DATA-DRIVEN APPROACHES FOR WASTEWATER MODELING

DEVELOPMENT OF DATA-DRIVEN APPROACHES FOR WASTEWATER MODELING

By PENGXIAO ZHOU, B.Eng., B.Mgt., M.A.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

Doctor of Philosophy (2023) McMaster University

(Civil Engineering) Hamilton, Ontario, Canada

TITLE: Development of Data-Driven Approaches for Wastewater Modeling

AUTHOR: Pengxiao Zhou

M.A.Sc. (McMaster University)

B.E., B.Mgt. (Jilin University)

SUPERVISOR: Zhong Li

NUMBER OF PAGES: XIII, 120

Lay Abstract

Ensuring appropriate treatment and recycling of wastewater is vital to sustain life. Wastewater treatment plants (WWTPs), which have complicated processes that include several intricate physical, chemical, and biological procedures, play a significant role in the water recycling. Due to stricter regulations and complex wastewater composition, the wastewater treatment system has become increasingly complex. Therefore, it is crucial to use simplified versions of the system, known as wastewater modeling, to effectively operate and manage the complex system. The aim of this thesis is to develop data-driven approaches for wastewater modeling.

Abstract

To effectively operate and manage the complex wastewater treatment system, simplified representations, known as wastewater modeling, are critical. Wastewater modeling allows for the understanding, monitoring, and prediction of wastewater treatment processes by capturing intricate relationships within the system. Process-driven models (PDMs), which rely on a set of interconnected hypotheses and assumptions, are commonly used to capture the physical, chemical, and biological mechanisms of wastewater treatment. More recently, with the development of advanced algorithms and sensor techniques, data-driven models (DDMs) that are based on analyzing the data about a system, specifically finding relationships between the system state variables without relying on explicit knowledge of the system, have emerged as a complementary alternative. However, both PDMs and DDMs suffer from their limitations. For example, uncertainties of PDMs can arise from imprecise calibration of empirical parameters and natural process variability. Applications of DDMs are limited to certain objectives because of a lack of high-quality dataset and struggling to capture changing relationship. Therefore, this dissertation aims to enhance the stable operation and effective management of WWTPs by addressing these limitations through the pursuit of three objectives: (1) investigating an efficient data-driven approach for uncertainty analysis of process-driven secondary settling tank models; (2) developing data-driven models that can leverage sparse and imbalanced data for the prediction of emerging contaminant removal; (3) exploring an advanced data-driven model for influent flow rate predictions during the COVID-19 emergency.

Acknowledgments

Four years have passed in the blink of an eye. This thesis marks the end of this precious four-year journey. The completion of this thesis is attributed to the support of many individuals.

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Zhong Li, for her invaluable guidance and support. Throughout my entire graduate career, she has carefully directed my research direction, provided me constructive feedbacks, and helped me to continuously improve. Her expertise, dedication, patience, and willingness to help have made this journey enjoyable. I would also like to thank my committee members, Dr. Sekerinski Emil and Dr. Spencer Snowling, for their insightful comments and suggestions, which have improved the quality of my work. Their expertise has been a great source of inspiration to me, and I am grateful for the opportunity to have worked with them. I am also grateful to the dedicated faculty and staff at McMaster University, who have provided me with invaluable assistance.

My parents' unwavering love and support have been a constant source of strength for me. I consider myself fortunate to have them in my life, and their presence has greatly influenced my personal development. I owe a great deal of gratitude to my friends, groupmates and colleagues for their support and encouragement. I would like to express my special thanks to my partner Qiulin Ma, who has been by my side for ten years. Her love and understanding have been the greatest source of strength throughout this journey.

As this four-year journey draws to a close, I extend my heartfelt thanks to everyone who has played a part in making this journey rich and fulfilling.

Declaration of Academic Achievement

This thesis is a "sandwich" thesis which contains two published peer-reviewed journal articles and two publishable manuscripts.

Article I: Zhou, P. and Li, Z., 2023. Arbitrary polynomial chaos expansion for uncertainty analysis of the one-dimensional hindered-compression continuous settling model. Journal of Water Process Engineering, 52, p.103489.

Article II: Zhou, P., Li, Z., El-Dakhakhni, W. and Smyth, S.A., 2022. Prediction of bisphenol A contamination in Canadian municipal wastewater. Journal of Water Process Engineering, 50, p.103304.

Article III: Zhou, P., Li, Z., Snowling, S., and Barclay, J., 2023. Unraveling the impact of COVID-19 lockdowns on Canadian municipal sewage. (Submitted).

Article IV: Zhou, P., Li, Z., Zhang, Y., Snowling, S., and Barclay, J., 2023. Adapting to a new normal: online machine learning for stream wastewater influent flow rate prediction in the era of COVID-19. (Submitted).

This thesis does not include two additional finished works on the applications of data-driven models in environmental fields that were produced during the Ph.D. study.

Article V: Zhou, P., Li, Z., Snowling, S., Goel, R. and Zhang, Q., 2022. Multi-step ahead prediction of hourly influent characteristics for wastewater treatment plants: a case study from North America. Environmental Monitoring and Assessment, 194(5), p.389.

Article VI: Zhou, P., Li, C., Li, Z. and Cai, Y., 2022. Assessing uncertainty propagation in hybrid models for daily streamflow simulation based on arbitrary polynomial chaos expansion. Advances in Water Resources, 160, p.104110.

Table of Content

List of Figures

List of Tables

Major Notations and Acronym

WWTP(s)	Wastewater treatment plant(s)
PDM(s)	Process-driven model(s)
SST	Secondary Settling Tank
ASM(s)	Activated sludge model(s)
MCS	Monte Carlo Simulation
DDM(s)	Data-driven model(s)
aPCE	Arbitrary polynomial chaos expansion
COVID-19	Coronavirus disease 2019

Chapter 1 – Introduction

Water on Earth is limited and goes through a continuous cycle. To maintain life, it is essential to ensure that wastewater is properly treated and reused. Wastewater treatment plant (WWTP) plays a vital role in the water cycle, as it removes pollutants from wastewater and facilitate the reuse of resources. Wastewater treatment, involving a combination of intricate physical, chemical, and biological processes, has become an increasingly complex system due to stricter regulations and more complicated wastewater composition. To effectively operate and manage the complex system, the utilization of simplified representations of the system, known as wastewater modeling, is crucial.

Wastewater modeling is important as it helps to understand, monitor, predict the behavior of wastewater treatment processes. Specifically, by creating an accurate model of a treatment process, it captures a complex relationship in the system, such as how different operational conditions will affect the process performance. The captured relationship can be used to keep track of treatment performance, and process control and optimization can be achieved by making adjustments accordingly. Predictions based on the models about how the process performance will respond to different scenarios can also help to make informed decisions and minimize the risk of process failures. The significant benefits of wastewater modeling, which comprises two distinct categories (i.e., process-driven and data-driven models), make it a vital tool in the field of wastewater management.

1.1 Process-driven wastewater models

Process-driven models (PDMs), or process-based/mechanistic models, refer to the mathematical representations that capture the fundamental functions of well-delimited processes or systems. PDMs often rely on a set of interconnected hypotheses and assumptions that are

tailored to capture behaviors of complex systems. Typically, PDMs are usually formulated using ordinary and partial differential equations, which describe how the system changes over time and space (Clark et al., 2011).

Traditional wastewater models are mostly PDMs developed based on the physical, chemical, and biological mechanisms of wastewater treatment. As an example, Secondary Settling Tank (SST) model is a commonly used PDM based on physical mechanisms in the field of wastewater modeling. SST is a unit process that enables the settling of biomass through the force of gravity, and it is the most frequently utilized solid-liquid separation system in the wastewater treatment industry (David et al., 2009; Li and Stenstrom, 2014). Clarification and thickening are two distinct functions of SST: clarification aims to remove the dispersed solids from the liquid; thickening is the process of compressing sludge for recycling or disposing. To adequately depict the clarification-thickening process, a process-driven SST model was firstly proposed by Kynch (Kynch, 1952). The sludge settling process was expressed by a one-dimensional mass balance partial differential equation. On the basis of experimental observations, further studies adding improvements were conducted: Petty (1975) extended the Kynch theory for continuous simulation; Takács et al. (1991) simulated a SST layer by layer, and the model has been widely used in commercial modeling tools such as GPS-X; Bürger et al. (2011) improved the model by facilitating reliable simulations.

Activated sludge model (ASM) based on chemical and biological mechanisms is another representation of PDMs in wastewater modeling field. Activated sludge systems have been used for a long time for municipal wastewater treatment, and researchers have been trying to develop various models for describing activated sludge processes over the last few decades. In 1982, to unify the research groups that work independently at an international level, the task group on

mathematical modeling for design and operation of biological wastewater treatment was established. Their joint work unites the concept, terminology, and notation that were used by researchers who work on activated sludge modeling and thus greatly accelerates the development of activated sludge models (ASMs). Models like ASM1, ASM2, ASM2D, and ASM3, that were systematically summarized by the task group, laid a solid foundation for future development and application of these PDMs (Henze et al., 2006).

While PDMs aid in comprehending various wastewater treatment processes, such as settling dynamics, potential challenges can arise from uncertainties due to imprecise calibration of empirical parameters and natural process variability. The industrial and academic communities have identified the need for uncertainty quantification. The study of uncertainty quantification in the wastewater modeling field is far less advanced in comparison with other fields, although there were researchers who tried to narrow the gap (Belia et al., 2009). Plósz et al. (2011) assessed the uncertainty originating from the SST model structure using Monte Carlo Simulation (MCS), and the influence of the SST sub-models on the plant-wide model performance was evaluated. Ramin et al. (2014) conducted a global sensitivity analysis to an SST model, and it was illustrated that the settling parameters were as influential as the biokinetic parameters on the uncertainty of the plant-wide WWTP model. Li and Stenstrom (2016) provided the sensitivity analysis of a one-dimensional SST model, and the uncertainty was quantified by MCS after selecting parameters based on sensitivity analysis. Sin et al. (2009) conducted uncertainty analysis for an ASM1 model and discussed the interpretation of uncertainty analysis results. Three different scenarios incorporating different types of uncertainty were used, and the results indicated that both biokinetic and hydraulic parameters induced significant uncertainty. Although MCS was again used, the sampling method was changed from random sampling to Latin hypercube sampling. Mannina et

al., 2010) conducted an uncertainty analysis of a membrane bioreactor using the Generalised Likelihood Uncertainty Estimation (GLUE) method, which is based on MCS. The same GLUE method was used again by Mannina et al. (2012) for the uncertainty analysis at a larger wastewater treatment plant. A modified model based on ASM1 and ASM2 was studied, and the study identified that model results strongly depended on the parameter ranges and the selected parameters. Mannina et al. (2018) continued to implement a sensitivity and uncertainty analysis for an ASM2D model. The sensitivity analysis was conducted by the Standardized Regression Coefficients (SRC) method, and 45 of the 122 model parameters were selected for uncertainty quantification based on MCS.

Currently, MCS remains the dominant method for uncertainty analysis of PDMs in the wastewater modeling field. MCS is a type of numerical simulation method. It originated in the 1940s and was first systematically proposed and applied in the Manhattan Project for the development of atomic bombs (Ditlevsen and Madsen, 1996). MCS has been widely used in many fields because it is simple and convenient to implement. Specifically, it has no special requirements on the form, dimension, and distribution of input variables; when there are enough random samples, it can guarantee the high accuracy of the estimation results (Rubinstein and Kroese, 2016). MCS usually entails the following steps: (1) define a domain of possible inputs; (2) generate inputs randomly from a probability distribution over the domain; (3) perform a deterministic computation on the input samples derived from the probability distribution; (4) aggregate the results. To achieve a certain level of precision, the computational complexity of MCS would grow exponentially with the increase in the number of inputs. Meanwhile, MCS assumes an exact probability density function for each uncertain variable and parameter in the modeling system, which is often unknown in real-world engineering applications. As a result, a more efficient and advanced method

for uncertainty analyses is desired for PDMs. Although there are already some data-driven approaches for uncertainty analysis in the literature that demonstrate advantages over Monte Carlo simulation, it is necessary to further investigate their performance on wastewater models.

1.2 Data-driven wastewater models

Although PDMs allow engineers to better understand the complex physical, chemical and biological processes in a wastewater treatment system, its real-world application can be challenging in certain situations. For instance, the target system could be too complicated to build a PDM for; the parameters of PDMs could be extremely hard to calibrate or validate; the system might be under human interference that cannot be address by PDMs. With the development of advanced algorithms and sensor techniques, data-driven models (DDMs) can be a feasible alternative to address these challenges. DDM is based on analyzing the data about a system, specifically finding relationships between the system state variables (input, internal and output variables) without relying on explicit knowledge of the physical behaviour of the system. A machine-learning algorithm using a representative training data set that contains all the behaviour found in the system is usually used to determine the relationship between a system's inputs and outputs.

Wastewater influent flow prediction is a typical example where DDMs can outperform PDMs. Building a PDM to simulate the municipal sewer system and predict the influent characteristics is theoretically feasible. However, several factors limit the application of PDMs. For instance, urban hydrological processes such as snowmelt and infiltration are often too difficult to simulate, domestic water usage patterns that rely on human activities are challenging to illustrate, and the aging pipes and sewer connections can make the physical parameter measurements extremely difficult (Zhang et al., 2019). Over the past two decades, DDMs have been successfully

applied to predict the influent characteristics as a substitute for traditional PDMs, due to their ability to overcome the abovementioned difficulties. El-Din and Smith (2002) proposed an artificial neural network model for short-term influent flow rate prediction at a WWTP in Canada. Rainfall data and historical influent flow data were used as the predictors. Satisfactory prediction results were obtained through the validation, and their results showed that the influent flow rate at the WWTP increased rapidly during extreme events. Wei et al. (2012) used the multi-layer perceptron model for influent flow rate prediction. Their study for the first time included the spatial information of weather data, and the results showed that the model could generate satisfactory predictions up to 150 minutes ahead. Kusiak et al. (2013) developed four data-mining algorithms, including multi-layer perceptron, classification and regression tree, multivariate adaptive regression spline, and random forest, for the prediction of daily influent carbonaceous biochemical oxygen demand (CBOD). The results illustrated that the performance was better when the CBOD values were high. In addition, the k-nearest neighbor (KNN) method was used by Kim et al. (2016) to predict the influent flow rate, chemical oxygen demand, suspended solid, total nitrogen, and total phosphorus in dry weather and wet weather separately. It was found that the KNN method was reliable for influent flow rate and water qualities prediction in dry weather, while the results suggested that the prediction should be made with caution in wet weather. Zhou et al. (2019) developed random forest regression models for daily influent flow rate prediction and a feature importance measurement were introduced for a further understanding of the wastewater influent mechanisms.

DDMs have been proven effective in different areas of wastewater modeling, due to their exceptional ability to capture highly nonlinear relationships. However, there are several challenges associated with the further application of DDMs in the wastewater modeling field. For example,

conventional DDMs require a significant amount of high-quality data. As a result, there have been limited studies investigating the application of DDMs for subjects without adequate data such as emerging contaminants. Because collecting emerging contaminant data consumes significant labour resources, there is a lack of high-quality dataset or balanced dataset on emerging contaminants. Thus, the development of advanced DDMs that can tolerate sparse and imbalanced datasets for subjects like emerging contaminants is valuable. Additionally, DDMs struggle to capture dynamic relationships. As the relationship between a system's inputs and outputs changes, the DDMs must be re-trained, which could dramatically increase computational burden. Specifically, despite the existence of numerous wastewater influent characteristic prediction models, the current applications are unable to effectively handle influent data streams with changing targeted relationships caused by emergency such as COVID-19 pandemic.

1.3 Objectives and organization

The aim of this dissertation is to enhance the stable operation and effective management of WWTPs through the development of data-driven approaches, which will be achieved by pursuing the following three objectives:

(1) Investigating an efficient data-driven approach for uncertainty analysis of SST models.

(2) Developing data-driven models that can leverage sparse and imbalanced data for the prediction of emerging contaminant removal.

(3) Exploring an advanced data-driven model for stream influent flow rate predictions, which involves two tasks: (a) assessing the impact of COVID-19 lockdowns on influent flow rate using data-driven models. (b) developing adaptive data-driven models that can adapt to changing patterns in streaming wastewater data

In Chapter 2, an arbitrary polynomial chaos expansion approach which involves data-driven procedures, is developed for parameter uncertainty analysis of the process-driven one-dimensional continuous settling model.

In Chapter 3, three data-driven models, the Multitask Shared Layer Neural Network (MLT-NN), Genetic Programming (GP), and Extra Trees (ET), are proposed for predicting the removal of Bisphenol A. These models aim to address data imbalance, enhance model interpretability, and evaluate feature importance, respectively.

In Chapter 4, the impact of the COVID-19 lockdowns on Canadian sewage is examined by analyzing the influent flow rates at two wastewater treatment plants located in Ontario. A thorough comparison of weekly patterns and daily average flow rates before and during lockdowns is conducted. Additionally, the observed influent flow rates are also compared with predicted no-lockdown scenario data, which are generated by random forest models.

In Chapter 5, a set of online learning models, including Adaptive Random Forest (aRF), Adaptive K-Nearest Neighbors (aKNN), and Adaptive Multi-Layer Perceptron (aMLP), are developed for wastewater influent prediction under the drastic changes caused by the COVID-19 pandemic. Their performance is compared with that of conventional batch learning models, including Random Forest (RF), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) at two Canadian WWTPs.

In chapter 6, the conclusions and contributions of this dissertation are summarized, and suggestions for future research are offered.

Reference

Belia, E., Amerlinck, Y., Benedetti, L., Johnson, B., Sin, G., Vanrolleghem, P.A., Gernaey, K. v, Gillot, S., Neumann, M.B., Rieger, L., 2009. Wastewater treatment modelling: dealing with uncertainties. Water Science and Technology 60, 1929–1941.

Bürger, R., Diehl, S., Nopens, I., 2011. A consistent modelling methodology for secondary settling tanks in wastewater treatment. Water Res 45, 2247–2260. https://doi.org/10.1016/J.WATRES.2011.01.020

David, R., Saucez, P., Vasel, J.L., vande Wouwer, A., 2009. Modeling and numerical simulation of secondary settlers: A Method of Lines strategy. Water Res 43, 319–330. https://doi.org/10.1016/J.WATRES.2008.10.037

Ditlevsen, O., Madsen, H.O., 1996. Structural reliability methods. Wiley New York.

El-Din, A.G., Smith, D.W., 2002. A neural network model to predict the wastewater inflow incorporating rainfall events. Water Res 36, 1115–1126.

Henze, M., Gujer, W., Mino, T., van Loosedrecht, M., 2006. Activated sludge models ASM1, ASM2, ASM2d and ASM3.

Kim, M., Kim, Y., Kim, H., Piao, W., Kim, C., 2016. Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. Front Environ Sci Eng 10, 299–310. https://doi.org/10.1007/s11783-015-0825-7

Kusiak, A., Verma, A., Wei, X., 2013. A data-mining approach to predict influent quality. Environ Monit Assess 185, 2197–2210. https://doi.org/10.1007/S10661-012-2701-2

Kynch, G.J., 1952. A theory of sedimentation. Transactions of the Faraday Society 48, 166–176. https://doi.org/10.1039/TF9524800166

Li, B., Stenstrom, M.K., 2016. Practical identifiability and uncertainty analysis of the one-dimensional hindered-compression continuous settling model. Water Res 90, 235–246. https://doi.org/10.1016/J.WATRES.2015.12.034

Li, B., Stenstrom, M.K., 2014. Dynamic one-dimensional modeling of secondary settling tanks and design impacts of sizing decisions. Water Res 50, 160–170. https://doi.org/10.1016/J.WATRES.2013.11.037

Mannina, G., Cosenza, A., Viviani, G., 2012. Uncertainty assessment of a model for biological nitrogen and phosphorus removal: Application to a large wastewater treatment plant. Physics and Chemistry of the Earth, Parts A/B/C 42, 61–69.

Mannina, G., Cosenza, A., Viviani, G., Ekama, G.A., 2018. Sensitivity and uncertainty analysis of an integrated ASM2d MBR model for wastewater treatment. Chemical Engineering Journal 351, 579–588. https://doi.org/10.1016/j.cej.2018.06.126

Mannina, G., di Bella, G., Viviani, G., 2010. Uncertainty assessment of a membrane bioreactor model using the GLUE methodology. Biochem Eng J 52, 263–275.

Petty, C.A., 1975. Continuous sedimentation of a suspension with a nonconvex flux law. Chem Eng Sci 30, 1451–1458. https://doi.org/10.1016/0009-2509(75)85022-6

Plósz, B.Gy., de Clercq, J., Nopens, I., Benedetti, L., Vanrolleghem, P.A., 2011. Shall we upgrade one-dimensional secondary settler models used in WWTP simulators? – An assessment of

model structure uncertainty and its propagation. Water Science and Technology 63, 1726–1738. https://doi.org/10.2166/wst.2011.412

Ramin, E., Flores-Alsina, X., Sin, G., Gernaey, K. v., Jeppsson, U., Mikkelsen, P.S., Plósz, B.G., 2014. Influence of selecting secondary settling tank sub-models on the calibration of WWTP models – A global sensitivity analysis using BSM2. Chemical Engineering Journal 241, 28–34. https://doi.org/10.1016/J.CEJ.2013.12.015

Rubinstein, R.Y., Kroese, D.P., 2016. Simulation and the Monte Carlo method. John Wiley & Sons.

Sin, G., Gernaey, K. v, Neumann, M.B., van Loosdrecht, M.C.M., Gujer, W., 2009. Uncertainty analysis in WWTP model applications: a critical discussion using an example from design. Water Res 43, 2894–2906.

Takács, I., Patry, G.G., Nolasco, D., 1991. A dynamic model of the clarification-thickening process. Water Res 25, 1263–1271. https://doi.org/10.1016/0043-1354(91)90066-Y

Wei, X., Kusiak, A., Sadat, H.R., 2012. Prediction of Influent Flow Rate: Data-Mining Approach. Journal of Energy Engineering 139, 118–123. https://doi.org/10.1061/(ASCE)EY.1943-7897.0000103

Zhang, Q., Li, Z., Snowling, S., Siam, A., El-Dakhakhni, W., 2019. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. Water Science and Technology 80, 243–253. https://doi.org/10.2166/WST.2019.263

Zhou, P., Li, Z., Snowling, S., Baetz, B.W., Na, D., Boyd, G., 2019. A random forest model for inflow prediction at wastewater treatment plants. Stochastic Environmental Research and Risk Assessment 2019 33:10 33, 1781–1792. https://doi.org/10.1007/S00477-019-01732-9

Chapter 2 – Data-driven Approach for Uncertainty Analysis

Parameters in widely used secondary settling tank (SST) models are associated with significant uncertainties. A novel and efficient approach for addressing such uncertainties is proposed. The new approach is as effective as the benchmark Monte Carlo simulation method. The approach dramatically reduces computational time compared with the benchmark.

This chapter has been published: Zhou, P. and Li, Z., 2023. Arbitrary polynomial chaos expansion for uncertainty analysis of the one-dimensional hindered-compression continuous settling model. Journal of Water Process Engineering, 52, p.103489. (DOI: https://doi.org/10.1016/j.jwpe.2023.103489) Copyright (2023) Elsevier.

Pengxiao Zhou was responsible for Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, and Writing – Review & Editing under Dr. Zhong Li's supervision.

