based on toxicological calculations and studies for uptake in different plants and animals. If a generated NASM is sampled and found to exceed any of the concentration limits for CM2, the material cannot be land applied within Ontario unless approved by the Director (a person responsible for the administration of a provision of NMA). If the CM1 limits are exceeded for any heavy metal, the NASM will need to be applied at a lower rate and appropriate setback distances will need to

**Table 2** CM1 and CM2 regulated metal concentration limits for NASM obtained from O. Reg. 267/03

Regulated heavy metal under O. Reg. 267/03	CM1 concentration limit in aqueous material (mg/L)	CM2 concentration limit in aqueous material (mg/L)
Arsenic	0.13	1.7
Cadmium	0.03	0.34
Cobalt	0.34	3.4
Chromium	2.1	28.0
Copper	1.0	17.0
Lead	1.5	11.0
Mercury	0.008	0.11
Molybdenum	0.05	0.94
Nickel	0.62	4.2
Selenium	0.02	0.34
Zinc	5.0	42.0

be followed. Additionally, O. Reg. 267/03 has concentration limits for non-aqueous material, which contains 1% or more total solids, in milligrams per kilogram of total solids dry weight for both CM1 and CM2 standards. These values are simply a multiple of 100 times greater than the values shown for aqueous standards. For example, the CM2 limit for zinc in dry material would be 4200 mg/kg.

Furthermore, under O. Reg. 267/03, there are maximum addition limits to the soil outlined for each heavy metal that cannot be exceeded for any 5-year period to ensure protection for Ontario soils. No person shall then apply NASM to the land at a rate that surpasses this maximum addition to the soil. The calculations to determine an appropriate application rate will then be based on the completed metal analysis and the features that are at the site location. The maximum limit for addition to the soil for a 5-year period for each regulated heavy metal under O. Reg. 267/03 are summarized in Table 3.

Additionally, for pathogen testing, Category 2 NASM is assumed to be “CP1.” Therefore, no sampling or analysis for pathogens is required. When used in reference to NASM, CP1 means that the “content of pathogens named in Column 1 of Tables 1 or 2 of Schedule 6 within O. Reg. 267/03 do not exceed the concentrations that are outlined” (O. Reg. 267/03, 2003). Category 3 NASM, except for sewage biosolids and other material containing human body waste, will be initially assumed as a “CP2” pathogen standard. Sampling and analysis can be completed if the generator wishes to determine pathogen levels to confirm it meets CP1 standards. Sources of NASM from sewage biosolids or other materials containing human body wastes have further constraints, and additional sampling and analysis can be performed to determine if it meets CP1 standards.



**Table 3** Maximum addition to soil for 5-year period for regulated metals obtained from O. Reg. 267/03

Regulated heavy metal under O. Reg. 267/03	Maximum addition to soil (kg/ha/5-years)
Arsenic	1.4
Cadmium	0.27
Cobalt	2.7
Chromium	23.3
Copper	13.6
Lead	9.0
Mercury	0.09
Molybdenum	0.8
Nickel	3.56
Selenium	0.27
Zinc	33

NASM can also be required to test for sodium, fats, oils, and grease (FOG) and boron under the O. Reg. 267/03. These additional parameters are regulated as repeated applications on the same field of NASM with high levels of these compounds could harm the long-term productivity of the field. The concentration of pathogens, heavy metals, and other regulated parameters that the NASM contains will then lead to different land application constraints as mentioned previously. For example, there are different setback standards to sensitive features such as depth to bedrock or distance to surface waters based on the content of pathogens and heavy metals. All of these requirements for land application based on the NASM can be found in Part VI, Land Application Standards, within O. Reg. 267/03.

Overall, O. Reg. 267/03 is a very effective regulatory framework that manages the land application of different sources of materials to ensure protection for the environment and the public. Additionally, the Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) has developed software that can be used to assist the agri-food processor when developing their required paperwork. Furthermore, amendments have continuously been made to the land application framework to help improve the process for NASM generators and applicators. However, after following the requirements set out in O. Reg. 267/03, agri-food processors falling into Category 2 and Category 3 believe that the regulations are still overly burdensome and areas can be further refined to increase their economic competitiveness, while still protecting public interest.

Presently, agri-food processors falling into Category 2 and Category 3 NASM have stringent requirements that they must follow to ensure safe application of their material. Processors are required to sample for a variety of parameters including nutrients, heavy metals, and pathogens. As such, these processors need to continually sample for the required 11 heavy metals, which is a repetitive cost to their facility. Although processors understand the need to ensure environmental due diligence for land application, after continually analyzing for heavy metals within their NASM,

they have found little evidence occurring. As such, processors are questioning the need for ongoing metal analysis and if any improvements can be made to the land application framework. Additionally, processors have discussed issues associated with the current land application timeframe, as increasing temperatures may allow for longer application timeframes prior to the ground freezing.

Currently, the Ontario government is committed to reducing the regulatory burden for all businesses to help improve Ontario's economic competitiveness. The agri-food sector is no different. As a result, previous changes have already been made to the NMA and O. Reg. 267/03. For example, amendments re-categorized lower risk manures from non-farm herbivorous animals such as zebras from Category 3 to Category 1 NASM [3]. Amendments were also made to remove the automatic cessation to the NMA after five years, decreasing repetitive paperwork [3]. The total expected benefits from these changes are expected to provide cost savings as well as additional business opportunities totaling an average of \$1.2 million per year, for all phased-in-farms [3].

For Ontario to continue its prosperous growth within the agri-food sector, the industry needs to continue to adopt sustainable and economically viable resource management or food-based wash-waters and solid residuals that are being generated. Therefore, this project was initiated through OMAFRA's Special Initiative Project (SIP) program, which allows funding for projects to be completed that respond to specific ministry priorities. For this project, emphasis was on looking at the heavy metal sampling requirements under O. Reg. 267/03. Additionally, time was spent investigating the benefit and risks of adding NASM to soil, to help improve the long-term soil health in Ontario.

## 2 Approach

A detailed approach was created prior to the start of this SIP to determine how the data would be obtained and how the research objectives would be accomplished. The overall scope of the SIP was on highlighting areas in O. Reg. 267/03, where regulatory burden could be reduced and regulatory oversight is still required. The initial focus was on heavy metal requirements; however, a full review of O. Reg. 267/03 was conducted to determine if any other improvements could be made. To complete this goal, the project was broken into three phases.

First, a thorough review was conducted on available OMAFRA data and peer-reviewed literature to identify any trends or limitations within any previously conducted studies. For this step, available NASM data was organized and evaluated from past projects completed on wash-water and residual management by OMAFRA. Additionally, a comprehensive review was completed on scientific studies of wash-water treatment and final disposal for the food processing industry to identify any other limitations within the available literature.

Next, a jurisdictional scan was completed to understand how food processing wash-waters and solid residuals are being regulated in other provinces and US states within North America. Completing this scan would allow Ontario to evaluate where it stands and determine if there were any other improvements that could be made to the framework based on these other regulatory systems. The initial focus was on how other jurisdictions were handling sampling and analysis procedures for heavy metals. However, additional regulatory parameters such as groundwater setback distances, sodium limits, and SOM requirements were also investigated to determine if any other improvements could be highlighted. To obtain the data, regulations and documentation associated with land application were searched online. A survey was also created and sent to various jurisdictions to obtain additional information on their regulatory framework if any contacts could be made.

Finally, a life cycle assessment (LCA) study was completed on a representative NASM land application process to quantify the environmental impacts associated with NASM disposal to help support the development of any improvement BMPs to Ontario's framework. There have been a limited number of LCA studies within the agri-food area that focus on land application of NASM for final disposal. Therefore, this LCA study was initiated to highlight the environmental impacts associated with NASM land application. The data obtained from the previous OMAFRA projects was used as the initial data set for the LCA and was supplemented by additional information collected from literature. Furthermore, the LCA was carried out under the guidelines of the International Organization for Standardization (ISO) 14,040.

Results from the OMAFRA and literature data evaluation and the jurisdiction scan will be highlighted within the next section in this paper. Additionally, potential improvements to Ontario's land application framework will be discussed. However, LCA results will not be shown in this paper, as they are not part of the scope for this conference. Although, all results from the project can be found within the thesis document through the University of Guelph's online library.

### **3 Results and Discussion**

#### ***3.1 Results from OMAFRA Wash-Water Evaluation***

The first step of the project after completing a review of O. Reg. 267/03 was focused on examining previous project and literature data on wash-water and solid residuals to identify any gaps and limitations within the samples and peer-reviewed studies. For this step, data was made available from OMAFRA projects that were conducted on agri-food wash-water operations. The data for these projects was sampled from a variety of agri-food facilities including fruits, vegetables, abattoirs, cheese, and dairy.

The wash-water data obtained from these projects was initially scattered in different reports and files, so it first had to be organized for samples that could be used for this project. The wash-water data was collected from various sampling locations through the facilities, before and after different treatment processes. The wash-water data was then sorted according to nutrient parameters such as biochemical oxygen demand (BOD) or total phosphorus, to try to determine optimal treatment processes based on treatment effectiveness. However, some samples were also analyzed for heavy metals, but these samples were limited which was assumed to be due to cost. Therefore, the wash-water data had to be sorted for samples after treatment processes, as this would be the material that would be land applied, and samples that contained heavy metal data.

With the heavy metal wash-water data set organized, the next step was to analyze for any trends or limitations within the samples. The focus on this step was comparing heavy metal requirements, as all these sources of NASM must test for 11 regulated heavy metals as they fall into Category 2 and Category 3 NASM. A summary of the wash-water samples that were collected are given in Table 4 and are compared to the requirements for CM1 and CM2 outlined in O. Reg. 267/03.

Based on the low occurrence of heavy metals within the wash-water samples, it is no surprise that agri-food processors have questioned the need for ongoing metal analysis. The repeated sampling incurs additional costs for the operator and can hinder economic expansion for smaller wash-water operations. Additionally, if landfilling is more cost efficient, this may be the option that the operation uses for final disposal, which reduces the application of beneficial SOM on Ontario soils. Therefore, based on the evidence shown here, Ontario may want to review their regulatory sampling frequencies for consistent agri-food processing material, if there is evidence that there is low heavy metal occurrence. However, it is also noted that additional heavy metal sampling should be completed first, as the current small data set is a limitation for these results. A limited number of samples could be obtained due to confidentiality requirements at facilities.

Table 4 shown above contains all Category 2 NASM material. Category 3 NASM then includes sewage biosolids and some agri-food processing wash-waters such as materials originating from dairy, egg, and meat processing facilities. When comparing food processing wash-water-based NASM to sewage biosolid sources of NASM, the influent material contributing to the sewage biosolid NASM will fluctuate based on industries that are contributing to the municipal wastewater treatment plant. As such, evidence was found through literature where sewage biosolid samples exceeded the standards for CM2 within Ontario.

Puddephatt [4] summarizes biosolid samples obtained from the Kitchener and Guelph wastewater treatment plants within Ontario for heavy metals and additional nutrients. This thesis study was conducted to determine the impacts of land applying biosolid materials to agricultural lands using relevant terrestrial biota. Within these samples, there was evidence that the material exceeded the CM2 standards for land application. Therefore, this material would not be able to be applied based on Ontario's regulatory standards, unless approved by the Director. As mentioned in the study, the content of heavy metals within the biosolids will depend on what the

**Table 4** NASM heavy metal data compared to CM1 and CM2 standards

NASM commodity	Total # of samples	# of samples > CM1 As value	# of samples > CM1 Cd value	# of samples > CM1 Cr value	# of samples > CM1 Co value	# of samples > CM1 Cu value
Apples	11	0	0	1	0	0
Carrots	22	0	0	0	0	0
Potatoes	26	0	2	0	0	0
Cheese	45	0	0	0	1	0
# of Samples > CM1 Standard	104	0	2	1	1	0
# of Samples > CM2 Standard	104	0	0	0	0	0

NASM commodity	Total # of samples	# of samples > CM1 Hg value	# of samples > CM1 Mo value	# of samples > CM1 Ni value	# of samples > CM1 Pb value	# of samples > CM1 Se value	# of samples > CM1 Zn value
Apples	11	0	0	1	0	0	0
Carrots	22	0	0	0	0	0	0
Potatoes	26	0	2	0	0	0	0
Cheese	45	1	0	0	0	0	0
# of samples > CM1 standard	104	1	2	1	0	0	0
# of samples > CM2 standard	104	0	0	0	0	0	0

industrial, commercial, and residential activities are in the surrounding community. Therefore, appropriate sampling protocols must be followed for sewage biosolids to ensure protection for the environment and public, as there is strong evidence of elevated levels of heavy metals.

However, compared to an agri-food wash-water facility, the influent material here is relatively constant, unless a completely new product is being produced at the facility. Therefore, one major recommendation based on the evidence shown here could be to create an additional Category 4 NASM for sewage biosolid sources, where the sampling and analysis protocols would be similar to the current framework. Although for agri-food facilities, the sampling and analysis frequency could potentially be decreased for heavy metals, if there is strong evidence of low heavy metal occurrence after conducting more samples. For example, it was found through the jurisdiction scan that Quebec separates agri-food biosolids from food processing sources and omits all heavy metal sampling from agri-food sources [5]. However, not all sampling should be reduced, to help ensure environmental due diligence within Ontario.

With the results summarized from the evaluation of heavy metal data, the next step of the project was to complete a review of land application regulatory frameworks across North America. This review would help determine evaluate where Ontario stood compared to other provinces and states, and if any improvements could be highlighted for the land application framework. The results from the North American jurisdiction scan are discussed in the following section.

### ***3.2 Results from North American Jurisdiction Scan***

Overall, the results from the North American evaluation revealed that Ontario was one of the most advanced jurisdictions for regulations associated with the land application of various materials. For all the jurisdictions that were evaluated, Ontario had one of the most well-documented and easiest to access land application frameworks used to protect the public and the environment. Additionally, Ontario has established standards for many land application criteria such as sodium, SOM, and FOG limits, where other jurisdictions have not created any. However, improvements were still identified that could be made to Ontario's framework based on the jurisdiction evaluation.

For example, a study was found that was completed by the Canadian Council of Ministers of the Environment (CCME) [6] on biosolid application regulations. From this study, it was identified that Ontario's CM2 limits for land application are some of the highest allowable levels of concentrations for each heavy metal. Although Ontario's CM1 limits are similar standards compared to other provinces, the land application of NASM in Ontario is still allowed up to the limits outlined for CM2, and these are some of the highest concentrations within Canada. The regulated heavy metal concentrations for Ontario and other provinces have been summarized in Table 5.

**Table 5** Regulated metal concentrations for various jurisdictions [6]

Jurisdiction	Allowable concentration in dry matter for heavy metals (mg/kg)										
	As	Cd	Co	Cr	Cu	Pb	Hg	Mo	Ni	Se	Zn
Ontario (CM1 limits)	75	3	34	210	100	150	0.8	5	62	2	500
Ontario (CM2 limits)	170	34	340	2800	1700	1100	11	94	420	34	4200
British Columbia (class A)	13	3	34	100	400	150	2	5	62	2	500
British Columbia (class B)	75	20	150	1060	2200	500	15	20	180	14	1850
Alberta	Metal limits are based on nitrogen to phosphorus ratio										
Saskatchewan	75	20	150	1060	760	180	5	20	180	14	1850
Manitoba	Metal limits are based on cumulative weight per hectare										
Quebec (category C1)	13	3	34	210	400	150	0.8	5	62	2	700
New Brunswick	13	3	34	210	400	150	0.8	5	62	2	700
PEI (EQ)	41	39	–	1200	1500	300	17	–	420	100	2800
PEI (A and B)	75	85	–	–	4300	840	57	75	420	100	7500
Nova Scotia (class A)	13	3	34	210	400	150	0.8	5	62	2	700
Nova Scotia (class B)	75	20	150	1060	760	500	5	20	180	14	1850
Newfoundland	41	39	–	–	1500	300	17	75	420	100	2800

Therefore, Table 5 above shows evidence that Ontario has some of the higher allowable limits for heavy metals within NASM for land application compared to other provinces within Canada. However, it is noted that if the NASM exceeds the CM1 standards, application rates will be decreased to a calculated rate to mitigate any risks to the environment and the public. Therefore, a simple heavy metal comparison between regulatory limits can be difficult due to additional constraints that can be used such as application rates. However, this summary does provide evidence that Ontario could establish lower heavy metal limits for NASM sources, to maintain protection for the environment and the public.

The jurisdiction scan within the USA then further revealed that Ontario had one of the best-documented systems for land application in North America. It was found that within the USA, the land application of biosolids from municipal wastewater treatment plants will be regulated under the United States Environmental Protection Agency (USEPA) Part 503 Biosolids Rule. The standards in Part 503 developed by the USEPA outline all of the requirements for the land application of biosolids, similar to O. Reg. 267/03. However, the individual states can then modify their regulatory framework for the land application of other sources of materials, such as for agri-food sources. Although, it was found that most states will follow the standards outlined within the USEPA Part 503 Biosolids Rule. The heavy metal standards have also been outlined in USEPA Part 503, and these limits are summarized in Table 6 and compared to Ontario's CM2 limits.

As given in Table 6, Ontario's CM2 limits are higher for Arsenic, Molybdenum, and Lead when compared to the concentrations outlined in the USEPA framework. However, Ontario's standards are lower for all the other heavy metals, excluding

**Table 6** CM2 concentration limits in Ontario compared to USEPA part 503 limits for biosolids

Regulated heavy metal	Maximum concentration in NASM for CM2 limits (mg/kg)	Maximum concentration in biosolids under USEPA Part 503.13 (mg/kg)
Arsenic	170	75
Cadmium	34	85
Cobalt	340	–
Chromium	2800	3000
Copper	1700	4300
Lead	1100	840
Mercury	11	57
Molybdenum	94	75
Nickel	420	420
Selenium	34	100
Zinc	4200	7500

Nickel. Furthermore, similar to Ontario, the maximum application rate for land application within the USA for municipal biosolids will then be restricted to not exceed the outlined concentrations added to soil. Although, for the USEPA framework, the outlined concentrations are provided on a per-year basis. As such, to compare to the USEPA standards, Ontario's limits were calculated for annual rates based on the 5-year maximums and are summarized in Table 7.

Thus, as given in Table 7 above, Ontario has also established much lower annual loading rates for all heavy metals when compared to the USEPA framework. As such,

**Table 7** Annual loading rates for heavy metals in Ontario compared to USEPA Part 503 loading rates

Regulated heavy metal	Ontario annual loading rates (kg/ha/year)	USEPA part 503 annual loading rates (kg/ha/year)
Arsenic	0.28	2.0
Cadmium	0.05	1.9
Cobalt	0.54	–
Chromium	4.66	–
Copper	2.72	75
Lead	1.8	15
Mercury	0.018	0.85
Molybdenum	0.16	–
Nickel	0.72	21
Selenium	0.054	5.0
Zinc	6.6	140



the USA may want to review their limits for annual heavy metal additions and create more conservative standards. However, it was also found through the jurisdiction scan that the states can also implement flexible land application frameworks through National Pollutant Discharge Elimination System (NPDES) documentation or other regulatory frameworks, for other sources than municipal sewage biosolids.

Overall, it was found that most states will follow the USEPA standards for land application, including for food processing wash-waters, but some states have enacted other systems for land application that can be specific to the agriculture operation. For example, the state of Nebraska uses several different regulations and writes conditions as part of each facility's NPDES. Additionally, the state of Kentucky has established landfarming and composting of special wastes under their 401 KAR 45:100 framework. Here, wastes will need to be analyzed for only five heavy metals, which include Cadmium, Copper, Zinc, Lead, and Nickel.

Finally, when investigating additional regulatory parameters for land application that included application criteria such as sodium, SOM, FOG and groundwater setback limits, it was difficult to find any information for these parameters. It was found that some provinces and states will monitor for groundwater contamination and for sodium limits; however, most jurisdictions have not established any set restrictions. As such, Ontario was found to be one of the only jurisdictions that has established criteria for land application for additional parameters. However, it is noted that other standards could potentially be used through NDPES forms or other site-specific documentation, which may not be publically released.

## 4 Conclusions and Next Steps

Overall, O. Reg. 267/03 is one of the most advanced regulatory frameworks across North America used for managing the land application of wastes generated from agriculture and non-agriculture operations. The regulatory framework is well-documented, easy to access, and allows for the application of many different sources of NASM. Additionally, OMAFRA has created many documents that can be used to assist NASM generators when developing their appropriate paperwork. Furthermore, Ontario also requires qualified personnel to supervise and document the NASM operation, which further strengthens the environmental and public protection. However, processors still believe that the framework can be relaxed in areas to strengthen their economic competitiveness while still protecting public interest.

As shown through the wash-water evaluation from previous projects, samples from food processing sources were found to have a low occurrence of any heavy metals. Out of 104 total samples, only two samples were found to exceed the CM1 standards for Cadmium and Molybdenum; however, no samples ever exceeded the CM2 standards. When compared to other jurisdictions such as Quebec or Kentucky, the regulatory framework allows for heavy sampling to be omitted for agri-food biosolid sources. Therefore, Ontario may want to allow for decreased heavy metal sampling frequency, if there is a strong evidence of low heavy metal occurrence

within the samples. However, it is noted that the next step should be to conduct more sampling first, to understand the quality of the produced NASM. Additionally, it is noted that not all sampling should be reduced, especially for larger facilities, to ensure environmental due diligence.

Furthermore, it was found through the jurisdiction scan that Ontario has some of the higher allowable levels of heavy metal concentrations that can be land applied. Although Ontario does have maximum limits for addition of heavy metals to the soil for a 5-year period that are similar to other provinces, Ontario may want to review its standards for allowable heavy metal concentration. This would further ensure that environmental risks are properly managed and public health safeguards remain in place.

Finally, it was also found that Ontario was one of the most advanced for jurisdictions that have established additional criteria for land application. When investigating additional criteria including groundwater setback restrictions, sodium constraints, SOM requirements, and FOG limits, it was found that most jurisdictions have not created any standards. Although, it is noted that many jurisdictions within the states can utilize the NDPES documents; therefore, site-specific standards are potentially being used. As such, another next step would be to have further discussions with other regulators, to determine if any other additional criteria have been created.

**Acknowledgements** The authors would like to thank OMAFRA for the funding through their Special Initiatives Program and the consultation and guidance provided by Peter Doris, Environmental Specialist, OMAFRA. Additionally, the authors are grateful for the efforts provided by the co-op students Liam Dunnett and Amin Costas, and research associate Zachary Kanmacher. Without their help, the project objectives could not have been completed.

## References

1. Nutrient Management Act (2002) S. O. 2002, Chapter 4. <https://www.ontario.ca/laws/statute/02n04>
2. OMAFRA (2016) Non-Agricultural Source Materials (NASM). <http://www.omafra.gov.on.ca/english/nm/nasm.html>
3. Devos G (2019) Proposed regulatory amendments to Ontario regulation 267/03 under the nutrient management act. <https://www.ontariocanada.com/registry/view.do?postingid=28106>
4. Puddephatt KJ (2013) Determining the sustainability of land-applying biosolids to agricultural lands using environmentally relevant terrestrial biota. Ryerson University Theses and Dissertations. Paper 1579
5. Environment Quebec (2004) Guidelines for the beneficial use of fertilizing residuals. Guidelines for the beneficial use of fertilizing residuals-Reference criteria and regulatory standards (banq.qc.ca)
6. Canadian Council of Ministers of the Environment. Biosolids Task Group (2010) A review of the current Canadian legislative framework for wastewater and biosolids. Canadian Electronic Library
7. Ontario Regulation 267/03, 2003. General. <https://www.ontario.ca/laws/regulation/030267>

# Settling and Rising Hydrodynamics of Microplastic Pollutants: A Numerical Study



Zihe Zhao and Shooka Karimpour

**Abstract** With ever-growing plastic production, the pollution of microplastics (MPs) has become a threatening environmental problem in the twenty-first century. It is crucial to investigate MPs' hydrodynamics due to their widespread pollution in aquatic environments. MPs are particularly difficult to characterize because they not only have a wide range of sizes and densities and are also found with highly variable shapes. So far, no numerical investigation exists that has sufficiently accounted for MPs' complex shapes. To accurately predict the fate, transport, and mobility of aquatic MPs, formulating settling hydrodynamics of MPs with complex shapes is essential. In the present study, a finite volume-based three-dimensional numerical model is employed to investigate the settling trajectories of MP particles with a wide variety of shapes and densities in a quiescent fluid. Seven test runs are performed utilizing the present model, with negatively buoyant MP particles whose densities range from 1100 to 2000 kg/m<sup>3</sup>. The particle Reynolds numbers of these settling particles fall in the nonlinear range of 100–1300. Our initial results have proven the present model's robustness in simulating the dynamics of MP-sized particles in a quasi-static fluid. Preliminary results with MP particles of regular shapes are shown to be consistent with previous empirical formulations derived from particle settling experiments. With the present model, complex shaped MPs, e.g., thin cylinders, resembling MP films, are also tested. The settling dynamics of thin cylindrical MP particles obtained from the present model match well with the settling patterns observed from experiments. Results from the present model will be used to parameterize MP particles with irregular shapes, enabling the prediction of MPs' transport history through large-scale models.

**Keywords** Hydrodynamics • Microplastic pollutants

---

Z. Zhao (✉) • S. Karimpour  
York University, Toronto, Canada  
e-mail: [zihezhaoyorku.ca](mailto:zihezhaoyorku.ca)

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_80](https://doi.org/10.1007/978-3-031-34593-7_80)

1263

## 1 Introduction

Microplastics (MPs), typically defined as plastic particles smaller than 5 mm in diameter [1], are originated either from direct production, or from the fragmentation of larger plastic particles. Pollution of MPs in aquatic ecosystems has become a threatening environmental problem in the twenty-first century [2]. This problem arises partly due to increasing production rates of plastics ever since they have been introduced into the industry since the 1950s [3–5]. More importantly, MP particles can be transported far and wide by water current, leading to their widespread pollution in aquatic environments [6, 7]. For this reason, investigating the hydrodynamics of MPs is crucial.

MPs in natural aquatic environments are found with highly varying densities, sizes, and shapes. This happens because plastics are originally produced with different compositions, and then modified through different weathering processes. Different MP particles can exhibit drastically different transportability and sinking/floating behaviors due to their differences in density. This density difference primarily arises from the difference in source material of plastic polymers, with polystyrene having a density of less than  $< 0.01 \text{ g/cm}^3$ , and Teflon with a density up to  $2.3 \text{ g/cm}^3$  [8].