Arbitrary Polynomial Chaos Expansion for Uncertainty Analysis of the One-Dimensional Hindered-Compression Continuous Settling Model

Abstract

Secondary settling tank (SST) models play a significant role in the simulation of a wastewater treatment system. They can estimate effluent and underflow quality and thus help with the design, management, and optimization of wastewater treatment systems. SST modeling consists of an empirical settling velocity function where parameter uncertainty could raise. The performance of an SST model could suffer from parameter uncertainty, which makes parameter uncertainty assessment valuable for SST modeling. Monte Carlo simulation (MCS) is a classical technique for assessing uncertainty, but it requires parameter distribution information and is computationally expensive. To overcome these limitations, arbitrary polynomial chaos expansion (aPCE), a novel approach has been adopted for the first time in this study. The well-recognized Bürger-Diehl SST model is used and the uncertainties originating from five essential model parameters are assessed by the novel aPCE method with the MCS technique being used as a benchmark. Probabilistic estimations of the model output, i.e., sludge blanket height (SBH), are generated by both aPCE and MCS. The comparison results between aPCE and MCS suggest that the aPCE approach can be as effective as MCS in quantifying the uncertainties associated with SST model parameters, while significantly reducing approximately 90% computational requirements. This study explicitly quantifies the uncertainties associated with SST model parameters in an efficient manner, which can provide robust support for the design, management, and optimization of wastewater treatment systems.

Keywords: secondary settling tank, arbitrary polynomial chaos expansion, parameter uncertainty, Monte Carlo simulation, sludge blanket height simulation.

2.1 Introduction

Wastewater treatment that enables the reduction of water pollution is essential for ensuring a sustainable future for human society. Secondary settling tank (SST), which is the most common solid-liquid separation facility, plays a crucial role during the treatment of wastewater (Li and Stenstrom, 2014a; Ramin et al., 2014). SST is a vital part of the activated sludge process as it serves as both a clarifier and a thickener. It allows biomass or solid particles in the treated wastewater to settle to the tank bottom, while the clear water leaves the tank from the upper level (David et al., 2009). Sludge will escape with the clear water if the settling tank fails as either clarifier or thickener. In addition to delivering an effluent of poor quality, loss of sludge could alter the behavior of the biological process by uncontrollably reducing the sludge age to values below those necessary for appropriate plant performance (Ekama et al., 1997). Therefore, it is essential to unravel the behavior of SST.

Mathematical models have been frequently employed as effective tools for understanding and analyzing various wastewater treatment processes (Goodarzi et al., 2022, 2020). To adequately depict the clarification-thickening process, a number of SST models have been developed over the past few decades, and the one-dimensional secondary settling tank (1D SST) model is the most widely used one in the wastewater industry due to its computational efficiency (Li and Stenstrom, 2014b; Plósz et al., 2011). The sludge settling process was first expressed as a one-dimensional mass balance partial differential equation (PDE) by Kynch (1952). Based on experimental observations, further studies were conducted: Petty (1975) extended the Kynch theory for continuous simulation; Takács et al. (1991) simulated the SST layer by layer, and their model has been widely used in commercial modeling tools till today; Bürger et al. (2011) improved the model by considering compression settling, diffusion effects and facilitating reliable simulations.

Although these 1D SST models help with recognizing the settling characteristics of sludges, model uncertainties could arise because of the imperfect calibration of empirical parameters and the natural variability of settling processes. For example, calibration of the parameters in empirical settling velocity functions can be highly uncertain; and the settling characteristics of sludges vary radically depending on constituents in the influent and conditions imposed on the biological reactors.

To ensure the effective and reliable use of 1D SST models, it is necessary to determine the scope and sources of uncertainty associated with model simulations. This will help with understanding the simulated systems, increasing the accuracy of model simulations, and defining realistic values for potential risk assessments (Clausnitzer et al., 1998; Højberg and Refsgaard, 2005). Monte Carlo simulation (MCS), a time-tested and brutal force method for uncertainty analysis, is the dominant method for estimating the uncertainty of 1D SST models in previous research. For instance, Li and Stenstrom (2016) applied MCS to analyze uncertainties of non-identifiable parameters in a 1D SST model. The MCS approach typically consists of the following steps: (1) static model generation, (2) input distribution identification, (3) random variable generation, and (4) analysis and decision-making (Raychaudhuri, 2008). MCS is simple and easily programable, but it is inadequate when the static model is complex (Kroese et al., 2014; Oladyshkin and Nowak, 2012). To achieve a certain level of precision, the computational complexity of MCS would grow exponentially with the increase in the number of inputs or parameters. Meanwhile, MCS assumes an exact probability density function for each uncertain variable and parameter in the modeling system, which is often unknown in real-world engineering applications. As a result, a more efficient and advanced method for uncertainty analyses is desired for complex SST models.

Recently, a number of alternative methods for uncertainty analysis have been developed for diverse applications (Donnelly et al., 2022; Ghiasi et al., 2022). Particularly, arbitrary polynomial chaos expansion (aPCE) has attracted much attention and shown a superior efficiency. aPCE is based on generalized polynomial chaos expansion (PCE) and it decomposes the distribution of a variable into multiple distributions of independent variables (Xiu and Karniadakis, 2003). aPCE is a mathematically optimal way to construct and obtain a model response surface in the form of a high dimensional polynomial in uncertain model parameters (Oladyshkin and Nowak, 2012). It can be regarded as a surrogate model which captures the relationship between a distribution of original model output and distributions of original model parameters. In comparison with MCS, aPCE does not require many repeated runs of original model, which makes it more efficient for the uncertainty analysis of large and complex models. More importantly, aPCE creates polynomials from raw statistical moments of model parameters, which means it is also suitable for models where parameter distributions are arbitrarily distributed (Oladyshkin and Nowak, 2012; Wan et al., 2020). aPCE has been successfully adopted in various disciplines as an efficient approach for assessing uncertainty propagation (Laowanitwattana and Uatrongjit, 2022; Yin et al., 2018). It has a significant potential for the uncertainty analysis of 1D SST models, which are often complex models with empirical and hard-to-determine parameter distributions. However, the potential of aPCE for the uncertainty analysis of 1D SST models has not been investigated.

Therefore, the objective of this study is to, for the first time, apply aPCE to assess the model parameter uncertainty of a 1D SST model. The proposed method will improve computing efficiency and address the issue of unknow parameter distributions while assessing the uncertainty of a 1D SST model. This entails the following tasks: (1) construct a state-of-the-art 1D SST model (i.e., the Bürger-Diehl model) with hypothetical data; (2) define the model parameters and assess

their uncertainty using the aPCE technique; (3) compare the results from the aPCE method and the benchmark MCS method. This work will discuss how aPCE can be used as a novel, efficient, and reliable uncertainty analysis method for SST models.

2.2 Methods and study system

2.2.1 Arbitrary polynomial chaos expansion

The core notion of PCE is that the composition of independent variables described by orthogonal polynomials can be used to express the distribution of a random variable (Ghaith et al., 2021; Zhou et al., 2022). The homogeneous function in the Wiener theory serves as the basis for the polynomial chaos expansion (PCE) approach (Wiener, 1938). Assume $Y = f(\xi)$ is a model, and $\xi = (\xi_{i_1}, \dots, \xi_{i_n})$ are uncertain model parameters in the format of random variables. The random output variable Y can be represented by a multivariate polynomial expansion as follows (Xiu and Karniadakis, 2003):

$$Y = a_0 P_0 + \sum_{i_1=1}^{\infty} a_{i_1} P_1(\xi_{i_1}) + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{i_1} a_{i_1 i_2} P_2(\xi_{i_1}, \xi_{i_2}) + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{i_1} \sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3} P_3(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) + \cdots \quad (2.1)$$

where $P_n(\xi_{i_1}, \dots, \xi_{i_n})$ is the polynomials (e.g. Hermite orthogonal polynomials) of order n in terms of the multi-dimensional independent standard random variables $\xi = (\xi_{i_1}, \dots, \xi_{i_n})$, and $a_{i_1,\dots,i_r}$ represents the PCE coefficients.

In practice, the m^{th} order truncated polynomial expansion of Y with respect to the d-dimension vector $\xi = (\xi_1, \xi_2, \dots, \xi_d)$ can be expressed as Eq. (2.2) and the number of the PCE coefficients can be counted as eq. (2.3).

$$Y \approx \sum_{j=0}^{m} \hat{a}_j \Psi_j(\xi) \quad (2.2)$$

$$r = \frac{(d+m)!}{d!m!} \tag{2.3}$$

where $\hat{a}_j$ and $a_{i_1,\dots,i_\mathrm{r}}$ have a one-to-one correspondence, and $\Psi_j(\xi)$ and $P_d(\xi_{i_1},\dots,\xi_\mathrm{d})$ do as well.

Obtaining polynomials $P_n(\xi_{i_1},\dots,\xi_{i_\mathrm{n}})$ of PCE depends on known distributions of the random variables (Xiu and Karniadakis, 2006). Thus, applying the PCE approach with arbitrary or unknown distributions of the random variables can be challenging, which can be addressed by the adoption of aPCE. Instead of requiring complete knowledge of a probability density function, aPCE generates polynomials from the existence of a finite number of moments of the variables, making it ideal for modelling systems where the random variables are arbitrarily distributed (Guo et al., 2019; Wan et al., 2020). Applying aPCE, the polynomials $P_n(\xi)$ in eq. (2.1) can be rewritten as (Oladyshkin and Nowak, 2012):

$$P_n(\xi) = P^{(n)}(\xi) = \sum_{i=0}^{n} P_i^{(n)} \xi^i \tag{2.4}$$

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-1} \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} P_0^{(n)} \\ P_1^{(n)} \\ \vdots \\ P_{n-1}^{(n)} \\ P_n^{(n)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{2.5}$$

where $P_i^{(n)}$ are coefficients of aPCE, μ_{2k-1} represents the $2k-1^{th}$ raw moment of ξ. To establish the aPCE model, raw moment of ξ is the only required information. A m^{th} order aPCE model can be expressed as eq. (2.6):

$$Y \approx \sum_{i=0,\mathrm{j}=0}^{\mathrm{m}} \hat{a}_j \, P_i^{(m)} \xi^i \tag{2.6}$$

And $\hat{a}_j$ estimation in terms of N samples which could be achieved by a non-intrusive method (i.e., least square regression) is shown as eq. (2.7):

$$Q = argmin \sum_{i=1}^{N} (\tilde{Y}_i - Y_i)^2 \tag{2.7}$$

where $\tilde{Y}_i$ is an observation and Y_i is a simulation output.

2.2.2 1D SST model

The Bürger-Diehl model (excluding hydrodynamic dispersion) was used as an example to show the applicability of aPCE for uncertainty analysis (Bürger et al., 2011). The model was selected for its reliability, flexibility, and availability (Bürger et al., 2013; Li and Stenstrom, 2016). Assume an ideal one-dimensional SST schema as shown in Fig. 2.1.

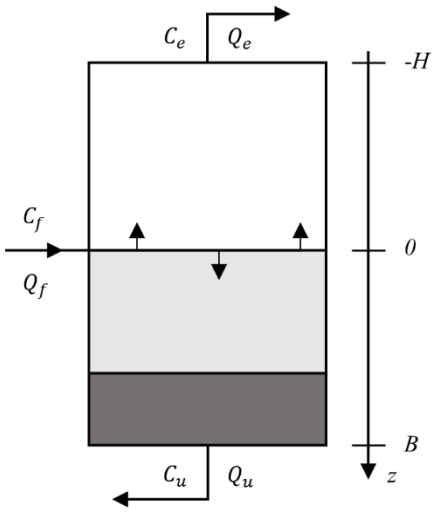

Fig. 2.1 Ideal one-dimensional SST

The Bürger-Diehl model can be expressed as follows:

$$\frac{\partial C}{\partial t} + \frac{\partial}{\partial z} F(C, z, t)) = \frac{\partial}{\partial z}(d_{comp}(C)\frac{\partial C}{\partial z}) + \frac{Q_f(t)C_f(t)}{A}\delta(z) \tag{2.8}$$

where t is time, z is the depth from feed level in SST, A is the cross-sectional area of SST, C is the solid concentration in SST, F is the convective flux function, d_{comp} is the compression function,

Q_f is the feed volumetric flow, C_f is the solid concentration of the feed flow, δ is the Dirac delta distribution.

The convective flux function for four different zones in SST can be expressed as eq. (2.9):

$$F(C,z,t) = \begin{cases} -\frac{Q_e(t)}{A}C_e(t) & for\ z < -H \quad effluent\ zone \\ v_{hs}(C)C - \frac{Q_e(t)}{A}C & for -H < z < 0 \quad clarification\ zone \\ v_{hs}(C)C + \frac{Q_u(t)}{A}C & for\ 0 < z < B \quad thickening\ zone \\ \frac{Q_u(t)}{A}C_u(t) & for\ z > B \quad underflow\ zone \end{cases} \tag{2.9}$$

where the Q_e is the effluent flow rate and C_e is the effluent solid concentration, the Q_u is the underflow flow rate and C_u is the underflow solid concentration, v_{hs} is the hindered settling velocity which is described by the Vesilind formula that incorporates v_0 and r_h which are the maximum theoretical settling velocity and the empirical parameter, respectively (Vesilind, 1968):

$$v_{hs}(C) = v_0 e^{-r_h C} \tag{2.10}$$

The compression function developed by Bürger et al. (2013, 2012) can be expressed as eq. (2.11):

$$d_{comp}(C) = \begin{cases} 0 & for\ 0 \le C \le C_C \\ \frac{\rho_s \alpha v_{hs}(C)}{g(\rho_s - \rho_f)(\beta + C + C_c)} & for\ C > C_C \end{cases} \tag{2.11}$$

where α and β are empirical parameters, ρ_s is the solid mass density, ρ_f is the fluid mass density, g is the gravity of acceleration, and C_C is a threshold concentration at which solid particles begin to physically contact one another.

As part of the 1D SST model output, sludge blanket height (SBH) is the most commonly used indicator of sludge concentration profiles (Narnoli and Mehrotra, 1997). In this study, SBH is determined based on two commonly used sludge concentration threshold values, i.e., 8 kg/m^3

and 10 kg/m^3 (Zinatizadeh et al., 2019). SBH_8 and SBH_{10} are defined as the layer number index of the layer that has a sludge concentration of 8 and 10 kg/m^3, respectively.

2.2.3 Uncertainty analysis

In this study, the goal of uncertainty analysis is to analyze parameter uncertainties originating from five selected parameters in the 1D SST model. The five parameters (i.e., v_0, r_h, C_c, α, β) are selected because they govern the two most important processes (settling velocity function and compression function) in the model. These five parameters are assumed independent, and the parameter distributions used in this study are listed in Table 2.1. The parameter distributions are estimated based on the literature (Li and Stenstrom, 2016; Plósz et al., 2011; Ramin et al., 2014). In practice, the exact parameter distributions are not expected to be available. Instead, only a set of calibrated model parameter values can be obtained. The proposed aPCE method that does not require complete knowledge of the parameter distributions is built based on statistical moments of the parameters, which can be calculated from a set of calibrated model parameter values. The parameter distributions are pre-defined in this study only because real-world calibrated model parameter values are not available from the literature. The pre-defined parameter distributions are used to produce a set of hypothetical model parameter values, which are reasonable substitutes for real-world data.

Table 2.1 Distributions of model parameters

Symbol	Definition	Distributions
v_0	maximum theoretical settling velocity (m/h)	U (3.47, 9.71)
r_h	hindered settling parameter (m3/kg)	U (0.15, 0.63)
C_c	gel concentration (kg/m3)	U (5.06, 15.27)
α	compression settling parameter (Pa)	U (0, 20)
β	compression settling parameter (kg/m3)	U (1, 10)

Fig. 2.2 shows the procedures for uncertainty analysis for the 1D SST model using the aPCE approach. After acknowledging the parameter distributions, 1,000 combinations of the five parameters are randomly sampled based on their distributions. Then, all samples are iteratively fed to the hypothetical 1D SST model, and the output results at a specific time step are labeled as SBH results from the MCS framework. As for the aPCE framework, the first 100 SBH results from MCS are obtained and transformed to a standard normal distribution by quantile transformation. The distribution transformation is done to improve the performance of aPCE (Patro and Sahu, 2015; Shalabi and Shaaban, 2006). Following this treatment, the 100 transformed SBH results with their corresponding model parameter samples are used to estimate the aPCE coefficients by regression (Oladyshkin and Nowak, 2012). Once the coefficients are calculated, the aPCE equation for this specific time step are constructed. The model parameter samples 101 to 1,000 are then fed to the constructed aPCE equation for validation purposes, and the corresponding inversed outputs of the aPCE model are labeled as SBH results from aPCE. Lastly, distributions of SBH results from both MCS and aPCE are compared and analyzed. For the aPCE part, the first 100 SBH results from MCS could be regarded as training samples, while the rest 900 SBH results from MCS are validation samples. All the results comparison shown below are on the validation samples. In this study, the aPCE is constructed with the 3^{rd} order truncated polynomial expansion and the 5-dimension random variables. It is implemented in *Julia* and the code used is based on the work of Oladyshkin and Nowak (2012).

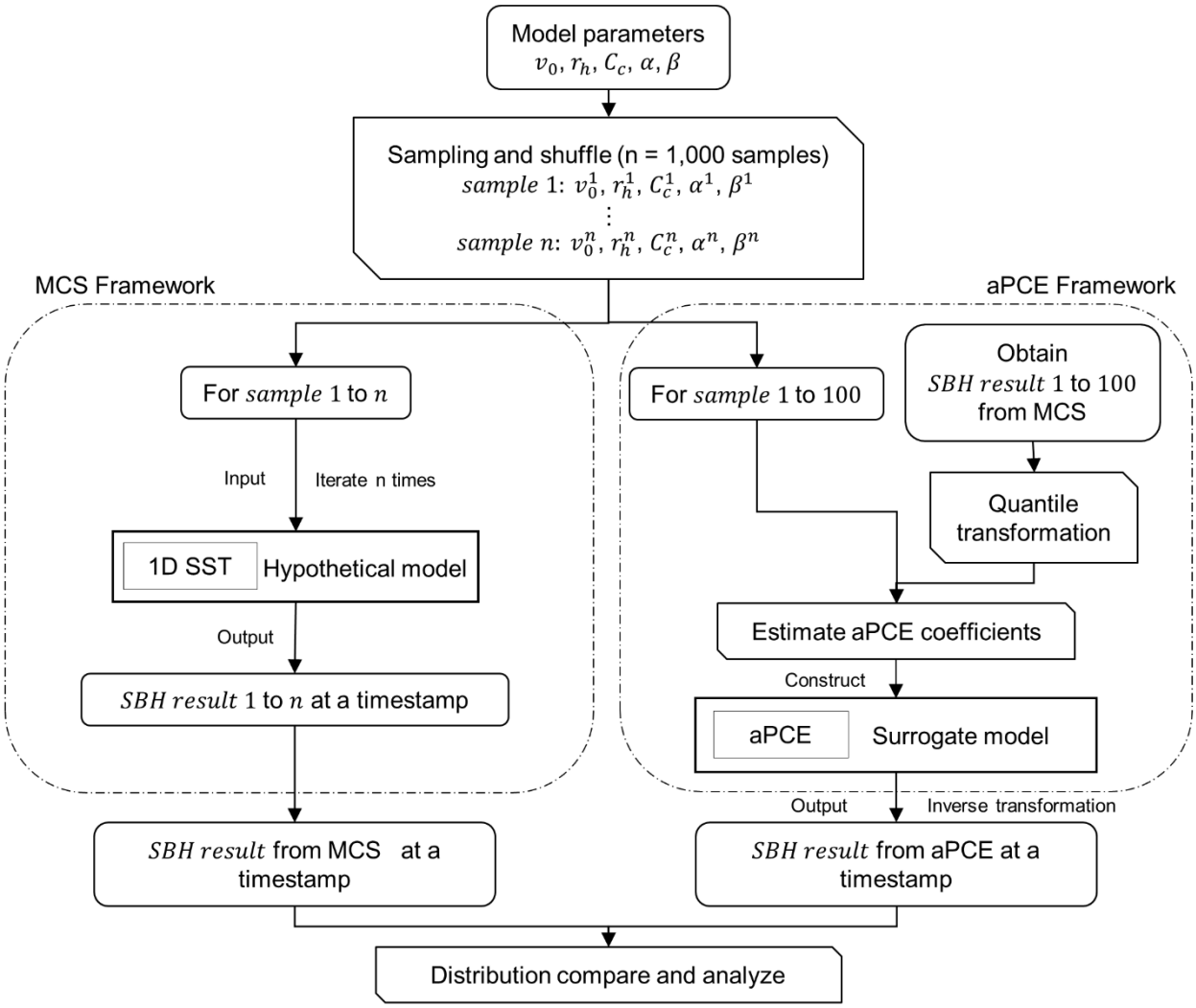

Fig. 2.2 Schema of uncertainty analysis for 1D SST

2.2.4 SST model setup

To demonstrate the applicability of the proposed aPCE method, a hypothetical case of the 1D SST model, proposed by Bürger et al. (2011) is used. The model parameters and boundary conditions are as follows: $H = 1m,\ B = 3m,\ A = 400m^2$; at time $t = 0$, the SST is full of sludge at the concentration $C = 2.0\ \mathrm{kg}/m^3$; the volumetric flow rates $Q_f = 250\ m^3/h$ and $Q_e = 170\ m^3/h$ and $Q_u = 80\ m^3/h$. The total simulation period is 400 hours and the change of feed concentration in time is as below:

$$C_f = \begin{cases} 4.0\ \mathrm{kg}/m^3, & 0 < t \leq 50\ hour \\ 3.7\ \mathrm{kg}/m^3, & 50 < t \leq 250\ hour \\ 4.1\ \mathrm{kg}/m^3, & 250 < t \leq 400\ hour \end{cases} \tag{2.12}$$

Additionally, the tank is divided into 40 layers. The 'layer j' refers to the interval $[z_{j-1}, z_j]$, where:

$$z_j = j\Delta z - H, \quad j = 0, \dots, 40 \tag{2.13}$$

$$\Delta z = \frac{B+H}{40} \tag{2.14}$$

At the top and bottom of the tank, which correspond to the effluent and underflow zones, another two layers were added, respectively (Bürger et al., 2013, 2012). Thus, the computational domain of the tank has a total of 44 layers.

2.3 Results and discussion

2.3.1 Data generation and preparation

To construct the aPCE model for a specific time step, the first 100 sets of valid SBH results from MCS at the time step were obtained. To provide an example of the 100 sets of SBH results, Fig. 2.3 shows the histograms of raw SBH_8 and SBH_{10} values at the 400th simulation hour. The mean values of SBH_8 and SBH_{10} are approximately 26 and 37 as shown in Figs. 2.3a and 2.3b, respectively. With a layer height of 0.1 m, this indicates an average of 1.1 m thickness between the two concentrations (8 kg/m^3 and 10 kg/m^3). The results validate the settling rule in SST: at a lower position of the tank, the solid concentration is supposed to be higher. Furthermore, SBH_8, spanning most layers of the tank as in Fig. 2.3a, has a more extensive distributional range than SBH_{10}. It implies that SBH_8 is more susceptible to model parameters that are associated with more uncertainty.

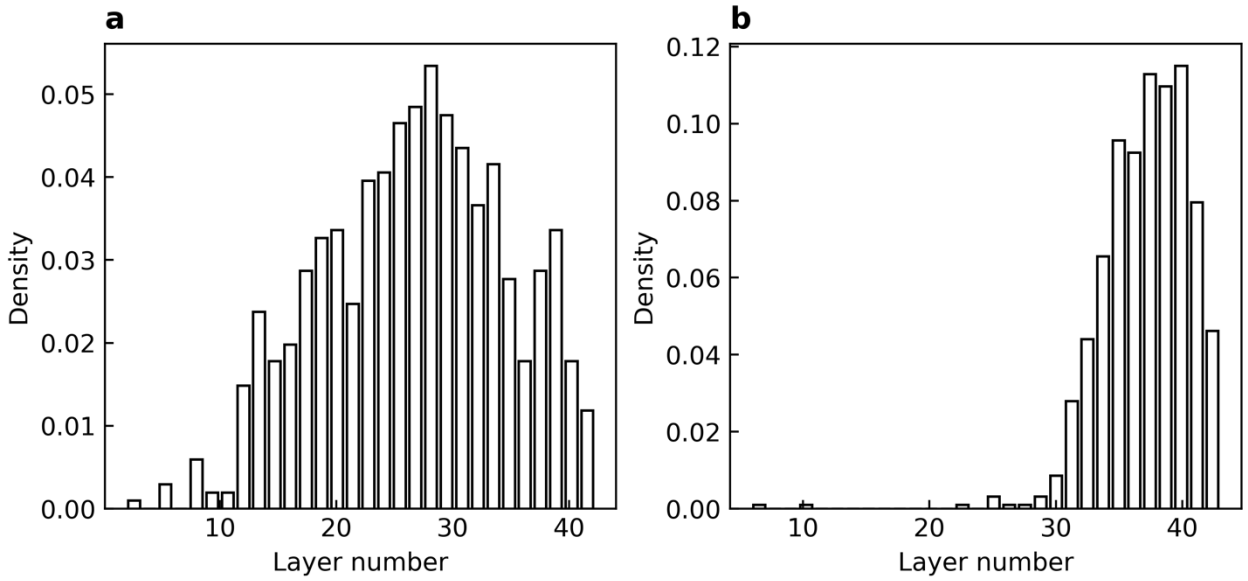

Fig. 2.3 Distributions of SBH_8 (a) and SBH_{10} (b)SBH at the 400th hour

2.3.2 Comparison of aPCE and MCS results at a single time step

The 100 sets of SBH results in Section 2.3.1 were transformed to fit standard normal distribution and then used along with their corresponding parameter values to estimate the aPCE coefficients for each time step. A set of 100 equations was generated to determine the aPCE coefficients and construct the aPCE equation for each time step after running 1D SST at the chosen 100 parameter samples. Then, a temporal series of aPCE equations was developed to quantify output uncertainties and produce probabilistic outputs. Fig. 2.4 compares the aPCE and MCS results at the 400th simulation hour. For both SBH_8 and SBH_{10}, the histogram produced by aPCE (marked as red) and that produced by MCS (marked as grey) are highly identical as shown in Figs. 2.4a and 2.4b. This indicates that aPCE in this study can well replicate the uncertainty analysis results from MCS. The probabilistic SBH outputs at the 400th simulation hour from aPCE were also compared with benchmark MCS results based on their mean and standard deviation values. In Figs. 2.4c and 2.4d, the notch of the box represents the median, and the lower and upper of the

box are the first quartile ($Q1$) and third quartile ($Q3$), respectively. IQR is the interquartile range which equals $Q3 - Q1$. The lower whisker extends to the first datum greater than $Q1 - 1.5 \cdot IQR$, while the upper whisker extends to the last datum less than $Q3 + 1.5 \cdot IQR$. The boxplots show that the medians and standard deviations obtained from MCS (marked as grey) are well replicated by aPCE (marked as red). For both SBH_8 and SBH_{10}, the differences between median values (MCS versus aPCE) are about only 1 layer, which indicate a $0.1\ m$ difference. The lower whiskers of aPCE are slightly (with a difference of less than $0.3\ m$) below that of MCS. The above comparisons at the 400^{th} simulation hour imply that the probabilistic results from aPCE, which requires 100 simulation runs, are consistent with those from MCS, which consists of 1,000 simulation runs.

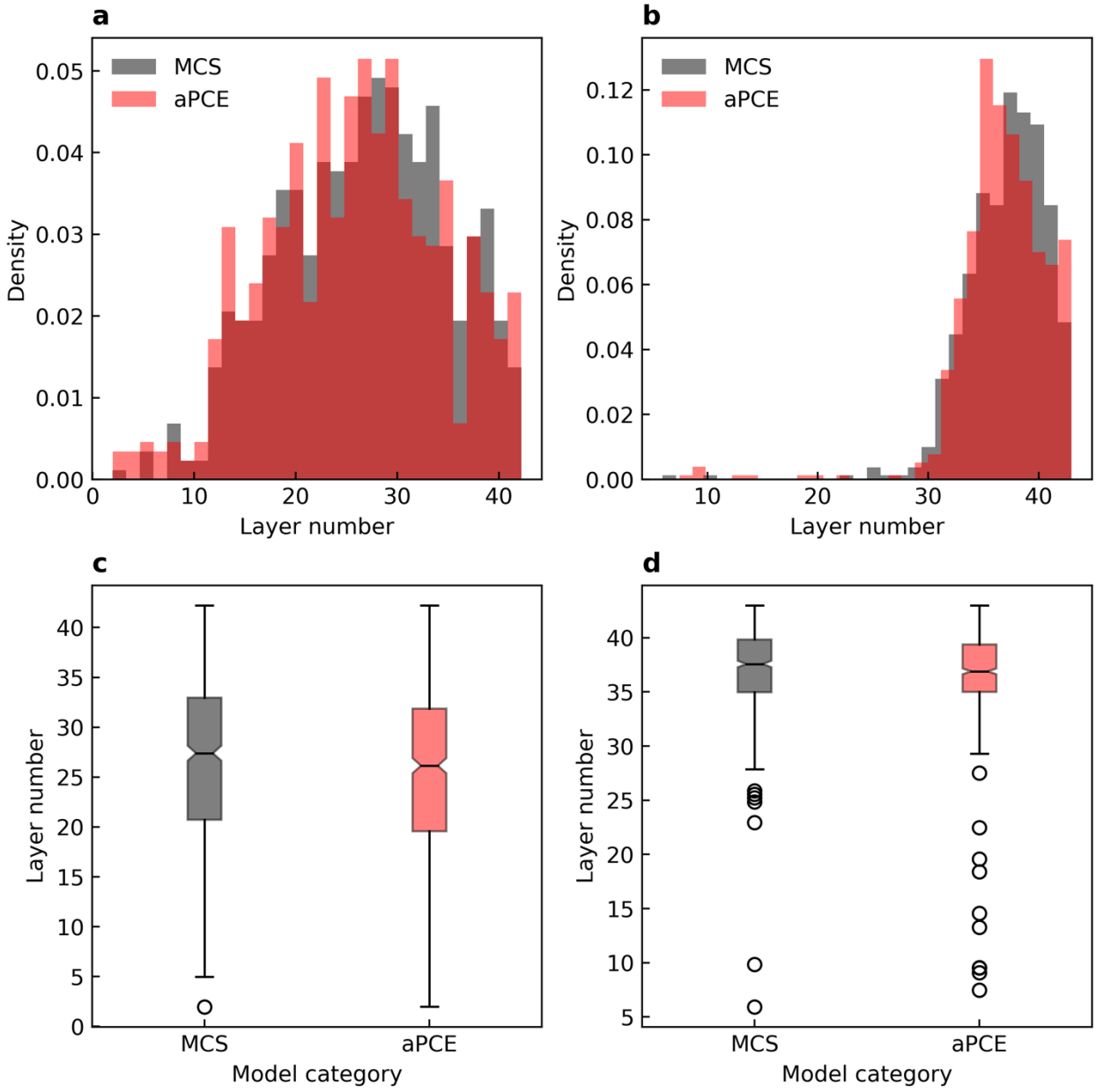

Fig. 2.4 Comparison of aPCE results and MCS results at the 400th simulation hour: a. histograms of SBH_8; b. histograms of SBH_{10}; c. boxplots of SBH_8; and d. boxplots of SBH_{10}

2.3.3 Comparison of aPCE and MCS at multiple time steps

The aPCE results were also compared with MCS at more time steps (50th, 150th, 250th, and 350th simulation hour) to further demonstrate its performance throughout the whole simulation period. Fig. 2.5 shows the comparison results at the four timestamps, and the colored plots are for

aPCE, while the grey plots are for MCS. The boxplots in Fig. 2.5 use the same legend as that in Fig. 2.4. It is observed that the distributions generated by aPCE are overall highly similar to that produced by MCS at all four timestamps. Figs. 2.5c and 2.5d show that the SBH_8 median values from aPCE are slightly smaller than those from MCS, while the SBH_{10} median values of aPCE results are almost the same as those from MCS. The SBH_8 and SBH_{10} median values show a rising trend followed by a downward trend throughout the simulation timeline, which reflects the dynamic settling process. Specifically, SBH_8 and SBH_{10} median values at the 150^{th} and 250^{th} simulation hours are a little higher than those at the 50^{th} and 350^{th} simulation hours. This trend is consistent with the common settling dynamics described in the literature. The solids in wastewater settle to the bottom of the tank over time, resulting in a rising trend for SBH median values. With the accumulation of sediment and an increase of feed concentration, the thickened sediment layer then leads to a downward trend for SBH median values. These results imply that aPCE can reproduce the probabilistic outputs from MCS at all four timestamps; it can also capture the dynamic changes during the settling process.

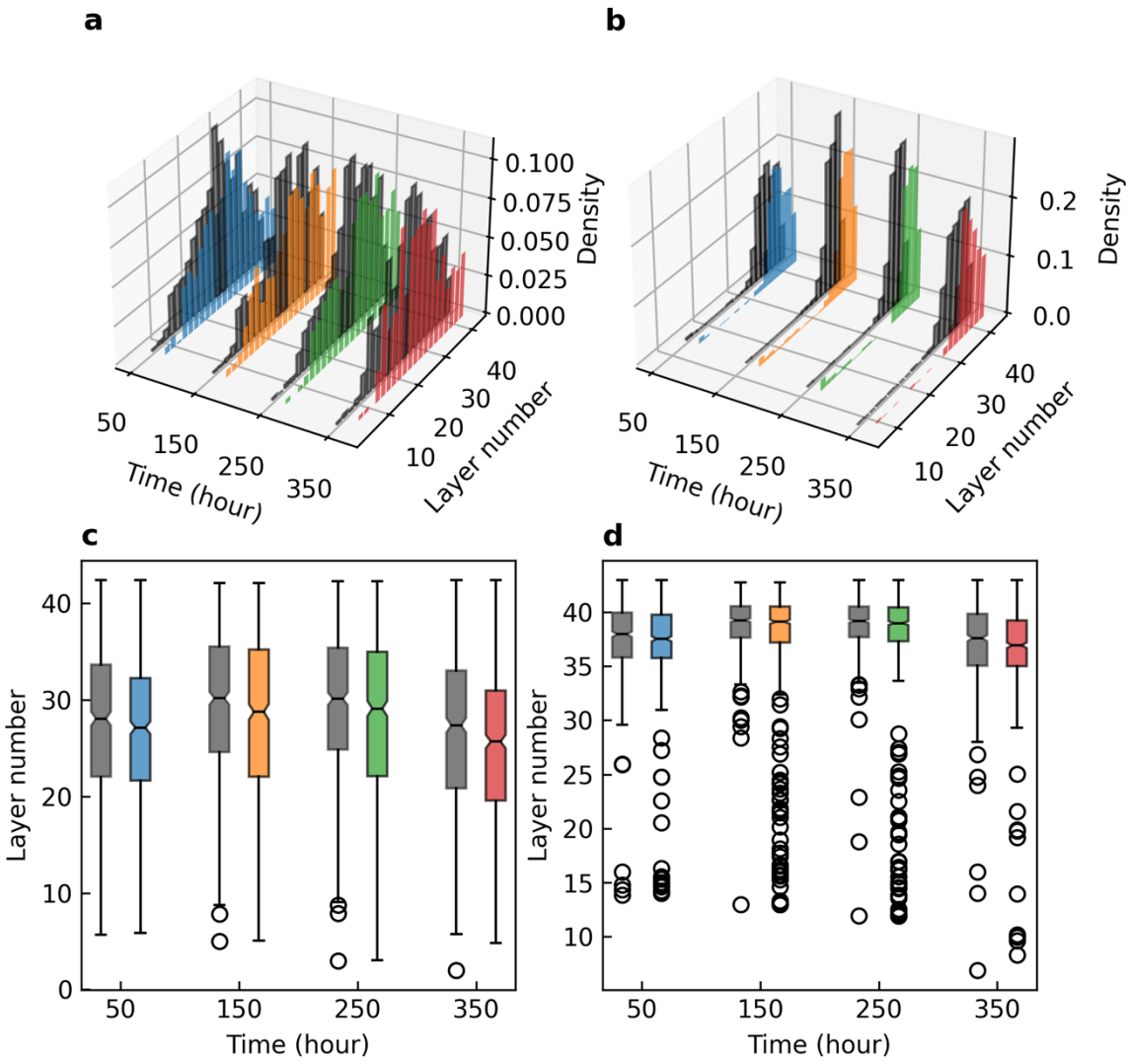

Fig. 2.5 Comparison of aPCE results and MCS results at 50th, 150th, 250th, and 350th hour: a. histograms of SBH_8; b. histograms of SBH_{10}; c. boxplots of SBH_8; and d. boxplots of SBH_{10}

To statistically compare the probabilistic outputs produced by aPCE with those produced by MCS at every time step of the whole simulation period, a nonparametric two-sample Kolmogorov-Smirnov test was implemented (Massey, 1951). In the Kolmogorov-Smirnov test, the null hypothesis is the statistical identity between the two distributions, and the test statistic is

the largest absolute distance between the two cumulative distribution functions. Fig. 2.6 shows the *p*-values of the Kolmogorov-Smirnov test at every time step. It can be observed that p-values are greater than 0.01 at all timestamps and are greater than 0.05 at most timestamps. These results suggest that there is no strong evidence to reject the null hypothesis that the two distributions generated by aPCE and MCS are statistically identical. It also implies that these distributions generated by aPCE and MCS are highly identical and aPCE can replace MCS at most time steps.

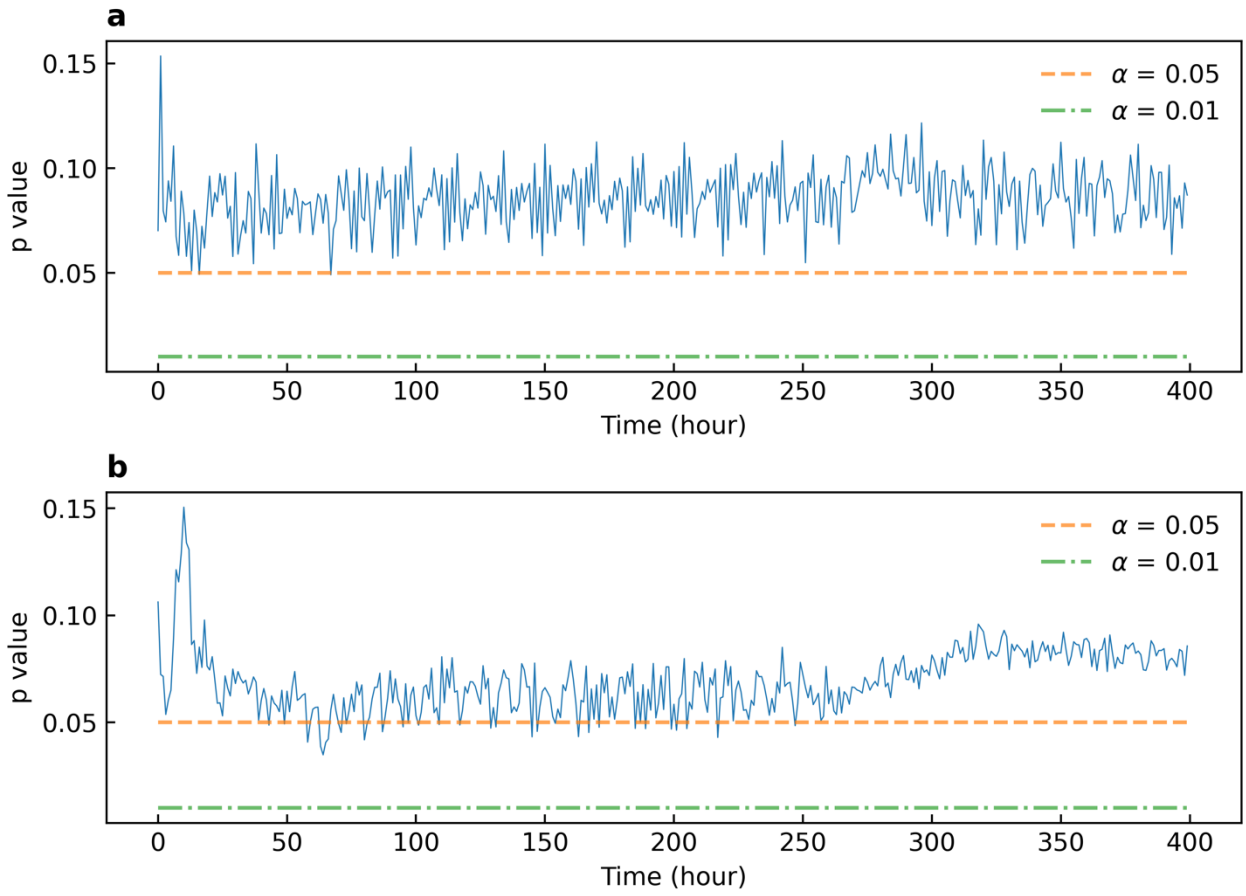

Fig. 2.6 The p-values of the Kolmogorov-Smirnov tests for: a. SBH_8; and b. SBH_{10}

Fig. 2.7 shows the probability of SBH_8 and SBH_{10} over time from aPCE results. The darker the color in Figs. 2.7a and 2.7b indicate a higher probability. It is observed that SBH_{10} spreads a smaller range. It may be because sediment layers with high concentrations are fixed at the relative

bottom of the tank due to greater gravity. Figs. 2.7c and 2.7d show the quantile for SBH_8 and SBH_{10} locations from aPCE results, respectively. The results verify that SBH_{10} spreads a smaller range. Additionally, there are clear changes after the 50th and 250th simulation hours due to the change in influent flow solid concentration (Equation 2). It implies that influent solid concentration could significantly affect the SBH location.

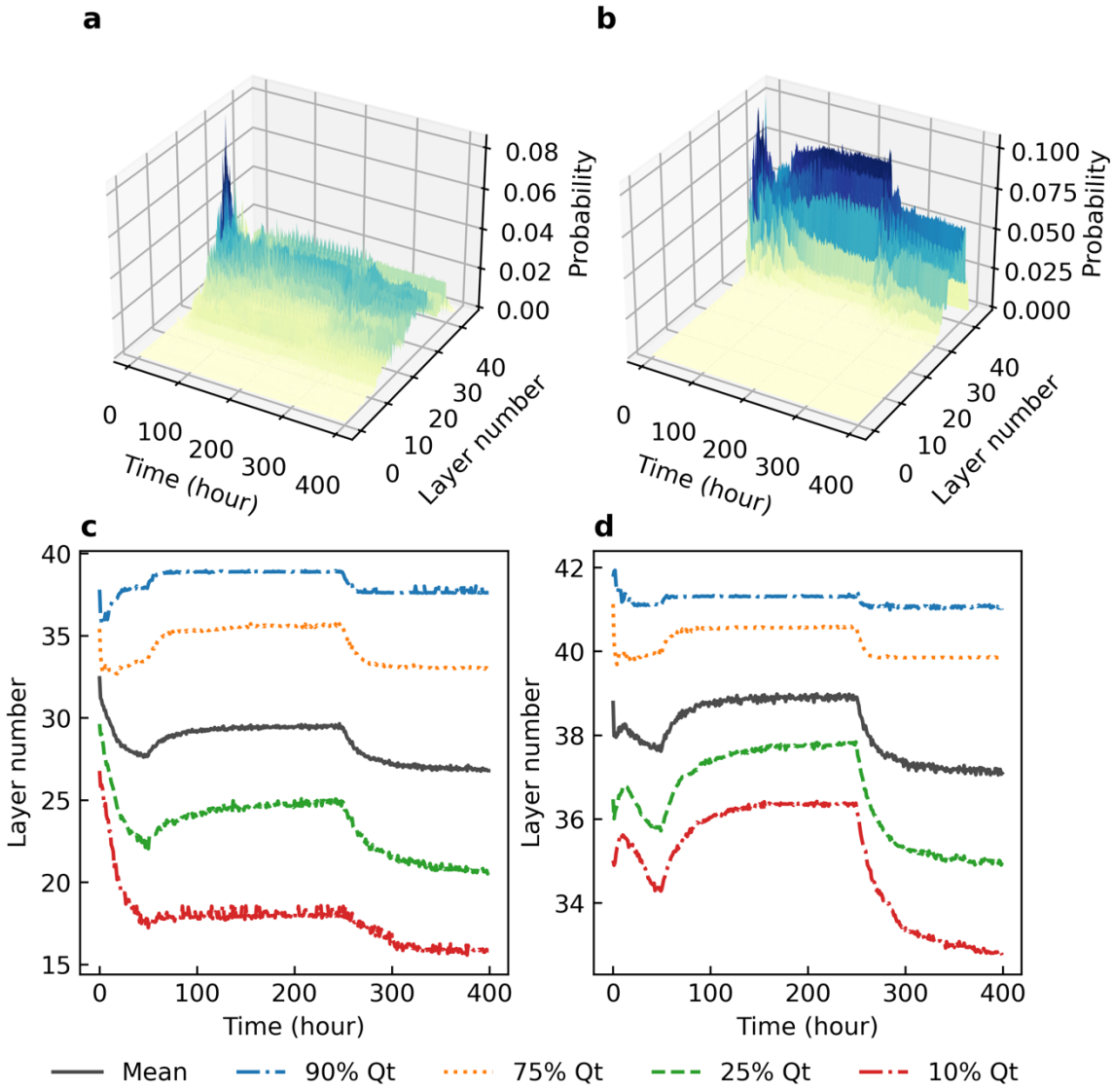

Fig. 2.7 Distributions of SBH location and its quantile from aPCE results: a. probability of SBH_8; b. probability of SBH_{10}; c. quantiles of SBH_8; and d. quantiles of SBH_{10}

These findings illustrate that the proposed aPCE approach, which requires only 100 simulation runs, can generate probabilistic outputs that are highly close to those of 1,000 MCS simulation runs. This suggests that aPCE can be as successful as MCS simulation in estimating the uncertainties associated with parameters while greatly reducing the computing time (by 90% in this study).

2.4 Conclusion

In this study, an arbitrary polynomial chaos expansion (aPCE) method is developed to assess model parameter uncertainty for a one-dimensional secondary settling tank (1D SST) model. The conventional PCE approach has been proven to be very efficient in quantifying parameter uncertainty; however, it can only be used when the distributions of model parameters are known, which is not always a valid assumption in wastewater modeling. In this improved aPCE approach, polynomials are built based on the raw moments of model parameters, which makes aPCE more applicable when prior distributions are not available.

The results demonstrate that the aPCE approach can be as effective as Monto Carlo simulation (MCS) in quantifying the uncertainties associated with 1D SST model parameters, while significantly reducing the computational loads (90% in this study). Comparing the probabilistic distributions obtained from aPCE and MCS, it is found that they have similar median, mean, and variance. In addition, the Kolmogorov-Smirnov test results suggest that the two distributions generated by aPCE and MCS are very likely to be statistically identical.

The proposed aPCE approach provides a reliable, efficient, and promising alternative for analyzing the uncertainty of model parameters in SST models. It could provide valuable technical support for wastewater risk assessment and management. In this study, aPCE is used based on the assumption that all uncertain parameters are independent variables. For future work, aPCE can be modified and tested for the analysis of dependent parameters and for more complex wastewater models.