Weathering of MPs happens as the sizes, shapes, and physical properties of MPs are altered during their transport, resulting in mechanical fragmentation and chemical modification of MP particles [7]. Mechanical fragmentation happens through collision and abrasion between particle to particle and particle to flow, resulting in the breakage of plastic particles. Plastic particles will undergo a size reduction and an increase in surface to volume ratio after mechanical weathering. Chemical modification of MPs may happen abiotically, or with the aid of microorganisms. These chemical modifications typically happen at the surface of the particle, with the aid of oxygen, sunlight, etc. Therefore, an increased surface to volume ratio and more access to oxygen and sunlight will typically increase the rapidity of chemical and biological weathering. A noticeable example of biological weathering is biofouling, in which biofilms form at the surface of MPs. These biofilms typically have a density greater than the surrounding fluid. With the formation of biofilms, the density and surface properties of MPs are altered [7].

The shapes of MPs found in natural aquatic environments are particularly difficult to categorize. This happens due to the wide varieties of manufactured plastic shapes and weathering processes plastics have been through from their source to where they are sampled. As biofouling alters the density of MPs and may cause originally buoyant particles to sink, the surface to volume ratio of the particles is particularly important. Based on this concern, MPs can be categorized according to the relation of their mutually perpendicular longest axis ( $a$ ), intermediate axis ( $b$ ), and shortest axis ( $c$ ) into three-dimensional (3D), two-dimensional (2D), and one-dimensional (1D) particles [9]. 3D particles (pellets and fragments) have three principal axes of the same order of magnitude ( $a \sim b \sim c$ ), 2D particles (thin films and disks) have the longest and intermediate axes much longer than the shortest axis ( $a \sim b \gg c$ ),

whereas 1D particles (fibers and fishing lines) have the longest axis much longer than the other two axes ( $a \gg b \sim c$ ).

The shapes of MPs are very importantly hydrodynamically, affecting the fluid drag force and drag force distribution on the particles. This in turn changes MPs' settling velocity and pattern. The settling and rising behaviors of particles are vital indicators of particles transport characteristics. Studying the settling behaviors is crucial, as it will provide insights to assess the pollution of MPs to certain aquatic ecosystems, design approaches and predict results for MPs' treatments.

Investigation on the hydrodynamics of MPs flourished since the 2000s, due to our increasing awareness of MPs' pollution. In particular, the settling of MP particles has been studied experiments by many researchers. The experiment conducted by Khatmullina [10] has used virgin MP particles with three regular shapes: 3D spheres, 3D isometric cylinders, and 1D fishing lines. The terminal settling velocities obtained from their experiments were compared with previous empirical formulas. While results from 3D MPs matches with previous formulas, results 1D MPs deviate greatly from the curve predicted by previous formulas. Another settling experiment [5] investigated the effect of MPs shape irregularity on the terminal settling velocities. The results from [5] also deviate from previous empirical formulas, and new formulas were subsequently proposed. The new formulas were tested in a later experiment by Waldschläger et al. [7] using 3D and 2D MPs. It was found that the formulas were only compatible with 3D MPs, but not with 2D MP firms. Recently, [11] also conducted a series of experiments to investigate the settling of MPs with many irregular shapes. However, they only used the Corey shape factor  $CSF = c/\sqrt{ab}$  [12] to differentiate MPs' shape. This approach is clearly too crude to differentiate MP shapes, as  $CSF$  is nearly zero for 2D and 1D MPs.

As MPs with different shapes exhibit different settling behaviors, their hydrodynamic parameters are very difficult to characterize for numerical models [13]. Among limited models, Jalón-Rojas et al. (2019) utilized large-scale transport models to predict the transport history and fate of MPs. The model used by Jalón-Rojas et al. (2019) is based on parameters of regularly shaped particles. These parameters are obtained from settling experiments of sediments and may not be appropriate for MPs. Particle's shape plays a critical role in altering the particle's hydrodynamic parameters. Notably, the correlation between drag coefficient  $C_D$  and particle Reynolds number  $Re_p$  for irregularly shaped particles deviates significantly from that of the spherical particles for particles with  $Re_p$  of around 1 to 1000, according to previous experimental studies [14]. To the authors' knowledge, no model has yet sufficiently accounted for MPs' complicated shapes. In order to accurately predict the fact and transport history of MP pollutants in aquatic environments, a model is crucially needed to parameterize MPs with variable shapes.

## 2 Methods

### 2.1 Theoretical Background

Settling of a particle submerged in quiescent fluid happens as the particle is subject to the gravitational force  $F_G$ , the buoyancy force  $F_B$ , and the drag force  $F_D$  (Fig. 1). For incompressible flow with constant properties, the gravitational force  $F_G$  and the buoyancy force  $F_B$  are constant,

$$F_G = -\rho_s g V \quad (1)$$

$$F_B = \rho g V, \quad (2)$$

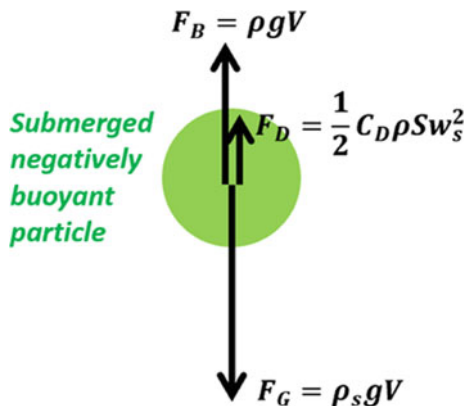
whereas the drag force  $F_D$  is dependent on the settling velocity  $w_s$  of the particle. Assuming that the particle is settling, drag force  $F_D$  that always oppose particle's motion is upward:

$$F_D = \frac{1}{2} C_D \rho S w_s^2, \quad (3)$$

where  $S$  is the projected area of the particle to the settling direction,  $\rho$  is the density of the ambient fluid,  $\rho_s$  is the density of the particle,  $V$  is the volume of the particle, and  $g$  is gravitational acceleration.

For a negatively buoyant submerged particle, after the onset of particle settling, the particle accelerates downward. With increasing downward velocity, drag force  $F_D$  increases, retarding downward acceleration. At a point, the gravitational force  $F_G$ , the buoyancy force  $F_B$ , and the drag force  $F_D$  reaches a balance, and the particle settling at a constant velocity, called terminal settling velocity. The force balance at this stage is presented as

**Fig. 1** Forces received by a settling submerged particle



$$\frac{1}{2}C_D\rho Sw_s^2 = (\rho_s - \rho)gV \quad (4)$$

The present study employs the computational fluid dynamic (CFD) software ANSYS Fluent, which simulates flow dynamics through solving the mass and momentum conservation equations with a finite volume approach. For incompressible flow with constant properties, the mass and momentum conservation equations are represented according to the Einstein summation notation in Cartesian coordinates as

$$\frac{\partial \rho}{\partial t} + \frac{\partial u_i}{\partial x_j} = 0 \quad (5)$$

$$\rho \left( \frac{\partial u_i}{\partial t} + u_k \frac{\partial u_i}{\partial x_k} \right) = -\frac{\partial p}{\partial x_i} + \mu \frac{\partial^2 u_i}{\partial x_k \partial x_k} + \rho g_i, \quad (6)$$

where  $i, j$ , and  $k = 1, 2$ , and  $3$  represent  $x, y$ , and  $z$ , respectively.

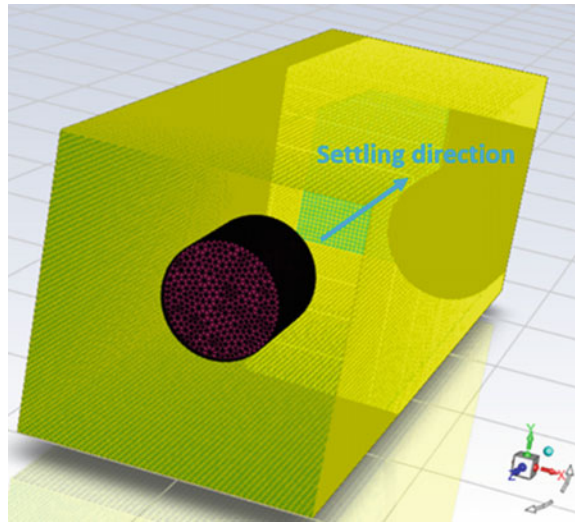
## 2.2 Model Setups

The model utilized in the present study simulates the settling of a MP-sized particle submerged in quiescent fluid. With a three-dimensional numerical scheme, the time evolution of MP particle position and its surrounding flow can be computed. The particle is defined in the model as a cavity with a certain shape, density, and moment of inertia. This enables the model to simulate the settling of MP particles with complex shapes, wide densities, and various density distributions. The six-degree-of-freedom (6DOF) methods are employed to simultaneously simulate particle–fluid interaction and particle settling due to the gravitational force  $F_G$ , the buoyancy force  $F_B$ , and the drag force  $F_D$ . A no-slip boundary condition is set at the outer boundary of the particle, so that the effect of drag on the particle is numerically computed without predefining drag coefficient  $C_D$ .

No turbulent modeling is introduced into the present model, as a mesh as fine as 50–100  $\mu\text{m}$  is utilized around the particle to fully resolve the flow momentum. For pressure–velocity coupling, the SIMPLE scheme is used. The second-order upwind method is used for flux interpolation in the momentum equations, and the first-order implicit approach is employed for time discretization. Using the PRESTO method, spatial pressure is approximated. The Geo-Reconstruct method is utilized for volume fraction calculation.

To fully resolve the flow around the MP-sized particle, 20–80 numerical grids are set in each primary dimension. For computational efficiency, coarse meshes are used outside the zone of interest, i.e., away from the MP particle. This is done with the overset meshing method (Fig. 2), which combines a coarse background mesh zone and a fine refinement mesh zone. The background mesh zone has uniform coarse

**Fig. 2** Overset meshing method employed in the present study. The background mesh zone is colored in yellow, encompassing the whole domain. The refinement mesh zone is colored in deep red, which surrounds the particle



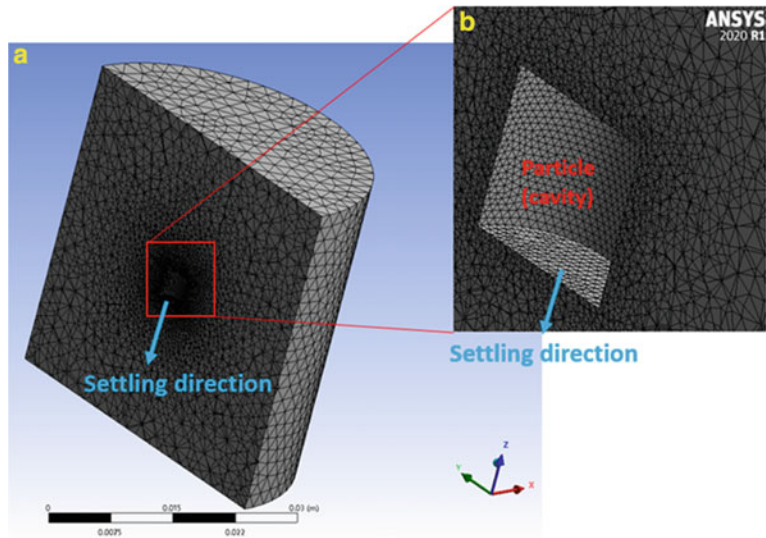
meshes, whereas the refinement mesh zone has grids refining progressively toward the particle (Fig. 3). As the particle settles, the refinement mesh zone moves with the particle by following the particle's center of gravity position. With this approach, the numerical domain adjacent to the particle is always calculated according to the refinement meshes, while areas far away from the particle is computed with coarse background meshes.

### 2.3 Test Run Setups

Seven test runs are conducted to test the model for simulating the settling of MP-sized particles (summarized in Table 1). The tested particles have three different shapes—perfect sphere ( $D = 0.004$  m), isometric cylinder ( $D = L = 0.004$  m), and thin cylinder ( $D = 0.004$  m,  $L = 0.001$  m). Negatively buoyant particles are used with densities ranging from 1100 to 2000 kg/m<sup>3</sup>. The refinement mesh zone has a length or diameter of 0.04 m, which equals ten times of the longest dimension of the particle. For spherical and isometrical cylindrical particles, the shape of refinement mesh zone mimics the shape of the particle. For thin cylindrical particles, a spherical refinement mesh zone is selected. For test runs 6 and 7, an initial tilt angle in  $y$  direction  $\theta_y = 10^\circ$  or  $30^\circ$  is introduced to investigate the settling patterns of thin cylindrical particles.

The numerical domain is filled with water with constant density  $\rho = 998.2$  kg/m<sup>3</sup> and a dynamic viscosity  $\mu = 0.0013$  Pa·s. For all seven test runs, the background mesh zone, which covers the full computational domain, has a length  $L_x =$  width  $L_y = 0.06$  m, and a height  $L_z = 0.2$  m. Preliminary testing suggests that these





**Fig. 3** Sliced view of the refinement mesh zone **a** and its zoom-in view **b** note that the meshes refine progressively toward the particle

**Table 1** Test run setups

Test run number	Particle density $\rho_s$ (kg/m <sup>3</sup> )	Particle shape	$a = b$ (m)	$c$ (m)	Corey shape factor $CSF$	Initial tilt angle $\theta_y$ (°)	Refinement mesh zone shape	Refinement mesh size at the particle $dl_{ir}$ (m)
1	1100	Sphere	0.004	0.004	1	0	Sphere	0.0001
2	1400	Sphere	0.004	0.004	1	0	Sphere	0.0001
3	2000	Sphere	0.004	0.004	1	0	Sphere	0.0001
4	1100	Isometric cyl	0.004	0.004	1	0	Isometric cyl	0.0001
5	1800	Isometric cyl	0.004	0.004	1	0	Isometric cyl	0.0001
6	1100	Thin cyl	0.004	0.001	0.25	30	Sphere	0.00005
7	1100	Thin cyl	0.004	0.001	0.25	10	Sphere	0.00005

dimensions are appropriate, because the domain boundaries are sufficiently far away from the flow zone, and the computation time is acceptable. As initial conditions of the model, no flow is present in the domain, and the center position of the particle and the refinement mesh zone is  $x = 0.03$  m,  $y = 0.03$  m, and  $z = 0.16$  m. The background mesh zone consists of perfectly cubical cells with a size  $dl_b = 0.001$  m. The cells in refinement mesh zone are tetrahedron. These tetrahedron cells are autogenerated by the ANSYS meshing software, with a predefined mesh size at the particle wall  $dl_{ir} = 0.0001$  or  $0.00005$  m, and a mesh size at the refinement mesh’s outer boundary  $dl_{or} =$

0.001 m. Although the test runs performed in the present study used particles with relatively regular shapes, the meshing system used in the present model is capable of imitating the outer wall of MP-sized particles with complex shapes.

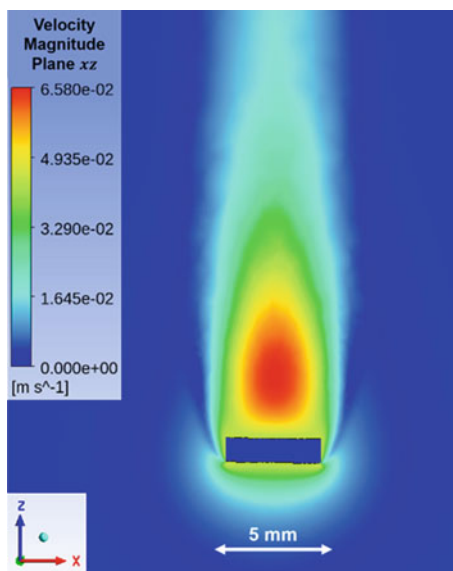
### 3 Results

With the 6DOF solver of ANSYS Fluent, the center of gravity position of the settling particle is recorded. It is observed that the particle first accelerates downward with decreasing acceleration, then reaches a quasi-steady state, where the particle settles at a roughly constant velocity. This matches well with observations from previous particle settling experiments. As indicated in the flow velocity contour below (Fig. 4), the flow around the smallest particle among all the test runs (Test Run 6 and 7) are well resolved.

#### 3.1 Result Parameterizations

Terminal settling velocity  $w_s$  is obtained by plotting the linear regression trend line of the downward change in particle vertical position over the change in time, after the settling has reached quasi-steady state. The terminal settling velocity  $w_s$  is the slope of that trend line.

**Fig. 4** Flow velocity magnitude contours on  $x - z$  plane around the settling MP-sized thin cylinder after reaching quasi-steady state (test run 7, at 2.045 s)



Knowing the terminal settling velocity  $w_s$ , drag coefficient  $C_D$  can be calculated from Eq. (4):

$$C_D = \frac{2g'V}{Sw_s^2}, \tag{7}$$

where  $g' = g(\rho_s - \rho)/\rho$  is reduced gravitational acceleration.

Particle Reynolds number  $Re_p$  is presented by the following eq. (7):

$$Re_p = \frac{\rho w_s D_n}{\mu} \tag{8}$$

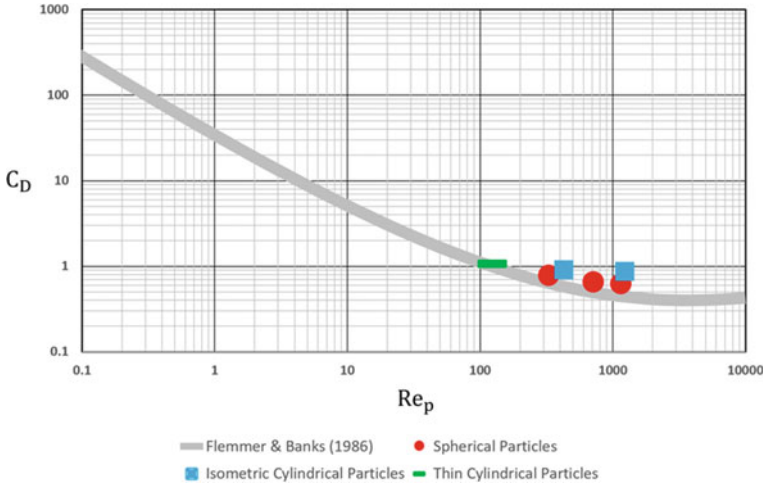
where  $D_n$  is the nominal diameter of the particle, which equals the diameter of a sphere with equivalent volume [12, 15].

Results obtained from the test runs are summarized in Table 2. With a  $D_n$  of 2.88–4.58 mm and a  $g'$  of 1–9.85, the particle Reynolds number  $Re_p$  from all seven test runs range from 123 to 1230. These  $Re_p$  all fall into the nonlinear drag range. Notably, for  $Re_p$  of around 100, drag coefficient  $C_D$  of an irregularly shaped particle deviates most from the  $C_D$  of a perfect sphere with the same  $Re_p$  [14]. Although MPs encompass smaller particles whose  $Re_p$  is smaller 1, falling into the linear drag range, the hydrodynamic interaction of particle settling—flow turbulence is most intricate in the nonlinear drag range [16, 17].

Parameterization of the settling MP-sized particles is done by plotting the obtained drag coefficient  $C_D$  with particle Reynolds number  $Re_p$  (Fig. 5). On that same figure, the  $C_D$ – $Re_p$  relationship obtained from the present test runs is compared with the expression provided by Flemmer and Bank [18] derived from settling experiment results of spherical particles. For spherical particles, it is shown that the results from

**Table 2** Test run results

Test run number	Particle shape	Particle nominal diameter $D_n$ (m)	Reduced gravitational acceleration $g'$	Initial tilt angle $\theta_y$ (°)	Terminal settling velocity $w_s$ (m/s)	Particle reynolds number $Re_p$	Drag coefficient $C_D$
1	Sphere	0.004	1.00	0	0.083	330	0.77
2	Sphere	0.004	3.95	0	0.18	717	0.65
3	Sphere	0.004	9.85	0	0.29	1154	0.62
4	Isometric cyl	0.00458	1.00	0	0.094	428	0.91
5	Isometric cyl	0.00458	7.88	0	0.27	1230	0.86
6	thin cyl	0.00288	1.00	30	0.0434	123	1.08
7	Thin cyl	0.00288	1.00	10	0.0435	126	1.03



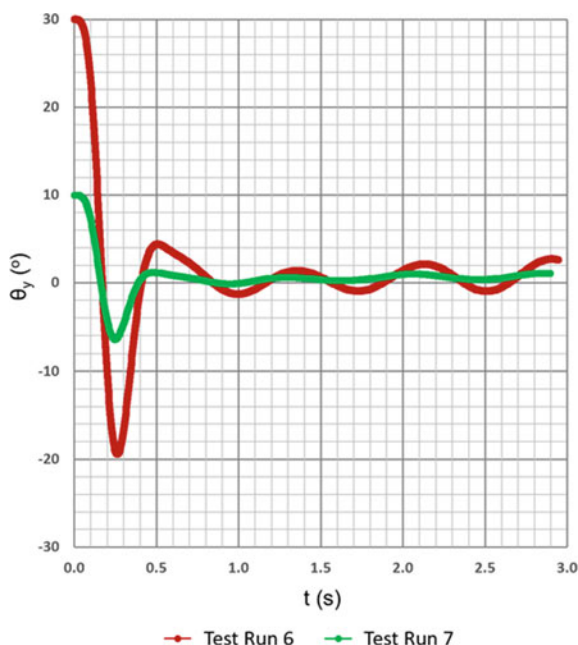
**Fig. 5** Drag coefficient  $C_D$ –particle Reynolds number  $Re_p$  relationship from the present test runs, compared with the empirical formula of [18]

the present test runs agrees well with previous experimental data. For isometrical cylindrical particles, a greater  $C_D$  is obtained with similar  $Re_p$  compared to spherical particles.

### 3.2 Oscillatory Settling of MP Thin Cylinders

The settling of particles with shapes different from perfect spheres has been investigated in experimental studies. Notably, [19] investigated the settling of disk-shaped particles with different particle Reynolds number  $Re_p$ . In their experiment, it was found that the particles settle in the direction perpendicular to the round faces of the disks. Oscillatory motions are also observed, with particles having greater  $Re_p$  exhibit greater oscillation. For the present test runs, oscillatory settling patterns are also observed for thin cylindrical particles. With initial tilt angles introduced only in  $y$  direction, tilt angles remain very small ( $<1^\circ$ ) in  $x$  and  $z$  directions. Both test runs yield very close terminal settling velocities, suggesting that initial tilt angle has little influence on the terminal settling velocity and thus  $C_D$ – $Re_p$  relationship. For Test Run 6 with a greater initial  $\theta_y = 30^\circ$  (compared to Test Run 7 with an initial  $\theta_y = 10^\circ$ ), oscillations with a greater magnitude are observed during the particle settling. For both Test Run 6 and 7, a similar oscillation period  $T_{os}$  of 0.7 s is observed (Fig. 6).

**Fig. 6** Tilt angles in  $y$  direction  $\theta_y$  of thin cylindrical particles during settling, with initial tilt angles in  $y$  direction of  $30^\circ$  (test run 6) and  $10^\circ$  (test run 7)



## 4 Discussion and Conclusion

The present study aims to investigate the settling hydrodynamics of MP-sized particles with various irregular shapes. A three-dimensional model is employed for the computation of particle and its surrounding fluid motion. Six-degree-of-freedom (6DOF) and overset meshing methods are applied for numerical mesh update during the setting of the particle. This ensures that the flow around the particle is always well resolved, and that the computational cost is acceptable. Seven test runs are conducted with the present numerical approaches. These test runs encompass MP-sized particles with different shapes (sphere, isometric cylinder, and thin cylinder), different nominal diameter  $D_n$  (2.88–4.58 mm), and different reduced gravitational acceleration  $g'$  (1–9.85). Particle Reynolds number  $Re_p$  calculated from the obtained terminal settling velocities falls in the range of 123–1230, which is the nonlinear drag range for particle settling.

For the parameterization of MP-sized particles, drag coefficient  $C_D$ –particle Reynolds number  $Re_p$  relationships obtained from the present test runs are plotted together with a well-established empirical formula. It is revealed that the present test results are compatible with previous experimental results for spherical particles. Isometric cylinder particles exhibit higher  $C_D$ , compared with spherical particles with a similar  $Re_p$ . The  $C_D$ – $Re_p$  of thin cylinders, whose nominal diameter  $D_n$  is selected as characteristic length, fall on the same curve as that of the spherical particle. It should be noted that for spheres and thin cylinders with the same  $D_n$ , thin cylinders

have a greater projected area the settling direction  $S$  (see Eq. 3), resulting in a greater drag force even if the drag coefficient is the same for both particles. In agreement with previous experimental observations of disk-shaped particles, MP-sized thin cylinder settles in the direction perpendicular to its wide circular faces; oscillatory settling patterns are also observed for thin cylinders. With a greater initial tilt angle, oscillation magnitude becomes greater. Difference in initial tilt angle is shown to have little effect on the terminal settling velocity and thus  $C_D-Re_p$  relationship.

The robustness and accuracy of the present numerical approaches have been validated through seven test runs. The present numerical approaches are being applied to parameterize MP-sized particles with different sizes, densities, and other irregular shapes. This parameterization effort is crucial to accurately predict the fate and transport history of aquatic MP pollutants through large-scale models.