References

Bürger, R., Diehl, S., Farås, S., Nopens, I., 2012. On reliable and unreliable numerical methods for the simulation of secondary settling tanks in wastewater treatment. Comput Chem Eng 41, 93–105. https://doi.org/10.1016/J.COMPCHEMENG.2012.02.016

Bürger, R., Diehl, S., Farås, S., Nopens, I., Torfs, E., 2013. A consistent modelling methodology for secondary settling tanks: a reliable numerical method. Water Science and Technology 68, 192–208. https://doi.org/10.2166/WST.2013.239

Bürger, R., Diehl, S., Nopens, I., 2011. A consistent modelling methodology for secondary settling tanks in wastewater treatment. Water Res 45, 2247–2260. https://doi.org/10.1016/J.WATRES.2011.01.020

Clausnitzer, V., Hopmans, J.W., Starr, J.L., 1998. Parameter Uncertainty Analysis of Common Infiltration Models. Soil Science Society of America Journal 62, 1477–1487. https://doi.org/10.2136/SSSAJ1998.03615995006200060002X

David, R., Saucez, P., Vasel, J.L., vande Wouwer, A., 2009. Modeling and numerical simulation of secondary settlers: A Method of Lines strategy. Water Res 43, 319–330. https://doi.org/10.1016/J.WATRES.2008.10.037

Donnelly, J., Abolfathi, S., Pearson, J., Chatrabgoun, O., Daneshkhah, A., 2022. Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model. Water Res 225, 119100. https://doi.org/10.1016/J.WATRES.2022.119100

Ekama, G., Ekama, G.A., Pitman, A.R., Smollen, M., Marais, G.V.R., 1997. SECONDARY SETTLING TANKS, in: London: International Association on Water Quality.

Ghaith, M., Li, Z., Baetz, B.W., 2021. Uncertainty Analysis for Hydrological Models With Interdependent Parameters: An Improved Polynomial Chaos Expansion Approach. Water Resour Res 57, e2020WR029149. https://doi.org/10.1029/2020WR029149

Ghiasi, B., Noori, R., Sheikhian, H., Zeynolabedin, A., Sun, Y., Jun, C., Hamouda, M., Bateni, S.M., Abolfathi, S., 2022. Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams. Scientific Reports 2022 12:1 12, 1–15. https://doi.org/10.1038/s41598-022-08417-4

Goodarzi, D., Abolfathi, S., Borzooei, S., 2020. Modelling solute transport in water disinfection systems: Effects of temperature gradient on the hydraulic and disinfection efficiency of serpentine chlorine contact tanks. Journal of Water Process Engineering 37, 101411. https://doi.org/10.1016/J.JWPE.2020.101411

Goodarzi, D., Mohammadian, A., Pearson, J., Abolfathi, S., 2022. Numerical modelling of hydraulic efficiency and pollution transport in waste stabilization ponds. Ecol Eng 182, 106702. https://doi.org/10.1016/J.ECOLENG.2022.106702

Guo, L., Liu, Y., Zhou, T., 2019. Data-driven polynomial chaos expansions: A weighted least-square approximation. J Comput Phys 381, 129–145. https://doi.org/10.1016/J.JCP.2018.12.020

Højberg, A.L., Refsgaard, J.C., 2005. Model uncertainty – parameter uncertainty versus conceptual models. Water Science and Technology 52, 177–186. https://doi.org/10.2166/WST.2005.0166

Kroese, D.P., Brereton, T., Taimre, T., Botev, Z.I., 2014. Why the Monte Carlo method is so important today. Wiley Interdiscip Rev Comput Stat 6, 386–392. https://doi.org/10.1002/WICS.1314

Kynch, G.J., 1952. A theory of sedimentation. Transactions of the Faraday Society 48, 166–176. https://doi.org/10.1039/TF9524800166

Laowanitwattana, J., Uatrongjit, S., 2022. Probabilistic Power Flow Analysis Based on Partial Least Square and Arbitrary Polynomial Chaos Expansion. IEEE Transactions on Power Systems 37, 1461–1470. https://doi.org/10.1109/TPWRS.2021.3099110

Li, B., Stenstrom, M.K., 2016. Practical identifiability and uncertainty analysis of the one-dimensional hindered-compression continuous settling model. Water Res 90, 235–246. https://doi.org/10.1016/J.WATRES.2015.12.034

Li, B., Stenstrom, M.K., 2014a. Research advances and challenges in one-dimensional modeling of secondary settling Tanks – A critical review. Water Res 65, 40–63. https://doi.org/10.1016/J.WATRES.2014.07.007

Li, B., Stenstrom, M.K., 2014b. Dynamic one-dimensional modeling of secondary settling tanks and design impacts of sizing decisions. Water Res 50, 160–170. https://doi.org/10.1016/J.WATRES.2013.11.037

Massey, F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. J Am Stat Assoc 46, 68–78. https://doi.org/10.1080/01621459.1951.10500769

Narnoli, S.K., Mehrotra, I., 1997. Sludge blanket of UASB reactor: Mathematical simulation. Water Res 31, 715–726. https://doi.org/10.1016/S0043-1354(97)80987-6

Oladyshkin, S., Nowak, W., 2012. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. Reliab Eng Syst Saf 106, 179–190. https://doi.org/10.1016/J.RESS.2012.05.002

Patro, S.G.K., Sahu, K.K., 2015. Normalization: A Preprocessing Stage. IARJSET 20–22. https://doi.org/10.48550/arxiv.1503.06462

Petty, C.A., 1975. Continuous sedimentation of a suspension with a nonconvex flux law. Chem Eng Sci 30, 1451–1458. https://doi.org/10.1016/0009-2509(75)85022-6

Plósz, B.Gy., de Clercq, J., Nopens, I., Benedetti, L., Vanrolleghem, P.A., 2011. Shall we upgrade one-dimensional secondary settler models used in WWTP simulators? – An assessment of model structure uncertainty and its propagation. Water Science and Technology 63, 1726–1738. https://doi.org/10.2166/wst.2011.412

Ramin, E., Flores-Alsina, X., Sin, G., Gernaey, K. v., Jeppsson, U., Mikkelsen, P.S., Plósz, B.G., 2014. Influence of selecting secondary settling tank sub-models on the calibration of WWTP models – A global sensitivity analysis using BSM2. Chemical Engineering Journal 241, 28–34. https://doi.org/10.1016/J.CEJ.2013.12.015

Raychaudhuri, S., 2008. Introduction to monte carlo simulation. Proceedings - Winter Simulation Conference 91–100. https://doi.org/10.1109/WSC.2008.4736059

Shalabi, L. al, Shaaban, Z., 2006. Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, in: 2006 International Conference on Dependability of Computer Systems. IEEE, pp. 207–214. https://doi.org/10.1109/DEPCOS-RELCOMEX.2006.38

Takács, I., Patry, G.G., Nolasco, D., 1991. A dynamic model of the clarification-thickening process. Water Res 25, 1263–1271. https://doi.org/10.1016/0043-1354(91)90066-Y

Vesilind, P.Aarne., 1968. Design of prototype thickeners from batch settling tests. Water Sewage Works 115, 302–307.

Wan, H.P., Ren, W.X., Todd, M.D., 2020. Arbitrary polynomial chaos expansion method for uncertainty quantification and global sensitivity analysis in structural dynamics. Mech Syst Signal Process 142, 106732. https://doi.org/10.1016/J.YMSSP.2020.106732

Wiener, N., 1938. The Homogeneous Chaos. American Journal of Mathematics 60, 897. https://doi.org/10.2307/2371268

Xiu, D., Karniadakis, G.E., 2006. The Wiener--Askey Polynomial Chaos for Stochastic Differential Equations. http://dx.doi.org/10.1137/S1064827501387826 24, 619–644. https://doi.org/10.1137/S1064827501387826

Xiu, D., Karniadakis, G.E., 2003. Modeling uncertainty in flow simulations via generalized polynomial chaos. J Comput Phys 187, 137–167. https://doi.org/10.1016/S0021-9991(03)00092-5

Yin, S., Yu, D., Luo, Z., Xia, B., 2018. An arbitrary polynomial chaos expansion approach for response analysis of acoustic systems with epistemic uncertainty. Comput Methods Appl Mech Eng 332, 280–302. https://doi.org/10.1016/J.CMA.2017.12.025

Zhou, P., Li, C., Li, Z., Cai, Y., 2022. Assessing uncertainty propagation in hybrid models for daily streamflow simulation based on arbitrary polynomial chaos expansion. Adv Water Resour 160, 104110. https://doi.org/10.1016/j.advwatres.2021.104110

Zinatizadeh, A.A., Rahimi, Z., Younesi, H., 2019. Sludge Blanket Height (SBH) as a Process Stability Indicator in UASFF Reactor: Relationship Between SBH and Sludge Concentration at Different Operating Conditions. Waste and Biomass Valorization 2019 11:8 11, 4003–4012. https://doi.org/10.1007/S12649-019-00708-8

Chapter 3 – Data-driven Approach for Emerging Contaminant Predictions

A framework for Bisphenol A (BPA) modeling at wastewater plants is proposed. Data from 12 plants are used to develop data-driven models for BPA prediction. Influencing factors of BPA removal are studied using network theory. The results imply that BPA can hardly be removed through primary treatment. Important factors for BPA removal at wastewater treatment plants are identified.

This chapter has been published: Zhou, P., Li, Z., El-Dakhakhni, W. and Smyth, S.A., 2022. Prediction of bisphenol A contamination in Canadian municipal wastewater. Journal of Water Process Engineering, 50, p.103304. (DOI: https://doi.org/10.1016/j.jwpe.2022.103304)

Pengxiao Zhou was responsible for Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, and Writing – Review & Editing under Dr. Zhong Li's supervision.

Prediction of Bisphenol A Contamination in Canadian Municipal Wastewater

Abstract

Bisphenol A (BPA) is one of the most common contaminants of emerging concerns (CECs), which pose a threat to human health. Conventional wastewater treatment plants (WWTPs) are considered as the major pathway of BPA entering the aqueous environment. To control and mitigate BPA contamination in the aquatic environment, predicting BPAs fate at WWTPs is critical. In this study, three machine learning models, including shared layer multi-task neural network (MLT-NN), genetic programming (GP), and extra trees (ET) are used to predict the effluent BPA concentration at twelve municipal WWTPs across Canada. Additionally, the theory of networks is adopted to analyze the interdependencies among the influencing factors of BPA removal. It is found that the proposed models can provide reasonable BPA effluent concentration predictions. They have advantages in alleviating data sparsity and imbalance, improving model interpretability, and measuring predictor importance, which is valuable for the modeling of BPA and many other CECs. The network analysis results imply there are moderate interdependencies among various influencing factors of BPA removal. Factors that significantly affect BPA effluent concentration and are thus important for BPA removal are identified. The results also show that BPA is unlikely to be removed at primary treatment plants, while BPA removal could be achieved through secondary or tertiary treatment. This study presents an integrated framework for the modeling and analysis of BPA at WWTPs, which can provide direct and robust decision support for the management of BPA as well as other emerging contaminants in municipal wastewater.

Keywords: bisphenol A, contaminants of emerging concerns, machine learning, theory of networks

3.1 Introduction

Contaminants of emerging concerns (CECs), such as pharmaceuticals and personal care products (PPCPs), endocrine-disrupting compounds (EDCs), flame retardants (FRs), pesticides, and artificial sweeteners (ASWs), and their metabolites, are considered as a growing threat to the aqueous environment and public health (K'oreje et al., 2020; Patel et al., 2020). Conventional wastewater treatment plants (WWTPs) are designed to remove more commonly seen pollutants, such as organic pollutants, phosphorus, and nitrogen, not CECs (Oliveira et al., 2020). As a result, WWTPs become one of the main pathways of the inductive release of CECs into the environment (Salimi et al., 2017). Understanding and predicting CECs' fate at WWTPs is of great importance to the mitigation of CEC-related risks.

Bisphenol A (BPA) is one of the most common CECs due to its massive use around the world. For the past few decades, BPA has been widely used as a raw material for manufacturing polycarbonate plastics and epoxy resins, which are used to produce daily consumer products such as water bottles, thermal paper, dental sealants, and medical equipment (Guerra et al., 2015; Pookpoosa et al., 2015). BPA has been found to have an adverse impact on human health, being responsible for an increase in incidences such as cancer and hormonal imbalance (Kitamura et al., 2005). Although various treatment methods (e.g., adsorption on activated carbon, ultrafiltration, biodegradation, and ozonation) have been proven to be effective for BPA removal, they are not available at most conventional WWTPs due to limited budget (Brugnera et al., 2010; Xu et al., 2018). Therefore, predicting BPA concentration in WWTP effluents is important for estimating the amount of BPA discharged into the aquatic environment. However, previous research on BPA modeling at WWTPs is very limited. Most previous studies on BPA are based on sample-by-sample analysis (Cao et al., 2022; Dong et al., 2021; Kang et al., 2021; Wang et al., 2021; Zhang

et al., 2022). Lee and Peart (2000) analyzed 36 Canadian wastewater influent/effluent sample pairs and reported that BPA in the influent can be eliminated during the treatment process at a median reduction rate of 68%. Guerra et al. (2015) investigated how parameters affect BPA occurrence, removal, and fate. More recently, several attempts have been made to investigate BPA distribution and modeling at large scales. For example, Gewurtz et al. (2021) used a multimedia approach to assess spatial and temporal trends of BPA in a Canadian environment. Tong et al. (2022) proposed a hybrid approach for BPA prediction in a reservoir that harvests rainfall water and acts as drinking and recreational water resources. To our knowledge, there are no previous studies on the prediction of BPA fate during municipal wastewater treatment processes. The decay and removal of BPA in wastewater are complex processes and it is hard to simulate such processes using conventional wastewater simulation models, which are typically process-based models (PBM) with limited capacity to capture complex relationships and are usually influenced by uncertain variables, such as pH and salting-out effects (Jhones dos Santos et al., 2021; Murugananthan et al., 2008).

Data-driven models (DDMs) have attracted much attention recently and have been successfully used as alternatives for conventional process-based models (PBMs) in the field of wastewater modeling (Dürrenmatt and Gujer, 2012; Newhart et al., 2019; Zhou et al., 2019). DDMs have advantages over PBMs in capturing highly complex and nonlinear relationships, but the lack of data may be a major obstacle to developing DDMs (Natarajan et al., 2020; Xue et al., 2014). In the past decade, a national wastewater monitoring program in Canada that monitors chemical substances has made it possible to obtain a certain amount of laboratory data for emerging contaminants prediction. However, using normal data-driven modeling techniques to predict BPA in wastewater is still a daunting challenge due to the data imbalance and sparsity caused by limited laboratory resources, as well as the poor interpretability of traditional DDMs.

To address such concerns, three well-customized DDMs including multitask shared layer neural network (MLT-NN), genetic programming (GP), and extra trees (ET) are introduced in this study. MLT-NN that can leverage useful information from other related learning tasks is used to alleviate the data sparsity and imbalance problem. Closed-form functions generated by GP and variable importance measures (VIM) derived from ET are used to better interpret the DDM results and investigate the impacts of different wastewater features on BPA effluent concentrations. In addition, theory of networks, which is known for visualizing interdependencies among features and has been widely applied in many fields, is adopted to study the interdependencies among different WWTP features and provide an insight into the influencing factors of BPA effluent concentration (Narayanan et al., 2021; Sharan et al., 2007).

The overall objective of this study is to establish an integrated framework for the prediction and evaluation of BPA removal at Canadian municipal WWTPs. This entails the following three tasks: (1) integrate data from different WWTPs and build data-driven models for the prediction of effluent BPA; (2) analyze how wastewater features (e.g., temperature, influent flow rate) affect the BPA effluent concentration; (3) assess the interdependencies among wastewater treatment features and further analyze the influencing factors of BPA effluent concentration. This study is the first attempt to develop DDMs for municipal wastewater BPA prediction. It can provide direct decision support for the removal and management of BPA through municipal WWTPs. It also provides an example of how existing challenges (i.e., data sparsity and imbalance, model interpretability, and feature importance measurement) can be addressed to predict contaminants of emerging concerns at WWTPs; the developed methodology can be extended to predict and manage various emerging contaminants at WWTPs in future.

3.2 Methodology

An integrated framework as shown in Fig. 3.1 is proposed for predicting BPA effluent concentration at municipal WWTPs in this study. The framework consists of two major parts: (1) DDMs for interval predictions of BPA effluent concentration; and (2) networks for feature independence analysis. For the DDM part, the selected MTL-NN, GP, and ET are introduced to solve the data imbalance problem associated with CEC data, generate a closed-form function for model interpretability, and measure variable importance, respectively. As for the network part, a network of wastewater treatment features is developed for assessing the interdependences among these features.

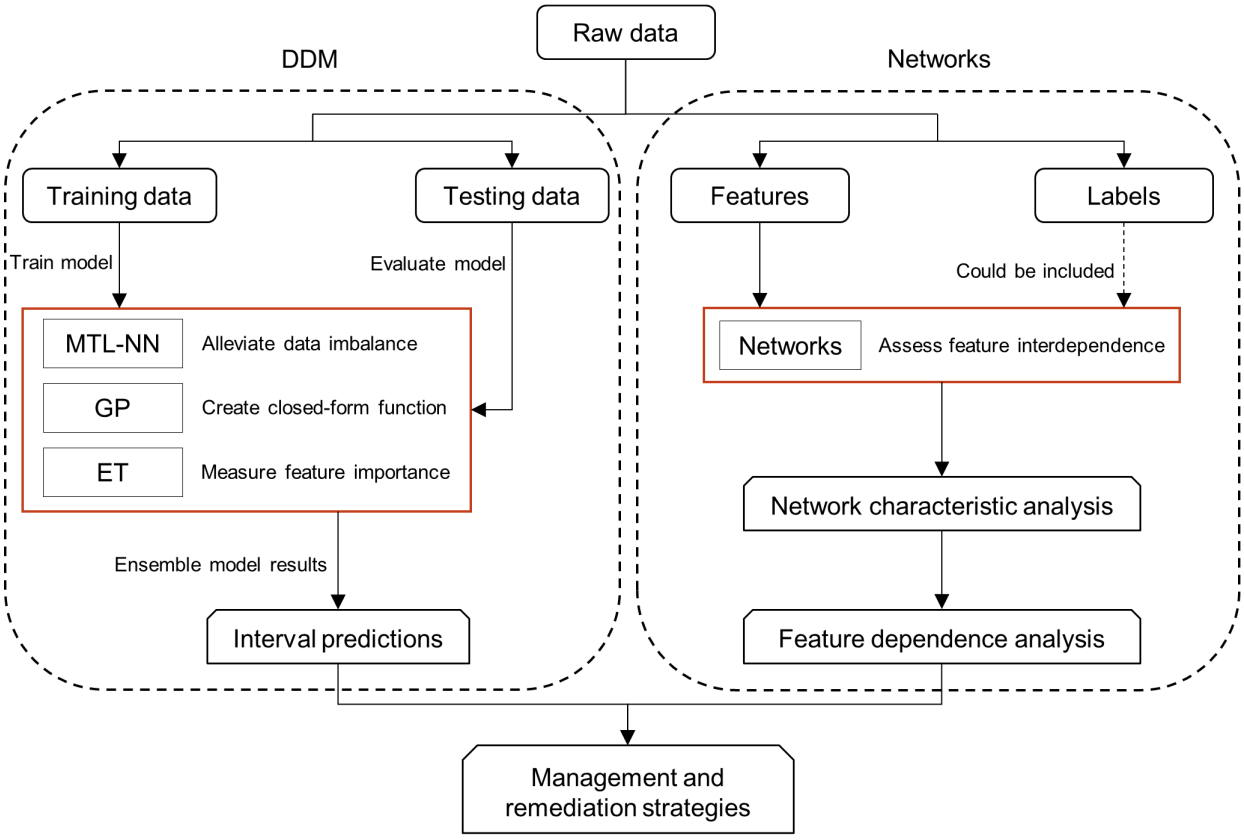

Fig. 3.1 Scheme of the proposed integrated framework

3.2.1 Deterministic prediction models

3.2.1.1 Multi-task shared layer neural network

Neural network models are known for their capability to capture complex relationships and have been widely applied in many fields (Almeida, 2002; Goh, 1995). However, BPA prediction at WWTPs is more challenging than common neural network modeling problems. Specifically, the BPA samples are expected to come from many WWTPs, which requires one generalized prediction platform. More importantly, the data collected from different WWTPs are imbalanced and sometimes sparse. For some WWTPs, data are available at as few as six points. To address these challenges, MTL-NN is introduced to this study. Multi-task learning is a good solution because it exploits useful information from other related learning tasks to help alleviate the data sparsity problem (Dorado-Moreno et al., 2020; Zhang and Yang, 2018). While a traditional neural network with a feed-forward architecture minimizes a cost function with respect to the learnable parameters defined in the architecture, MTL-NN minimizes a global cost function defined as a linear combination of the task-specific cost functions with weights (Michelucci and Venturini, 2019). The MTL-NN architecture used in this study is shown in Appendix (Fig. A-1). The network consisting of twelve tasks is composed of two common hidden layers and two task-specific hidden layers. Each task in the network represents the prediction of BPA effluent concentration at a specific WWTP.

3.2.1.2 Genetic programming

GP model uses biological evolutionary thinking and requires fewer data comparing with other DDMs, which makes it advantageous when dealing with data sparsity (Anand, 2012; O'Neill et al., 2010; Vladislavleva et al., 2010). Additionally, a closed-form function can be generated and thus help improve the interpretability of the model. The workflow of GP can be described as

follows: (1) GP generates some random initial solutions (individuals); (2) decodes and solves the target value of the current individual, and makes a selection according to fitness; (3) crosses the selected individuals with a certain probability to obtain offspring; (4) mutates offspring with a certain probability; (5) goes back to step (2) until the optimization target value meets requirements.

3.2.1.3 Extra trees

Tree structure models have been successfully used in many disciplines and extra trees (ET) present an ensemble structure of decision trees (Ahmad et al., 2018; Kingsford and Salzberg, 2008). After the forest that consists of a certain number of trees is constructed and finalized, each tree can generate one predicted value and the average predicted values are used as the output of the ET model. In comparison with other tree-structural meta estimators such as the random forest model, ET uses the entire original sample set instead of bootstrapping samples and splits nodes randomly instead of choosing the optimum criteria. These differences can help ET reduce biases and variances. The representativeness of input variables (also called predictors or features) is critical to a data-driven model. Thus, the importance of input variables should be carefully evaluated and analyzed. The variable importance measure (VIM) is another advantage of the tree structure model. VIM is based on the calculation of impurity decreases while splitting the nodes in the tree, which makes it an effective tool to assess variable importance.

3.2.2 Theory of networks

The theory of networks could unravel the nature and extent of connections in complex systems (Barabási and Albert, 1999; Watts and Strogatz, 1998). A network is a set of points connected by lines, where the points are referred to as vertices or nodes and the lines are referred to as edges or links (Sivakumar, 2014). The nodes could be recognized as factors, while the links are their connections. In the theory of networks, interdependence among factors could be described

mathematically by an adjacency matrix (A). When two factors in a network are independent, the corresponding element in matrix (A) is zero; otherwise, the corresponding element has a value of one. The interdependence can be unidirectional, which is represented by a directed link, or bidirectional, which is represented by an undirected link. The characteristics of a network can be evaluated by several criteria: density (D) is the degree of closeness among factors, average degree (k) is the average direct influence among nodes, clustering coefficient (C_{coe}) is the extent of interactions between a node's neighbors and itself, closeness centrality (C_c) is the closeness of a node to the rest of nodes in the network, betweenness centrality (B_c) of a node is the frequency of the node appearing in the shortest paths between two other nodes, and eigenvector centrality (E_c) is the influence of a node in the network.

3.3 Study area and data collection

In this study, wastewater data at twelve anonymous WWTPs across Canada were collected from the database provided by Canada's Chemical Management Plan. The WWTPs can be classified into different categories (i.e., primary, secondary, and tertiary) based on the treatment processes. Different WWTPs may have different treatment units and thus lead to different treatment efficiency. The characteristics of the twelve WWTPs are shown in Table 3.1. Sewage samples of raw influent and final effluent were collected both in summer and winter. BPA concentration of the samples was manually measured in the laboratory, and readers are referred to references for details of the data generation (Gewurtz et al., 2021; Guerra et al., 2015). Table 3.1 also summarizes the number of BPA samples that were collected at different WWTPs. It can be observed that the number of BPA samples is limited, with a total of 112 samples from the 12 WWTPs, and half of the WWTPs have as few as 6 samples. Each sample presented consists of two parts of data: predictors and predictands. The predictors include BPA raw influent

concentration (ng/L), winter (November to April) or summer (May to October), temperature (◦C), influent flow rate (m^3/d), and WWTP types. The predictand is the final BPA effluent concentration (ng/L). Fig. 3.2 shows all the features used in the developed DDMs for effluent BPA predictions.

Table 3.1 WWTPs Characteristics and the Number of BPA samples at different WWTPs

WWTP code	Number of BPA samples	Treatment type
E	14	Advanced, biological nutrient removal
HG	6	Secondary, extended aeration
J	6	Lagoon, facultative
MH	6	Secondary, activated sludge
N	6	Primary, chemical assist
OX	7	Secondary, membrane bioreactor
Q	21	Secondary, activated sludge
R	7	Lagoon, aerated with primary treatment
TB	15	Lagoon, aerated
U	12	Primary, chemical assist
WF	6	Lagoon, aerated
Y	6	Lagoon, facultative

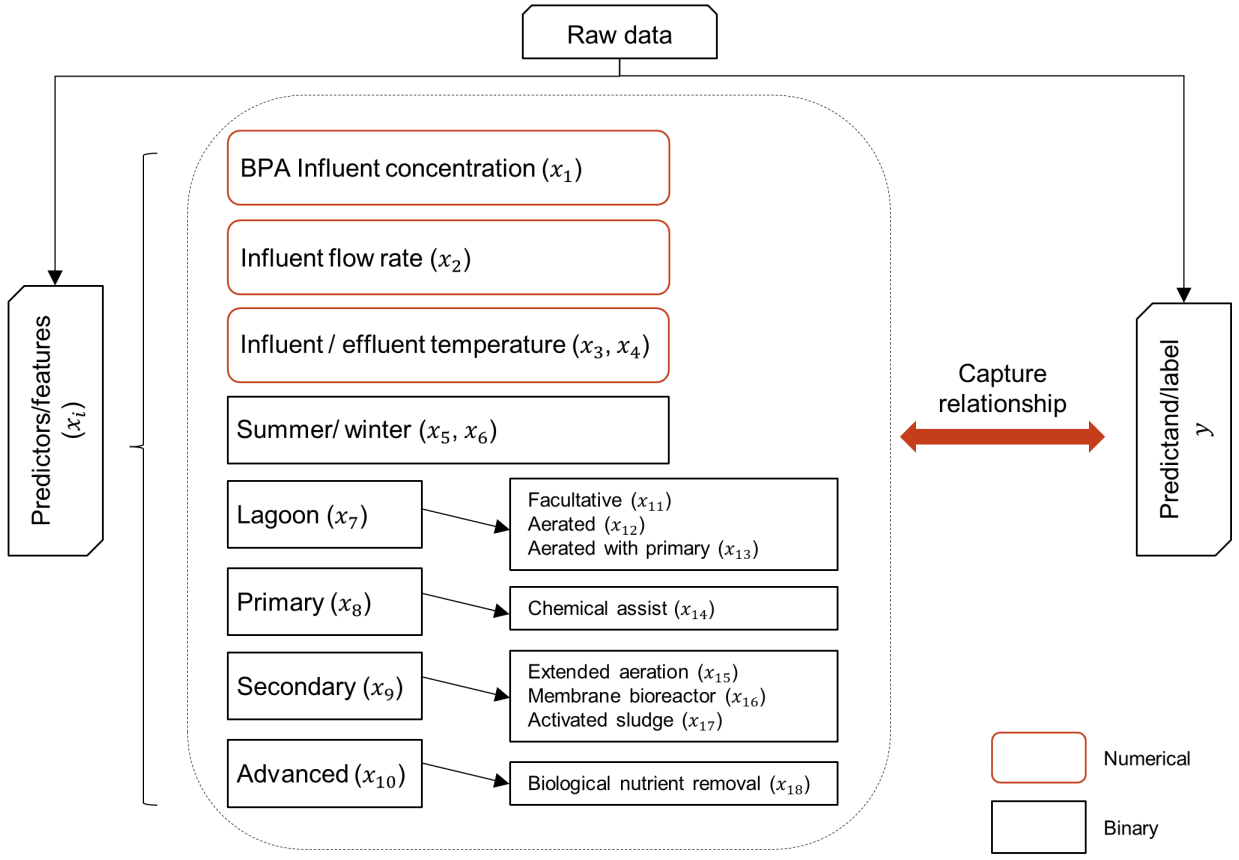

Fig. 3.2 All features used for effluent BPA prediction

3.4 Results

3.4.1 Deterministic prediction and model comparison

In this study, three DDMs (i.e., MLT-NN, ET, and GP) are utilized to predict effluent BPA concentration across the twelve selected wastewater treatment plants. One random sample from each WWTP is reserved for model validation due to the sample sparsity, while the rest of the samples are used for training. The model performance is evaluated by mean absolute percentage error (*MAPE*), root mean square error (*RMSE*), and coefficient of determination (R^2). Table A-1 in the Appendix presents the overall results of the evaluation criteria for the validation performance. It is found that all three models could provide satisfactory BPA effluent concentration predictions.

The ET model has the lowest *MAPE* of 0.26 and the highest R^2 of 0.859, the GP model reaches the lowest *RMSE* of 148.08, while the MLT-NN maintains a medium performance overall.

To further demonstrate the performance of the models, scatter plots of observed versus predicted BPA effluent concentration on the validation samples are presented in Fig. 3.3. These scatterplots show moderately strong and positive associations between the predicted and observed values with very few outliers. For instance, it can be observed that the extremely high observed value (1,030 ng/L) from station U is overestimated by all three models. It is also noteworthy that ET and MLT-NN perform better on samples with values under 200 ng/L, which may be the reason why GP shows an overall higher *MAPE*.

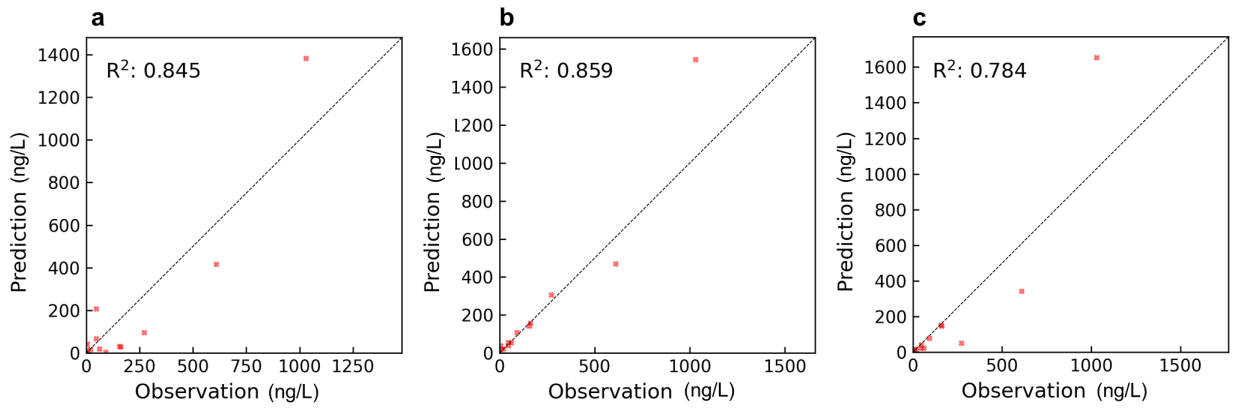

Fig. 3.3 Scatter plots of observed versus predicted BPA effluent concentration on validation samples for (a): genetic programming (GP) model, (b): extra trees (ET) model, and (c): shared-layer neural network (MLT-NN).

3.4.2 Interval prediction and station comparison

In addition to deterministic predictions, interval predictions were generated by integrating the deterministic predictions from the training of the three models mentioned above. Fig. 3.4 shows the interval predictions for the 12 wastewater treatment stations. Most of the validation samples fall into the predicted intervals except plants N and U, which are both primary treatment plants with chemical assist. BPA effluent concentrations at these two plants are significantly higher than the other wastewater treatment plants. It implies BPA is not likely to be removed through primary treatment with chemical assist. It is noted that some of the BPA effluent concentrations at N and U are higher than their corresponding influent concentrations, which could be caused by a mismatch of influent and effluent time stamps. On the other hand, BPA concentrations at secondary and tertiary treatment plants are reduced, which indicates BPA removal through secondary and/or tertiary treatment. It is also worth mentioning that the prediction performance at plants N and U are different: the BPA concentrations were underestimated and overestimated, respectively. It may be because that DDMs tend to generate predictions with the least errors for all samples (i.e., converge to the average value of all samples) from primary plants. Having most of the samples used to train the three DDMs collected from plants with secondary and/or tertiary treatment may be the reason why there is bias while estimating BPA effluent concentration at plants N and U. It is thus possible that the model performance would be improved when excluding samples from plants N and U.

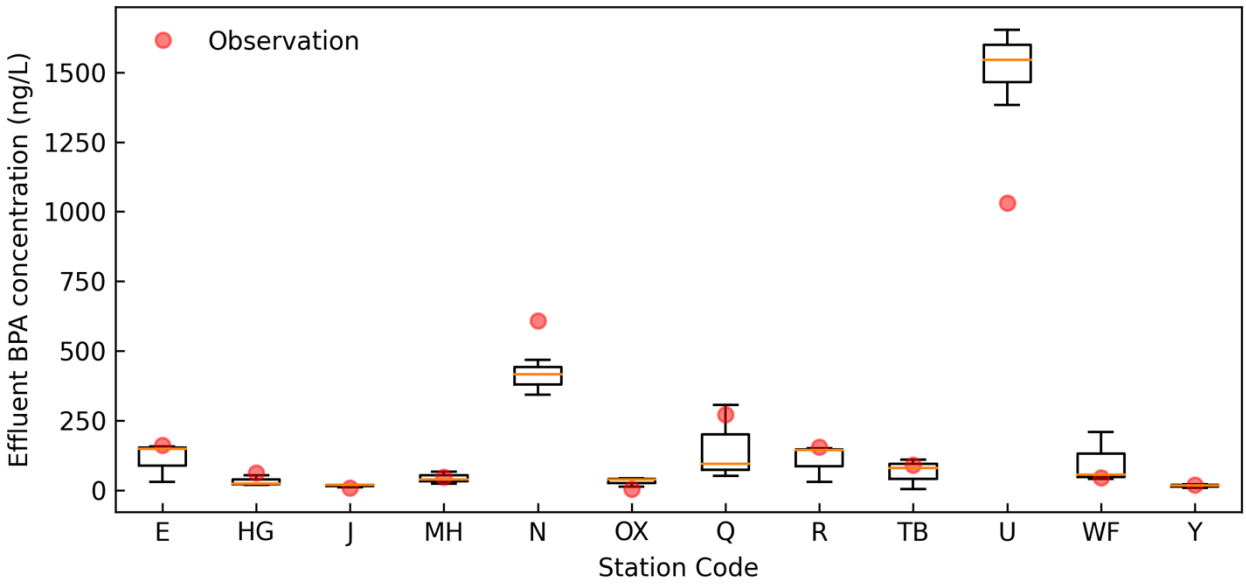

Fig. 3.4 Interval predictions and station comparison.

3.4.3 Closed-form function and feature importance

To further analyze the mechanisms of BPA removal, one of closed-form functions for the estimation of BPA effluent concentration was obtained using the GP model (Equation 3.1):

$$y = x_1 \cdot (x_{13} + x_{14}) + 2x_4 + x_6 \quad (3.1)$$

where y, x_1, x_4, x_6, x_{13}, and x_{14} represent BPA effluent concentration, BPA influent concentration, effluent temperature, winter, aerated with primary treatment, and chemical assist, respectively (Fig 3. 2). It is implied that these features are important when the GP model makes its predictions. To further understand how the identified features in the equation (i.e., x_1, x_4, x_6, x_{13}, and x_{14}) affect BPA effluent concentration, Sobol's sensitivity analysis, which generates samples by Saltelli's extension of Sobol's sequence, is adopted (Saltelli, 2002; Sobol, 2001). The first-order sensitivities of the five features are shown in Fig. 3.5a. It illustrates that BPA influent concentration, aerated with primary treatment, and chemical assist exhibit first-order sensitivities

but winter and effluent temperature appear to have no first-order effects. BPA influent concentration is the most sensitive feature among the five.

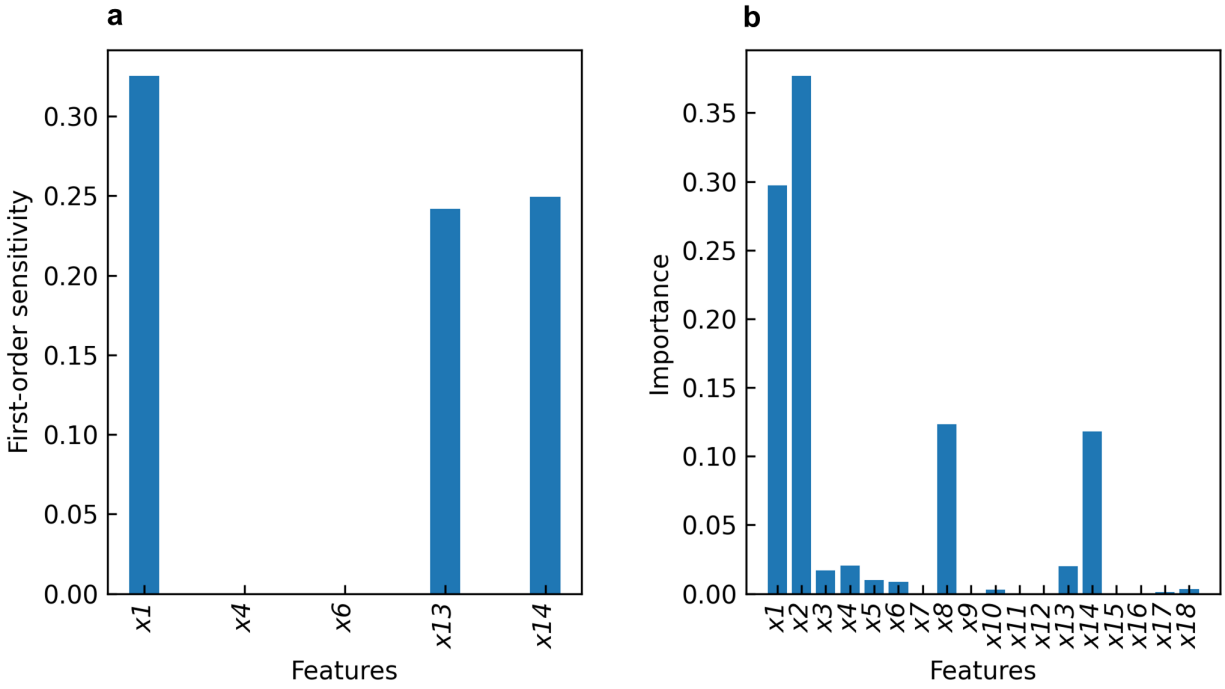

Fig. 3.5 Feature importance results: (a) first-order sensitivities of identified features; (b) VIM results from the ET model

Additionally, feature importance was calculated using the ET model and the results are shown in Fig. 3.5b. Similar to the closed-form function analysis results, BPA influent concentration, winter, effluent temperature, aerated with primary treatment, and chemical assist are found to be important. It is shown that the influent flow rate has the highest importance in terms of VIM. It is worth mentioning that the influent flow rate is not an important feature of the GP model. This may be because that ET is the decision tree-based algorithm while GP is a regression algorithm. Influent flow rate thus plays different roles in splitting a tree and elementary

arithmetic. BPA influent concentration, which is the most important feature identified through the GP model, is the second most important feature of the ET model.

3.4.4 Network analysis results

While the closed-form function generated by the GP model and the VIM results produced by the ET model can provide some insights into the importance of different features, the interdependencies among these features are yet to be investigated. To address feature interdependencies, the adjacency matrix representing the interdependencies between the features influencing BPA effluent concentration was defined by field experts and the best of our knowledge for network analysis. The adjacency matrix and its corresponding directed network of features (NoF) are shown in Table 3.2 and Fig. 3.6a, respectively. Another new feature (i.e. collection date) has been included in the network analysis. In Table 3.2, when two features are independent, the corresponding element in the matrix is zero; otherwise, the element has a value of one. In Fig. 3.6a, the nodes (N) represent the features, while the links (L) represent the interdependence between two features. The NoF developed is an unweighted-directed network with $N = 19$ and $L = 35$, where a link directed from feature 1 to feature 2 indicates feature 2's dependency on feature 1.

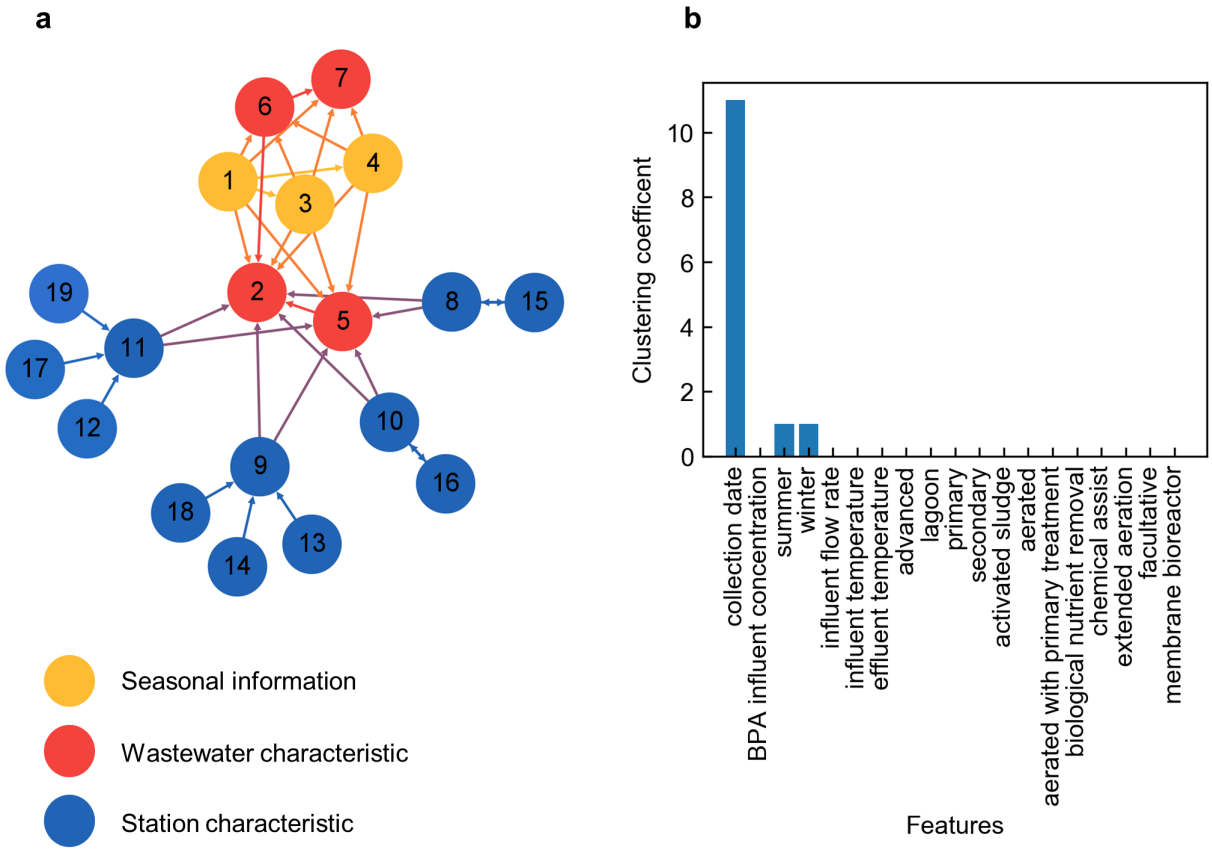

Fig. 3.6 The Network of Features: all features are numbered (1-collection date, 2-BPA influent concentration, 3-summer, 4-winter, 5-influent flow rate, 6-influent temperature, 7-effluent temperature, 8-advanced, 9-lagoon, 10-primary, 11-secondary, 12-activated sludge, 13-aerated, 14-aerated with primary treatment, 15-biological nutrient removal, 16-chemical assist, 17-extended aeration, 18-facultative, 19-membrane bioreactor)

Table 3.2 Matrix representing the interdependencies among the features

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
F1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
F2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
F4	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
F5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F8	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F9	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F10	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
F11	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F12	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
F14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
F15	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
F16	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
F17	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F18	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
F19	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

F1-collection date
F2-BPA influent concentration
F3-summer
F4-winter
F5-influent flow rate
F6-influent temperature
F7-effluent temperature
F8-advanced
F9-lagoon
F10-primary
F11-secondary
F12-activated sludge
F13-aerated
F14-aerated with primary treatment
F15-biological nutrient removal
F16-chemical assist
F17-extended aeration
F18-facultative
F19-membrane bioreactor

In comparison to the networks provided by Narayanan et al. (2021) and Gao et al. (2017), which were a dense network (D = 0.21) and a sparse network (D = 0.06) respectively, the NoF in this study implies moderate interdependencies among its features with D equal to 0.11. It is indicated that the influencing factors of BPA effluent concentration are moderately interdependent, thus identifying the dominant features by the network analysis is necessary. The k_{in} and k_{out} of

the NoF are both 1.84, which implies that every feature in this study is affected by or affects an average of approximately two other features. This reiterates the need to study the interdependencies among these features. The criterion C_{coe} assesses how closely a node's neighbors interact with one another. Feature 1 (i.e., collection date) shows the highest C_{coe} value as shown in Fig. 3.6b, and it indicates that feature 1 and its neighbors (e.g., summer and influent temperature) are highly interconnected, and they form a local cluster that may have a collective impact on the BPA effluent concentration. Similarly, summer and winter with their respective neighbors form two local clusters, and these clusters may have a collective impact on the BPA effluent concentration.

To further analyze the characteristics of the network, centrality measures that reflect the proportional importance of individual nodes on the overall network are introduced. Features 2 and 5 (i.e., BPA influent concentration and influent flow rate) have the highest C_c values as shown in Fig. 3.7a, implying that they are linked to multiple features forming intricately connected sets. These connected sets might collectively affect the BPA effluent concentration. This supports the findings from Section 3.4.3 that BPA influent concentration and influent flow rate affect the BPA effluent concentration most. Features 3, 4, 10, and 16 (i.e., summer, winter, primary, and chemical assist) have negligible C_c values, which indicates a very low influence on BPA effluent concentration. Fig. 3.7b shows that the B_c values of features 9 and 11 (i.e., lagoon and secondary) are higher than the others. This means these features act as primary connectors in the NoF, which implies that they play important roles in the system. Similarly, as shown in Fig. 3.7c, factors 2, 5, 8, and 10 (i.e., BPA influent concentration, influent flow rate, advanced, and primary) have higher E_c values, which means they have predominant indirect connections to factors other than their close neighbors. These features are likely to impact the BPA effluent concentration significantly.