## References

1. Arthur C, Baker JE, Bamford HA (2009) Proceedings of the international research workshop on the occurrence, effects, and fate of microplastic marine debris. September 9–11, 2008, University of Washington Tacoma, Tacoma, WA, USA
2. Hengstmann E, Fischer EK (2019) Nile red staining in microplastic analysis—proposal for a reliable and fast identification approach for large microplastics. *Environ Monit Assess* 191(10):1–9
3. Barnes DK, Galgani F, Thompson RC, Barlaz M (2009) Accumulation and fragmentation of plastic debris in global environments. *Philos Trans Royal Soc B Biol Sci* 364(1526):1985–1998
4. Derraik JG (2002) The pollution of the marine environment by plastic debris: a review. *Mar Pollut Bull* 44(9):842–852
5. Waldschläger K, Schüttrumpf H (2019) Effects of particle properties on the settling and rise velocities of microplastics in freshwater under laboratory conditions. *Environ Sci Technol* 53(4):1958–1966
6. Shamskhany A, Karimpour S (2021) The role of microplastics' size and density on their vertical mixing and transport. In: The proceeding of the Canadian society of civil engineering annual conference
7. Waldschlaeger K, Born M, Cowger W, Gray A, Schuettrumpf H (2020) Settling and rising velocities of environmentally weathered micro-and macroplastic particles. *Environ Res* 191:110192
8. Chubarenko I, Esiukova E, Bagaev A, Isachenko I, Demchenko N, Zobkov M, Efimova I, Bagaeva M, Khatmullina AL (2018) Behavior of microplastics in coastal zones. In: *Microplastic contamination in aquatic environments*, Elsevier, 175–223
9. Chubarenko I, Bagaev A, Zobkov M, Esiukova E (2016) On some physical and dynamical properties of microplastic particles in marine environment. *Mar Pollut Bull* 108(1–2):105–112
10. Khatmullina L, Isachenko I (2017) Settling velocity of microplastic particles of regular shapes. *Mar Pollut Bull* 114(2):871–880
11. Wang Z, Dou M, Ren P, Sun B, Jia R, Zhou Y (2021) Settling velocity of irregularly shaped microplastics under steady and dynamic flow conditions. *Environ Sci Pollut Res* 28(44):62116–62132
12. Komar PD (1980) Settling velocities of circular cylinders at low reynolds numbers. *J Geol* 88(3):327–336
13. Shamskhany A, Li Z, Patel P, Karimpour S (2021) Evidence of microplastic size impact on mobility and transport in the marine environment: a review and synthesis of recent research. *Front Mar Sci* 8:760649

14. Loth E (2008) Drag of non-spherical solid particles of regular and irregular shape. *Powder Technol* 182(3):342–353
15. Dietrich WE (1982) Settling velocity of natural particles. *Water Resour Res* 18(6):1615–1626
16. Mei R (1994) Effect of turbulence on the particle settling velocity in the nonlinear drag range. *Int J Multiph Flow* 20(2):273–284
17. Stout JE, Arya SP, Genikhovich EL (1995) The effect of nonlinear drag on the motion and settling velocity of heavy particles. *J Atmos Sci* 52(22):3836–3848
18. Flemmer RL, Banks CL (1986) On the drag coefficient of a sphere. *Powder Technol* 48(3):217–221
19. Willmarth WW, Hawk NE, Harvey RL (1964) Steady and unsteady motions and wakes of freely falling disks. *Phys Fluids* 7(2):197–208

# Effects of Amendments on Bioretention Systems: The Field and Laboratory Investigations



Yihui Zhang, Anton Skorobogatov, Jianxun He, Caterina Valeo, Angus Chu, Bert van Duin, and Leta van Duin

**Abstract** In the face of urbanization and climate change, Canadian municipalities have intended to supplement the centralized drainage system with decentralized stormwater management practices, such as bioretention systems, to better management urban stormwater runoff. However, it is not uncommon to observe that bioretention systems leach nutrients (both phosphorus (P) and nitrogen (N)), particularly in the beginning of their operation. This study examined the effects of six amendments selected primarily for reducing P leaching of bioretention systems through monitoring six amendment and two control cells in the field in the growing season of 2020. In addition, two amendments were examined in the laboratory setting to investigate the influence of amendment mixture percentage on their performance using laboratory columns. The field results showed that all amendments had the capability of preventing or mitigating P leaching from bioretention systems, with the water treatment residual (WTR) outperforming all other amendments, followed by the sorptive MEDIA (SM) and activated aluminum (AA). In addition, some of the amendments (i.e., drywall (DRY), WTR, and SM) were also found to be beneficial in reducing the N leaching to a slight degree, whereas eggshell (EGG) introduced an extra source of N leached. Furthermore, the temporal evolution of the P leaching of the amendment cells was found to be different from that of the control cells, whereas same result was not observed in the N leaching. The laboratory results further confirmed the effectiveness of SM and AA in reducing P leaching. The results also revealed

---

Y. Zhang · A. Skorobogatov · J. He (✉) · A. Chu · B. van Duin

Civil Engineering, Schulich School of Engineering, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

e-mail: [jianhe@ucalgary.ca](mailto:jianhe@ucalgary.ca)

C. Valeo

Mechanical Engineering, University of Victoria, 3800 Finnerty Road, Victoria, BC V8P 5C2, Canada

B. van Duin

The City of Calgary, 625 - 25 Ave S.E., Calgary, AB T2G 4K8, Canada

L. van Duin

The Alberta Low Impact Development Partnership, PO Box 34267, Westbrook PO Calgary, AB T3C 3W2, Canada

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_81](https://doi.org/10.1007/978-3-031-34593-7_81)

1277



that the increase of the mixture percentage of amendment did not always improve the performance (in term of P removal rate) and reduce the degree of P leaching. The observed difference in the P leaching behavior between the bioretention cells and columns suggests that caution should be paid when translating laboratory knowledge into the field.

**Keywords** Amendments · Bioretention systems

## 1 Introduction

In the face of urbanization and climate change, Canadian municipalities have intended to supplement the centralized stormwater management systems with the small-scale and more flexible low impact development (LID) practices. The LID practices have been implemented in proximity to stormwater source to improve urban stormwater management. These practices maintain the pre-development hydrologic regime through infiltrating, filtering, storing, evaporating, and retaining stormwater runoff. One of very common LID practices is the bioretention system, which is a vegetated depression and often consists of ponding space, vegetation, occasionally an overflow system at its surface, a permeable growing media layer, a drainage layer, and an optional underdrain system. Bioretention systems have been demonstrated to have dual effects in controlling stormwater runoff quantity (e.g., reducing the runoff volume and peak flow) and improving its quality (i.e., removing various pollutants including suspended solids, heavy metals, nutrients, and many others) [1–5].

Irrespective of the documented benefits of bioretention systems in removing various pollutants, the existing body of knowledge has also revealed that bioretention systems can act as the source of nutrients (e.g., phosphorus (P) and nitrogen (N)), namely leaching nutrients from the system themselves and consequently leading to increased nutrient concentrations and/or loadings of stormwater runoff. In practice, it is common that nutrient-enriched media have been used to construct bioretention systems to support the establishment and survival of vegetation [6]. Compost has been commonly added to bioretention media to improve media aggregation and increase porosity and consequently improve the infiltration rate and the water retention capacity, absorb or aid in the degradation of some pollutants, and reduce the use of fertilizers by improving the soil nutrient holding capacity [5, 7]. However, the nutrient-enriched media is prone to leach nutrients, and thus, nutrient leaching from bioretention systems is not uncommon [8–10], especially at the beginning of their operation.

To mitigate the leaching or improve the removal efficiency of P, a number of amendments have been researched, in particular in laboratory settings. The amendments include water and wastewater treatment residual (WTR), biochar, coconut coir, fly ash, iron-based amendments and zeolite containing adsorptive metals (such as aluminum (Al), iron (Fe), and calcium (Ca)) (e.g., [8, 11–15]). It is not surprising that the performance of an amendment would be associated with the amendment

type and chemical characteristics [4]. The objective of this paper is thus to examine six amendments, which are potentially considered for the study region (i.e., the City of Calgary, Alberta) and identify the optimal amendment(s) for a designated bioretention media (Media 40) [16] using bioretention cells in the field. In addition, bioretention columns were applied to examine the optimal mixture percentage for two selected amendments in the laboratory settings.

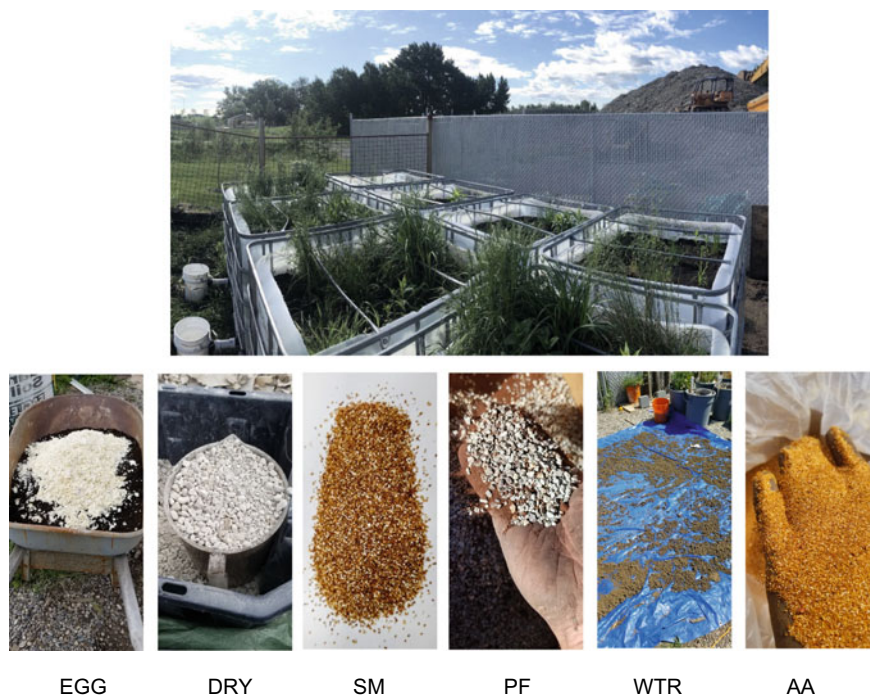
## 2 Study Materials and Method

### 2.1 Study Site and Bioretention Cells

A bioretention research facility, constructed in the Town of Okotoks, which is 45 km south of the City of Calgary, Alberta, consists of 24 bioretention mesocosms for examining the effect of media and vegetation types on bioretention performance and eight bioretention cells (Fig. 1) for investigating the effects of amendments in reducing P leaching. Six amendments of interest to the study were examined. Three amendments including WTR, drywall (DRY), and eggshell (EGG) are free, while the other three amendments including sorptiveMEDIA (SM), Phosfilter (PF), and activated aluminum (AA) are commercially available. The WTR, DRY, and EGG were mixed at 10% (in volume) with bioretention media, while the other three amendments were mixed at 5% (in volume). Two bioretention cells were used as control cells (without amendment, called C1 and C2 cells), and each of the rest of the cells was mixed with a type of amendment. The amendment cells are called WTR, DRY, EGG, SM, PF, and AA cells throughout this paper. All the bioretention cells were constructed with Media 40, sourced from a local supplier (Eagle Lake Landscaping) and planted with the same vegetation including 16 different native, deep-rooting types of grass and forbs. The dry bulk density and the average particle size of the media are 0.86 g/cm<sup>3</sup> and 0.58 mm, respectively. The media contains 16.4% of organic matter. Each bioretention cell is 1 m × 1 m in surface, and 1 m deep, and consists of a 0.2 m ponding area on the top, followed by a 0.6 m growing media layer and a 0.2 m drainage layer (fine sand) at the bottom.

### 2.2 Field Monitoring and Water Sample Analysis

The bioretention cells, which did not receive stormwater runoff (just receiving direct rainfall) in natural events, were monitored in simulated events in 2020. The simulated events having four different event magnitudes (5, 10, 15, and 25 mm) were designed to mimic the median precipitation during the growing season (from May to October) in this study region. A total of 20 events were simulated in the monitoring period. The hydrologic loading applied to each cell was determined according to the



**Fig. 1** Photos of the bioretention cells and six amendments examined in the study

event magnitude and the I/P ratio (15), which is the ratio of contributing catchment area (assumed to be impervious for simplicity) to the surface area of bioretention system (pervious) [16]. The events were simulated in approximately 5-day intervals to represent the local condition (but occasionally in 10-day intervals before 25 mm event). The inflow applied in the simulated events was withdrawn from a nearby stormwater pond. As the primary objective of this work was to investigate the effects of various amendments in reducing nutrient leaching from growing media, inflow was not spiked to the typical concentrations of N and P of stormwater runoff in the study region, whereas sediments (from street sweeping) were added to inflow to represent the typical total suspended solids concentration (400 mg/L) of stormwater runoff.

A water sample was collected, and the outflow volume was measured for each cell in each simulated event. According to the field observation, the gravity drainage often ceased in the next day of a simulated event; thus, a composite water sample was prepared using one sample collected from the day of a simulated event and one sample collected from the following day. The total outflow volume was the sum of the volumes measured on these two days. Three water samples of completely mixed inflow were collected for each simulated event. The nutrient species including reactive phosphorous (RP), total phosphorous (TP), nitrate, and total nitrogen (TN)

were analyzed using Hatch's methods. The pollutant concentrations were reported as the event mean concentrations (EMCs).

## 2.3 Laboratory Experiment

To investigate the impact of the mixture percentage of amendments on their performance, laboratory bioretention columns, each which is 20 cm in diameter and filled with Mix 40 media in 60 cm deep, were constructed. Tap water was applied as the inflow in the laboratory experiment. The columns were constructed to explore the impact of the mixture percentage on the efficiency of two commercially available amendments including SM and AA, which were identified to be more effective than the other commercially available amendment (PF) in the field experiment. The amendments of no-cost were not included into the laboratory experiment. A large number of studies on the effect of amendments have been conducted at various mixture percentages, but in general ranging from 3 to 17% in volume (e.g., [2, 11, 17, 18]). Thus, three mixture percentages, namely 5, 10, and 15%, were examined herein. The control columns (without amendments) were also constructed. The water samples of inflow and outflow were tested for RP, TP, nitrate, and TN. The hydrologic loading of approximately 22 m water was applied to each column. The hydrologic loading is equivalent to about 4 years of annual precipitation of the City of Calgary (considering the I/P ratio of 15). A total of five columns were constructed for SM. Duplicate control columns (LC1 and LC2) and duplicate columns for the mixture percentage of 10% (SM2 and SM3) were constructed. The columns for the mixture percentages of 5% and 15% are called SM1 and SM4, respectively. For AA, four columns were constructed including control (LC3), AA1 (5% mixture percentage), AA2 (10% mixture percentage), and AA3 (15% mixture percentage).

## 2.4 Data Analysis

Apart from using the event-based EMCs to characterize the nutrient leaching and examine the temporal variations of nutrient discharge from the cells, two assessment metrics, the pollutant removal rate (PRR in %) of a pollutant constituent and the water retention rate (WWR in %) of a cell were calculated to examine their pollutant removal efficiency and hydrologic performance, respectively. The PRR and WRR per event or per season are calculated by the following equations:

$$\text{WRR (\%)} = \frac{V_{\text{in}} - V_{\text{out}}}{V_{\text{in}}} \times 100 \quad (1)$$

$$\text{PRR(\%)} = \frac{L_{\text{in}} - L_{\text{out}}}{L_{\text{in}}} \times 100, \quad (2)$$

where  $V_{in}$  (L) and  $V_{out}$  (L) are the volumes of inflow and outflow, respectively;  $L_{in}$  (mg) and  $L_{out}$  (mg) are the pollutant loading of inflow and outflow, respectively. A negative PRR means that pollutant leaching occurs and the cell acts as a source of the pollutant, while a positive PRR suggests that the cell acts as the sink of the pollutant, namely removing the pollutant from inflow.

As the datasets used in the analysis were not normally distributed in general, nonparametric statistical tests were adopted herein. To compare the medians of two samples and more than two samples, the Mann–Whitney and Kruskal–Wallis tests were employed, respectively. In addition, the non-parametric Mann–Kendall test was used to investigate the temporal trend in the event-based EMCs of nutrients. All the statistical tests were conducted at the significant level of 5%.

### 3 Results and Discussion

#### 3.1 Results from Field Monitoring

##### 3.1.1 Effects of Amendments on Hydrologic and Water Quality Performance

The seasonal WRRs of all bioretention cells are presented in Table 1. The amendment cells had similar seasonal WRRs (in the range of 35–39%) with the exception of the EGG cell. The seasonal WRR of the EGG cell was found to be slightly higher than that of other cells. As the cells were exposed to the same environment and hydrologic loading, the difference in the seasonal WRR might reflect the differences in vegetation growth and moisture conditions during the field season.

The seasonal PRRs of all the cells are also presented in Table 1. Compared with the seasonal PRRs of RP and TP of both control cells, which leached P, the seasonal PRRs

**Table 1** Seasonal water retention rate (WRR) and pollutant removal efficiency (PRR) of the bioretention cells

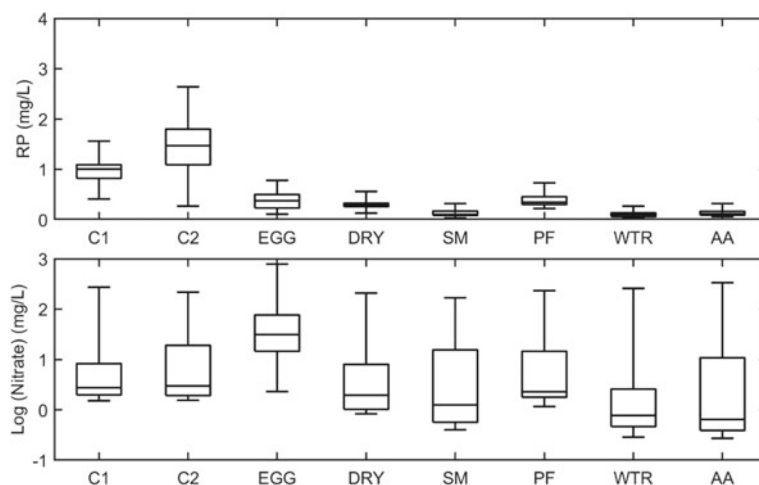
Bioretention cell	Seasonal WRR (%)	Seasonal PRR (%)			
		RP	TP	Nitrate	TN
C1	44	– 464	– 32	– 1956	– 576
C2	32	– 999	– 122	– 2865	– 845
EGG	46	– 125	19	– 8058	– 2442
DRY	37	– 101	35	– 1533	– 429
SM	41	44	55	– 1344	– 352
PF	35	– 125	3	– 2769	– 861
WTR	38	53	60	– 1810	– 498
AA	36	28	45	– 4317	– 1304

of all amendment cells were significantly improved. It was observed that all amendment cells behaved as the sink of TP, while SM, WTR, and AA cells removed RP from inflow. These results demonstrated that all amendments examined are capable of effectively mitigating or preventing P leaching from bioretention systems. Among the amendments, WTR appeared to outperform all other amendments in reducing the P leaching from bioretention system over the field season. It is not surprising that the amendments in general were not able to mitigate the N leaching, although a slight reduction of N leaching was observed from DRY, SM, and WTR cells. Increased N leaching (relative to the control cells) was observed from EGG and AA cells. EGG would introduce additional source of N to bioretention systems. Eggshell has been known to be capable of enhancing vegetation growth as shell membrane is rich in protein and therefore N [19], whereas the AA, the Al-rich amendment, the Al ions would accelerate the dissociation of water molecules in bioretention systems and consequently accelerate the nitrification of N in media and cause the leaching of N [20]. Overall, WTR and SM outperformed all other amendments, as they prevented the P leaching as well as did not increase the N leaching from the cells.

### 3.1.2 Effects of Amendments on Event-Based Nutrient Discharge

To further evaluate the effectiveness of the amendments, the EMCs of the nutrients, which indicate the degree of their leaching/discharge, were examined at the event-based scale. It is worth mentioning that similar results were observed for RP and TP and for nitrate and TN for all cells. Thus, the box plots of the EMCs of RP and nitrate of the cells are shown in Fig. 1. As shown in this figure, the EMCs of RP of all amendment cells were lower than those of C1 and C2. The medians of the EMCs of RP of all amendment cells were also detected to be significantly lower than those of control cells in the Kruskal–Wallis test. In addition, as given in Table 1, relative to the control cells, all amendment cells reduced the RP loading and mitigated or prevented the RP leaching. These results demonstrate that the amendments were effective in reducing P loading and mitigating the degree of P leaching (Fig. 2).

As for the event-based EMCs of nitrate, its median of the EGG cell was significantly higher than that of all other cells (including the two control cells). However, the nitrate EMCs of all other amendment cells appeared to be similar to or slightly lower than those of the control cells. Statistically, the medians of nitrate EMCs of the PF and DRY cells were not significantly different from those of the control cells, whereas the medians of nitrate EMCs of the WTR, SM, and AA cells were significantly lower than those of the control cells. These results suggest that EGG would be the extra source of N leached, while WTR, SM, and AA were also beneficial in reducing the N leaching from bioretention media to some degree. When selecting an optimal amendment to reduce P leaching or improve P removal, consideration of whether the amendment would introduce more leachable N or reduce N leaching is desirable.



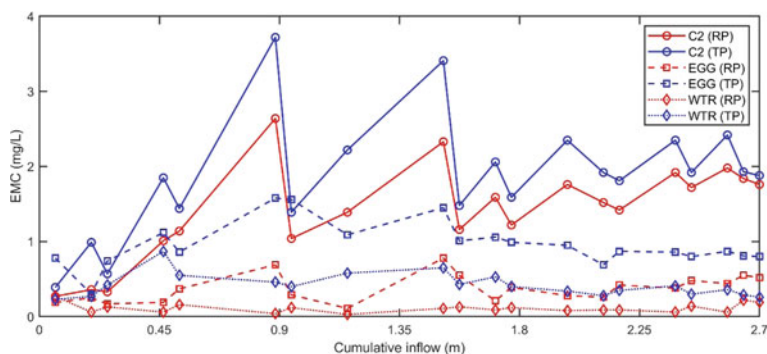
**Fig. 2** Whisker-box plots of the event-based EMCs of RP and nitrate for all bioretention cells

### 3.1.3 Temporal Variation in Event-Based Nutrient Discharge

Figure 3 shows the temporal variations in the EMCs of RP and TP of three selected cells, including one control cell (C2), one cell that prevented P leaching (WTR), and one cell that mitigated P leaching (EGG). It is apparent that different variation patterns in the EMCs of RP and TP were observed among the three cells. In general, the event-based EMCs of P species of the C2 cell increased with the increase of the cumulative inflow (also with time) in the field season. In particular, significant upward trends in the event-based EMCs of TP and RP of the C2 cell were detected in the Mann–Kendall test. However, the upward trends in the TP and RP EMCs were not detected for the WTR and EGG cells (including all other amendment cells). The different temporal variations in the EMCs of P species between control and amendment cells further confirmed the effectiveness of the amendments. In addition, the results also suggest that the effectiveness of the amendments became more prominent along with the time/the increase of the cumulative inflow in the study period.

Considering the setup of the field experiment, bioretention media can be considered as the primary source of P leached. As for the control cells, the EMCs of P species were generally increased temporally, and the significant of upward trends were detected in the Mann–Kendall test. In addition, the time interval with the previous simulated events of Events 6 and 9 were 10 days (not 5 days). Higher EMCs of RP and TP were observed from these two events. All these results suggest that the degree of the P leaching might be associated with the decomposition of compost in media. More leachable P in media is available with the progress of the experiment in the study time period and with the increase of the inter-event period.

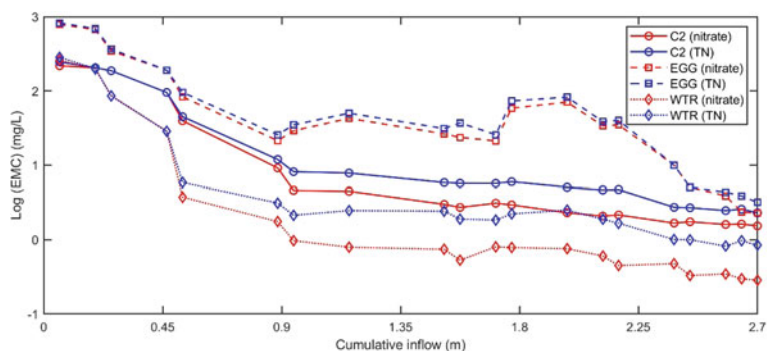




**Fig. 3** Temporal variations in the EMCs of TP and RP of the C2, WTR, and EGG cells

Figure 4 shows the temporal variations in the EMCs of nitrate and TN of three selected cells including C2, WTR, and EGG cells. Differing from the temporal variations in the EMCs of TP and RP, the EMCs of nitrate and TN of both control and amendment cells (except the EGG cell) continuously decreased with time, namely with the increase of the cumulative inflow. The EMCs of N species declined at a faster rate in a number of initial events and then slightly decreased afterward in general. The similar patterns of the temporal variations in the EMCs of N species between the control and amendment cells (except EGG cell) suggest that the amendments in general did not alter the N leaching from bioretention systems, whereas the EMCs of nitrate and TN of the EGG cell declined initially but did not continuously decline in the middle of the field season (corresponding to the cumulative inflow in the range of 0.9–2.25 m). This difference between the EGG cells and all other cells argues that eggshells introduced extra source of N leached and consequently increased N leaching.

Differing from the P leaching, overall similar behavior of N leaching was observed in both the control and amendment cells (except EGG cell), namely the EMCs of N



**Fig. 4** Temporal variations in the EMCs of nitrate and TN of the C2, WTR, and EGG cells



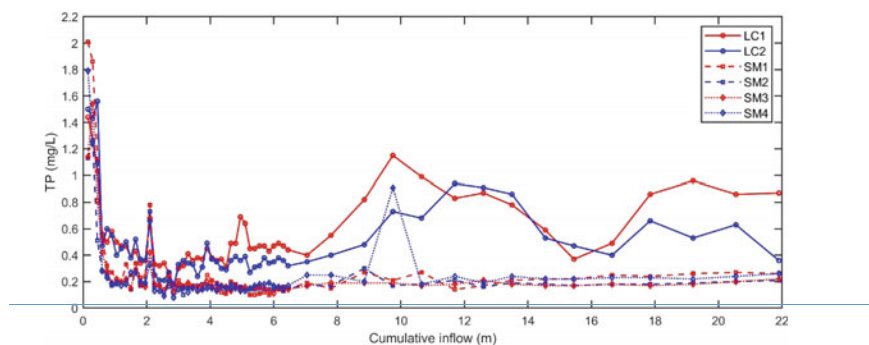
species continuously decreased. Majority leachable N was available at the commence of the experiment. The decrease in the EMCs of N species might be ascribed to the decrease of leachable N in media with the progress of the field experiment. The observed different behavior in the N and P leaching demonstrates that the P leaching would be more persistent than the N leaching from the bioretention media (without amendments). The effectiveness of the amendments in preventing or mitigating P leaching was observed during the whole field season. As the leachable P in bioretention media might continue increase (shown by the control cells), further investigation of the effectiveness of the amendments over a longer time period is desired. On the other hand, the service life the amendments, apart from the efficiency evaluated in the field season, should also been taken into consideration when selecting optimal amendment(s) for real practice.