Overall, BPA influent concentration, influent flow rate, primary, and advanced are identified as important features in the network analysis. When making BPA management and remediation strategies, these features require more attention.

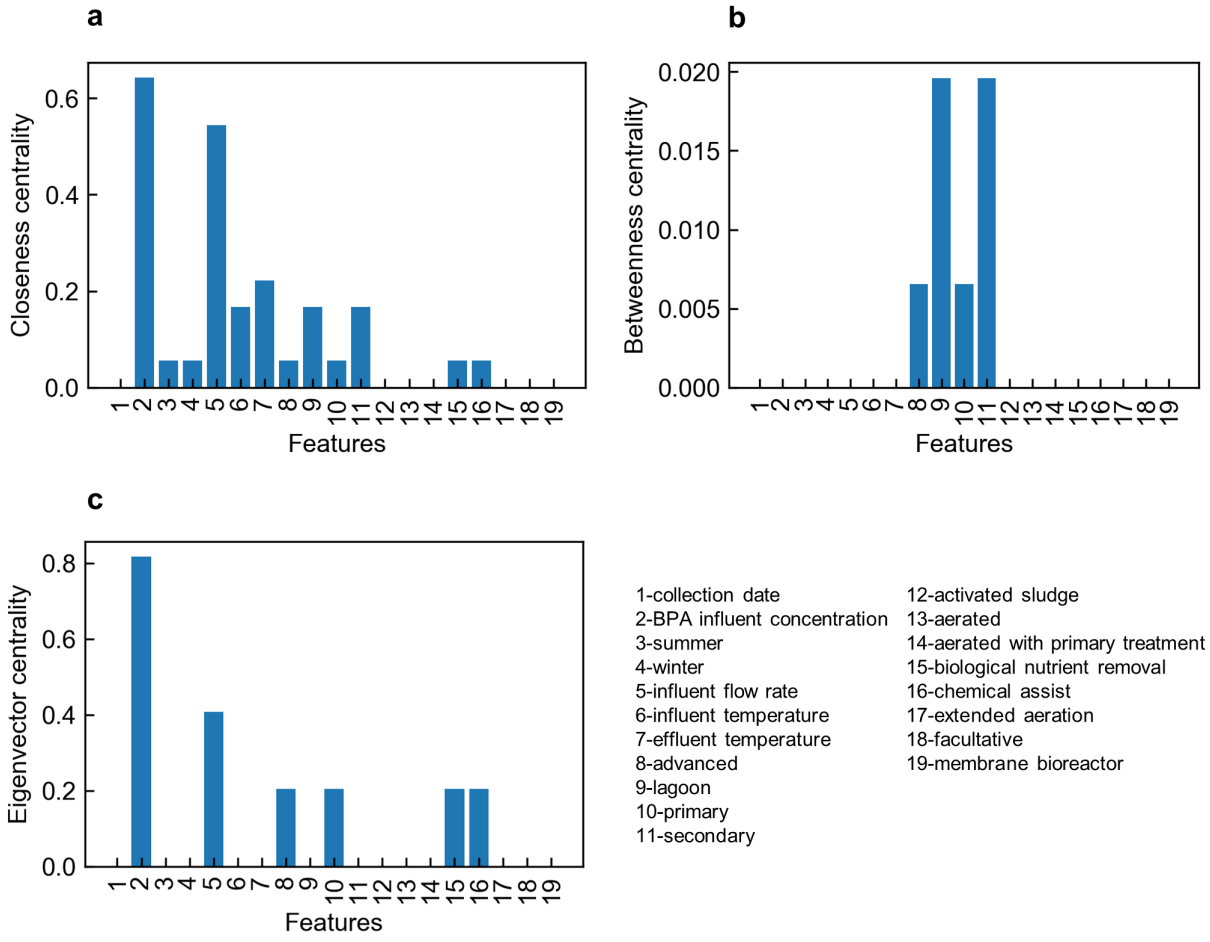

Fig. 3.7 Centrality measures

3.5 Conclusions

In this study, an integrated framework was proposed for the prediction of BPA effluent concentration at Canadian municipal WWTPs. The framework consists of two major parts: (1) DDMs for effluent BPA prediction; and (2) a network for feature dependencies analysis. Specifically, MLT-NN, GP, and ET models were applied to address the data sparsity problem and

generate effluent BPA predictions. BPA influent concentration, seasonal information, influent flow rate, influent temperature, effluent temperature, and characteristics of wastewater treatment plants were used for predictions of BPA effluent concentration. The performance of the proposed models was evaluated by *MAPE*, *RMSE*, and R^2. The results showed that the models could alleviate the data sparsity and imbalance problem and provide fair predictions.

In addition, to address the lack of analysis of features' impact on BPA effluent concentration and its interdependencies, a closed-form function from GP, the variable importance measure (VIM) method, and the theory of networks were proposed in this study. The results of both closed-form function and VIM imply that BPA influent concentration and primary treatment with chemical assist are the most important features for the prediction of effluent BPA. The results of network analysis demonstrated that the interdependencies among the input features are moderate. It was also found that BPA influent concentration, influent flow rate, primary, and advance are important influencing factors of effluent BPA concentration and thus are important for enhancing BPA removal at WWTPs.

This research proposed a new framework for effluent BPA prediction and the analysis of its influencing factors. This framework could be leveraged to study many other emerging contaminants for their removal during wastewater treatment. Collecting samples of emerging contaminants from scratch and making the proposed framework capable of generating reasonable predictions for extreme values might be major potential challenges.

Data Availability

The data are available at https://open.canada.ca/data/en/dataset/417e59c8-340e-4f6c-b139-1287cd5bd9d9/resource/e3b34fde-138d-4cb2-9ba6-3a532cb2ec00.

Acknowledgments

This study is supported by the MacDATA Institute at McMaster University, Canada. We would like to express appreciation to fellows at MacDATA institute for their advice.

Appendix

Table 3A-1 Performance of different models

Criteria / Models	*MAPE*	*RMSE (ng/L)*	R^2
GP	3.15	148.08	0.845
ET	0.26	155.02	0.859
MLT-NN	0.89	206.19	0.784

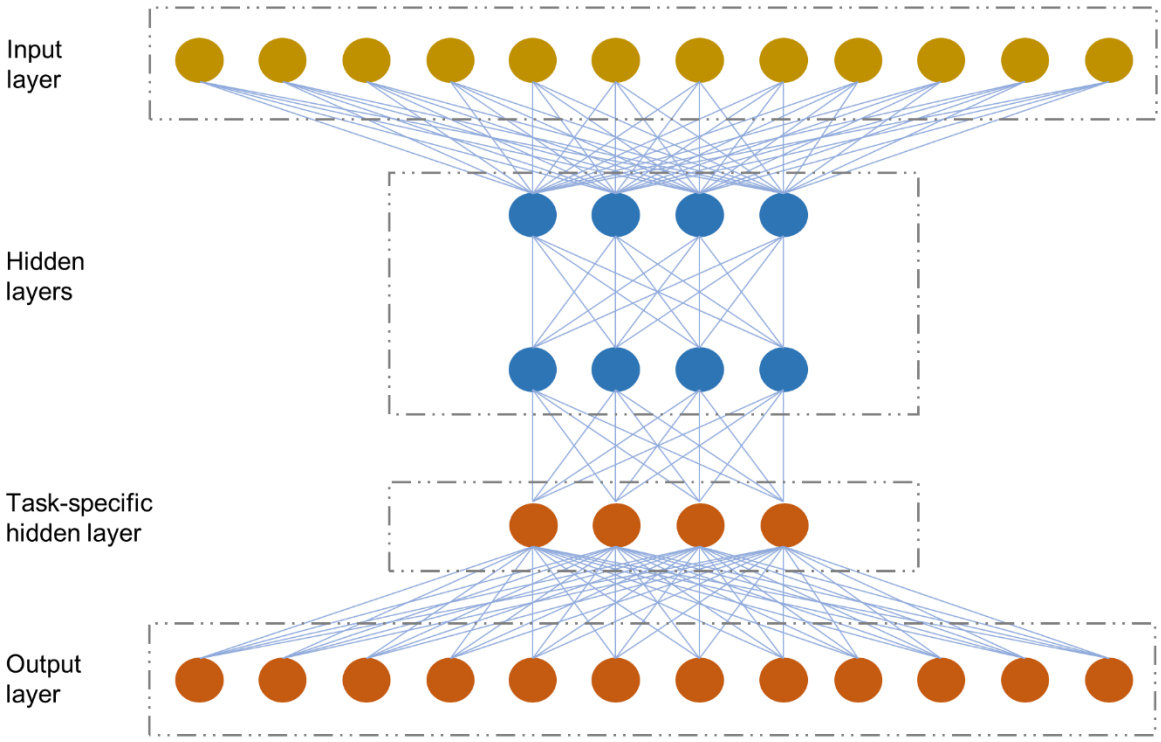

Fig. A-1 The MTL-NN structure adopted in this study

References

Ahmad, M.W., Reynolds, J., Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J Clean Prod 203, 810–821. https://doi.org/10.1016/J.JCLEPRO.2018.08.207

Almeida, J.S., 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol 13, 72–76. https://doi.org/10.1016/S0958-1669(02)00288-4

Anand, D., 2012. Feature Extraction for Collaborative Filtering: A Genetic Programming Approach.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science (1979) 286, 509–512. https://doi.org/10.1126/SCIENCE.286.5439.509

Brugnera, M.F., Rajeshwar, K., Cardoso, J.C., Zanoni, M.V.B., 2010. Bisphenol A removal from wastewater using self-organized TIO2 nanotubular array electrodes. Chemosphere 78, 569–575. https://doi.org/10.1016/J.CHEMOSPHERE.2009.10.058

Dorado-Moreno, M., Navarin, N., Gutiérrez, P.A., Prieto, L., Sperduti, A., Salcedo-Sanz, S., Hervás-Martínez, C., 2020. Multi-task learning for the prediction of wind power ramp events with deep neural networks. Neural Networks 123, 401–411. https://doi.org/10.1016/J.NEUNET.2019.12.017

Dürrenmatt, D.J.Ô., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. Environmental Modelling & Software 30, 47–56. https://doi.org/10.1016/J.ENVSOFT.2011.11.007

Gao, S., Zhen, Z., Li, Z., Zhao, Y., Qin, X., 2017. Complex Network Model for Characterizing Hazards and Risks Associated with Mine-tailings Facility. Geo-Resources Environment and Engineering (GREE) 2, 101–107. https://doi.org/10.15273/GREE.2017.02.019

Gewurtz, S.B., Tardif, G., Power, M., Backus, S.M., Dove, A., Dubé-Roberge, K., Garron, C., King, M., Lalonde, B., Letcher, R.J., Martin, P.A., McDaniel, T. v., McGoldrick, D.J., Pelletier, M., Small, J., Smyth, S.A., Teslic, S., Tessier, J., 2021. Bisphenol A in the Canadian environment: A multimedia analysis. Science of The Total Environment 755, 142472. https://doi.org/10.1016/J.SCITOTENV.2020.142472

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering 9, 143–151. https://doi.org/10.1016/0954-1810(94)00011-S

Guerra, P., Kim, M., Teslic, S., Alaee, M., Smyth, S.A., 2015. Bisphenol-A removal in various wastewater treatment processes: Operational conditions, mass balance, and optimization. J Environ Manage 152, 192–200. https://doi.org/10.1016/J.JENVMAN.2015.01.044

Jhones dos Santos, A., Sirés, I., Brillas, E., 2021. Removal of bisphenol A from acidic sulfate medium and urban wastewater using persulfate activated with electroregenerated Fe2+. Chemosphere 263, 128271. https://doi.org/10.1016/J.CHEMOSPHERE.2020.128271

Kingsford, C., Salzberg, S.L., 2008. What are decision trees? Nature Biotechnology 2008 26:9 26, 1011–1013. https://doi.org/10.1038/nbt0908-1011

Kitamura, S., Suzuki, T., Sanoh, S., Kohta, R., Jinno, N., Sugihara, K., Yoshihara, S., Fujimoto, N., Watanabe, H., Ohta, S., 2005. Comparative Study of the Endocrine-Disrupting Activity of Bisphenol A and 19 Related Compounds. Toxicological Sciences 84, 249–259. https://doi.org/10.1093/TOXSCI/KFI074

K'oreje, K.O., Okoth, M., van Langenhove, H., Demeestere, K., 2020. Occurrence and treatment of contaminants of emerging concern in the African aquatic environment: Literature review and a look ahead. J Environ Manage 254, 109752. https://doi.org/10.1016/j.jenvman.2019.109752

Lee, H.-B., Peart, T.E., 2000. Bisphenol A Contamination in Canadian Municipal and Industrial Wastewater and Sludge Samples. Water Quality Research Journal 35, 283–298. https://doi.org/10.2166/WQRJ.2000.018

Michelucci, U., Venturini, F., 2019. Multi-Task Learning for Multi-Dimensional Regression: Application to Luminescence Sensing. Applied Sciences 2019, Vol. 9, Page 4748 9, 4748. https://doi.org/10.3390/APP9224748

Murugananthan, M., Yoshihara, S., Rakuma, T., Shirakashi, T., 2008. Mineralization of bisphenol A (BPA) by anodic oxidation with boron-doped diamond (BDD) electrode. J Hazard Mater 154, 213–220. https://doi.org/10.1016/J.JHAZMAT.2007.10.011

Narayanan, B.L., Yosri, A., Ezzeldin, M., El-Dakhakhni, W., Dickson-Anderson, S., 2021. A complex network theoretic approach for interdependence investigation: An application to radionuclide behavior in the subsurface. Comput Geosci 157, 104913. https://doi.org/10.1016/J.CAGEO.2021.104913

Natarajan, Senthilselvan, Vairavasundaram, S., Natarajan, Sivaramakrishnan, Gandomi, A.H., 2020. Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data. Expert Syst Appl 149, 113248. https://doi.org/10.1016/J.ESWA.2020.113248

Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. Water Res 157, 498–513. https://doi.org/10.1016/J.WATRES.2019.03.030

Oliveira, M. de, Frihling, B.E.F., Velasques, J., Filho, F.J.C.M., Cavalheri, P.S., Migliolo, L., 2020. Pharmaceuticals residues and xenobiotics contaminants: Occurrence, analytical techniques and sustainable alternatives for wastewater treatment. Science of The Total Environment 705, 135568. https://doi.org/10.1016/j.scitotenv.2019.135568

O'Neill, M., Vanneschi, L., Gustafson, S., Banzhaf, W., 2010. Open issues in genetic programming. Genetic Programming and Evolvable Machines 2010 11:3 11, 339–363. https://doi.org/10.1007/S10710-010-9113-2

Patel, N., Khan, MD.Z.A., Shahane, S., Rai, D., Chauhan, D., Kant, C., Chaudhary, V.K., 2020. Emerging Pollutants in Aquatic Environment: Source, Effect, and Challenges in Biomonitoring and Bioremediation- A Review. Pollution 6, 99–113. https://doi.org/10.22059/POLL.2019.285116.646

Pookpoosa, I., Jindal, R., Morknoy, D., Tantrakarnapa, K., 2015. Occurrence and efficacy of bisphenol A (BPA) treatment in selected municipal wastewater treatment plants, Bangkok, Thailand. Water Science and Technology 72, 463–471. https://doi.org/10.2166/WST.2015.232

Salimi, M., Esrafili, A., Gholami, M., Jonidi Jafari, A., Rezaei Kalantary, R., Farzadkia, M., Kermani, M., Sobhi, H.R., 2017. Contaminants of emerging concern: a review of new approach in AOP technologies. Environ Monit Assess 189, 414. https://doi.org/10.1007/s10661-017-6097-x

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput Phys Commun 145, 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1

Sharan, R., Ulitsky, I., Shamir, R., 2007. Network-based prediction of protein function. Mol Syst Biol 3, 88. https://doi.org/10.1038/MSB4100129

Sivakumar, B., 2014. Networks: a generic theory for hydrology? Stochastic Environmental Research and Risk Assessment 2014 29:3 29, 761–771. https://doi.org/10.1007/S00477-014-0902-7

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55, 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Tong, X., You, L., Zhang, J., He, Y., Gin, K.Y.H., 2022. Advancing prediction of emerging contaminants in a tropical reservoir with general water quality indicators based on a hybrid process and data-driven approach. J Hazard Mater 430, 128492. https://doi.org/10.1016/J.JHAZMAT.2022.128492

Vladislavleva, K., Veeramachaneni, K., Burland, M., Parcon, J., O'Reilly, U.M., 2010. Knowledge mining with genetic programming methods For variable selection in flavor design. Proceedings of the 12th Annual Genetic and Evolutionary Computation Conference, GECCO '10 941–948. https://doi.org/10.1145/1830483.1830651

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442. https://doi.org/10.1038/30918

Xu, L., Yang, L., Johansson, E.M.J., Wang, Y., Jin, P., 2018. Photocatalytic activity and mechanism of bisphenol a removal over TiO2−x/rGO nanocomposite driven by visible light. Chemical Engineering Journal 350, 1043–1055. https://doi.org/10.1016/J.CEJ.2018.06.046

Xue, A.Y., Qi, J., Xie, X., Zhang, R., Huang, J., Li, Y., 2014. Solving the data sparsity problem in destination prediction. The VLDB Journal 2014 24:2 24, 219–243. https://doi.org/10.1007/S00778-014-0369-7

Zhang, Y., Yang, Q., 2018. An overview of multi-task learning. Natl Sci Rev 5, 30–43. https://doi.org/10.1093/NSR/NWX105

Zhou, P., Li, Z., Snowling, S., Baetz, B.W., Na, D., Boyd, G., 2019. A random forest model for inflow prediction at wastewater treatment plants. Stochastic Environmental Research and Risk Assessment 2019 33:10 33, 1781–1792. https://doi.org/10.1007/S00477-019-01732-9

Chapter 4 – Impact of COVID-19 Lockdowns through Data-driven

Pandemics have posed new challenges to wastewater modeling as they change old patterns, such as residents' water consumption patterns. To address such challenges, assessing the impact of COVID-19 is crucial for improving wastewater modeling. In this study, a comparison of influent flow rates before and during lockdowns was conducted. No-lockdown scenario data were generated by random forest models. Weekly patterns of influent flow exhibited differences before and during lockdowns. There is less variability of influent flow rate during lockdowns compared to before lockdowns. A spike in influent flow rates is observed after the easing of provincial emergency state.

This chapter has been submitted for publication consideration.

Pengxiao Zhou was responsible for Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, and Writing – Review & Editing under Dr. Zhong Li’s supervision.

Unraveling the Impact of COVID-19 Lockdowns on Canadian Municipal Sewage

Abstract

The COVID-19 pandemic and resulting lockdowns have had significant impacts on various aspects of society, including municipal sewage. This study investigates changes in Canadian municipal sewage during COVID-19 lockdowns by examining influent flow rates at two wastewater treatment plants in Ontario, Canada. A comparison of weekly patterns and daily average flow rates before and during lockdowns was conducted. The observed influent flow rates were also compared with predicted no-lockdown scenario data, which were generated by random forest models. The results showed that weekly patterns of influent flow exhibited differences before and during lockdowns, and there is less variability of influent flow rate during lockdowns. Additionally, both plants experienced a decrease in influent flow rates during the lockdowns, and a spike after the easing of provincial emergency state. This knowledge can be used to improve wastewater management strategies and inform policy decisions during times of crisis in the future.

Keywords: COVID-19, lockdowns, influent flow rates.

4.1 Introduction

The COVID-19 pandemic has had a profound impact on the daily lives of people around the world (Gautam and Hens, 2020; Khan et al., 2020). In an effort to slow the spread of the virus, governments of over 100 countries have implemented various measures, including lockdowns that restrict the movement and activities of billions of people (Alfano and Ercolano, 2020; Dunford et al., 2020; Hien Lau et al., 2020). For example, a provincial state of emergency was declared on March 17th, 2020 in Ontario, Canada, and schools, non-essential services, and recreational facilities were closed since the same time. These lockdowns have resulted in significant changes to the way people live and work and have had far-reaching impacts on various aspects of society.

One area that has been significantly affected by the COVID-19 lockdowns is municipal sewage (Abu-Bakar et al., 2021; Hillary et al., 2021; Nemati and Tran, 2022; Wurtzer et al., 2020). As people spend more time at home and engage in different activities, their water usage patterns may change. This could lead to an impact on the amount of sewage produced by households and affect the operations and management of wastewater treatment plants. Specifically, if water usage increases during the lockdowns, it results in an increase in the volume of generated sewage. This would put additional strain on sewage treatment and disposal systems, which could make it challenging to maintain desired treatment levels (Boyd et al., 2019; Zhou et al., 2019). On the other hand, if water usage decreases during the lockdowns, it may also result in other challenges, such as the need to adjust sewage treatment processes to handle lower volumes of wastewater (Sperling, 2015). Therefore, understanding how these lockdowns affect the amount of sewage is important and valuable.

In this study, we aim to investigate the changes in Canadian municipal sewage during the COVID-19 lockdowns by examining the inflow rate data at municipal wastewater treatment plants.

The influent flow rate of a wastewater treatment plant refers to the rate at which wastewater is brought from homes and small businesses into the plant for treatment. It is typically measured in units of volume per unit time, such as liter per day. The influent flow rate is an important factor in the design and operation of wastewater treatment plants, as it determines not only the capacity of the plant and the size of treatment equipment but also the control of chemical dosing and aeration rate (Andreides et al., 2022; Zhou et al., 2022). Influent flow rate data can provide an insight into how much water is being used in a given area, as well as how the patterns of water usage have been affected by lockdown measures. For example, we can see if there has been a change in the overall amount of water being used, as well as if there have been shifts in the times of day when water is being used the most. This knowledge is beneficial for wastewater simulations and can provide valuable insights into the impact of lockdowns on water management, and can also be used to inform policy decisions and improve water management strategies during times of crisis in the future.

4.2 Materials and methods

Two anonymous wastewater treatment plants (Plants A and B) located in Ontario, Canada were selected for this study. The influent flow rate at Plant A was measured on a hourly basis from November 1, 2016 to August 3, 2021. While that at Plant B was measured on a 15-minute basis from January 1, 2019 to November 30, 2021. There were a small number of missing data points, and simple linear interpolation was adopted to fill in the gaps by estimating the values based on the surrounding known data. Wastewater collection systems of both Plant A and B were designed to gather wastewater from homes, and small businesses. Plant A has separate storm and sanitary sewers, while Plant B has a small portion of its sewers that combine both types of water. However, because of the downspouts and sump pumps illegally connected to the sanitary system, both plants

experience increased inflow during rainfall events. To distinguish the impacts of meteorological conditions from those of lockdowns, hourly meteorological data, including precipitation, snow depth, air temperature, humidity, pressure, wind direction, and wind speed, were collected and analyzed for both plants in the same time periods.

The data before March 14th, 2020 were labelled as before lockdowns, while data from that date forward were labelled as during lockdowns. March 14th, 2020 was selected as an important date break here because it marked the start of the first lockdown in Ontario, during which schools, non-essential services, and recreational facilities were closed. To understand how the patterns of water usage have been affected by lockdown measures, we firstly compared weekly influent flow rate patterns before and during lockdowns. All collected influent flow rate data were converted to hourly frequency by calculating the mean of all sample values in that hour. The data were organized by day of the week, and the pattern of the first week during lockdowns (from March 14th, 2020 to March 20th, 2020) and the pattern of the one that two weeks before lockdowns (from February 29th, 2020 to March 6th, 2020) were compared. To provide more benchmark data, the weekly pattern from the same period last year (From March 16th, 2019 to March 22th, 2019) and the average weekly pattern from year 2019 were added. Further, to investigate if there was a change in the overall amount of water being used during lockdowns, daily average influent flow rates before and during lockdowns were examined. A daily average influent flow rate was calculated as the average of all the instances on that day.