### 3.2 Laboratory Results

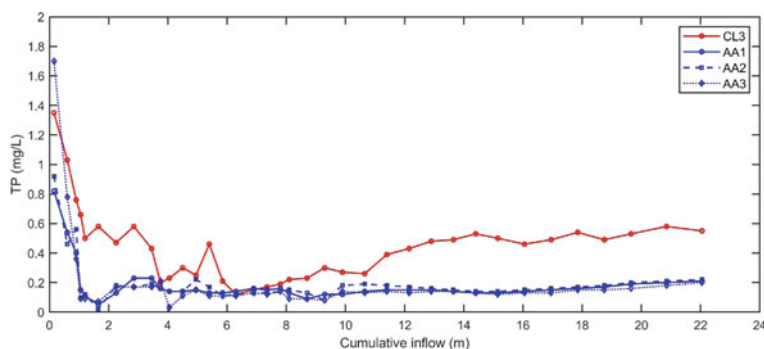
Similar to the results from the field experiment, both SM and AA columns (at all three mixture percentages) were shown to be effectively reducing the P leaching in the laboratory experiment, while SM was slightly more effective in reducing P leaching compared to AA (results not shown), whereas AA columns leached more N compared to the control cells, which is also consistent with the results from the field observations (Table 1). Comparing the removal efficiency of P among the different mixture percentages, 10% and 15% mixture percentages appeared to be slightly better than other mixture percentages for SM and AA, respectively. This suggests that the efficiency of amendments did not always increase with the increase of the mixture percentage.

Figures 5 and 6 further present the variations in TP concentrations with the cumulative inflow applied to the columns. These figures demonstrate that there were no apparent differences in the variations in TP concentrations at the three different mixture percentages for both SM and AA. In addition, the significant differences in the medians of TP concentrations were not detected from both SM and AA columns of different mixture percentages, respectively. Moreover, the TP concentration of the amendment columns appeared to increase with the increase of the cumulative inflow gradually from a time point to the end of the laboratory experiment. This suggests that the efficiency of the amendments in reducing P leaching slightly deteriorated after applying 5 m (for SM) or 10 m (for AA) of inflow. The TP concentration of the control columns was initially decreased and then started to increase after applying 5 m of the inflow. Therefore, the decreased of the amendment efficiency in reducing P leaching might be ascribed to the increased leachable P in media and/or the decrease of the capacity of amendments, whose capacity is limited.

The P leaching from the control cells (in the field) and the control columns (in the laboratory) appeared to be different. In the field, overall P concentration continued increased over the time/ with the increase of the cumulative inflow; whereas in the laboratory, P concentration initially decreased and then started to increase. This



**Fig. 5** Temporal variations in TP concentration of the SM columns (SM1, SM2, SM3, and SM4) and the control columns (LC1 and LC2)



**Fig. 6** Temporal variations in TP concentration of the AA columns (AA1, AA2, and AA3) and the control column (LC3)

implies that the P leaching might be sensitive to the environmental condition and the application of inflow or hydrologic loading. Therefore, special cautions need to be paid when translating laboratory knowledge to the field.

## 4 Conclusions

In this paper, six candidate amendments (including EGG, DRY, SM, PF, WTR, and AA) for reducing P leaching from bioretention systems were examined in the field setting in the growing season of 2020. Comparing with the control cells, all six amendments demonstrated to be capable of mitigating the P leaching, in particular WTR, SM, and AA can effectively prevent the P leaching from bioretention systems during the field season. Among these three amendments, AA appeared to enhance the N leaching slightly. Therefore, overall WTR and SM outperformed other

amendments in reducing P leaching without increasing N leaching from bioretention systems. The observed different temporal behaviors between P and N discharges of the control and amendment cells suggested that different processes govern the leaching of P and N from bioretention systems. Although bioretention media were considered as the primary source of nutrient leached in the setup of the field experiment, extra N source might be introduced due to the addition of amendment (i.e., EGG) or the N leaching can be increased/accelerated (e.g., AA). The results also implied that long-term observation (multiple years) is required to select optimal amendment(s) for bioretention systems as the control cells demonstrated that the P leaching would persist in multiple years, and thus the service life of the amendments should be also taken into consideration. In the laboratory experiment, although the efficiency of the SM and AA in reducing P leaching was confirmed, no significant improvement in the efficiency of the amendments was observed with the increase of the mixture percentage. In addition, the difference in the temporal evolution of P EMCs observed in the field and the laboratory setting may call the attention to the knowledge translation from the laboratory to the field for real practice.

**Acknowledgements** This work was funded by Natural Science and Engineering Research Council of Canada (NSERC) Collaborative Research and Development (CRD) program. The authors also would like to acknowledge the financial support from the Town of Okotoks, the City of Calgary, the Bow River Basin Council, and Source2Source.

## References

1. Muhammad S (2016) A Review of the bioretention system for sustainable storm water management in urban areas. *Mater Geoenviron* 63(4):227–236
2. Yan Q, James BR, Davis AP (2018) Bioretention media for enhanced permeability and phosphorus sorption from synthetic urban stormwater. *J Sustain Water Built Environ* 4(1):04017013
3. Chapman C, Horner RR (2010) Performance assessment of a street-drainage bioretention system. *Water Environ Res* 82(2):109–119
4. O'Neill SW, Davis AP (2012) Water treatment residual as a bioretention amendment for phosphorus. II: long-term column studies. *J Environ Eng* 138(3):328–337
5. Sun Y, Chen SS, Lau AY, Tsang DC, Mohanty SK, Bhatnagar A, Rinklebe J, Lin KY, Ok YS (2020) Waste-derived compost and biochar amendments for stormwater treatment in bioretention column: co-transport of metals and colloids. *J Hazard Mater* 383:121243
6. Dagenais D, Brisson J, Fletcher TD (2018) The role of plants in bioretention systems; does the science underpin current guidance? *Ecol Eng* 120(532–545):532–545
7. Hurley S, Shrestha P, Cording A (2017) Nutrient leaching from compost: implications for bioretention and other green stormwater infrastructure. *J Sustain Water Built Environ* 3(3):04017006
8. Tirpak RA, Afroz AN, Winston RJ, Valenca R, Schiff K, Mohanty SK (2021) Conventional and amended bioretention soil media for targeted pollutant treatment: a critical review to guide the state of the practice. *Water Res* 189:116648
9. Brown RA, Birgand F, Hunt WF (2013) Analysis of consecutive events for nutrient and sediment treatment in field-monitored bioretention cells. *Water Air Soil Pollution* 224:1581
10. Shrestha P, Hurley SE, Wemple BC (2018) Effects of different soil media, vegetation, and hydrologic treatments on nutrient and sediment removal in roadside bioretention systems. *Ecol Eng* 112:116–131

11. Balch GC, Broadbent H, Wootton BC, Collins SL (2013). Phosphorus Removal Performance of Bioretention Soil Mix Amended with Imbrium®Systems Sorbtive®Media. Retrieved from
12. Marvin J, Passeport E, Drake J (2020) State-of-the-art review of phosphorus sorption amendments in bioretention media: a systematic literature review. *J Sustain Water Built Environ* 6(1):03119001
13. Han C, Lalley J, Iyannac N, Nadagoudad MN (2017) Removal of phosphate using calcium and magnesium-modified iron-based adsorbents. *Mater Chem Phys* 198:115–124
14. Hunt WF (2003) Pollutant removal evaluation and hydraulic characterization for bioretention stormwater treatment devices. Retrieved from <https://www.proquest.com/docview/304248502?pq-origsite=gscholar&fromopenview=true>
15. Liu J, Davis AP (2014) Phosphorus speciation and treatment using enhanced phosphorus removal bioretention. *Environ Sci Technol* 48(1):607–614
16. City of Calgary (2016) Low impact development guidelines: module 2—bioretention and bioswales. Final Report
17. Malcolm EG, Reese ML, Schaus MH, Ozmon IM, Tran LM (2014) Measurements of nutrients and mercury in green roof and gravel roof. *Ecol Eng* 73:705–712
18. Goh H, Zakaria N, Lau T, Foo K, Chang C, n Leow C (2015) Mesocosm study of enhanced bioretention media in treating nutrient rich stormwater for mixed development area. *Urban Water J* 14(2):134–142
19. Khairnar MD, Nair SS (2019) Study on eggshell and fruit peels as a fertilizer. In: International conference on sustainable development (ICSD 2019), New York, USA
20. Rosenband V, Gany A (2010) Application of activated aluminum powder for generation of hydrogen from water. *Int J Hydrogen Energy* 35(20):10898–10904

# **Cold-Regions Specialty**

# Cold Temperature Effects on Reinforced Concrete Structural Behavior



William T. Riddell, Douglas B. Cleary, Gilson R. Lomboy, Shahriar Abubakri, Benjamin E. Watts, Danielle E. Kennedy, Brian Berry, Amelia Chan, Nicholas Giagunto, Joseph Goodberlet, Maximilian Husar, Joseph Kayal, and Christopher McCormick

**Abstract** The design and construction of reinforced concrete structures in the arctic is of increased interest. Concrete and steel properties change as temperatures drop below freezing. Change in concrete properties such as compressive and tensile strengths and elastic modulus can be large. The changes in steel properties are smaller over this temperature range, but could still be important. These changes in properties due to arctic temperatures will impact the structural behavior of reinforced concrete elements as well as frames. In this study, representative elements of reinforced concrete frames are analyzed to determine how changes in element stiffness over a range of temperatures influences moment demand in frame elements based on variations in material properties with temperature that are reported in the literature. The temperature-affected strengths and stiffness of these elements were incorporated into a simple, analytical frame model to assess the impact of reduced temperature on frame behavior. In particular, the changes in demand on elements and connections for service load conditions as the temperature decreases are evaluated.

**Keywords** Cold temperature · Reinforced concrete · Structural behavior

## 1 Introduction

The effects of climate change are increasing the importance of construction in arctic environments. Research on the mechanical behavior of concrete suggests that both strength and stiffness of concrete increase with decreasing temperatures. Increased

---

W. T. Riddell · D. B. Cleary (✉) · G. R. Lomboy · S. Abubakri · B. Berry · A. Chan · N. Giagunto · J. Goodberlet · M. Husar · J. Kayal · C. McCormick  
Civil and Environmental Engineering, Henry M. Rowan College of Engineering, Rowan University, 201 Mullica Hill Road, Glassboro, NJ 08028, USA  
e-mail: [cleary@rowan.edu](mailto:cleary@rowan.edu)

B. E. Watts · D. E. Kennedy  
U.S. Army Engineer Research and Development Center (ERDC), Cold Regions Research and Engineering Laboratory (CRREL), 72 Lyme Road, Hanover, NH 03755-1290, USA

strength and stiffness of reinforced concrete structural components such as beams and columns have also been observed with decreasing temperatures. A first order analysis of these results suggest that it is conservative to use nominal room temperature behavior to design reinforced concrete structures that could be exposed to cold temperatures. However, structural behavior is affected by section properties that are, in turn, affected by changing material properties for both reinforcing steel and concrete. Changes in the stiffness of structural elements will affect the way that load flows through a structure. Prior research has not investigated how changes in material properties resulting from cold will affect load flow through a reinforced concrete structure. It is important to ensure that changes in section properties related to cold temperature does not result in significant changes in the load flow through a structure, as this could affect the required strength of specific members or connections.

The purpose of this paper is to present a sensitivity study on reinforced concrete structures that will allow general conclusions regarding the effect of changes in material properties resulting from cold temperatures on the overall flow of service loads through a reinforced concrete structure to be made. Changes in maximum service load moments in beams and at connections for various beam dimensions, reinforcing steel, and beam span are considered.

## **2 Literature Review**

A review of pertinent literature is presented in this section. Previous efforts to characterize mechanical behavior of concrete and reinforcing materials, as well as structural components are discussed.

### ***2.1 Concrete in Cold Weather***

The properties of concrete that are most applicable to this study are compressive strength, tensile strength, and modulus. Research efforts on these fronts are described in the following sections.

#### **2.1.1 Compressive Strength**

The compressive strength of concrete increases with decreasing temperature [1, 2]. Montejó et al. [3] surveyed work back to the 1980's on concrete behavior in cold weather. Several proposed equations to relate the strength increase to temperature and moisture content were identified. In particular, the works of Browne and Bamforth [4], Goto and Miura [5], and Okada and Iguro [6] were cited. The increased strength is larger at higher moisture contents, suggesting the strength increase comes from water expansion as it freezes and fills capillary pores and closes microcracks through

internal prestressing [7]. Montejo et al. [3] concluded an equation proposed by [4]

$$f'_c(T) = f'_c(20) - \frac{T w}{12} \quad (1)$$

with  $f'_c$  in MPa and valid over the range 0 to  $-120^\circ\text{C}$ , provided the best fit to the available data. In the equation,  $T$  is the temperature of interest in  $^\circ\text{C}$  and  $w$  is the moisture content of the concrete which is taken as 3% in typical air dry conditions.

### 2.1.2 Tensile Strength

The tensile strength of concrete also increases with decreasing temperatures as measured by split cylinder tests [1, 8, 9] and flexural tests [10]. In some studies, the increase in tensile or flexural strength was larger in comparison with the compressive strength. According to Rostàs and Wiedemann [11] and Sloan [12], at about  $-40^\circ\text{C}$ , tensile strength increased by nearly 180% compared to tensile strength at room temperature. However, there was a decrease in tensile strength when the concrete reached  $-160^\circ\text{C}$ . Nasser and Evans [9] suggest the increase in tensile or flexural strength can be attributed to adhesive forces in the ice within the capillary pore structure of the concrete.

Montejo et al. [3] used data from Lee et al. [1, 2], Kasami et al. [8], and Nasser and Evans [9] and developed the relationship

$$\sigma_t(T) = (1 - 0.0105T)k_{1T}\sqrt{f'_c(T)} \quad (2)$$

for  $T$  between 0 and  $-50^\circ\text{C}$ , where  $k_{1T}$  is the ratio between the tensile strength of the concrete and the square root of the compressive strength of concrete at room temperature. In this equation, the compressive strength at the temperature of interest is used under the radical.

### 2.1.3 Modulus of Elasticity

Montejo et al. [3] extracted data from Lee et al. [1], Kasami et al. [8], Marshall [13], and Filiatrault and Holleran [14] to show the strength and modulus of elasticity of concrete increases with decreasing temperatures down to  $-40^\circ\text{C}$ . Montejo's review found that the ACI 318 Building Code [15] relationship between the elastic modulus and compressive strength of the concrete holds true as long as the corresponding compressive strength of the concrete at the target low temperature is used in the relationship. This relationship is defined for normal weight concrete as

$$E_c = 4700\sqrt{f'_c} \quad (3)$$

where  $f'_c$  is the compressive strength of the concrete in MPa and  $E_c$  is in MPa.



Rostàs and Wiedemann [11] noted that concrete becomes more brittle at extremely low temperatures. They also observed an increase in the modulus of elasticity at  $-170^{\circ}\text{C}$  compared to  $20^{\circ}\text{C}$  and also observed an increase in the elastic modulus as moisture content increased in the concrete.

## ***2.2 Reinforcing Material in Cold Weather***

The mechanical properties of metallic rebar materials at cold temperatures have been investigated by several researchers. Of primary concern to this study is the effect of temperature on elastic modulus. Most researchers have either not reported or not found that the modulus of elasticity of steel was affected by temperature [3, 14, 16]. Yan and Xie [17] reported slight changes (both increasing and decreasing) in elastic modulus with decreasing temperature and believed the results support neglecting this minor observation.

## ***2.3 Reinforced Concrete Structural Components in Cold Weather***

There are limited data available from testing of reinforced concrete structural components in cold temperatures. Sloan [12] observed that, depending on the circumstance, an increase in strength can be either beneficial or detrimental to seismic behavior of a structure.

DeRosa et al. [18] studied crack width behavior of reinforced concrete (RC) tension specimens in cold conditions. The stiffness, strength, creep, and crack widths of reinforced concrete beams at cold temperatures were also examined. Findings were that cracks in reinforced decreased in width at lower temperatures, which may increase the shear capacity of a section through improved interlock along the crack. No significant increase in flexural capacity was found for the beams tested at  $-20^{\circ}\text{C}$ . However, an increase in shear capacity was noted. No difference in stiffness of the beams was noted.

Mirzazadeh et al. [19] performed 4-point load testing on reinforced concrete beams including a period under sustained load, following by cyclic loading, and final increased loading to failure. Four beams in total were tested, two at  $15^{\circ}\text{C}$  and two at  $-25^{\circ}\text{C}$ . One beam tested at each temperature included stirrups and one tested at each temperature did not. The two beams with stirrups failed in a ductile manner. The strength and ultimate deflection of these beams increased by 8 and 34%, respectively, at the low temperature. A stiffness increase was not seen in the elastic portion of the load deflection curve while increased ductility at failure was observed. The beams without shear reinforcement experienced brittle failure and the shear capacity increased by 13% at the cold temperature.

Mirzazadeh et al. [20] published results from a study that investigated the fatigue performance of reinforced concrete beams. The beams were the same configurations as for the previous study [19]. Beams with stirrups failed due to bar rupture, regardless of temperature, which is typical. The beam tested at  $-20^{\circ}\text{C}$  had an 11% longer fatigue life compared to the beam tested at room temperature. The increased fatigue life was attributed to the higher concrete strength and lower stress in the steel that occurs in lower temperatures. The reinforced concrete beam at room temperature suffered greater stiffness degradation as a result of cyclic loading than the beam at  $-20^{\circ}\text{C}$ .

Yan and Xie [21] discussed experimental results and finite element simulations of reinforced concrete beams at low temperatures. Numerical models were calibrated from beam tests conducted in a cooling chamber at 20,  $-40$ ,  $-70$ , and  $-100^{\circ}\text{C}$ . The reinforcement and concrete materials were tested at these temperatures to develop material stress-strain behavior. Concrete samples used for material characterization and beam testing were prepared at controlled room temperature and tested at the various temperatures. The compressive strength and elastic modulus of the concrete increase with decreasing temperature while the ultimate strain achieved in the concrete dropped with decreasing temperature. Temperature had no effect on the elastic modulus of the reinforcing steel. As the temperature decreased, the yield strength of the reinforcing steel increased while the ductility decreased. Despite the reduction in ductility, the reinforcement steel continued to satisfy the 0.5% strain requirement for flexural members found in ACI 318. All of the beams failed in a flexural mode and the load carrying capacity increased with decreasing temperature with increases of 15 and 26% with the decreased test temperature, respectively.

Sloan [12] evaluated reinforced concrete “column type” structural components under simulated earthquake conditions at temperatures ranging from 23 to  $-40^{\circ}\text{C}$ . Sloan observed that flexural strength of the members increased by 7.9% at  $-20^{\circ}\text{C}$  and 13% at  $-40^{\circ}\text{C}$  compared to room temperature behavior. However, displacement capacity, measured by drift prior to first rupture, was reduced from 5.97 to 4.86% as temperature decreased from room temperature to  $-40^{\circ}\text{C}$ . Sloan observed that temperature did not significantly affect the damping behavior of the components.

Montejo et al. [3] studied the cyclic response of reinforced concrete circular columns at temperatures ranging from  $-40$  to  $20^{\circ}\text{C}$ . Two different design strategies were investigated: reinforced concrete-filled steel tube columns and circular concrete columns with spiral reinforcing bars. The steel tubes provide shear strength and confinement, but do not directly contribute to axial or flexural capacity. All columns failed in longitudinal buckling. The ductility of the columns decreased as temperature decreased leading to lower displacements at failure. The study concluded that as temperature decreases, columns undergo an increase in flexural strength and a decrease in displacement capacity. With elastic stiffness defined as the ratio of lateral force to lateral displacement at first yield of the longitudinal reinforcement, elastic stiffness was 90% larger at  $-40^{\circ}\text{C}$  and 40% larger at  $-20^{\circ}\text{C}$  than the elastic stiffness at room temperature.

In a follow-up study, Montejo et al. [22] investigated the seismic behavior of ordinary reinforced concrete columns and reinforced concrete-filled steel tube columns.

Results were similar to the first study. Low temperatures increased the flexural capacity and initial stiffness of the columns and reduced the displacement capacity.

Shih and Lee [23] and Shih et al. [24] studied the effects of temperature ranging from  $+20$  down to  $-70$  °C on the bond behavior between concrete and deformed rebars under monotonic, repeated, and cyclic loading using pullout specimens. Kivekas and Korhonen [25] subjected reinforced concrete beams made from high early strength concrete with both hot-rolled deformed and cold-formed smooth reinforcing bars, in both notched and unnotched condition, to impact loads at temperatures ranging from  $+20$  to  $-70$  °C and concluded that the Charpy-V test is not a good representation of the transition temperature for reinforced concrete structures, as the loading rates and mechanism are different from what occurs in reinforced concrete beams.

### 3 Approach

Single bay frames with pin-pin or fixed connections and distributed loads acting on the top beam, shown in Fig. 1, were chosen for this study. These configurations resemble aspects of real structures, which are statically indeterminate, yet simple enough to allow for an analytical solution. In all cases considered in this paper, the height of the columns,  $L_c$  was 3700 mm (12 feet). The length of the beam  $L_b$  was either 4900 (16 ft) or 7300 mm (24 feet).

A schematic of the cross section for both beams and columns is shown in Fig. 2. Three different combinations of  $b$  and  $h$  were considered for the beams, resulting in 305 mm  $\times$  610 mm (12 in  $\times$  24 in) “tall,” 460 mm  $\times$  460 mm (18 in  $\times$  18 in) “square,” or 610 mm  $\times$  305 mm (24 in  $\times$  12 in) “wide” cross sections. The reinforcing steel was the same for both compression and tension sides of the beams. Specific bars were not identified, but three ratios of  $\rho = 0.010$ , 0.015 and 0.020 were considered. The value of  $d$  was  $h - 64$  mm ( $h - 2.5$  inches), and  $d'$  was 64 mm (2.5 inches). Between the permutations in aspect and steel ratios, nine different beam designs

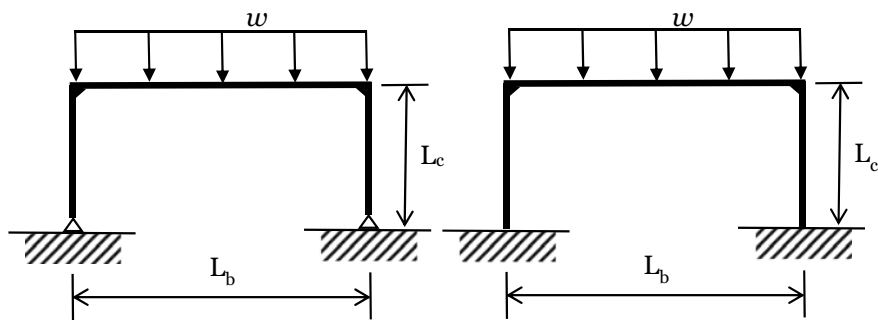


Fig. 1 Portal frame with pinned and fixed supports

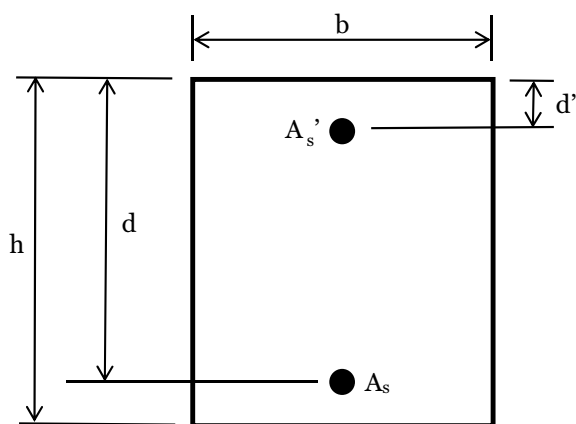
were evaluated. A single column design was used throughout the study. Columns were 610 mm  $\times$  610 mm (24 in  $\times$  24 in), with  $\rho = 0.015$ .

Temperature effects on concrete were modeled based upon the findings of Brown and Bamforth [4] (Eq. 1), and the ACI 318 Building Code [15] relationship between strength and modulus (Eq. 3). The resulting strength and moduli values are summarized in Table 1. The Young's modulus for reinforcing steel is  $E = 200$  GPa (29,000 ksi) at all temperatures.

Once the section geometry and material properties are defined, uncracked transformed moment of inertia, and cracked transformed moment of inertia were calculated for each of the nine beam sections and the column section at 0,  $-20$ ,  $-40$ , and  $-60$  °C. In addition, the nominal moment capacity,  $M_n$  at 0 °C was found for each of the nine beam sections. This study is intended to reflect service loads. Therefore, the distributed load,  $w$ , was chosen such that a simply supported span moment was  $2/3$  of the nominal moment capacity, as shown in Eq. (4).

$$w = \frac{8\left(\frac{2}{3}M_n\right)}{L_b^2} \quad (4)$$

**Fig. 2** Schematic cross section



**Table 1** Assumed compression strength and Young's modulus for concrete

Temperature (°C)	$f'_c$ (MPa)	$E$ (GPa)
0	27.6	24.8
$-20$	33.1	27.2
$-40$	37.9	29.2
$-60$	42.7	31.0

## 4 Results

The nominal moment capacities at 0 °C, and the resulting distributed loads,  $w$ , are summarized in Table 2 for 18 configurations (nine beam sections for two different beam lengths). The values for  $w$  were the same for corresponding pin and fixed configurations. While the distributed load was based on the nominal strength of each section at 0 °C, the same  $w$  was applied for the analyses at 0, −20, −40, and −60 °C.

Linear analyses of the frames presented in Fig. 1 result in the reactions shown in Fig. 3, where the reactions  $R$  and  $M$  depend on the relative lengths, section properties, and fixity as

$$R = \frac{wL_b\beta^2}{8\alpha + 12\beta} \quad (\text{pinned case})$$

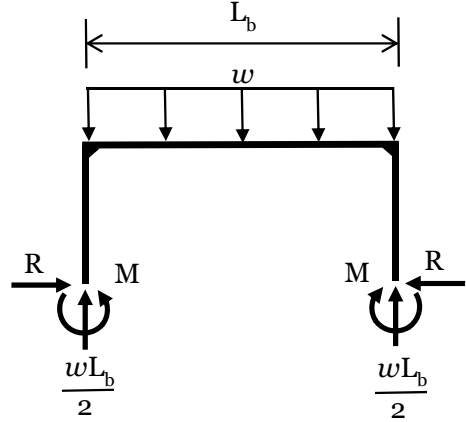
$$R = \frac{wL_b\beta^2}{4\alpha + 8\beta} \quad (\text{fixed case}) \quad (5)$$

$$M = 0 \quad (\text{pinned case})$$

$$M = \frac{wL_b^2\beta}{12\alpha + 24\beta} \quad (\text{fixed case}) \quad (6)$$

**Table 2** Nominal strength of beams and applied  $w$

$L_b$ (mm)	$b \times h$ (mm <sup>2</sup> )	$\rho$	$M_n$ (kN/m)	$w$ (kN/m)
4900	305 × 610	0.010	253	76.2
		0.015	489	108
		0.020	619	137
	460 × 460	0.010	267	59.4
		0.015	381	84.6
		0.020	482	107
	610 × 305	0.010	134	29.8
		0.015	191	42.5
		0.020	242	53.7
7300	305 × 610	0.010	253	33.9
		0.015	489	48.3
		0.020	619	61.1
	460 × 460	0.010	267	26.4
		0.015	381	37.7
		0.020	482	47.6
	610 × 305	0.010	134	13.3
		0.015	191	18.8
		0.020	242	23.9

**Fig. 3** Reactions for frames

where

$$\alpha = \frac{(EI)_b}{(EI)_c} \quad (7)$$

and

$$\beta = \frac{L_b}{L_c} \quad (8)$$

The resulting moments at the beam column connections and the beam midspan are given by

$$\begin{aligned} M_{\text{connection}} &= \frac{-wL_b^2\beta}{8\alpha + 12\beta} \quad (\text{pinned case}) \\ M_{\text{connection}} &= \frac{-wL_b^2\beta}{6\alpha + 12\beta} \quad (\text{fixed case}) \end{aligned} \quad (9)$$

$$\begin{aligned} M_{\text{span}} &= \frac{wL_b^2}{8} - \frac{wL_b^2\beta}{8\alpha + 12\beta} \quad (\text{pinned case}) \\ M_{\text{span}} &= \frac{wL_b^2}{8} - \frac{wL_b^2\beta}{6\alpha + 12\beta} \quad (\text{fixed case}) \end{aligned} \quad (10)$$

The results for the connection moment under pinned connection condition (Eq. 9) are plotted for nine beam cross sections for the  $L_b = 4900$  (16 ft) and 7300 mm (24 ft) in Figs. 4 and 5, respectively. The results for the midspan moment (Eq. 10) are plotted for  $L_b = 4900$  (16 ft) and 7300 mm (24 ft) in Figs. 6 and 7, respectively. For all four plots, the moments are scaled by  $wL_b^2/8$ , which is the moment that would correspond to the midspan of a simply supported beam. The analysis was performed

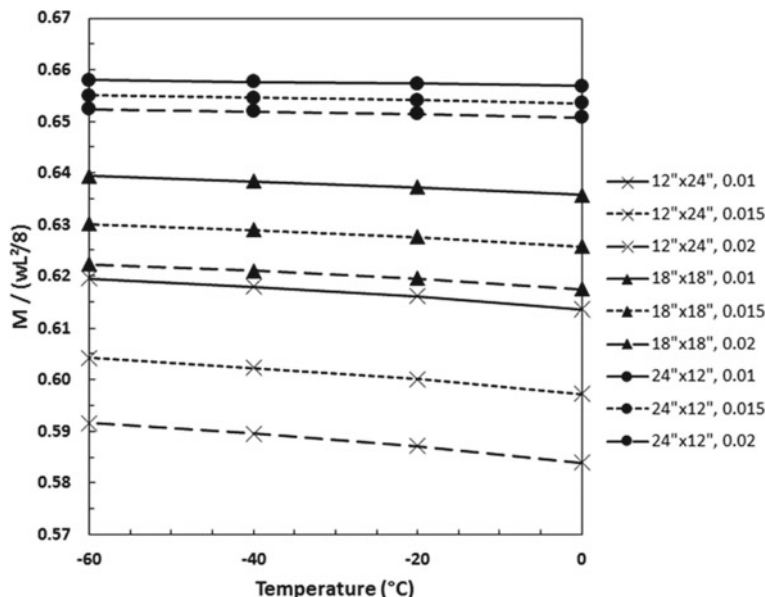


Fig. 4 Moments at connection for 4900 mm span

using the cracked transformed moment of inertia of the beam and the uncracked transformed moment of inertia for the columns. After the analysis, the assumption that the columns were uncracked was validated. The fixed-base portal frame was also evaluated. Similar trends were found with slightly less effect from temperature change.

## 5 Summary and Conclusions

Simple structurally indeterminate reinforced concrete frames were used to conduct sensitivity studies on the effect of cold temperatures on load flow through a structure: two different frame fixity conditions, two different structural geometries, three different beam cross sections, and three different reinforcing ratios resulted in thirty-six configurations being considered.

Under vertical service loads, there is relatively small effect of temperature on the moments for the frame considered. The greatest effect was for the deep beams, for which moments at the connections and midspan changed <2% as temperature decreased from 0 to -60 °C. In this study, the columns and beam had the same concrete compressive strength. In instances where the column is made with higher compressive strength concrete a larger change might be seen as indicated by the presence of  $\alpha$  in the denominator of the connection moment (Eq. 9).

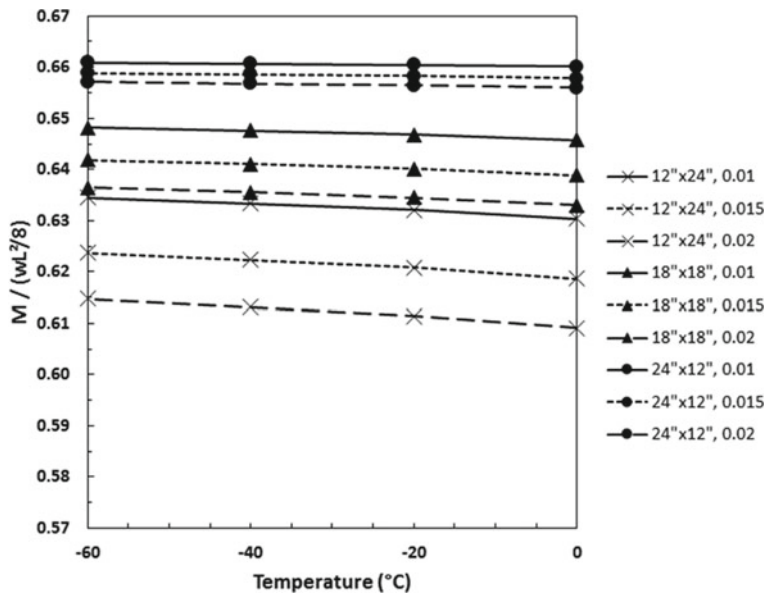


Fig. 5 Moments at connection for 7300 mm span

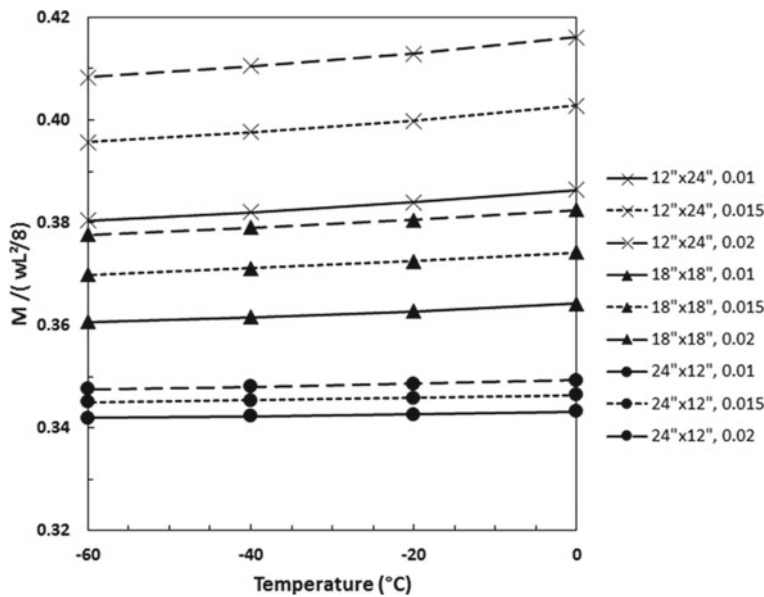


Fig. 6 Moments at midspan for 4900 mm span



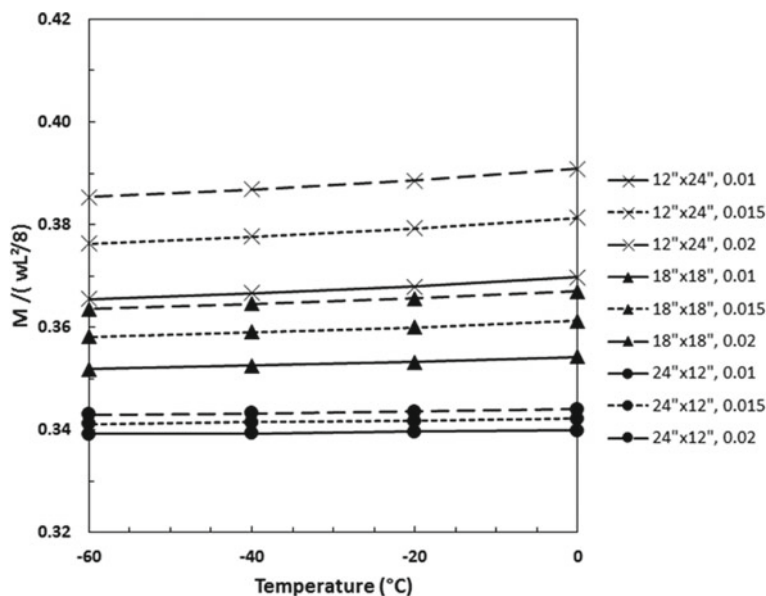


Fig. 7 Moments at midspan for 7300 mm span

A relatively heavy column compared to the beam was considered in this study. The effect of a greater variation column to beam stiffness should be considered in future work. Additional efforts to consider lateral loading, increasing loads into the failure regime, and the behavior of a structure that has temperature differentials between floors and columns are also needed. These future studies will be guided by the insight developed from this study.

## References

1. Lee GC, Shih TS, Chang KC (1988) Mechanical properties of concrete at low temperature. *J Cold Reg Eng* 2(1):13–24
2. Lee GC, Shih TS, Chang KC (1988) Mechanical properties of high-strength concrete at low temperature. *J Cold Reg Eng* 2(4):169–178
3. Montejo L, Sloan J, Kowalsky M, Hassan T (2008) Cyclic response of reinforced concrete members at low temperatures. *J Cold Reg Eng* 22(3):79–102
4. Browne RD, Bamforth PB (1981) The use of concrete for cryogenic storage: a summary of research, past, and present. In: *Proceedings of the 1st international conference on cryogenic concrete*. The Concrete Society, Newcastle Upon Tyne, pp 135–166
5. Goto Y, Miura T (1979) Experimental studies on properties of concrete cooled to about minus 160 °C. *Technol Rep Tohoku Univ* 44(2):357–385
6. Okada T, Iguro M (1978) Bending behavior of prestressed concrete beams under low temperatures. *J Jpn Prestressed Concr Eng Assoc* 20(8):15–17
7. Elices M (1987) Cryogenic prestressed concrete: fracture aspects. *Theor Appl Mech* 7:51–63

8. Kasami H, Tanaka Y, Kishima Y, Yamane S (1981) Properties of concrete at very low temperatures. In: Proceedings of the 1st international conference on cryogenic concrete. The Concrete Society, Newcastle Upon Tyne, pp 123–134
9. Nasser KW, Evans GA (1973) Low temperature effects on hardened air entrained concrete. In: Behavior of concrete under temperature extremes, ACI SP-39, pp 79–90
10. Cantin R, Pigeon M (1998) Durability and mechanical properties at low temperatures of steel-fiber reinforced concrete. In: Proceedings of the 2nd international conference on concrete under severe conditions. E & FN SPON, London, pp 1761–1770
11. Rostásy FS, Wiedemann G (1980) Stress strain behavior of concrete at extremely low temperature. *Cem Concr Res* 10(4):565–572
12. Sloan JE (2005) The seismic behavior of reinforced concrete members at low temperature. MS thesis, Department of Civil Construction and Environmental Engineering, North Carolina State Univ., Raleigh, NC
13. Marshall AL (1982) Cryogenic concrete. *Cryogenics* 22(11):555–565
14. Filiatrault A, Holleran M (2001) Stress-strain behavior of reinforcing steel and concrete under seismic strain rates and low temperatures. *Mater Struct* 34(5):235–239
15. ACI Committee 318 (2019) Building code requirements for structural concrete (ACI 318-19), ACI Committee 318. American Concrete Institute
16. Levings J, Sritharan S (2012) Effects of cold temperature and strain rate on the stress-strain behavior of ASTM A706 grade 420(60) steel reinforcement. *J Mater Civ Eng* 24(12):1441–1449
17. Yan J-B, Xie J (2017) Experimental studies on mechanical properties of steel reinforcements under cryogenic temperatures. *Constr Build Mater* 151:661–672
18. DeRosa D, Hoult NA, Green MF (2015) Effects of varying temperature on the performance of reinforced concrete. *Mater Struct* 48:1109–1123
19. Mirzazadeh MM, Noël M, Green MF (2016) Effects of low temperature on the static behaviour of reinforced concrete beams with temperature differentials. *Constr Build Mater* 112:191–201
20. Mirzazadeh MM, Noël M, Green MF (2017) Fatigue behavior of reinforced concrete beams with temperature differentials at room and low temperature. *J Struct Eng* 143(7):04017056
21. Yan JB, Xie J (2017) Behaviours of reinforced concrete beams under low temperatures. *Constr Build Mater* 141:410–425
22. Montejo L, Kowalsky M, Hassan T (2009) Seismic behavior of flexural dominated reinforced concrete bridge columns at low temperatures. *J Cold Reg Eng* 23(1):18–42
23. Shih TS, Lee GC (1987) Local concrete-steel bond behavior at low temperature. *J Struct Eng* 113(11):2278–2289
24. Shih TS, Lee GC, Chang KC (1988) High-strength concrete-steel bond behavior at low temperature. *J Cold Reg Eng* 24:157–168
25. Kivekäs L, Korhonen CJ (1986) Brittleness of reinforced concrete structures under arctic conditions. In: CRREL TR-86-2. Cold Regions Research and Engineering Laboratory

# Development of Effective and Low-Cost Water Treatment Method for First Nations and Rural Communities in British Columbia, Canada



Zawad Abedin, Jianbing Li, and Sayed Mohammad Nasiruddin

**Abstract** In this study, an effective and low-cost water treatment system is developed. A critical water contamination problem, experienced by the rural, remote, First Nation communities was identified, which is Manganese. Then a treatment system was developed to effectively remove Manganese from the source water. Greensand plus was used as a filtration media, and the prototype was designed to best suit the need of the community residents to remove Manganese in raw water. The laboratory scale experiments were designed using Design Expert software; results were analyzed using RSM method, and the test apparatus was able to reach a removal efficiency 96.50% which can effectively treat source water levels found in the raw water sample. Therefore, a fit-for-purpose solution is developed to remove Manganese from raw water, which is cost-effective, easy to use, and maintain.

**Keywords** Water treatment method · Micro systems · British Columbia

---

Z. Abedin (✉) · S. M. Nasiruddin  
NRES, University of Northern British Columbia, Prince George, BC, Canada  
e-mail: [abedin@unbc.ca](mailto:abedin@unbc.ca)

J. Li  
School of Engineering, University of Northern British Columbia, Prince George, BC, Canada

© Canadian Society for Civil Engineering 2023  
R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2022*, Lecture Notes in Civil Engineering 363,  
[https://doi.org/10.1007/978-3-031-34593-7\\_83](https://doi.org/10.1007/978-3-031-34593-7_83)

1307

## 1 Introduction

Despite the common belief that Canada has abundant fresh water resources, water insecurity is a significant concern for residents in many communities across different provinces in Canada, including British Columbia (BC). Canada ranks third in the world in terms of freshwater resources, accounting for nearly 7% of the world's renewable fresh water [1]. Despite Canada's abundance of freshwater resources, six million Canadians are at risk of waterborne disease. Furthermore, the current level of reinvestment in linear assets (transmission lines, pipes, distribution network) and nonlinear assets (treatment plants, pumping stations, and reservoirs) will jeopardize Canadian water supplies over time [2]. More than 25% of these drinking water assets, particularly in small communities, do not meet the "good condition" criterion [2]. Water insecurity is significant in remote communities, especially among First Nations. In Canada, the National Assessment of First Nations Water and Wastewater Systems (2011) shows that 73% of water systems were at medium to high levels of risk because the facilities could not meet the design protocols or substandard water quality [3]. Many drinking water systems of remote, rural indigenous communities are under long-term drinking water advisories over a year to decades. All public drinking water systems on reserve were supposed to be free from advisories by March 2021; however, the current water governance failed to fulfill the commitment.

### *1.1 Water Insecurity in First Nations and Rural Communities*

Indigenous communities are known to have the poorest water quality in Canada [4]. In 2001, Indian and Northern Affairs Canada (INAC) reported that the quality and safety of three-quarters of the water systems in Indigenous communities were at high risk. Overall, the national freshwater quality remained relatively stable between 2003–2005 and 2010–2012. The federal government's audits over two decades show a pattern of overpromising and underperforming on water and sanitation on reserves. It should be noted, there are considerably more First Nations and small/microwater systems in BC than in other provinces. BC has 198 First Nations and 290 systems that were reviewed for the National Assessment. Ontario, with the next largest numbers, has 121 First Nations, and 158 systems were assessed. Having assessed the risk level of each system the financial cost to meet AANDC's departmental protocols for safe water is estimated for BC systems is \$324 million [3]. Out of the 334 community and public water systems that the FNHA reports on as of January 26, 2021, 57 long-term drinking water advisories on public systems on reserves are in effect. But, as

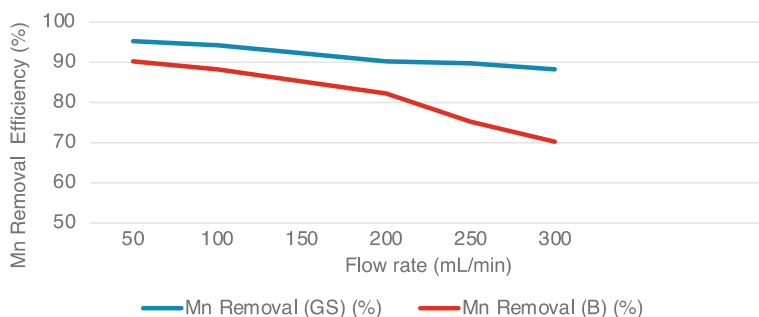
of April 9, 2021, among the public systems on reserves, 61 water systems are still under long-term drinking water advisories [5], indicating that providing clean water to remote communities remains a major challenge in Canada. This includes water systems with five or more connections (CWS). The data for microsystems is not captured in these statistics. Boil water advisories are generally issued when there are microbial contamination and boiled water, hence, can be consumed. On the contrary, do not consume advisories are issued when the water is not safe for consumption, and in most cases, boiling makes the water even worse. For example, when water is contaminated by metal elements such as Manganese, do not consume advisory is issued [6].

## 2 Experimental Design and Methodology

### 2.1 *Selection of Experimental Parameters and Apparatus Design*

A laboratory scale testing was initiated to evaluate Manganese removal performance of commercially available filtration media. A bench top water treatment system was assembled to treat water prepared in the lab, simulating the Manganese concentration of the raw sample water.

The filtration material used in the test bench is Greensand Plus. As a commercial product, the detailed product information of Greensand Plus can be found on the supplier's website. BIRM® was not selected since it needs specific oxygen level in the water which is not achievable without additional equipment and not recommended for gravity filter [7]. A preliminary laboratory test result also confirms that BIRM® is not a suitable material for Manganese removal from the local source water without addition of oxidant (Fig. 1). A cylindrical vessel was used, in similar configuration as described in Sect. 2.2. The experiment was carried for same volume of filtration material and raw water condition; Greensand Plus outperforms BIRM®. For same amount of pollutant concentration, the Greensand Plus removed Manganese more efficiently than BIRM®. This laboratory test was performed only to verify the manufacturer's datasheet information, particular to the scope of this research. Greensand is also widely available, and accessible for rural and first nation residents, which makes greensand an excellent choice as well.



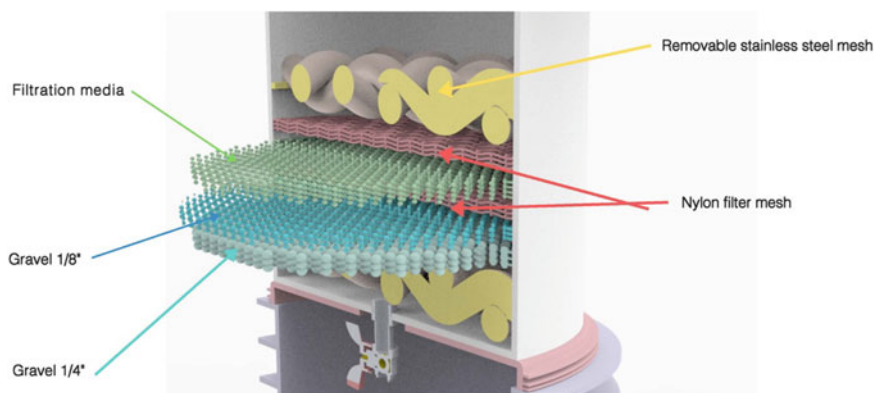
**Fig. 1** Manganese removal efficiency comparison: greensand plus (GS) versus BIRM (B)

## 2.2 Filtration Media, Flow Rate, Diameter-to-Depth Ratio, and Influent Manganese Concentration

Greensand Plus was selected for Manganese removal, based on literature review and the preliminary laboratory test. Greensand plus has an effective size of 0.30–0.35 mm. From the available literature of previous research [8–11] and the manufacturer datasheet [12], the service flowrate ( $q$ ) requirement is approximately 11 L/min per square foot, when bed depth ( $h$ ) is 24–30 inch in municipal type water treatment plant. In this research, the parameters were scaled down and tested for removal efficiency.

The flow rate range calculation was done as follows:

- With the maximum bed depth ( $h$ ) of 76.2 cm (2.5 ft), the service flow rate ( $q$ ) can be up to 11 L/min for a bed volume of  $0.07 \text{ m}^3$  ( $2.5 \text{ ft}^3$ ).
- For a cylindrical vessel with a radius ( $r$ ) of 15 cm (0.5 ft), the total surface area  $= \pi r^2 = 0.07 \text{ m}^2$  ( $0.8 \text{ ft}^2$ ). With a range of depth of 5–15 cm (2–6 inch), considering design feasibility and previous research recommendations as stated, bed volume range can be calculated as  $0.004\text{--}0.011 \text{ m}^3$  ( $0.13\text{--}0.4 \text{ ft}^3$ ). Combining and performing unitary calculation, the allowable flow rate lies within the range of 50–600 mL/min.
- Diameter-to-depth ratio is an important parameter as related with retention time. The diameter-to-depth ratio is selected to be from 2 to 6, based on the requirement to ensure maximum retention time calculation.



**Fig. 2** Apparatus cross section

- Influent Manganese concentration is another key parameter. The range of Manganese concentration was determined based on the raw water sampling data. The average Manganese concentration was about 0.25 mg/L. Therefore, the range was selected based on this information, available literature and published report [13], to ensure the range satisfies reported range of contamination of Manganese.
- The design of apparatus was done using Solidworks (version 2020). The apparatus was prepared using cylindrical HDPE plastic container. The cross-sectional view is shown in Fig. 2.

### 2.3 Experimental Design

Response surface methodology (RSM) is a statistical, theoretical, and mathematical technique for model building in order to optimize the level of independent variables [14]. RSM was used to investigate the effect of independent variables, including diameter-to-depth ratio, flow rate, and influent Manganese concentration on response variable, the removal efficiency. RSM allows to estimate interaction and even quadratic effects. RSM designs allows to find improved or optimal process settings; troubleshoot process problems and weak points; make a process relatively insensitive to external and non-controllable influences [14]. RSM design along with coded and uncoded levels are presented in Table 1. Central composite design and quadratic model was used to design this experiment. Central composite design consists of twenty sets of experiment. That includes six axial points, eight fractional

**Table 1** Independent variables and their corresponding levels for removal efficiency

Independent variable	Symbol	Coded levels				
		$-\alpha^*$	$-1$	$0$	$+1$	$+\alpha$
Diameter-to-depth ratio	A	1	2	4	6	7
Flow rate (mL/min)	B	37	50	325	600	787
Influent manganese concentration (mg/L)	C	0.01	0.15	0.35	0.55	0.69

\* Alpha ( $\alpha$ ) is the distance of each axial point (also called star point) from the center in a central composite design

factorial points, and six central points. The overall experimental setup according to central composite design, is summarized in Table 1. Coded level indicates the boundary values for different parameters, used in central composite design.

3 Results and Discussion

Experimental data were statistically analyzed using Design Expert (version 13). Several statistical parameters (lack-of-fit, predicted and adjusted multiple correlation coefficients, and coefficient of variation) of different polynomial models were compared to select the best fitting polynomial model. To understand the effect of independent variables on response variables, response surface plots were generated using Design Expert. All these experiments were performed in triplicate.

3.1 Fitting the Model

The effects of independent variables on removal efficiency are given in Table 2. A polynomial regression equation obtained from RSM is presented as Eq. 1. Coefficients of the polynomial equation were computed from experimental data to predict the values of the response variable. This equation is in terms of actual variables and can be used to make predictions about the response for given levels of each variable. Here, the levels should be specified in the original units for each variable. Removal efficiency is defined as the percentage of Manganese concentration removed from the raw water.