The variations for observed influent flow rates can be attributed to the combined effects of changes in both residents' water usage patterns and meteorological conditions. Random forest (RF) models, which have been proved to effectively predict influent flow rate under normal conditions (i.e., no lockdowns), were developed to exclude the meteorological impacts during the

implementation of lockdown measures (Breiman, 2001; Zhao et al., 2020). Hourly influent flow rates during lockdowns were compared with the same period no-lockdown scenario predictions, which were generated by RF models. The RF models were trained based on influent flow rate data before January 1st, 2020. The relationships between variables (i.e., collected meteorological variables) and target (i.e., influent flow rates) were captured. The model configurations and performance are detailed in the Appendix. Upon exclusion of the meteorological impacts modeled by RF models from the observed overall changes in influent flow rates, the impacts of residents' water usage on amounts of influent flow rates were isolated.

4.3 Results and discussion

4.3.1 Changes in weekly pattern

Fig. 4.1 and Fig. 4.A1 illustrate the weekly patterns of wastewater influent flow at wastewater treatment plants A and B during four different periods. As shown in the figure, all periods exhibit similar overall flow patterns, with higher flows during daytime and lower flows at night. However, there are some notable differences between the periods before and during lockdowns. In particular, the weekly pattern from two weeks prior to lockdowns (Fig. 4.1a) exhibits a noticeable lag between weekdays (grey and yellow) and weekends (blue and red) at around 8-10am in the morning, which disappears in the weekly pattern from the first week of lockdowns (Fig. 4.1b). Additionally, the influent flow rates before lockdowns on Friday and Saturday nights (between 9-12pm) used to be the lowest, but this trend becomes less apparent during lockdown measures were implemented. It is clear that the patterns of influent flow rate are influenced by the lockdown measures. In the average weekly pattern from 2019 (Fig. 4.1c), the influent flow rates show a lag between weekdays and weekends in the morning and are lowest on

Friday and Saturday nights. Similarly, these two characteristics can also be observed in Fig. 4.1d, although there are some fluctuations.

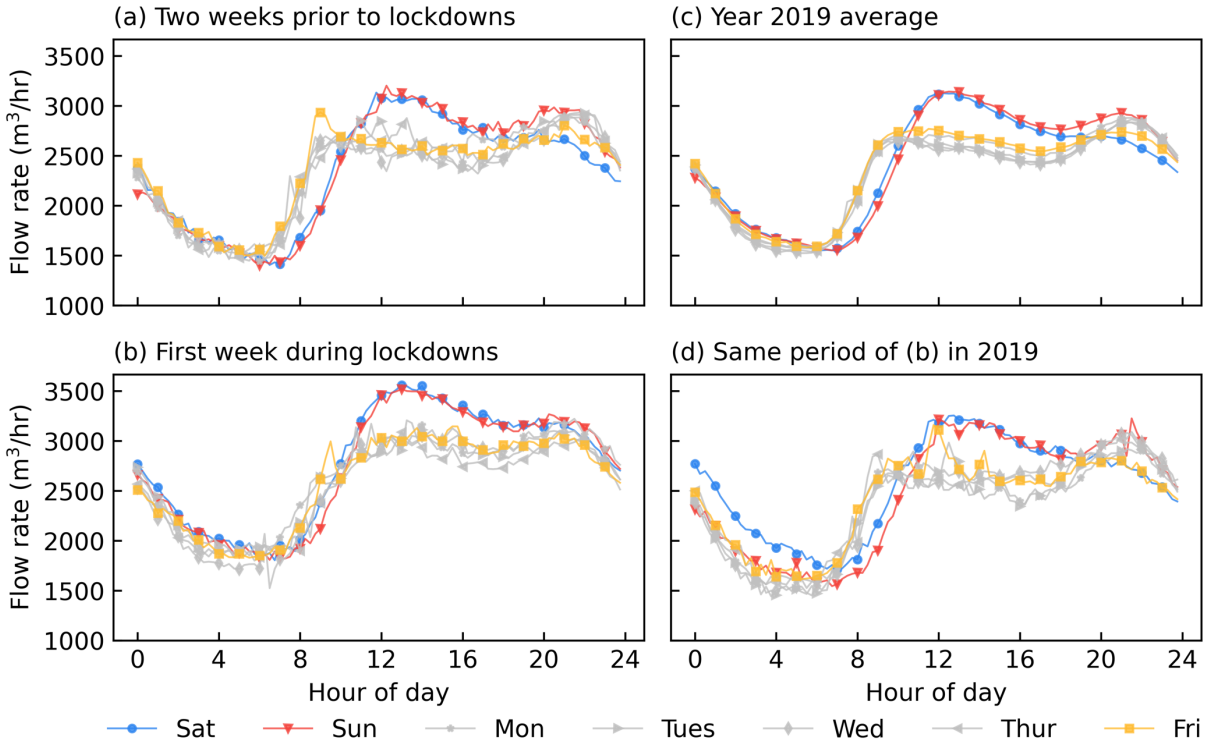

Fig. 4.1 Weekly patterns of wastewater influent flow at Plant A: (a) two weeks prior to the first lockdown, from February 29th, 2020 to March 6th, 2020; (b) the first week during lockdowns, from March 14th, 2020 to March 20th, 2020; (c) 2019 average; (d) same period of (b) in 2019

These findings suggest that there are differences in the patterns of wastewater influent flow before and during lockdowns. This may be caused by the change in household water consumption patterns, as the influent flow is mainly from sanitary sewers. It is possible that the disappearance of the lag between weekday and weekend mornings is due to the increase in remote work, which saves commuting time and thus postpones the time for residents to get up and wash. Additionally, the trend of lower influent flow rates on Friday and Saturday nights is less pronounced during lockdowns, which may be due to a reduction in evening activities and residents behaving more

like it is a workday. To gain a more thorough understanding of the factors influencing the differences in weekly influent flow patterns, further surveys with local residents about their water consumption habits across different times of the week could be useful. This information could help to shed light on how wastewater treatment plants can cope with the changes in influent flow patterns.

4.3.2 Changes in amount of influent flow

It is important to understand changes in the volume of influent flow, as it can impact the effectiveness of sewage treatment and disposal systems. Fig. 4.2 shows a comparison of daily average influent flow rates before and during the lockdowns. The data in Figs. 4.2a and 4.2b reveal a trend of regular fluctuations in influent flow rates over time, which may be due to seasonal variations. The highest influent flow rate was recorded in the spring, possibly due to snow melting.(Zhang et al., 2019) Both treatment plants show a decrease in highest influent flow rates in spring during lockdowns. This may be due to a reduction in industrial and commercial activity during lockdowns.

To further visualize and compare the changes, boxplots were generated for both plants. In Figs 4.2c and 4.2d, the notch of the box represents the median, and the lower and upper of the box are the first quartile ($Q1$) and third quartile ($Q3$), respectively. IQR is the interquartile range which equals $Q3 - Q1$. The lower whisker extends to the first datum greater than $Q1 - 2 \cdot IQR$, while the upper whisker extends to the last datum less than $Q3 + 2 \cdot IQR$. The results suggest that there are more outliers in the influent flow rate data collected before lockdowns (Fig 4.2c) than during lockdowns (Fig 4.2d). This indicates a difference in daily influent flow rate between the two periods. Although the median daily influent flow rate during lockdowns was similar with these before lockdowns for both plants, the range ($Q3 - Q1$) of water usage during lockdowns was

narrower than before lockdowns. This suggests that there is less variability and lower volumes of influent flow rate during lockdowns compared to before.

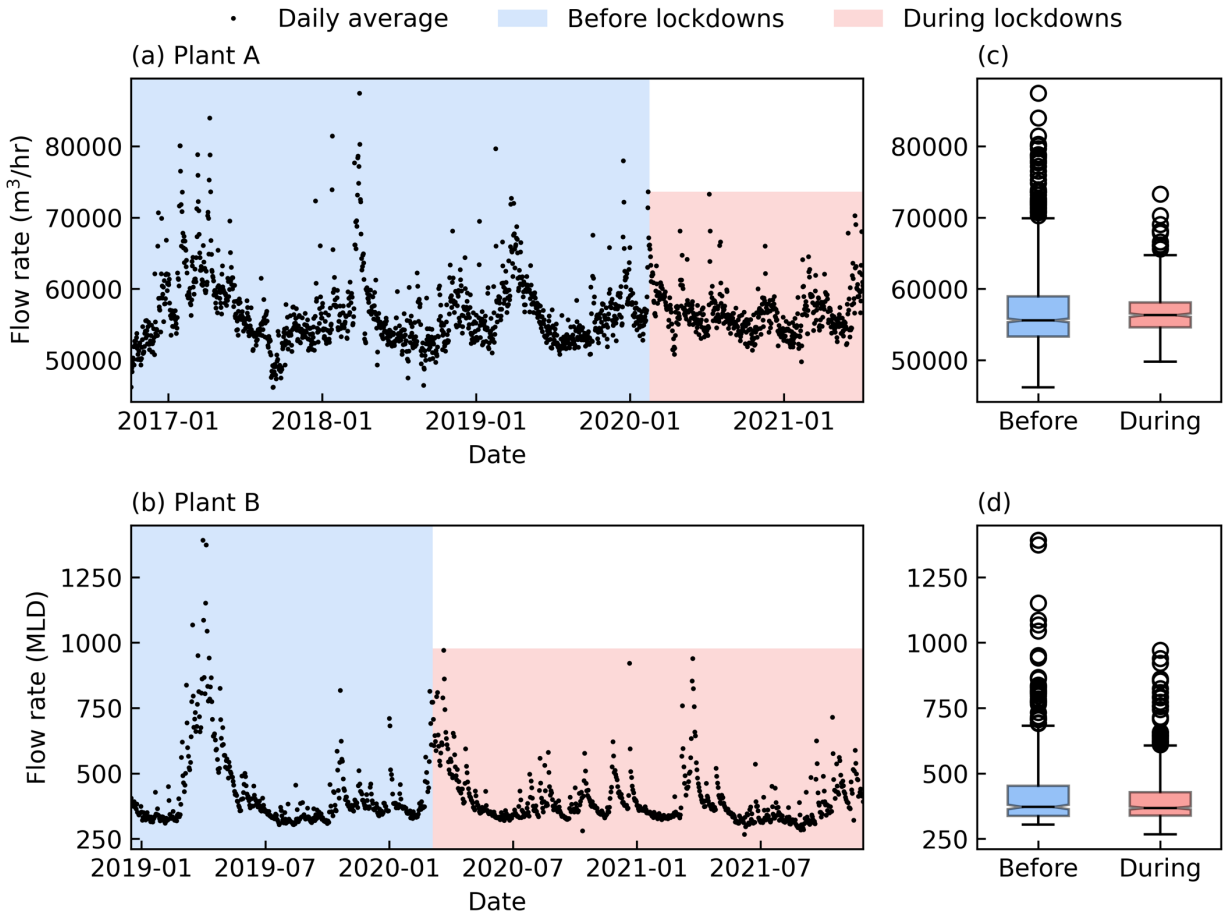

Fig. 4.2 Changes in the volume of influent flow: (a) daily average influent flow rate at plant A, (b) daily average influent flow rate at plant B, (c) a comparison of influent flow rate at Plant A before and during lockdowns, (d) a comparison of influent flow rate at Plant B before and during lockdowns

4.3.3 Impact of lockdown measures

To exam the impact of lockdown measures on wastewater production, the observed influent flow rates at Plants A and B were compared with historical data and predicted no-lockdown scenario data in Fig. 4 3. The historical data consisted of influent flow rates in previous years. The predicted no-lockdown scenario data, on the other hand, was based on RF models. The differences between observed and predicted no-lockdown scenario series can be treated as the impacts of lockdown measures. It was found that the influent flow rates during the provincial state of emergency were significantly lower than both the historical data and predicted no-lockdown scenario data. This indicates that the lockdown measures had a remarkable effect on wastewater production. Specifically, the decrease in influent flow rates suggests that the lockdown measures resulted in a reduction in the amount of wastewater being produced. There could be several reasons for this decrease. For example, the lockdown measures may have resulted in a reduction in the number of people going out, leading to less wastewater being produced possibly from showers and flushing toilets. The lockdowns may also have resulted in a decrease in industrial and commercial activities, which could have contributed to the lower influent flow rates. Interestingly, the easing of provincial state of emergency could result in a spike in the influent flow rates at both water treatment plants. This could be due to more people returning to work or resuming normal daily activities and an increase in industrial and commercial activities.

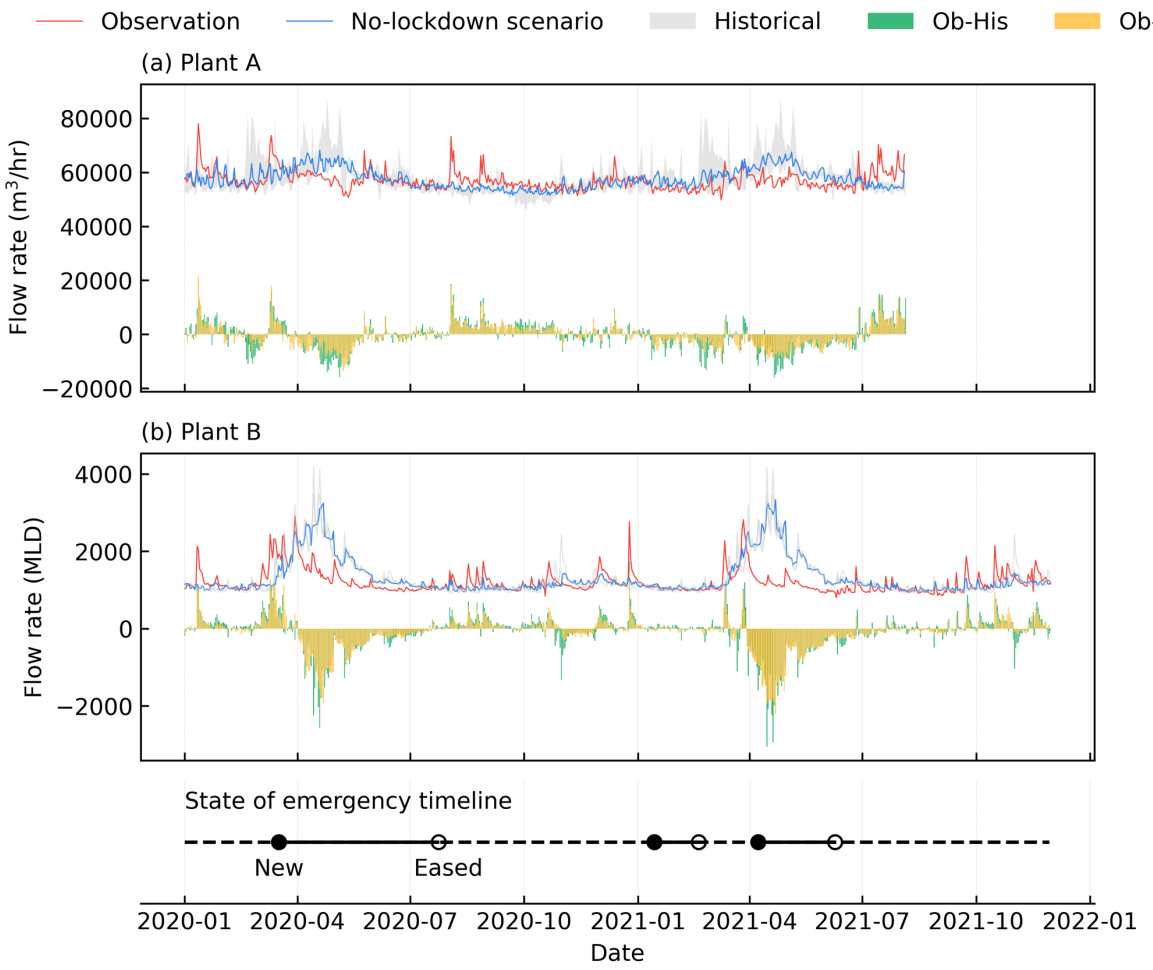

Fig. 4.3 The observed influent flow rates (Ob) at Plants A and B comparing to historical data (His) and predicted no-lockdown scenario data (NoL)

This study analyzed the impact of lockdown measures on wastewater production in two cities by comparing current influent flow rates with historical data and predicted data. The results showed that the lockdown measures had a significant impact on residents' water consumption habits and wastewater production: (1) the difference in influent flow rates between weekdays and weekends in the morning disappeared during lockdowns, and the trend of lower influent flow rates on Friday and Saturday nights was less pronounced; (2) there was less variability in influent flow rates during lockdowns compared to before lockdowns; (3) Both plants saw a decrease in influent

flow rates during lockdowns, followed by a spike in influent flow rates after the easing of the provincial emergency state. This analysis was conducted on the basis of the monitoring data from wastewater treatment plants. Further research and survey on local residents' water consumption habits could help to understand the factors influencing the changes in influent flow patterns.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Availability of Data and Materials

The data that support the findings of this study are available from Hatch Ltd., and restrictions apply to the availability of these data. Data are available on request with the permission of Hatch Ltd.

Appendix

Predictions of influent flow rate under the no-lockdown scenario were generated by random forest (RF) models. The predictor variables included precipitation, snow depth, air temperature, humidity, pressure, wind direction, wind speed, hour of day, day of the week, and month of year. The target variable was the hourly influent flow rate at the current time step. The collected data for Plant A was sequentially split into a training set (from November 1, 2016 to December 31, 2019) and a testing set (from January 1, 2020 to August 3, 2021), while that for Plant B was also sequentially split into a training set (from January 1, 2019 to December 31, 2019) and a testing set (from January 1, 2020 to November 30, 2021).The random forest model for each plant was trained on the training set. The predictions made on the testing sets were regarded as no-lockdown scenario influent flow rates. The models were implemented using the scikit-learn library in Python. To improve the models' performance, hyperparameter tuning was performed using

cross-validation on the training set. The hyperparameters that were tuned included the maximum depth of each tree. Root mean square error (RMSE) defined in scikit-learn library was used as a criterion to evaluate the models' performance (Fig. 4.A2).

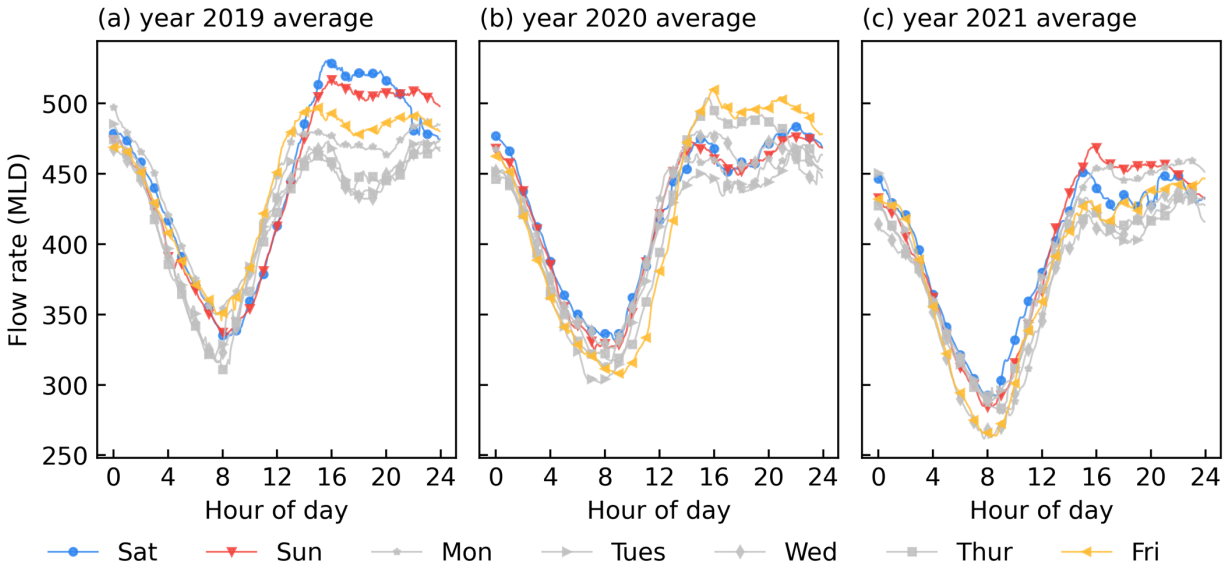

Fig. 4.A1 Weekly pattern of Plant B: (a) year 2019 average before lockdowns, (b) year 2020 average mostly during lockdowns, (c) year 2021 average during lockdowns

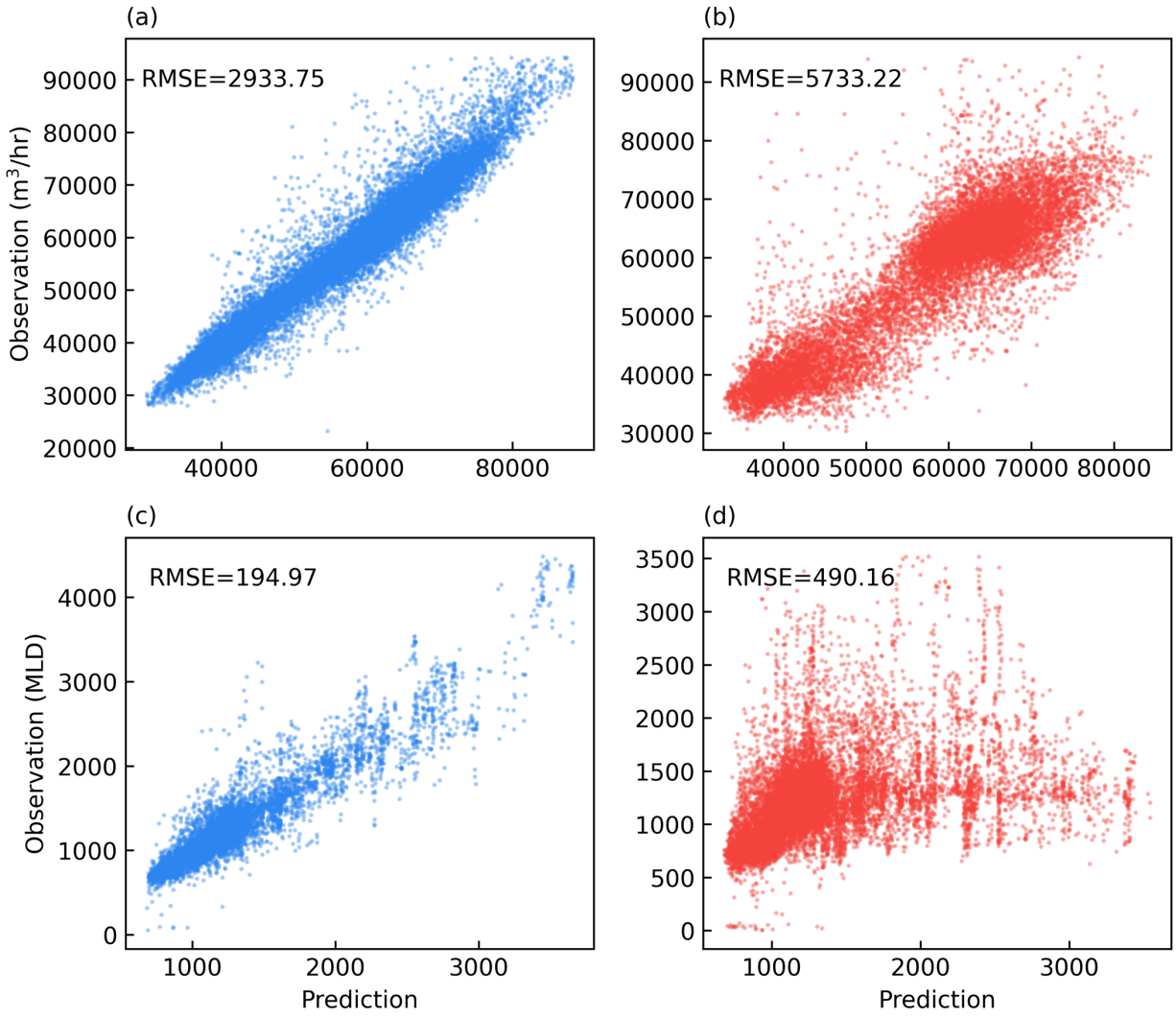

Fig. 4.A2 Scatter plots of RF models: (a) predictions made on training set for Plant A, (b) predictions made on testing set for Plant A, (c) predictions made on training set for Plant B, (d) predictions made on testing set for Plant B

References

Abu-Bakar, H., Williams, L., Hallett, S.H., 2021. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. npj Clean Water 2021 4:1 4, 1–9. https://doi.org/10.1038/s41545-021-00103-8

Alfano, V., Ercolano, S., 2020. The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel Analysis. Applied Health Economics and Health Policy 2020 18:4 18, 509–517. https://doi.org/10.1007/S40258-020-00596-3

Andreides, M., Dolejš, P., Bartáček, J., 2022. The prediction of WWTP influent characteristics: Good practices and challenges. Journal of Water Process Engineering 49, 103009. https://doi.org/10.1016/J.JWPE.2022.103009

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q., Zhou, P., 2019. Influent forecasting for wastewater treatment plants in North America. Sustainability (Switzerland) 11. https://doi.org/10.3390/su11061764

Breiman, L., 2001. Random Forests 45, 5–32.