Removal efficiency = 98.36244 + 1.45659A + 0.002347B – 7.00914C

– 0.003323AB – 0.530218AC – 0.003811BC – 0.256267A<sup>2</sup>

+ 5.27379 × 10<sup>–6</sup>B<sup>2</sup> + 10.43657C<sup>2</sup>

(1)



**Table 2** Experimental design for removal efficiency with independent variables, experimental, and predicted value of response

Run	Independent variables			Response value	
	Diameter-to-depth ratio	Flow rate (mL/min)	Influent manganese concentration (mg/L)	Removal efficiency (%)	
				Experimental	Predicted
1	6	50	0.15	95.61	95.69
2	1	325	0.35	99.12	98.09
3	2	50	0.01	98.89	99.97
4	6	600	0.15	87.2	87.58
5	4	325	0.35	96	94.74
6	7	325	0.35	84.3	85.58
7	2	50	0.55	98.85	98.66
8	6	600	0.55	87.15	85.51
9	4	325	0.35	94	94.74
10	2	600	0.55	97.29	97.03
11	4	325	0.35	94.5	94.74
12	4	325	0.01	99.1	96.95
13	4	325	0.35	94.3	94.74
14	4	325	0.69	93.2	94.89
15	6	50	0.55	95.59	94.46
16	4	325	0.35	94	94.74
17	4	787	0.35	91.52	91.77
18	2	600	0.15	97.3	98.25
19	4	37	0.35	96	96.05
20	4	325	0.35	95	94.74

\*Where A is diameter-to-depth ratio, B is the flow rate (mL/min), and C is the influent manganese content (mg/L)

Statistical analysis (ANOVA) results revealed that the experimental data were represented well with the quadratic polynomial model with coefficient of determination ( $R^2$ ) values for removal efficiency being 0.94 (Table 3).

The  $p$ -value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis. A regression model exhibits lack-of-fit when it fails to adequately describe the functional relationship between the experimental factors and the response variable.

For all variables, the lack-of-fit was non-significant, indicating that the model is statistically accurate ( $p \leq 0.05$ ). If the  $R^2$  number is closer to unity, it indicates that the model is fitting the data better. Lower  $R^2$  values, on the other hand, suggest that response factors were insufficient to explain behavior variance [15]. In this study,  $R^2$

**Table 3** Regression coefficients values for removal efficiency

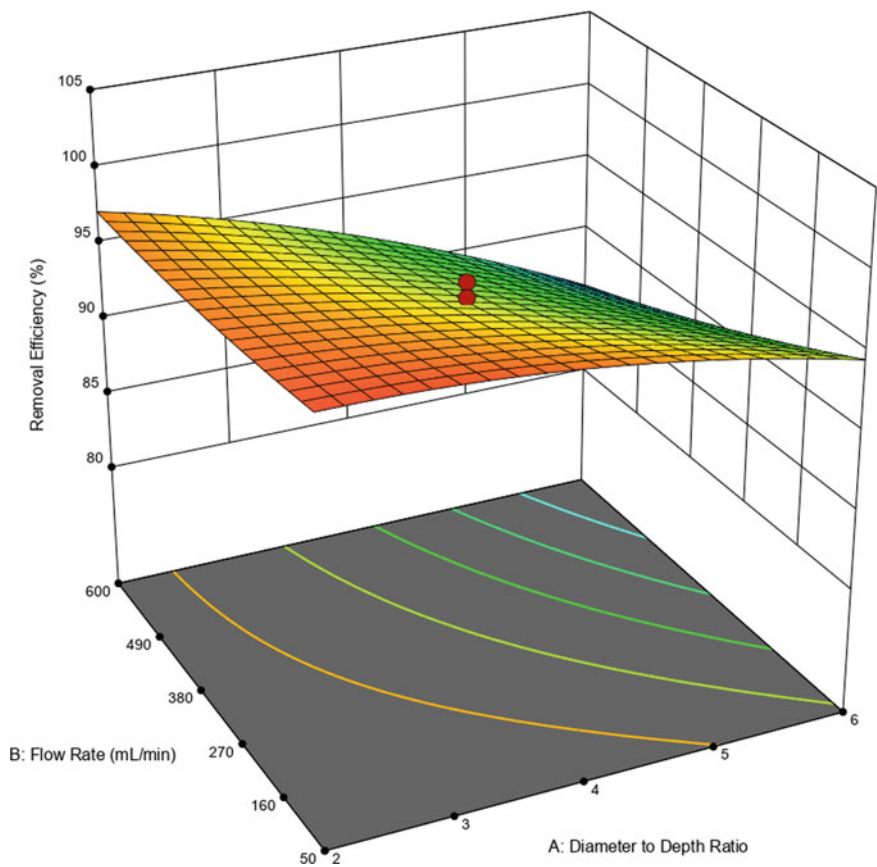
Regression coefficients	Value corresponding to response variable (dimensionless)	<i>p</i> -value
Intercept	94.74	
A-diameter-to-depth ratio	– 3.72	< 0.0001
B-flow rate	– 2.43	0.0004
C-influent concentration	– 0.6126	0.1249
AB	– 1.83	0.0045
AC	– 0.2121	0.6588
BC	– 0.2096	0.6625
$A^2$	– 1.03	0.0191
$B^2$	0.3988	0.4345
$C^2$	0.4175	0.2664
$R^2$	0.94	

value close to unity shows that a quadratic polynomial model can properly represent the effects of diameter-to-depth ratio, flow rate, and influent concentration on the response variable.

### 3.2 *Effect of Independent Variables on Response Variable*

Removal efficiency was successfully determined by using different levels of independent variables. The removal efficiency of Manganese depended on diameter-to-depth ratio due to its significant effect at linear ( $p < 0.0001$ ), quadratic ( $p < 0.05$ ), and interaction ( $p < 0.05$ ) levels. Diameter-to-depth ratio is the indicator of filtration bed volume that relates to the retention time of the raw water within the treatment system. Longer filtration material facilitates the oxidation-filtration process, which leads to better Manganese removal [16]. Another independent variable, which had a significant effect on Manganese removal efficiency, was flow rate ( $p < 0.001$ ).

The influence of diameter-to-depth ratio (*A*) and flow rate (*B*) on removal efficiency is illustrated in Fig. 3. Variable *A* exerts a quadratic effect on removal efficiency. At higher diameter-to-depth ratio, increase in removal efficiency was observed with the decrease of flow rate. This trend was observed due to the higher retention time and improved adsorption kinetics [17].



**Fig. 3** 3D graphic surface optimization of removal efficiency

**3.3 Verification of RSM Model**

Optimized filtration configurations were used to check the suitability of the model for prediction of response value. Optimized parameter conditions were validated by performing experiments. The predicted removal efficiency at optimized conditions was 95.19%. On the other hand, the experimental value at optimized conditions was 96.31% of removal efficiency. Experimental response value was well in agreement with predicted response value (Table 4).

**Table 4** Optimum conditions, experimental, and predicted values of optimized conditions

Optimum conditions		Actual levels
Diameter-to-depth ratio		4
Flow rate (mg/L)		255
Influent manganese concentration (mg/L)		0.5
Response	Predicted values	Experimental values
Removal efficiency (%)	95.19	96.31

## 4 Conclusion

In this study, the main outcome is a low-cost and effective Manganese removal system for the purpose of clean and safe drinking water. No specific solution is available to remove Manganese only; therefore, a detail investigation of removal of Manganese was performed to recommend a fit-for-purpose solution for the community residents. It was indicated that response surface methodology is a useful tool to optimize the design conditions of such gravity-based simple filtration systems. The laboratory experiments indicated that the diameter-to-depth ratio plays a key role in determination of the removal efficiency. The flow rate is also an important parameter, whereas the influent water Manganese concentration did not affect the removal efficiency significantly. This study also confirmed that the quadratic model was sufficient to describe and predict the response of removal efficiency with the change of independent variables. Optimized independent parameter levels were validated by performing experiments under optimized conditions in laboratory scale. The predicted response value at optimized conditions was 96.22% of removal efficiency while the experimental value at optimized conditions was 96.31% of removal efficiency. This research was intended to find a fit-for-purpose solution for a critical contaminant. However, there are certain advantage, disadvantage, limitations of the system and based on those, recommendations for future research. The advantages of this system include cost efficiency, easy to assemble, and easy to maintain. However, this system may not be suitable where other contaminants co-exist with Manganese. Also, regeneration requirements may vary based on the raw water quality. Limitations of this research are quite a few. Chemistry and waste management side of the proposed prototype was not investigated. The types and number of wastes to be generated and how to manage those wastes were not considered. Also, since the tests used water containing Manganese only, the effect on removal efficiency due to the coexistence of other parameters such as Calcium and Iron in water was not investigated.

**Acknowledgements** This research was funded by Interior Universities Research Coalition (IURC), and supported by Lheidli T'enneh First Nation, Stellat'en First Nation, Saiku'z First Nation and First Nation Health Authority.

## References

1. Lui E (2015) On notice for a drinking water crisis in Canada. The Council of Canadians, Ottawa
2. Canadian Infrastructure Report Card (2016) Informing the future. Ottawa
3. Department of Indian and Northern Affairs Canada (2011) National assessment of first nations water and wastewater systems. Department of Indian and Northern Affairs Canada, Orangeville
4. O'Connor DR (2002) Report of the walkerton inquiry: a strategy for safe drinking water, part one. Ontario Ministry of the Attorney General, Toronto
5. Indigenous Services Canada (ISC) (2020) Government of Canada. <https://www.sac-isc.gc.ca/eng/1506514143353/1533317130660>. Accessed 17 Feb 2020
6. HealthLinkBC (2019) Manganese in drinking water. HealthLink BC. <https://www.healthlinkbc.ca/healthlinkbc-files/manganese-drinking-water>
7. Clack Corporation (2021) Clack Corporation. Clack Corporation. [www.clackcorp.com](http://www.clackcorp.com). Accessed 01 Feb 2021
8. Knocke WR, Ramon JR, Thompson CP (1988) Soluble manganese removal on oxide-coated filter media. J Am Water Works Assoc 88:65–70
9. Markle PB, Knocke WR, Gallagher DL (1997) Method for coating filter media with synthetic manganese oxide. J Environ Eng 127:642–649
10. Tobiason JE et al (2008) Characterization and performance of filter media for manganese control. AWWA Research Foundation, Denver
11. Tobiason JE et al (2016) Manganese removal from drinking water sources. Curr Pollut Rep 2:168–177
12. Lenntech (2021) Water treatment and purification. Lenntech. <https://www.lenntech.com/Data-sheets/Clack-manganesegreensand-L.pdf>. Accessed 1 Feb 2021
13. Government of BC (2019) Guidance on manganese in drinking water. Government of BC, Victoria
14. Homayoonfal M, Khodaiyan F, Mousavi M (2015) Modelling and optimising of physico-chemical features of walnut-oil beverage emulsions by implementation of response surface methodology: effect of preparation conditions on emulsion stability. Food Chem 174(1):649–659
15. Myers RH, Montgomery DC, Anderson-Cook CM (2016) Response surface methodology: process and product optimization using designed experiments. John Wiley & Sons, Hoboken
16. Walkerton Clean Water Centre (2018) Evaluation of greensand filtration operation for the reduction of manganese. Government of Ontario, Hamilton
17. Coffey BM (1990) Removal of soluble iron and manganese from groundwater by chemical oxidation and oxide-coated multi-media filtration. Virginia Polytechnic Institute and State University, Blacksburg

# Biorefinery Paradigm in Wastewater Management: Opportunities for Resource Recovery from Aerobic Granular Sludge Systems



Oliver Terna Iorhemen and Sandra Ukaigwe

**Abstract** Aerobic granules, with a size range of 0.2–5 mm, are a dense colony of different strains of microorganisms held together by extracellular polymeric substances (EPS) secreted by the bacterial cells under special operational conditions. These granules are the principal components of the aerobic granular sludge (AGS) biotechnology. AGS has been widely and successfully applied for the treatment of municipal and different industrial wastewater streams in the past two decades and has the potential for successful application in cold regions. Moreover, in addition to efficient wastewater treatment, AGS also offers great opportunity for high-value resource recovery. AGS granules contain high phosphorus and high EPS contents. The high EPS content presents an opportunity to recover numerous high-value resources including xanthan and curdlan from AGS biosolids. Xanthan is widely used in the food, biomedical, and oil industries, as well as for soil strengthening in geotechnical applications. Curdlan is used in biomedical and pharmaceutical, food, cosmetic, as well as construction industries. These resources are currently in high demand globally and research efforts geared towards their recovery would be imperative to achieving the biorefinery concept in wastewater. Literature search shows that research is needed on the optimization of the biosynthesis of xanthan and curdlan in AGS bioreactors while maintaining the wastewater treatment capability of the granules. In addition, the market potentials of xanthan and curdlan needs to be determined in order to make their recovery from waste aerobic granules lucrative. This paper provides an overview of resources recoverable from waste aerobic granules.

**Keywords** Biorefinery paradigm • Wastewater management • Resource recovery • Aerobic granular sludge system

---

O. T. Iorhemen (✉)

School of Engineering, University of Northern British Columbia, Prince George, BC, Canada

e-mail: [oliver.iorhemen@unbc.ca](mailto:oliver.iorhemen@unbc.ca)

S. Ukaigwe

Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada

e-mail: [ukaigwe@ualberta.ca](mailto:ukaigwe@ualberta.ca)

© Canadian Society for Civil Engineering 2023

R. Gupta et al. (eds.), *Proceedings of the Canadian Society of Civil Engineering*

*Annual Conference 2022*, Lecture Notes in Civil Engineering 363,

[https://doi.org/10.1007/978-3-031-34593-7\\_84](https://doi.org/10.1007/978-3-031-34593-7_84)

1319

## 1 Introduction

Biological wastewater treatment is an established cost-effective treatment option to remove wastewater pollutants [1]. ASP has been the conventional biological municipal wastewater technology in the last century. However, ASP suffers from many drawbacks including, solids–liquid separation issues, large space requirement for secondary clarifiers, production of excess sludge, and limitations with removal of recalcitrants. The solids–liquid separation is particularly imperative to producing high-quality effluent [2]. To overcome these challenges, the aerobic granular sludge (AGS) biotechnology has emerged as a compelling alternative to the conventional ASP and currently considered one of the most promising biological municipal wastewater treatment technologies [3–5]. Since ASP has been successfully applied in cold regions, the AGS biotechnology can also be applied in cold regions for the treatment of different wastewater streams.

Aerobic granules, with a size range of 0.2–5 mm, are a dense colony of different strains of microbes held together by extracellular polymeric substances (EPS) secreted by the bacterial cells under special operational conditions [6, 7]. Granule formation occurs in four stages. The first stage of aerobic granule formation from activated sludge (as seed sludge) is the initial cell-to-cell attachment, stimulated by the physical motion within individual cells. This is followed by cell attachments and the formation of microaggregates. Subsequently, enhanced attachment by EPS production through quorum sensing (QS) occurs. And finally, granular sludge conformation and maturation by hydrodynamic shear forces ensures, allowing mass transfer behaviour in order to facilitate diverse metabolic pathways to obtain energy [8]. The lodgement of the self-immobilized AGS microorganisms in the EPS constitutes a tridimensional matrix without any supporting material. On the other hand, microorganisms are suspended in the wastewater in the ASP. While digesting the wastewater, the microorganisms collide with each other, forming large flocs, with larger capacity to degrade the biological components of the wastewater.

Compared to the conventional ASP, AGS exhibits better settleability, small footprints (compact plant), high biomass retention (hence less sludge production), ability to withstand high-strength wastewater and shock loadings [9], thus making the AGS biotechnology ideal for the treatment of different wastewater types. The biomass concentration in the ASP bioreactor is 2000–3000 mg/L, while AGS bioreactors are operated with biomass concentration above 8000 mg/L [4]. Moreover, AGS process simplify process design as the biological treatment and biomass separation occur in the same reactor, thus, secondary clarifiers are eliminated from the treatment process. However, the two major limitations reported for AGS biotechnology are long granulation time and granular instability for long-term operations [3]. Because AGS has uneven layered structure, diffusion restrictions cause microorganisms in the outer layers to consume a significant amount of nutrients, limiting nutrients and oxygen in the inner cells. Therefore, the metabolic activity of the inner layer cells deteriorates, causing the layer to disintegrate under the shearing force [10].

AGS has been successfully applied for the treatment of municipal wastewater, high-strength organic wastewater, high-strength ammonium wastewater, nutrients removal, sulphate and nuclear waste, landfill leachate, and bioremediation of xenobiotic compounds such as phenol, chlorinated phenols, pentachlorophenol, as well as biosorption of heavy metals [5–7, 11, 12].

In addition to efficient wastewater treatment, AGS biotechnology offers a great opportunity for resource recovery. This is mainly due to the high content of EPS in the granule matrix. Within the EPS components, there are numerous high-value resources including: Phosphorus (P), alginate-like exopolysaccharides (ALE), polyhydroxyalkanoates (PHA), and tryptophan that can be recovered, and whose recoverability has been extensively explored due to high demand and numerous industrial applications [13–16]. Moreover, aerobic granules also possess high capability to remove nutrients [4, 5, 17]. Furthermore, due to both microbial P removal and P precipitation, high P content is usually found within AGS granules [18].

Circular economy concept recently gained traction in the wastewater management industry. This concept seeks to keep materials and products in the economy as long as possible by promoting recycle and reuse. This implies waste should be treated as secondary raw materials that can be recycled to process and re-used [19]. In this context, wastewater reclamation and reuse offer an excellent option to increase our water resources. Reclaimed water finds numerous applications including: industrial agricultural, urban, and recreational applications as well as indirect potable reuse [20–22]. In addition to wastewater reclamation and reuse, other opportunities are also present such as nutrients (nitrogen and phosphorus) recovery [23–26] and renewable energy recovery [27–29].

The circular economy concept has given birth to the biorefinery concept in wastewater management. The biorefinery concept seeks to mine resources from wastewater, offering the opportunity of obtaining high-value products from wastewater. Thus, wastewater treatment plants are presently being designed as wastewater-resource factories inserted into circular cities; hence, wastewater is no longer considered a pollutant to be treated and discharged, rather it has become a valuable source of renewable energy, clean water, and more importantly raw material [16]. Studies have shown that waste AGS is one of the top valuable raw materials from biological wastewater treatment processes due to its significant reuse potentials [13].

Due to the unique structure of AGS, the application of the biorefinery concept to AGS-based wastewater treatment systems offers the possibility to recover P, ALE, PHA, tryptophan, polysaccharide-based biomaterial, low-cost adsorbents, and methane from the anaerobic digestion of waste AGS [13–16]. While additional resources have been recently identified in waste aerobic granules including; xanthan, curdlan, tyrosine, and phenylalanine [14, 15], there is however, no information in the literature on their recovery. This paper provides an overview of the potential to recover xanthan and curdlan from AGS waste granules.



## 2 EPS in the Granule Matrix

Aerobic granule structure exhibits high EPS content [4, 5]. EPS forms a hydrogel matrix as a dense network that contributes to the strength and stability of granules [16]. Studies indicate that the EPS matrix of AGS appears to depend on the wastewater composition, the types of granules, dominant microbial groups, and reactor operating conditions [4]. The high content of EPS in AGS offers a great opportunity to recover valuable resources from waste aerobic granules since EPS is a naturally occurring biopolymer that is both renewable and biodegradable [30]. EPS contains polysaccharides and proteins as the major components, the other components include lipids, humic, nucleic, and uronic acids [17].

## 3 Current State of Resource Recovery from AGS Biosolids

Current resource recovery options from AGS biosolids are outlined below.

### 3.1 Low-Cost Adsorbents

Waste aerobic granules have been widely applied as low-cost adsorbents for the biosorption of heavy metals and dyes [12]. This is made possible by the unique feature of aerobic granules, large surface area, high porosity, and high sorption efficiency [31, 32]. In addition, the different microbial populations in AGS granules can produce diverse surface functional groups that act as binding sites for heavy metals, thus making them excellent biosorbents for the removal of heavy metals [5, 12]. Heavy metals that have been successfully removed through biosorption using waste granules include zinc, lead, cadmium, nickel, chromium, copper, cobalt, iron, beryllium, manganese, strontium, antimony, and cerium [12]. The reuse of excess AGS granules as low-cost adsorbents would be an eco-friendly means of both heavy metal removal and AGS recycle. Biosorbents can also be converted into post-sorption added value products that can be applied to other uses. Additionally, heavy metal removal using biosorption not only serves to remove these toxic elements from the environment but also allows for their recovery, which increases interest in this process [33].

The underlying mechanisms involved in the biosorption of various heavy metals from wastewater using AGS granules has been widely studied [5]. For instance, [34] studied the kinetics of Zn(II) biosorption on AGS granule surface for removal of Zinc from industrial wastewater and obtained absorption capacity of 270 mg Zn(II)/g. Their results also showed that the kinetics of Zn(II) biosorption on the aerobic granule surface were related to both initial Zn(II) and AGS granule concentrations.

They concluded that AGS granules can be used as effective biosorbent for efficient removal of Zn(II) or other types of heavy metals from industrial wastewater. In another study, they investigated the biosorption ability of AGS granules for cadmium removal from industrial wastewater and found that the AGS granules have higher biosorption capacity of 566 mg Cd<sup>2+</sup>/g. The biosorption ability was also dependent on initial cadmium and AGS granules concentrations [35]. Wei et al. [36] also evaluated the applicability of AGS for Zn(II) removal from wastewater and obtained 80.12% total removal efficiency of Zn(II) through adsorption. In a different but related study, [37] compared the Zn<sup>2+</sup> adsorptive capacity of salinity-aided AGS granules with conventional AGS granules and reported high adsorptive capacity for conventional AGS; however, the adsorptive capacity for the salinity-aided AGS increased by 19.90%. Li et al. [38] evaluated the feasibility of using AGS as adsorbents for Ni(II) sorption from aqueous solution and concluded that AGS granules have 1.2 times higher adsorption capacity for Ni(II) than anaerobic granular sludge (AnGS) (65.77 vs. 54.18 mg/g) and significantly higher adsorption capacity than activated carbon (65.77 vs. 8.61 mg/g). Yang et al. [39] assessed the feasibility and compared the performance of algal–bacterial AGS and conventional bacterial AGS for treating Cr(VI)-containing wastewater and reported higher Cr(VI) biosorption capacity (9.60 vs. 8.49 mg/g) under the same test conditions for algal–bacterial AGS. However, both algal–bacterial and conventional bacterial AGS showed promising biosorbent for Cr(VI) removal despite the former displaying superiority in biosorption capacity. Other studies on the AGS adsorptive capacity for heavy metal removal include the removal of copper from industrial wastewater and leachates, arsenic removal from organic wastewater and lead removal from contaminated wastewater, and many more.

### 3.2 *Dewatering and Use as Manure*

With increased application of the AGS biotechnology, large amount of excess AGS wastes will definitely be generated. However, characterization of AGS wastes shows increased EPS content, a substance that can enhance the aggregation of soil particles and benefit plants by maintaining the moisture of the environment and trapping nutrients [40]. But more studies are required to evaluate the application of AGS wastes as manure including the economics of extraction and commercial viability. Currently, there is a general lack of understanding about the dewaterability of digested AGS and studies on the mechanisms of AGS dewatering are also sparse. Very few studies have also examined the impact of sludge dewatering procedures routinely employed in conventional WWTPs on the AGS dewatering performance, dewatering efficiency, and dewatering process optimization of AGS [41]. Thus, more investigations are also required in this area.

### 3.3 *Phosphorus*

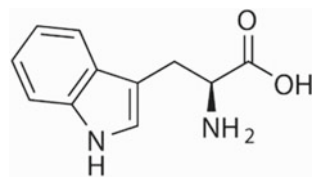
Biological wastewater treatment is currently a recognized recovery source for phosphorus. One of the key features of AGS systems is outstanding phosphorus removal [42–44]. The mechanisms of phosphorus removal in AGS systems include uptake by polyphosphate accumulating organisms (PAOs) through enhanced biological phosphorus removal (EBPR) and phosphorus precipitation in the granule matrix. It has been indicated that of the total phosphorus removed in AGS systems, uptake by PAOs constitute 55–73.7% [18, 45], while 45–70% is attributed to phosphorus precipitation in the aerobic granules [8, 45, 46].

To recover phosphorus via struvite crystallization in AGS bioreactors, Lu et al. [47] diverted granules to a side-stream batch reactor with a non-woven cloth filter and added a carbon source to enhance P release into the liquid fraction under anaerobic condition. About 93% of the total P contained in the wastewater was reportedly released as orthophosphate in the side-stream batch reactor. When the concentration of orthophosphate in the bulk liquid of the side-stream batch reactor became high, the phosphorus precipitated, was separated from the sludge granules, and returned to the main reactor [47]. This system allows for phosphorus recovery without disrupting the operation of the AGS bioreactor. The potential of thermally treated AGS granules to act as a high-quality, slow-release P fertilizer in agriculture has also been tested. Liu et al. [48] reported that slow P release would be beneficial when applied to crops being cultivated as it allows crops grow at a more uniform rate and prevent P leaching to the environment. However, issues of heavy metal contamination need to be addressed. Recovered phosphorus can be used as fertilizer [26, 30] and as raw material in the soap and detergent, leather, textile and rubber, ceramics, and metal industries [13].

### 3.4 *ALE*

AGS technology produces ALE in abundance [49, 50]. ALE is considered a valuable bio-based product which has been considered a raw material in many industries and can also be used as a soil-enhancer (manure) to improve water retention in semi-arid environments [11, 16]. Diverse ALE extraction methods have been applied in AGS systems including: centrifugation, sonication, ethylenediaminetetraacetic acid (EDTA), formamide with sodium hydroxide (NaOH), formaldehyde with NaOH and sodium carbonate ( $\text{Na}_2\text{CO}_3$ ) with heat and constant mixing, etc., depending on the type of sludge and the EPS of interest [14]. ALE has been variously extracted from AGS [49]. In comparison with activated sludge, the ALE recovered from AGS biosolids was two times the amount recovered from activated sludge;  $160 \pm 4$  mg/g [volatile suspended solids (VSS) ratio] and  $72 \pm 6$  mg/g (VSS ratio) [51]. The difference in the chemical structure of the ALE from these two sources was noted which impacted their hydrogel and mechanical properties. Although other studies

**Fig. 1** Chemical structure of tryptophan



also found similarities in hydrogel properties for ALE extracted from AGS, activated sludge, and commercial alginate [50, 52]. AGS cultivated with different substrates (wastewater sources) yield ALE with different properties. High concentration of organic matter has been indicated to result in high ALE content [53]. When the feed to the AGs bioreactor was propionate, the ALE yield was 10% w/w (VSS based) [53], with synthetic saline wastewater as feed, the ALE yield was 5% w/w [38, 54], and real wastewater yielded 16% w/w (VSS based) [49, 51]. This implies the substrate type (influent wastewater composition) has a strong influence on the microbial community and the quantity and composition as well as hydrogel properties of the ALE. ALE is used as a thickener (food and textile industries), film-forming agent (food and papermaking industries), and gelling agent for medicine encapsulation and dressing production (pharmaceutical and biomedical industries) [55, 56].

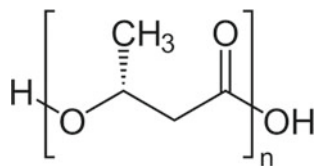
### 3.5 Tryptophan

Tryptophan is an amino acid that is widely used in numerous industries including: food (as preservative due to its antioxidant properties and as a supplement in animal feed), pharmaceutical (to treat depression, anxiety, and sleep disorder), and agricultural industries (foliar spray, seed priming, or soil application) [13]. Literature search shows that three methods have generally been applied in the production of tryptophan, microbial fermentation using glucose, xylose, galactose, etc., chemical synthesis from indole and biochemical synthesis from indole, pyruvate, and ammonia [13]. Numerous studies have reported the presence of tryptophan in the protein portion of EPS of aerobic granules [57–59]. In terms of tryptophan recovery, studies are scarce in the scientific literature. However, Rollemberg et al. [53] reported 0.048 g tryptophan/g VSS in the sludge discharged from AGS system (Fig. 1).

### 3.6 PHA

PHA are biopolymers which are applied in biomedicine to manufacture cardiovascular products and drug carriers, and offers potential application for treatment of injuries [60]. PHA in aerobic granules was reported to reach 40–70% of the cell dry weight [61]. AGS yields a high PHA content compared to activated sludge. 0.36 mg

**Fig. 2** Chemical structure of PHA



PHA/mg chemical oxygen demand (COD) [yield of 39% of cell dry weight (CDW)] was obtained in activated sludge [62], whereas 0.66 mg PHA/mg COD (yield of 68% of CDW) was obtained in AGS [63]. Presently, the prevailing PHAs production method is via pure microbial cultures. However, PHA production using mixed microbial cultures is increasingly being considered. AGS systems have great potential for PHAs recovery due to their high biomass retention capacity, which increases PHA productivity [13] (Fig. 2).

### 3.7 Biogas

Biogas production through the anaerobic digestion of waste AGS granules offers another opportunity for resource recovery from AGS systems. The energy recovered through this process would contribute to attaining a circular economy in the wastewater management industry. However, the major challenge may lie in the digestibility of aerobic granules. In this regard, [64] reported that the anaerobic biodegradability of biosolids from AGS systems achieved similar performance (33–49%) to that of waste activated sludge (30–50%). However, thermal pre-treatment was more effective (enhancement of up to 88%) at initially low biodegradability of the raw biosolids (~33%) than initially high biodegradability (49%) [64]. Val Del Río et al. [65] studied the feasibility of anaerobic digestion of raw biosolids from AGS systems, thermal pre-treated AGS (at 133 °C), and a mixture of thermal pre-treated AGS with primary sludge from a municipal wastewater treatment plant. They found that anaerobic digestion of biosolids from AGS systems achieved similar performance as that reported for waste activated sludge. Thermal pre-treated AGS had better performance by 32 and 47% in terms of biodegradability and solids reduction, respectively. The mixture of thermal pre-treated AGS with raw primary sludge achieved even better results, producing an increase of 17% of solids reduction with respect to the value obtained when only pre-treated AGS. Biodegradability, however, decreased from 58 to 53%.

Bernat et al. [66] conducted biogas potential tests under mesophilic conditions at different organic loading rates (OLRs) with biosolids from AGS systems, waste activated sludge, and their mixtures with primary sludge. The biosolids from AGS systems produced the lowest VS/TS ratio and hard-to-biodegrade lignin of about 54% of the fibrous materials. These characteristics resulted in low biogas yield (320–410 dm<sup>3</sup>/kg Total Solids), methane content 56.7–59.5% in the biogas [66]. A value 1.8 times lower than that of the waste activated sludge. The study also found that methane content in the biogas was higher when biosolids from AGS and waste activated sludge

were co-digested with primary sludge. Jahn et al. [41] investigated the biodegradation and methane yield of AGS under mesophilic conditions and hydraulic retention times (HRT) of 25 and 40 days. The study found the methane yield to be  $260 \text{ mL gVSS}^{-1}$  which was slightly more than that of activated sludge of  $240 \text{ mL gVSS}^{-1}$ .

Similarly, [67] studied the digestibility of biosolids from AGS system under mesophilic conditions. Results show that the sludge discharged after each cycle from an AGS system had biochemical methane potential (BMP) of  $296 \pm 15 \text{ mL CH}_4/\text{g VS}$  substrate, whereas the selectively discharged sludge from biomass growth achieved a BMP of  $194 \pm 10 \text{ mL CH}_4/\text{g VS}$  substrate. The BMP of the sludge discharged after each cycle was similar to that of primary sludge, while the BMP of the selectively discharged sludge due to biomass growth was lower than that of waste activated sludge ( $232 \pm 11 \text{ mL CH}_4/\text{g VS}$  substrate). Mechanical destruction of the compact structure of the selectively discharged sludge due to biomass growth only accelerated the methane production rate but did not significantly affect the BMP value. To increase methane production, steam explosion at  $170^\circ\text{C}$  was found to be an effective thermal pre-treatment for biosolids having low methane yield due to high mineral content [68].

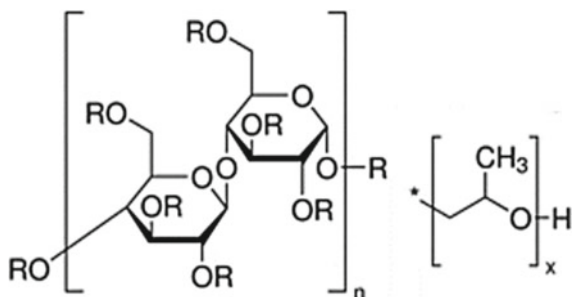
To increase the biodegradability of biosolids from AGS systems, some studies have adopted mechanical crushing of the granules. However, mechanical destruction only accelerates the rate of degradation of rapidly degradable organic matter and releases a fraction of the slowly biodegradable organics, but has no major effect on the BMP [65].

### 3.8 Digestate from Anaerobic Digestion

Digestate from the anaerobic digestion of waste activated sludge is typically stabilized and applied in agriculture. The digestate resulting from the anaerobic digestion of waste granules has potential application in agriculture. This constitutes another resource recovery option from AGS-based wastewater treatment systems. However, the main challenge in this area is the dewaterability of the digestate. There is not enough information available in the scientific literature on the fate of the digestate from the anaerobic digestion of biosolids from AGS reactors. Thus, it would be an interesting area to explore for biorefinery.

AGS processes have the potential to produce stabilized sludge due to selective sludge discharge that remove biomass with low sediment ability, while poor dewaterability has been reported for AS sludge. Selective sludge discharge overall prevents high biomass concentration within the reactor. Furthermore, AGS processes have lower biomass yield coefficient. The theoretical growth yields of AGS granules are estimated to be  $0.2\text{--}0.3 \text{ g VSS/g COD}_{\text{removed}}$  about 40% reduction in sludge production in comparison with conventional AS, which has a sludge growth yield of around  $0.45 \text{ g VSS/g COD}_{\text{removed}}$  and produces massive amounts of excess waste sludge produced [15]. Generally, the methods used for sludge handling in AGS process

**Fig. 3** Chemical structure of xanthan



are similar to those applied in AS processes with two major differences: significant reduction in polymeric substances required for the dewatering and reduced total volume of the excess sludge in comparison with AS [69].

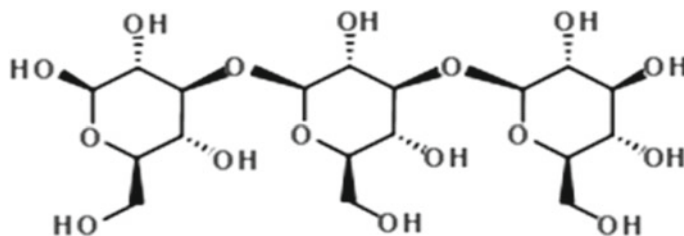
## 4 Potential for the High Production of Xanthan and Curdlan

### 4.1 Xanthan

Xanthan is produced by *Xanthomonas* sp. and has recently been identified in AGS [70, 71]. Xanthan is widely used in the food industry, in the biomedical industry for the production of hydrocolloid (waterproof dressing for wounds), in the oil industry for viscosity control (gelling and suspending agent), and for soil strengthening in geotechnical applications [72, 73]. Globally, the xanthan gum market was estimated at USD 987.7 million in 2020 [74]. Conventional xanthan gum production is costly because the process is energy-intensive, the raw materials (glucose or sucrose) are expensive, and the process requires specialized laboratories, as such, its production from wastes is particularly encouraged [75]. It is therefore, imperative to explore biosynthesis of xanthan in AGS bioreactors. The varying of operational factors such as feast-famine period ratio, organic loading rate, carbon-to-nitrogen ration, cycle time, and superficial up flow air velocity would present good starting point for studying Xanthan production from AGS (Fig. 3).

### 4.2 Curdlan

Curdlan is a water-insoluble linear homopolysaccharide polymer, produced by pathogenic bacteria such as *Agrobacterium* sp., *Pseudomonas* sp., and *Bacillus* sp. [76], with great significance in bacterial communities arising from its key role in inter-cellular communication and cellular coordination [77]. Curdlan is used in biomedical



**Fig. 4** Chemical structure of curdlan

and pharmaceutical industries for wound repair, anti-tumour, and anti-AIDS agents to inhibit HIV infection. In the food industries for thickening agent or fat-mimic substitute due to its unique rheological and thermal gelling properties [78], and in the cosmetic industry for moisturizing increase, as incense agents and for chemo-sensor preparations [76, 77]. Other uses of curdlan include pore-clogging (grouting) agent in soils, adsorbent for contaminated soil remediation, and superplasticizer in concrete mixtures due to its ability to absorb water [79]. Curdlan has also been reported to show no toxicity, carcinogenicity or any effect on animal reproduction [72]. With the approval of curdlan by the US Food and Drug Administration for use in the food industry in 2005, there is an expansion of demand for curdlan both in North America and Europe [80, 81]. However, the low conversion rate of glucose in curdlan production increases its overall production costs [80]. Thus, efforts need to be geared towards the production of curdlan in the granule matrix while maintaining wastewater treatment efficiency of the granules (Fig. 4).

## 5 Perspectives

While Sect. 4 focused on the technological aspect, the market potentials of xanthan and curdlan need to be determined in order to make their recovery from waste aerobic granules lucrative. In addition, the composition and concentration of each EPS extract are significantly impacted by the extraction and purification protocols [50]. As applied to xanthan and curdlan, their efficient extraction and identification in the EPS extract as well as the development of their purification methods are essential. It is not yet clear if the xanthan and curdlan extracted from waste granules would be purified to the extent of replacing conventionally produced xanthan and curdlan. Even in less stringent applications of xanthan (such as gelling and suspending agents for viscosity control in the oil industry and soil strengthening in geotechnical applications) and curdlan (such as grouting agents in soils, adsorbent for remediation of contaminated soils, and superplasticizer in concrete mixtures), the extraction and purification mechanism must be evaluated. Furthermore, studies are needed to understand the relationship between xanthan and AGS EPS as well as the relationship between curdlan and AGS EPS.



## 6 Conclusion

Current state of resource recovery from AGS systems has been reviewed. AGS has been successfully applied for the treatment of municipal and different industrial wastewater types, as a result of its many salient features. In addition to efficient wastewater treatment, AGS presents additional benefit of high-value resource recovery opportunities, owing to the high EPS content of the AGS granules. Research is at advanced stage on the recovery of high-value products including: phosphorus, alginate-like exopolysaccharides, polyhydroxyalkanoates, and tryptophan from AGS biosolids. Furthermore, xanthan and curdlan have recently been identified in waste aerobic granules as recoverable high-value products. Both products have myriad applications across many industries. However, despite the positive identification of xanthan and curdlan in AGS granules, their recovery mechanism is yet to be explored. Research is needed on the optimization of the biosynthesis of xanthan and curdlan in AGS bioreactors and the determination of market potentials of xanthan and curdlan in order to make their recovery from waste aerobic granules lucrative.

## References

1. Mittal A (2011) Biological wastewater treatment. *Water Today* 1:32–44
2. Hamza RA, Iorhemen OT, Tay JH (2016) Advances in biological systems for the treatment of high-strength wastewater. *J Water Process Eng* 10:128–142
3. Franca RDG, Pinheiro HM, Loosdrecht MCM, Lourenço ND (2018) Stability of aerobic granules during long-term bioreactor operation. *Biotechnol Adv* 36(1):228–246
4. Nanchaiaiah YV, Sarvajith M (2019) Aerobic granular sludge process: a fast growing biological treatment for sustainable wastewater treatment. *Curr Opin Environ Sci Health* 12:57–65
5. Purba LDA, Ibiyeye HT, Yuzir A, Mohamad SE, Iwamoto K, Zamyadi A, Abdullah N (2020) Various applications of aerobic granular sludge: a review. *Environ Technol Innov* 20:101045
6. de Sousa Rollemberg SL, Barros ARM, Firmino PIM, Dos Santos AB (2018) Aerobic granular sludge: cultivation parameters and removal mechanisms. *Bioresour Technol* 270:678–688
7. Zheng S, Lu H, Zhang G (2020) The recent development of the aerobic granular sludge for industrial wastewater treatment: a mini review. *Environ Technol Rev* 9(1):55–66
8. Sarma SJ, Tay J-H (2018) Carbon, nitrogen and phosphorus removal mechanisms of aerobic granules. *Crit Rev Biotechnol* 38(7):1077–1088
9. Adav SS, Lee D-J, Show K-Y, Tay J-H (2008) Aerobic granular sludge: recent advances. *Biotechnol Adv* 26(5):411–423
10. Chiu ZC, Chen MY, Lee DJ, Wang CH, Lai JY (2007) Oxygen diffusion and consumption in active aerobic granules of heterogeneous structure. *Appl Microbiol Biotechnol* 75(3):685–691
11. Nanchaiaiah YV, Reddy GKK (2018) Aerobic granular sludge technology: mechanisms of granulation and biotechnological applications. *Bioresour Technol* 247:1128–1143
12. Wang L, Liu X, Lee D-J, Tay J-H, Zhang Y, Wan C-L, Chen X-F (2018) Recent advances on biosorption by aerobic granular sludge. *J Hazard Mater* 357:253–270
13. Carvalho CA, Santos AF, Ferreira TJT, Lira VNSA, Barros ARM, Santos AB (2021) Resource recovery in aerobic granular sludge systems: is it feasible or still a long way to go? *Chemosphere* 274:129881
14. Felz S, Neu TR, van Loosdrecht MCM, Lin Y (2020) Aerobic granular sludge contains hyaluronic acid-like and sulfated glycosaminoglycans-like polymers. *Water Res* 169:115291

15. Ferreira TJJ, de Sousa Rollemberg SL, de Barros AN, de Lima JPM, Dos Santos AB (2021) Integrated review of resource recovery on aerobic granular sludge systems: possibilities and challenges for the application of the biorefinery concept. *J Environ Manag* 291:112718
16. Kehrein P, van Loosdrecht M, Osseweijer P, Posada J (2020) Exploring resource recovery potentials for the aerobic granular sludge process by mass and energy balances: energy, biopolymer and phosphorous recovery from municipal wastewater. *Environ Sci Water Res Technol* 6:2164–2179
17. Deng S, Wang L, Su H (2016) Role and influence of extracellular polymeric substances on the preparation of aerobic granular sludge. *J Environ Manage* 173:49–54
18. Huang W, Huang W, Li H, Lei Z, Zhang Z, Tay JH, Lee D-J (2015) Species and distribution of inorganic and organic phosphorus in enhanced phosphorus removal aerobic granular sludge. *Bioresour Technol* 193:549–552
19. Ghisellini P, Cialani C, Ulgiati S (2016) A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems. *J Clean Prod* 114:11–32
20. Ahmad NNR, Ang WL, Teow YH, Mohammad AW, Hilal N (2022) Nanofiltration membrane processes for water recycling, reuse and product recovery within various industries: a review. *J Water Process Eng* 45:102478
21. Jeffrey P, Yang Z, Judd SJ (2022) The status of potable water reuse implementation. *Water Res* 15:118198
22. Lahlou F-Z, Mackey HR, Al-Ansari T (2022) Role of wastewater in achieving carbon and water neutral agricultural production. *J Clean Prod* 339:130706
23. Chrispim MC, Scholz M, Nolasco MA (2019) Phosphorus recovery from municipal wastewater treatment: critical review of challenges and opportunities for developing countries. *J Environ Manag* 248:109268
24. Saliu TD, Oladoja NA (2021) Nutrient recovery from wastewater and reuse in agriculture: a review. *Environ Chem Lett* 19:2299–2316
25. Ye Y, Ngo HH, Guo W, Chang SW, Nguyen DD, Zhang X, Zhang J, Liang S (2020) Nutrient recovery from wastewater: from technology to economy. *Bioresour Technol Rep* 11:100425
26. Yuan Z, Pratt S, Batstone DJ (2012) Phosphorus recovery from wastewater through microbial processes. *Curr Opin Biotechnol* 23(6):878–883
27. Campana PE, Mainardis M, Moretti A, Cottes M (2021) 100% renewable wastewater treatment plants: techno-economic assessment using a modelling and optimization approach. *Energy Convers Manag* 239:114214
28. Paucar NE, Sato C (2021) Microbial fuel cell for energy production, nutrient removal and recovery from wastewater: a review. *Processes* 9(8):1318
29. Periyasamy S, Temesgen T, Karthik V, Beula Isabel J, Kavitha S, Rajesh Banu J, Sivashanmugam P (2022) Chapter 17: wastewater to biogas recovery. In: An A, Tyagi V, Kumar M, Cetecioglu Z (eds) *Clean energy and resource recovery*. Elsevier, Amsterdam, pp 301–314
30. Dall'Agnol P, Junior NL, Muller JM, Xavier JA, Domingos DG, da Costa RHR (2020) A comparative study of phosphorus removal using biopolymer from aerobic granular sludge: a factorial experimental evaluation. *J Environ Chem Eng* 8(2):103541
31. Ahn KH, Hong SW (2015) Characteristics of the adsorbed heavy metals onto aerobic granules: isotherms and distributions. *Desal Water Treat* 53(9):2388–2402
32. Liu Y, Tay J-H (2004) State of the art of biogranulation technology for wastewater treatment. *Biotechnol Adv* 22(7):533–563
33. Kanamarlapudi SLRK, Chintalapudi VK, Muddada S (2018) Application of biosorption for removal of heavy metals from wastewater. In: *Biosorption* JD, Vrana, B (eds) Biosorption. IntechOpen, London
34. Liu Y, Yang S-F, Tan S-F, Lin Y-M, Tay J-H (2002) Aerobic granules: A novel zinc biosorbent. *Lett. Appl Microbiol* 35(6):548–551
35. Liu Y, Yang S-F, Xu H, Woon K-H, Lin Y-M, Tay J-H (2003) Biosorption kinetics of cadmium (ii) on aerobic granular sludge. *Process Biochemistry*, 38(7):997–1001
36. Wei D, Li M, Wang X, Han F, Li L, Guo J, Ai L, Fang L, Liu L, Du B, Wei Q (2016) Extracellular polymeric substances for zn (ii) binding during its sorption process onto aerobic granular sludge. *J Hazard Mater* 301:407–415

37. Wu, X, Li W, Ou D, Li C, Hou M, Li H, Liu Y (2019) Enhanced adsorption of  $\text{Zn}^{2+}$  by salinity-aided aerobic granular sludge: Performance and binding mechanism. *J Environ Manage* 242:266–271
38. Li X, Luo J, Guo G, Mackey HR, Hao T, Chen G (2017) Seawater-based wastewater accelerates development of aerobic granular sludge: a laboratory proof-of-concept. *Water Res* 115:210–219
39. Yang X, Zhao Z, Yu Y, Shimizu K, Zhang Z, Lei Z, Lee D-J (2020) Enhanced biosorption of  $\text{Cr}(\text{VI})$  from synthetic wastewater using algal-bacterial aerobic granular sludge: Batch experiments, kinetics and mechanisms. *Sep Purif Technol* 251:117323
40. Pronk M, De Kreuk M, De Bruin B, Kamminga P, Kleerebezem RV, Van Loosdrecht M (2015) Full scale performance of the aerobic granular sludge process for sewage treatment. *Water Res* 84:207–217
41. Jahn L, Saracevic E, Svardal K, Krampe J (2019) Anaerobic biodegradation and dewaterability of aerobic granular sludge. *J Chem Technol Biotechnol* 94(9):2908–2916
42. de Kreuk MK, Heijnen JJ, van Loosdrecht MCM (2005) Simultaneous COD, nitrogen, and phosphate removal by aerobic granular sludge. *Biotechnol Bioeng* 90(6):761–769
43. Khan MZ, Mondal PK, Sabir S (2013) Aerobic granulation for wastewater bioremediation: a review. *Can J Chem Eng* 91(6):1045–1058
44. Yilmaz G, Lemaire R, Keller J, Yuan Z (2008) Simultaneous nitrification, denitrification, and phosphorus removal from nutrient-rich industrial wastewater using granular sludge. *Biotechnol Bioeng* 100(3):529–541
45. Mañas A, Biscans B, Spérandio M (2011) Biologically induced phosphorus precipitation in aerobic granular sludge process. *Water Res* 45(12):3776–3786
46. Filali A, Mañas A, Mercade M, Bessière Y, Biscans B, Spérandio M (2012) Stability and performance of two GSBs operated in alternating anoxic/aerobic or anaerobic/aerobic conditions for nutrient removal. *Biochem Eng J* 67:10–19
47. Lu Y-Z, Wang H-F, Kotsopoulos TA, Zeng RJ (2016) Advanced phosphorus recovery using a novel SBR system with granular sludge in simultaneous nitrification, denitrification and phosphorus removal process. *Appl Microbiol Biotechnol* 100(10):4367–4374
48. Liu Q, Wan J, Wang J, Li S, Dagot C, Wang Y (2017) Recovery of phosphorus via harvesting phosphorus-accumulating granular sludge in sequencing batch airlift reactor. *Bioresour Technol* 224:87–93
49. Lin Y, de Kreuk M, Van Loosdrecht M, Adin A (2010) Characterization of alginate-like exopolysaccharides isolated from aerobic granular sludge in pilot-plant. *Water Res* 44(11):3355–3364
50. Schambeck CM, Girbal-Neuhauser E, Böni L, Fischer P, Bessière Y, Paul E, da Costa RHR, Derlon N (2020) Chemical and physical properties of alginate-like exopolymers of aerobic granules and flocs produced from different wastewaters. *Bioresour Technol* 312:123632
51. Lin YM, Sharma PK, Van Loosdrecht MCM (2013) The chemical and mechanical differences between alginate-like exopolysaccharides isolated from aerobic flocculent sludge and aerobic granular sludge. *Water Res* 47:57–65
52. Sam SB, Dulekgurgen E (2016) Characterization of exopolysaccharides from floccular and aerobic granular activated sludge as alginate-like-exops. *Desalin Water Treat* 57(6):2534–2545
53. Yang Y-C, Liu X, Wan C, Sun S, Lee D-J (2014) Accelerated aerobic granulation using alternating feed loadings: alginate-like exopolysaccharides. *Bioresour Technol* 171:360–366
54. Meng F, Liu D, Pan Y, Xi L, Yang D, Huang W (2019) Enhanced amount and quality of alginate-like exopolysaccharides in aerobic granular sludge for the treatment of salty wastewater. *BioResources* 14:139–165
55. Moradali MF, Ghods S, Rehm BH (2018) Alginate biosynthesis and biotechnological production. *Alginates and their biomedical applications*. Springer, New York, pp 1–25
56. Szekalska M, Puciłowska A, Szymańska E, Ciosek P, Winnicka K (2016) Alginate: current use and future perspectives in pharmaceutical and biomedical applications. *Int J Polym Sci* 2016:7697031
57. Hamza RA, Sheng Z, Iorhemen OT, Zaghoul MS, Tay JH (2018) Impact of food-to-microorganisms ratio on the stability of aerobic granular sludge treating high-strength organic wastewater. *Water Res* 147:287–298

58. Li Z, Lin L, Liu X, Wan C, Lee D-J (2020) Understanding the role of extracellular polymeric substances in the rheological properties of aerobic granular sludge. *Sci Total Environ* 705:135948
59. Rollemberg SLS, dos Santos AF, Ferreira TJT, Firmino PIM, dos Santos AB (2021) Evaluation of the production of alginate-like exopolysaccharides (ale) and tryptophan in aerobic granular sludge systems. *Bioprocess Biosyst Eng* 44(2):259–270
60. Mozejko-Ciesielska J, Kiewisz R (2016) Bacterial polyhydroxyalkanoates: still fabulous? *Microbiol Res* 192:271–282
61. Wang S, Shi W, Tang T, Wang Y, Zhi L, Lv J, Li J (2017) Function of quorum sensing and cell signaling in the formation of aerobic granular sludge. *Funct Quorum Sens Cell Sign Format Aerobic Granul Sludge* 16(1):1–13
62. Waller JL, Green PG, Loge FJ (2012) Mixed-culture polyhydroxyalkanoate production from olive oil mill pomace. *Bioresour Technol* 120:285–289
63. Gobi K, Vadivelu VM (2014) Aerobic dynamic feeding as a strategy for in situ accumulation of polyhydroxyalkanoate in aerobic granules. *Bioresour Technol* 161:441–445
64. Val del Río A, Morales N, Isanta E, Mosquera-Corral A, Campos JL, Steyer JP, Carrère H (2011) Thermal pre-treatment of aerobic granular sludge: impact on anaerobic biodegradability. *Water Res* 45(18):6011–6020
65. Val Del Río Á, Palmeiro-Sanchez T, Figueroa M, Mosquera-Corral A, Campos JL, Méndez R (2014) Anaerobic digestion of aerobic granular biomass: effects of thermal pre-treatment and addition of primary sludge. *J Chem Technol Biotechnol* 89(5):690–697
66. Bernat K, Cydzik-Kwiatkowska A, Wojnowska-Baryła I, Karczewska M (2017) Physicochemical properties and biogas productivity of aerobic granular sludge and activated sludge. *Biochem Eng J* 117:43–51
67. Guo H, van Lier JB, de Kreuk M (2020) Digestibility of waste aerobic granular sludge from a full-scale municipal wastewater treatment system. *Water Res* 173:115617
68. Liu Y, Nilsen PJ, Maulidiany ND (2019) Thermal pretreatment to enhance biogas production of waste aerobic granular sludge with and without calcium phosphate precipitates. *Chemosphere* 234:725–732
69. Zhou J, Zheng G, Zhang X, Zhou L (2014) Influences of extracellular polymeric substances on the dewaterability of sewage sludge during bioleaching. *PLoS ONE* 9(7):e102688
70. Weissbrodt D, Neu T, Kuhlicke U, Rappaz Y, Holliger C (2013) Assessment of bacterial and structural dynamics in aerobic granular biofilms. *Front Microbiol* 4:175
71. Zhang Z, Yu Z, Wang Z, Ma K, Xu X, Alvarezc PJJ, Zhu L (2019) Understanding of aerobic sludge granulation enhanced by sludge retention time in the aspect of quorum sensing. *Bioresour Technol* 272:226–234
72. Chang I, Lee M, Tran ATP, Lee S, Kwon Y-M, Im J, Cho G-C (2020) Review on biopolymer-based soil treatment (bpst) technology in geotechnical engineering practices. *Transp Geotech* 24:100385
73. Petri DFS (2015) Xanthan gum: a versatile biopolymer for biomedical and technological applications. *J Appl Polym Sci* 132:1–13
74. Rončević Z, Grahovac J, Dodić S, Vučurović D, Dodić J (2019) Utilisation of winery wastewater for xanthan production in stirred tank bioreactor: bioprocess modelling and optimisation. *Food Bioprod Process* 117:113–125
75. Palaniraj A, Jayaraman V (2011) Production, recovery and applications of xanthan gum by *Xanthomonas campestris*. *J Food Eng* 106(1):1–12
76. Yang M, Zhu Y, Li Y, Bao J, Fan X, Qu Y, Wang Y, Hu Z, Li Q (2016) Production and optimization of curdlan produced by *Pseudomonas* sp. Q1212. *Int J Biol Macromol* 89:25–34
77. Zhang R, Edgar KJ (2014) Properties, chemistry, and applications of the bioactive polysaccharide curdlan. *Biomacromol* 15(4):1079–1096
78. Jindal N, Singh Khattar J (2018) Chapter 4: microbial polysaccharides in food industry. In: Grumezescu AM, Holban AM (eds) *Biopolymers for food design*. Academic Press, New York, pp 95–123

79. Spicer EJJ, Goldenthal EI, Ikeda T (1999) A toxicological assessment of curdlan. *Food Chem Toxicol* 37(4):455–479
80. Yuan M, Fu G, Sun Y, Zhang D (2021) Biosynthesis and applications of curdlan. *Carbohydr Polym* 273:118597
81. Zhai W, Danjo T, Iwata T (2017) Synthesis and physical properties of curdlan branched ester derivatives. *J Polym Res* 25(3):181

# Effects of Microplastic Size on Oil Dispersion in Oceans



Min Yang, Baiyu Zhang, Hemeihui Zhao, Chushi Wang, and Bing Chen

**Abstract** Oil spills are a major concern in oceans. Chemical dispersants are widely used as effective oil spill treating agents in the marine environment, especially in cold regions. Microplastics (MPs) are widely observed in oil-polluted oceans. Understanding how MPs affect oil dispersion thus becomes essential. Recent studies presented the negative influence of MPs on oil dispersion effectiveness due to the existence of MP-oil-dispersant agglomerates (MODAs). However, it is still unclear how MPs with various particle sizes would interact with oils during dispersion using different dispersant-to-oil volumetric ratios (DORs). Our study explored the effects of MPs with five sizes on oil dispersion under two DORs. Polyethylene (PE) MPs (7, 13, 23, 40, 90  $\mu\text{m}$ ) were applied. Newfoundland offshore and low sulfur crude oils were used. Corexit EC9500A was selected as the chemical dispersant. Results indicated that MPs affected oil dispersion effectiveness under various conditions owing to MODA resurfacing. The dispersion effectiveness of low sulfur oil at DOR 1:100 was between  $21.43 \pm 3.72$  and  $35.47 \pm 3.06\%$ , with MP size rising. With Newfoundland offshore oil, the dispersion effectiveness was greater than Low Sulfur oil. It increased obviously from  $57.62 \pm 3.46$  to  $88.42 \pm 14.23\%$  under DOR 1:100 with MP size rising. The oil droplet size of Newfoundland offshore was smaller than that of low sulfur under both DORs. Findings from this research will provide fundamental data for future decision-making on oil spill response in cold regions in the presence of MPs.

**Keywords** Microplastics · Oil spill · Chemical dispersant · Oil droplet size · Oil dispersion effectiveness · Cold region

---

M. Yang · B. Zhang (✉) · B. Chen  
Northern Region Persistent Organic Pollutant Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL 1B 3X5, Canada  
e-mail: [bzhang@mun.ca](mailto:bzhang@mun.ca)

H. Zhao · C. Wang  
Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL 1B 3X5, Canada

## 1 Introduction

Microplastics (MPs) (plastic particles < 5 mm) have been widely distributed in the marine environment including the Arctic Ocean [1, 2]. A recent study estimated that ~24.4 trillion MP particles existed in the upper oceans [3]. Moreover, MPs generated from face masks might pose a new threat to the marine environment during the COVID pandemic [4]. Approximately, 1.56 billion face masks were estimated to enter our oceans in 2020 [5]. Owing to their small size, MPs could be easily bioaccumulated in marine organisms [6, 7]. Since MPs have a large specific surface area and surface roughness, they could adsorb organic contaminants such as polycyclic aromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs) [8, 9]. For instance, the PAH concentration adsorbed on plastic particles collected from the North Pacific Gyre was between 4 and 846 ng/g, while the level of PAHs in the seawater column was 0.3–9 ng/L [10]. MPs would carry organic pollutants in the oceans, thereby posing potential toxic effects on different marine species [9, 11, 12].

With the detection of MPs in oil-contaminated oceans, the interaction between MPs and oil gradually attracted researchers' attention [13]. Accidental oil spills are a rising concern due to offshore oil exploration and transportation [14]. In the past decades, chemical dispersants have been applied as one of the major treatment technologies during oil spill response operations [15]. Chemical dispersants can be efficiently applied in harsh environments such as cold water [15, 16]. They broke oil slick into tiny oil droplets, promoting oil biodegradation in oceans [17]. The dispersant efficiency was affected by mixing energy, salinity, dispersant-to-oil volumetric ratio (DOR), and suspended particles such as MPs [18, 19]. For example, by increasing MP concentration, the dispersion effectiveness of a heavy oil decreased 38.25% from  $58.38 \pm 3.01$  to  $20.13 \pm 6.36\%$  [13]. The reduction of oil dispersion effectiveness was caused by the formation and resurfacing of MP-oil-dispersant agglomerates (MODAs). Researchers found that MPs with different sizes would form two classes of MODAs, i.e., MODA-1 (MPs covered by oil) and MODA-2 (MPs and oil droplets embedded) [20]. Different MODAs may affect oil dispersion effectiveness differently. However, rare studies explored how different MP sizes would affect oil dispersion in oceans.

This study aimed at investigating the effects of MP size (7–90  $\mu\text{m}$ ) on oil droplet size and oil dispersion effectiveness under two DORs (1:25 and 1:100). The DORs of 1:25 and 1:100 were selected to simulate the surface and subsea dispersant application, respectively. This study provided fundamental data to assess the impact of MPs with various sizes on dispersant application during oil spill responses. Findings will facilitate researchers to explore MPs-oil interactions and their environmental risks in the marine environment.

## 2 Methodology

### 2.1 Chemicals and Materials

Different crude oils, i.e., light oil (Newfoundland offshore) and heavy oil (low sulfur), were used as example oils. Corexit EC9500A was adopted as the chemical dispersant. Polyethylene (PE) MPs with particle sizes of 7, 13, 23, 40, 90  $\mu\text{m}$  were acquired from Micro Powders Inc. (New York, USA). The PE MPs were selected as they were one of the most abundant MPs in oceans. Dichloromethane (DCM) and the sea salt were bought from Millipore Sigma (Ontario, Canada).

The synthetic seawater (salinity 34 g/L) was made by dissolving the sea salt in ultrapure water and passed through a 0.20  $\mu\text{m}$  membrane filter to remove suspended particles. The salinity was measured by a conductivity portable meter.

### 2.2 Experimental Design

Experiments regarding the effects of MP size on oil droplet size and oil dispersion effectiveness were run in the MP-oil dispersion system proposed by [13]. In general, the 100  $\mu\text{L}$  oil was released into 120 mL synthetic seawater in a baffled flask, followed by the addition of Corexit EC9500A to reach a DOR of 1:25 and 1:100. After that, 400 mg/L of MPs were added to the baffled flask. The baffled flask was shaken at 200 rpm for 10 min and kept still for another 10 min. Collecting 30 mL sample (sample-A) for oil dispersion effectiveness measurement and collecting another 5 mL sample (sample-B) for oil droplet size detection. Transferring sample-A to a separatory funnel and extracting oil in sample-A three times with 5 mL DCM. The extraction was adjusted to 50 mL to measure oil concentration through the GENESYS 10S Ultraviolet (UV)-VIS spectrophotometer (Thermo Fisher Scientific, Hillsboro, Oregon, USA). Sample-B was diluted 20 times with ultrapure water and released into the chamber of the Laser In Situ Scattering and Transmissometry (LISST-200X) (Sequoia Scientific, Bellevue, Washington, USA) for oil droplet size measurement. Triplicate experiments were conducted.

### 2.3 Analysis of Oil Droplet Size and Oil Dispersion Effectiveness

The oil droplet size was detected by LISST-200X. The detection limitation was between 1 and 500  $\mu\text{m}$ . It could generate both mean particle size and particle size distribution. This study targeted analyzing the mean oil droplet size.



The oil dispersion effectiveness was equal to the mass of oil dispersed divided by the total mass of oil added. The mass of oil dispersed was measured by the UV–VIS spectrophotometer at 340, 370, and 400 nm wavelengths. DCM was used as a background for UV measurement. Calculation details were as follows [21].

$$\text{Area} = \frac{(\text{Abs}_{340} + \text{Abs}_{370}) \times 30}{2} + \frac{(\text{Abs}_{370} + \text{Abs}_{400}) \times 30}{2} \quad (1)$$

$$\text{Oil dispersed (g)} = \frac{\text{Area}}{\text{Calibration curve slope}} \times V_{\text{DCM}} \times \frac{V_{\text{tw}}}{V_{\text{ew}}}, \quad (2)$$

where the  $\text{Abs}_{340}$ ,  $\text{Abs}_{370}$ , and  $\text{Abs}_{400}$  are absorbance at 340, 370, and 400 nm, respectively.  $V_{\text{DCM}}$  is the DCM extraction volume;  $V_{\text{tw}}$  is the total seawater volume (i.e., 120 mL) in the baffled flask;  $V_{\text{ew}}$  is the seawater volume (i.e., 30 mL) collected for oil extraction.

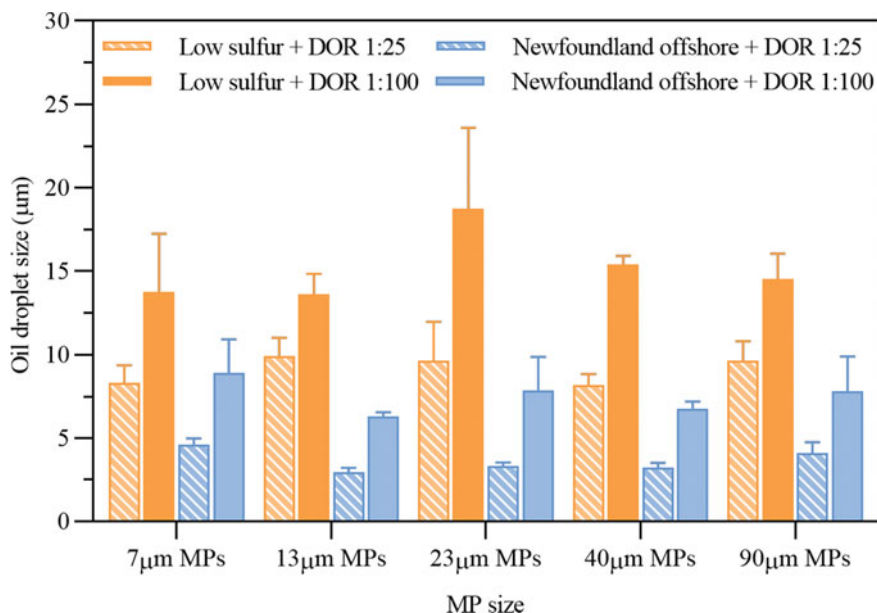
## 2.4 Quality Assurance and Quality Control (QA/QC)

QA/QC was followed to guarantee the reliability of the experimental results. The experiments on oil droplet size and oil dispersion effectiveness measurement used the standard and widely reported methodologies, and all experiments were conducted in triplicate [20, 21]. The coefficient of determination ( $R^2$ ) for the Newfoundland offshore oil calibration curve was 0.9987, and that of low sulfur oil was 0.9998, indicating the good fit of linear regression.

# 3 Results and Discussion

## 3.1 Effects of MP Size on Oil Droplet Size

Effects of the MP size on oil droplet size at DORs of 1:25 and 1:100 were investigated. In Fig. 1, by increasing MP size from 7 to 90  $\mu\text{m}$ , the variation of droplet size of Newfoundland offshore oil showed a similar trend under different DORs. The maximum oil droplet size under DOR 1:25 and 1:100 occurred at 7  $\mu\text{m}$  MPs, and the minimum value was at 13  $\mu\text{m}$  MPs. However, the decrease of DOR increased the oil droplet size. Under the DOR of 1:25, the droplet size of Newfoundland offshore oil was between  $2.94 \pm 0.26$  and  $4.60 \pm 0.37$   $\mu\text{m}$ . By reducing DOR to 1:100, the droplet size of Newfoundland offshore oil ranged from  $6.31 \pm 0.23$  to  $8.91 \pm 2.00$   $\mu\text{m}$ . This finding was similar to previous reports [15, 22, 23]. It suggested that dispersant dosage was essential in determining oil droplet size, regardless of the existence of MPs.



**Fig. 1** Oil droplet size under different MP sizes

In addition, the droplet size of low sulfur oil did not show a specific trend with MP size rising. Under the DOR of 1:25, the maximum droplet size appeared at 13  $\mu\text{m}$  MPs, and the minimum droplet size was at 40  $\mu\text{m}$  MPs. Under the DOR of 1:100, the maximum droplet size was at 23  $\mu\text{m}$  MPs, and the minimum value was at 13  $\mu\text{m}$  MPs. Nevertheless, the droplet size of low sulfur oil was larger than that of Newfoundland offshore oil. Under the DOR of 1:25, the oil droplet size kept stable between  $8.17 \pm 0.66$  and  $9.92 \pm 1.09$   $\mu\text{m}$ . Under the DOR of 1:100, the oil droplet size was between  $13.60 \pm 1.24$  and  $18.75 \pm 4.84$   $\mu\text{m}$ . This result was consistent with previous studies without MPs addition [18, 24, 25]. Dispersants were more effective on lower viscosity oils. Since the viscosity of low sulfur oil was higher than that of Newfoundland offshore oil, low sulfur oil was more difficult to be dispersed into small droplets than Newfoundland offshore oil [15].

### 3.2 Effects of MP Size on Oil Dispersion Effectiveness

Effects of the MP size on oil dispersion effectiveness at DORs of 1:25 and 1:100 was explored. Light oil and heavy oil performed differently under various MP sizes. The dispersion effectiveness of low sulfur oil kept stable with MP size rising under different DORs. The dispersion effectiveness of Newfoundland offshore oil kept steady under a high DOR, but it increased with MP size rising under a low DOR.

As shown in Fig. 2, the dispersion effectiveness of low sulfur oil under a DOR of 1:25 floated between  $45.58 \pm 6.51$  and  $62.02 \pm 4.72\%$  by increasing MP size from 7 to 90  $\mu\text{m}$ . The maximum dispersion effectiveness was at 40  $\mu\text{m}$  MPs. The minimum dispersion effectiveness was at 7  $\mu\text{m}$  MPs, indicating 7  $\mu\text{m}$  MPs have the more significant negative impact on oil dispersion. The decrease in oil dispersion effectiveness was caused by the resurfacing of 7  $\mu\text{m}$  MPs formed MODAs, which brought oil onto seawater surface and caused the reduction of oil concentration in seawater [20]. Meanwhile, under a DOR of 1:100, the dispersion effectiveness of low sulfur oil was between  $21.43 \pm 3.72$  and  $35.47 \pm 3.06\%$ . The minimum was at 23  $\mu\text{m}$  MPs, and the maximum was at 90  $\mu\text{m}$  MPs.

For Newfoundland offshore oil, its dispersion effectiveness was much greater than low sulfur oil. Owing to the lower viscosity of Newfoundland offshore oil, it was easier to be dispersed in seawater [13, 14]. Under a DOR of 1:25, the dispersion effectiveness of Newfoundland offshore oil ranged from  $107.97 \pm 4.31$  to  $122.72 \pm 4.65\%$ . The minimum dispersion effectiveness was at 13  $\mu\text{m}$  MPs, and the maximum was at 7  $\mu\text{m}$  MPs. Meanwhile, under a DOR of 1:100, the dispersion effectiveness of Newfoundland offshore oil increased obviously from  $57.62 \pm 3.46$  to  $88.42 \pm 14.23\%$ , with MP size rising. It suggested that MPs with a smaller size such as 7  $\mu\text{m}$  MPs had a greater negative effect on oil dispersion effectiveness owing to the formation and resurfacing of MODAs in seawater.

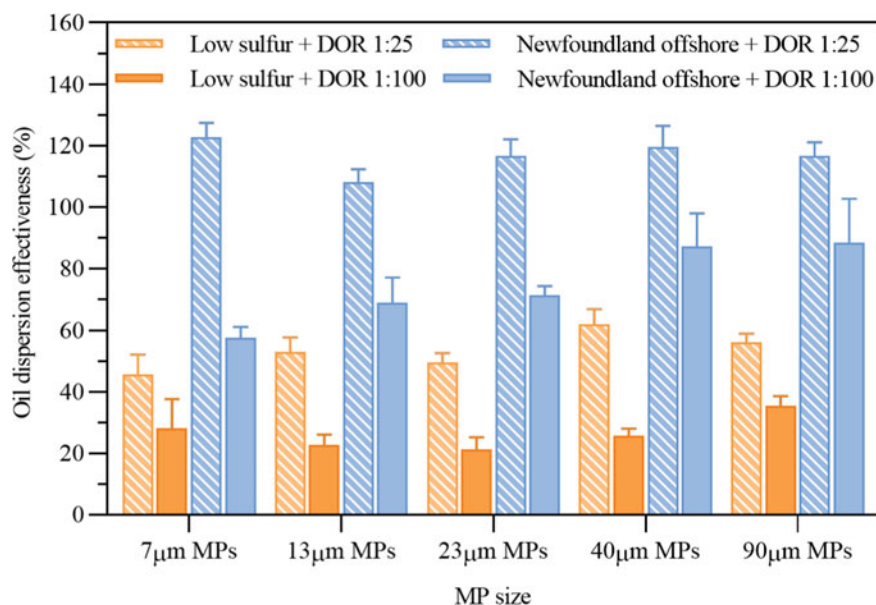


Fig. 2 Oil dispersion effectiveness under different MP sizes

## 4 Conclusion

This study investigated the effects of MP size on oil droplet size and oil dispersion effectiveness in the marine environment. The variation of droplet size of Newfoundland offshore oil showed a similar trend under different DORs, and the maximum oil droplet size was reached at the MP size of 7  $\mu\text{m}$ . Moreover, under the same DOR, the droplet size generated by low sulfur oil was larger than that of Newfoundland offshore oil in the presence of MPs. It suggested that dispersant has a better dispersion efficiency on Newfoundland offshore. In addition, the increase of MP size has less impact on the dispersion effectiveness of low sulfur under a low DOR (1:100), while its effects on Newfoundland offshore oil were more prominent. The decrease of oil dispersion effectiveness was caused by the formation and resurface of MODAs, which brought oil onto seawater surface and resulted in the reduction in oil concentration in seawater.

Although studies have explored the role of MPs on oil dispersion in oceans, rare studies discussed how MP size affected oil dispersion efficiency. By revealing the impact of MPs on offshore oil spill treatments, this study provides new ideas to develop predictive models to guide decision-making, which is crucial for marine environmental protection. Moreover, MP-oil interactions explored in this study would lay a foundation for the development of potential pollutant removal technologies in oceans. This study will fill the knowledge gaps in MP-oil interactions and inspire researchers to better understand the environmental risk assessment of MP-oil co-contaminant in oceans.

**Acknowledgements** This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Multi-Partner Research Initiative (MPRI)/Fisheries and Oceans Canada (DFO).

## References

1. Ross PS, Chastain S, Vassilenko E, Etemadifar A, Zimmermann S, Quesnel S-A, Eert J, Solomon E, Patankar S, Posacka AM (2021) Pervasive distribution of polyester fibres in the arctic ocean is driven by Atlantic inputs. *Nat Commun* 12(1):1–9
2. Thompson RC, Olsen Y, Mitchell RP, Davis A, Rowland SJ, John AW, McGonigle D, Russell AE (2004) Lost at sea: where is all the plastic? *Science* 304(5672):838–838
3. Isobe A, Azuma T, Cordova MR, C  zar A, Galgani F, Hagita R, Kanhai LD, Imai K, Iwasaki S, Kako SI (2021) A multilevel dataset of microplastic abundance in the world's Upper Ocean and the Laurentian Great Lakes. *Microplast Nanoplast* 1(1):1–14
4. Wang Z, An C, Chen X, Lee K, Zhang B, Feng Q (2021) Disposable masks release microplastics to the aqueous environment with exacerbation by natural weathering. *J Hazard Mater* 417:126036
5. Bondaroff TP, Cooke S (2020) Masks on the beach: the impact of COVID-19 on marine plastic pollution. *OceansAsia*, London
6. Cole M, Lindeque P, Fileman E, Halsband C, Goodhead R, Moger J, Galloway TS (2013) Microplastic ingestion by zooplankton. *Environ Sci Technol* 47(12):6646–6655

7. Moore R, Loseto L, Noel M, Etemadifar A, Brewster J, MacPhee S, Bendell L, Ross P (2020) Microplastics in beluga whales (*Delphinapterus leucas*) from the Eastern Beaufort Sea. *Mar Pollut Bull* 150:110723
8. Lee H, Shim WJ, Kwon JH (2014) Sorption capacity of plastic debris for hydrophobic organic chemicals. *Sci Total Environ* 470:1545–1552
9. Wang F, Wong CS, Chen D, Lu X, Wang F, Zeng EY (2018) Interaction of toxic chemicals with microplastics: a critical review. *Water Res* 139:208–219
10. Mendoza LMR, Jones PR (2015) Characterisation of microplastics and toxic chemicals extracted from microplastic samples from the North Pacific Gyre. *Environ Chem* 12(5):611–617
11. Lin W, Li X, Yang M, Lee K, Chen B, Zhang BH (2018) Brominated flame retardants, microplastics, and biocides in the marine environment: recent updates of occurrence, analysis, and impacts. *Adv Mar Biol* 81:167–211
12. Zarfl C, Matthies M (2010) Are marine plastic particles transport vectors for organic pollutants to the arctic? *Mar Pollut Bull* 60(10):1810–1814
13. Yang M, Chen B, Xin X, Song X, Liu J, Dong G, Lee K, Zhang B (2021) Interactions between microplastics and oil dispersion in the marine environment. *J Hazard Mater* 403:123944
14. Zhang B, Matchinski EJ, Chen B, Ye X, Jing L, Lee K (2019) World seas: an environmental evaluation. Elsevier, Amsterdam, pp 391–406
15. Merlin F, Zhu Z, Yang M, Chen B, Lee K, Boufadel MC, Isaacman L, Zhang B (2021) Dispersants as marine oil spill treating agents: a review on mesoscale tests and field trials. *Environ Syst Res* 10(1):1–19
16. Belore RC, Trudel K, Mullin JV, Guarino A (2009) Large-scale cold water dispersant effectiveness experiments with Alaskan crude oils and Corexit 9500 and 9527 dispersants. *Mar Pollut Bull* 58(1):118–128
17. Brakstad OG, Davies EJ, Ribicic D, Winkler A, Brønner U, Netzer R (2018) Biodegradation of dispersed oil in natural seawaters from Western Greenland and a Norwegian Fjord. *Polar Biol* 41(12):2435–2450
18. Pan Z, Zhao L, Boufadel MC, King T, Robinson B, Conmy R, Lee K (2017) Impact of mixing time and energy on the dispersion effectiveness and droplets size of oil. *Chemosphere* 166:246–254
19. Yang M, Zhang B, Chen Y, Xin X, Lee K, Chen B (2021) Impact of microplastics on oil dispersion efficiency in the marine environment. *Sustainability* 13(24):13752
20. Yang M, Zhang B, Xin X, Liu B, Zhu Z, Dong G, Zhao Y, Lee K, Chen B (2022) Microplastic-oil-dispersant agglomerates in the marine environment: formation mechanism and impact on oil dispersion. *J Hazard Mater* 426:127825
21. Venosa A, Holder E (2011) Laboratory-scale testing of dispersant effectiveness of 20 oils using the baffled flask test. In: US environmental protection agency, pp 600–699
22. Board OS, Council NR (2005) Oil spill dispersants: efficacy and effects, national. Academies Press, New York
23. Board OS, National Academies of Sciences Engineering, and Medicine (2020) The use of dispersants in marine oil spill response. National Academies Press, Washington, DC
24. POLARIS Applied Sciences I (2013) A comparison of the properties of diluted bitumen crudes with other oils
25. Song X, Chen B, Liu B, Lye LM, Ye X, Nyantekyi-Kwakye B, Zhang B (2022) Impacts of frazil ice on the effectiveness of oil dispersion and migration of dispersed oil. *Environ Sci Technol* 56(2):835–844