Dunford, D., Dale, B., Stylianou, N., Lowther, E., Ahmed, M., de la Torre Arenas, I., 2020. Coronavirus: The world in lockdown in maps and charts. BBC News 9, 462.

Gautam, S., Hens, L., 2020. COVID-19: impact by and on the environment, health and economy. Environment, Development and Sustainability 2020 22:6 22, 4953–4954. https://doi.org/10.1007/S10668-020-00818-7

Hien Lau, Veria Khosrawipour, Piotr Kocbach, Agata Mikolajczyk, Justyna Schubert, Jacek Bania, Tanja Khosrawipour, 2020. The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. J Travel Med 27, 37. https://doi.org/10.1093/jtm/taaa037

Hillary, L.S., Farkas, K., Maher, K.H., Lucaci, A., Thorpe, J., Distaso, M.A., Gaze, W.H., Paterson, S., Burke, T., Connor, T.R., McDonald, J.E., Malham, S.K., Jones, D.L., 2021. Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. Water Res 200, 117214. https://doi.org/10.1016/J.WATRES.2021.117214

Khan, I., Shah, D., Shah, S.S., 2020. COVID-19 pandemic and its positive impacts on environment: an updated review. International Journal of Environmental Science and Technology 2020 18:2 18, 521–530. https://doi.org/10.1007/S13762-020-03021-3

Nemati, M., Tran, D., 2022. The Impact of COVID-19 on Urban Water Consumption in the United States. Water 2022, Vol. 14, Page 3096 14, 3096. https://doi.org/10.3390/W14193096

Sperling, M. von, 2015. Basic Principles of Wastewater Treatment. Water Intelligence Online 6, 9781780402093–9781780402093. https://doi.org/10.2166/9781780402093

Wurtzer, S., Marechal, V., Mouchel, J., Maday, Y., Teyssou, R., Richard, E., Almayrac, J., Moulin, L., 2020. Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. medRxiv 2020.04.12.20062679. https://doi.org/10.1101/2020.04.12.20062679

Zhang, Q., Li, Z., Snowling, S., Siam, A., El-Dakhakhni, W., 2019. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. Water Science and Technology 80, 243–253. https://doi.org/10.2166/WST.2019.263

Zhao, Y., Zhang, K., Xu, X., Shen, H., Zhu, X., Zhang, Y., Hu, Y., Shen, G., 2020. Substantial Changes in Nitrogen Dioxide and Ozone after Excluding Meteorological Impacts during the COVID-19 Outbreak in Mainland China. Environ Sci Technol Lett 7, 402–408. https://doi.org/10.1021/ACS.ESTLETT.0C00304

Zhou, P., Li, Z., Snowling, S., Baetz, B.W., Na, D., Boyd, G., 2019. A random forest model for inflow prediction at wastewater treatment plants. Stochastic Environmental Research and Risk Assessment 2019 33:10 33, 1781–1792. https://doi.org/10.1007/S00477-019-01732-9

Zhou, P., Li, Z., Snowling, S., Goel, R., Zhang, Q., 2022. Multi-step ahead prediction of hourly influent characteristics for wastewater treatment plants: a case study from North America. Environmental Monitoring and Assessment 2022 194:5 194, 1–14. https://doi.org/10.1007/S10661-022-09957-Y

Chapter 5 – Online Machine Learning for Influent Flow Rate Prediction

The Online learning models accurately predict influent flow rate at wastewater plants. Models adapt to changing input-output relationships and are friendly to large data. Online learning models outperform conventional batch learning models. An optimal prediction strategy is identified through uncertainty analysis. The proposed models provide support for coping with emergencies like COVID-19.

This chapter has been submitted for publication consideration.

Pengxiao Zhou was responsible for Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, and Writing – Review & Editing under Dr. Zhong Li's supervision.

Adapting to a New Normal: Online Machine Learning for Stream Wastewater Influent Flow Rate Prediction in the Era of COVID-19

Abstract

Accurate influent flow rate prediction is important for operators and managers at wastewater treatment plants (WWTPs) as it is closely related to wastewater characteristics such as biochemical oxygen demand (BOD), total suspend solids (TSS), and pH. Previous studies have been conducted to predict influent flow rate, and it was proved that data-driven models are effective tools. However, most of these studies have focused on batch learning, which is inadequate for wastewater prediction in the era of COVID-19 as the influent pattern changed significantly. Online learning that has distinct advantages of dealing with stream data, large dataset, and changing data pattern, has a potential to address this issue. In this study, the performance of conventional batch learning models Random Forest (RF), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP), and their respective online learning models Adaptive Random Forest (aRF), Adaptive K-Nearest Neighbors (aKNN), and Adaptive Multi-Layer Perceptron (aMLP), were compared for predicting influent flow rate at two Canadian WWTPs. Online learning models achieved the highest R^2, the lowest MAPE, and the lowest RMSE compared to conventional batch learning models in all scenarios. The R^2 values on testing dataset for 24-hour ahead prediction of the aRF, aKNN, and aMLP at Plant A were 0.90, 0.73, and 0.87, respectively; these values at Plant B were 0.75, 0.78, and 0.56, respectively. The proposed online learning models are effective in making reliable predictions under changing data patterns, and they are efficient in dealing with continuous and large influent data streams. They can be used to provide robust decision support for wastewater treatment and management in the changing era of COVID-19 and also under other unprecedented emergencies that could change wastewater influent patterns.

Keywords: wastewater prediction, data stream, online learning, batch learning, influent flow rates

5.1 Introduction

Accurate prediction of influent flow rate at wastewater treatment plants (WWTPs) is crucial for the proper operation of treatment facilities (Boyd et al., 2019; Zhou et al., 2022b). This is because accurate influent flow rate prediction enables efficient use of resources, such as dosing chemicals, as influent flow rate is strongly correlated to wastewater characteristics such as biochemical oxygen demand (BOD), total suspend solids (TSS), and pH (Bechmann et al., 1999; Wei and Kusiak, 2014). Knowing influent flow rate in advance can also prevents overflows, which can lead to equipment damage and environmental pollution particularly for WWTPs with combined sewer systems during extreme weather events (Wei et al., 2012; Zhang et al., 2019).

Over the past several decades, numerous studies have been conducted to predict influent flow rate, and data-driven models have been proved to be an effective approach (Andreides et al., 2022b; Ansari et al., 2018; Ma et al., 2014; P Zhou et al., 2019; Zhu and Anderson, 2019). Despite the impressive amount of effort put on data mining of influent flow rate, most of the previous work focuses on supervised batch learning. The primary goal of supervised learning is to formulate a model that can predict a target y (e.g., influent flow rate) given a set of features X (e.g., precipitation and temperature) (Bzdok et al., 2018; Caruana and Niculescu-Mizil, 2006). Supervised batch learning can be succinctly described as a process consisting of three main steps: (1) loading and preprocessing of the data, (2) training a model on the processed data, and (3) evaluating the performance of the trained model on unseen data. Various batch learning data-driven models have been validated for influent flow rate prediction following the above process (Andreides et al., 2022b; Zhang et al., 2019). However, this method of proceeding has certain drawbacks. Particularly, the batch learning regime is not suitable for prediction problems where

there are significant changes in the input-output relationships. (Fontenla-Romero et al., 2013; Hoi et al., 2014). In the event that new data with new patterns becomes available, the model must be entirely retrained from scratch using the combination of the old and new data, which can be computationally expensive and time-consuming. This is particularly problematic in real-world influent flow rate prediction scenario where new data with new patterns is constantly arriving on an hourly or minute basis. The new patterns come with new data make the mapping function captured by the trained model disabled. During the COVID-19 pandemic, the drawbacks of batch learning approaches have become more evident. The COVID-19 pandemic has had a profound impact on the daily lives of people around the world (Gautam and Hens, 2020; Khan et al., 2020). Wastewater-Based Epidemiology has gained significant attention, and there is a growing amount of wastewater data available (Hillary et al., 2021). Additionally, to slow the spread of the virus, governments have implemented various measures, including lockdowns that close schools, non-essential services, and recreational facilities, result in restrict the movement and activities of billions of people (Alfano and Ercolano, 2020; Khan et al., 2020). As individuals spend more time at home and engage in various activities, their water consumption habits change (Abu-Bakar et al., 2021; Nemati and Tran, 2022). Correspondingly, the pattern of influent flow rate has also altered, and the mapping function inferred from pre-pandemic data is no longer valid. Conventional batch learning is effective only under situations where the entire dataset is accessible, there is an infinite amount of training time, and the underlying process of data generation does not change (Fontenla-Romero et al., 2013). To adapt changing patterns and large amount of new data in practice during the pandemic, a new learning regime for data-driven models is required.

Online machine learning that represents an important family of efficient and scalable machine learning algorithms has a potential to address these challenges. It has distinct advantages

in situations where data arrives continuously, application is large-scale, and the process underlying the data generation is changing (Fontenla-Romero et al., 2013; Hoi et al., 2014). Over the past years, a multitude of online learning algorithms have been developed (Hoi et al., 2021; Jain et al., 2014). Ensemble learner is often a preferred method for learning from data streams that are constantly changing, as it can achieve high performance without much optimization and its flexibility allows for learners to be added, updated, reset, or removed as needed (Gomes et al., 2017). Thus, adaptive Random Forests model (aRF) as a representative ensemble learner for online learning was firstly adopted to influent flow rate prediction in this study. Meanwhile, the K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) algorithms are widely used in the batch learning setting due to their effectiveness and efficiency (Ahmed et al., 2010; Taunk et al., 2019a). Both KNN and MLP were proved to be effective in influent flow rate prediction. In the challenging context of online learning, adaptive K-Nearest Neighbors (aKNN) and adaptive Multi-Layer Perceptron (aMLP) were also considered due to aKNN's simplicity and effectiveness and aMLP's ability to approximate any measurable function.

In this study, we aim to explore the applications of online learning algorithms in the prediction of wastewater influent flow rates at two Canadian municipal treatment plants. Three online learning algorithms including aRF, aKNN, and aMLLP were adopted, and they were also compared with their respective conventional batch learning algorithms RF, KNN, and MLP. This study entails the following objectives: (1) develop three online learning models and three batch learning models to predict influent flow rate at two wastewater treatment plants; (2) evaluate and compare the performance of the developed models; (3) conduct an uncertainty analysis on the developed models and find an optimal prediction strategy. The proposed models will provide robust and reliable predictions while the wastewater influent pattern changes due to COVID-19

and therefore be beneficial for wastewater operators and managers making quick responses to emergencies.

5.2 Methods

5.2.1 Batch learning models

The Random Forest (RF) that developed by Breiman is a classic ensemble method that aggregates multiple decision trees (Breiman, 2001b). It utilizes a combination of bootstrapping and random split selection to create a group of independent decision trees, which are then ensembled to make a prediction. The RF method has several advantages, such as the ability to accommodate both numerical and categorical data, being robust to changes in hyperparameters, having a lower risk of overfitting, and the ability to identify the importance of variables (Kovacs et al., 2022). This RF has been found to be effective in predicting wastewater influent flow rate (P Zhou et al., 2019a; P Zhou et al., 2019). The k-nearest neighbors (KNN) is a type of 'lazy learning' algorithm, which means it only stores a training dataset instead of undergoing a training stage. When a new sample is entered, the k (an integer value specified by the user) closest neighbors is selected to represent it based on the distances (e.g., Euclidean distance) between the new sample and its neighbors. KNN is known for its simplicity, and it has also been proven to be an effective method for influent characteristics prediction (Kim et al., 2016; Taunk et al., 2019b). Artificial neural networks are widely recognized as a powerful tool for predictive modeling across various disciplines (Agirre-Basurko et al., 2006; Zhang et al., 2019). Multi-layer perceptron (MLP) is a typical artificial neural network model, often used as a baseline model due to its ability to approximate any measurable function and its ease of construction and tuning. MLP is composed of interconnected layers of neurons that communicate through weights. Typically, it consists of an

input layer, one or more hidden layers, and an output layer. It is widely used in wastewater modeling and has been proved to be effective.

5.2.2 Online learning models

While the data streams of influent flow rate bring a dynamic pattern, methods that can learning from evolving data streams and the input-output relationship is desired. The Adaptive Random Forests (aRF) algorithm is an adaption of the traditional Random Forests (RF) algorithm, designed to handle data streams (Gomes et al., 2018, 2017). It combines the traits of batch algorithms with dynamic update methods to efficiently process data streams. The development of an adaptive Random Forest (aRF) consists of two major processes: (1) a base-tree creation algorithm and (2) a dynamic update method. The classical RF by Breiman creates a base-tree through algorithms such as CART (also called CART tree), which assume that all training examples can be stored simultaneously and limit the number of examples each base-tree can learn from. The base-tree of aRF is a Hoeffding tree introduced in (Domingos and Hulten, 2000), which overcomes the limitation of CART tree and can handle extremely large datasets. For the dynamic update method of aRF, it mainly relies on a drift detection method such as ADWIN that could detect warnings and drifts (Bifet and Gavalda, 2007). In aRF, when a warning is detected by ADWIN, “background” trees are initialized and trained alongside the ensemble without affecting its predictions. If a drift is detected confirmed for the tree that triggered the warning, it is then replaced with its corresponding background tree.

The adaptive k-nearest neighbors (aKNN) method is based on the classical KNN method. KNN has no training stage as it only stores a training dataset. However, using the traditional KNN method becomes infeasible when learning from data streams, as storing the data prior to learning is neither useful (old data may not represent the current concept) nor practical (data may surpass

available memory). aKNN includes a slide window to store last observed data instead of storing all data. The maximum size of the window storing the last observed samples is also an integer value specified by the user. The adaptive Multi-Layer Perceptron (aMLP) adopts a partial fit function, in contrast to the traditional Multi-Layer Perceptron (MLP). This means that when new samples are introduced, the aMLP model only partially fit to incorporate the new data, rather than retraining the entire model from scratch. These models were implemented using scikit-learn and scikit-multiflow python libraries (Montiel et al., 2018; Pedregosa et al., 2011).

5.2.3 Model comparison

Traditional batch learning models and online learning models were compared in this study. Fig. 5.1 shows the design of the experiment. For both batch learning and online learning models, the whole dataset was divided into the same training and testing datasets. In batching learning, the models were trained using the entire training dataset and the best performing models were selected through cross-validation. After selecting the best models, the testing set was used to evaluate the models' performance on unseen data. This was done by feeding the testing dataset into the models and comparing the predicted outcomes with the actual observations. In online learning, the adaptive models, the same as batch learning, were initially trained using the training dataset and the best performing models were acquired. These models were then used to make predictions on the next data sample t_{i+1}. When a new observation (e.g., t_{i+1}) became available, it was fed into the best performing models, which were then updated accordingly. The updated models were then used to make predictions on the next data sample t_{i+2}, and this model update process was repeated until predictions were made for all the data samples in the testing dataset.

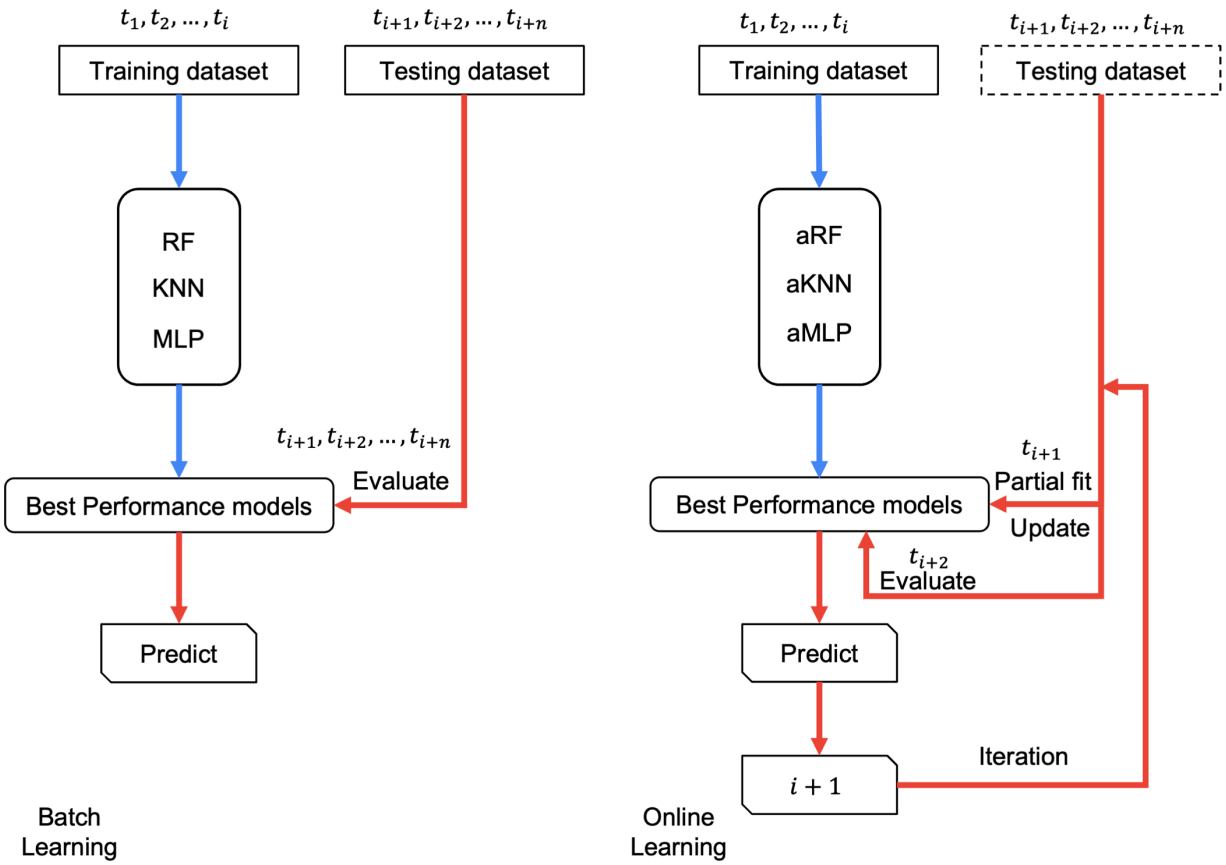

Fig. 5.1 Schema of the experiments (left: batch learning models; right: online learning models)

5.3 Study area and data

In this study, two wastewater treatment plants in Ontario, Canada (Plants A and B) were selected for case studies. The flow rate of the influent at Plant A was monitored on an hourly basis from November 1, 2016 to August 3, 2021, while the flow rate of Plant B was measured every 15 minutes from January 1, 2019 to November 30, 2021. Limited missing data points were filled using simple linear interpolation by estimating the values based on the surrounding known data. Both Plants A and B collect wastewater from homes and small businesses. Plant A has separate storm and sanitary sewers while Plant B has a small portion of its sewers that combine both types of water. Both plants experience increased inflow during rainfall events (for Plant A, this is mainly due to illegal connections of downspouts and sump pumps to the sanitary system). Therefore,

hourly meteorological data including precipitation, snow depth, temperature, humidity, pressure, wind direction and wind speed were collected and used as features. Additionally, timestamp information including hour of the day, day of the week, day of the month, and month of the year were also used as features in this study. For Plant A, the data before March 1st, 2020 were used as training dataset, while data from that date forward were testing dataset. For Plant B, the data before January 1st, 2020 was selected as training dataset, while data from that forward were testing dataset. The testing datasets for both plants cover at least one period of COVID-19 lockdown in Ontario, during which schools, non-essential services, and recreational facilities were closed.

5.4 Results and discussion

5.4.1 Overall model performance

Conventional batch learning RF, KNN, and MLP models as well as online learning aRF, aKNN, and aMLP models were created for influent flow rate prediction at two plants. Two modeling scenarios including 24-hour ahead prediction and no lead time prediction were considered. These models were tuned to achieve optimal performance while avoiding overfitting. Table 5.1 displays the performance metrics for each model on the testing dataset.

Table 5.1 Performance metrics for each model, by plant and scenario

	Plant A						Plant B					
	24-hour ahead			No lead time			24-hour ahead			No lead time		
	R^2	MAPE (%)	RMSE (m3/hr)	R^2	MAPE (%)	RMSE (m3/hr)	R^2	MAPE (%)	RMSE (MLD)	R^2	MAPE (%)	RMSE (MLD)
RF	0.79	7.59	5663.73	0.78	7.79	5764.24	0.17	32.35	522.34	0.17	30.32	498.00
KNN	0.51	11.84	8901.19	0.53	11.62	8721.91	0.07	27.78	383.11	0.09	26.45	372.18
MLP	0.75	9.68	6714.23	0.77	8.95	6296.41	0.24	29.84	483.29	0.25	29.96	466.62
aRF	**0.90**	7.35	4895.25	**0.90**	7.40	4905.84	0.75	14.82	206.32	**0.77**	14.59	201.58

aKNN	0.73	8.67	6342.32	0.73	8.68	6348.84	**0.78**	**10.99**	**181.02**	0.77	**11.48**	**184.46**
aMLP	0.87	**5.56**	**4424.91**	0.88	**5.17**	**4193.20**	0.56	22.83	252.12	0.68	17.96	221.03

The predictions for the no lead time scenario were overall slightly better than those for the 24-hour ahead scenario. The average R^2, MAPE, and RMSE metrics for Plant A in the scenario with no leading time are 0.765, 8.27%, and 6038.41 m^3/hr respectively, while for Plant B they are 0.455, 21.79%, and 323.98 million liter per day (MLD) respectively. The average R^2, MAPE, and RMSE metrics for Plant A in the scenario with 24-hour ahead are 0.758, 8.45%, and 6156.94 m^3/hr respectively, while for Plant B they are 0.428, 23.11%, 338.03 MLD respectively. For both Plant A and Plant B in the scenario with no leading time, a slightly higher R^2, lower MAPE, and lower RMSE were achieved in comparison with 24-hour ahead scenario. It is common to observe a prediction accuracy drop when the prediction horizon expanded, this is likely due to the fact that as the temporal distance between the features and targets increases, the correlation between them becomes weaker (Wei and Kusiak, 2014; P Zhou et al., 2019). Given that time leading predictions are more applicable and practical for wastewater management, the following analysis focus on predictions made with a 24-hour ahead.

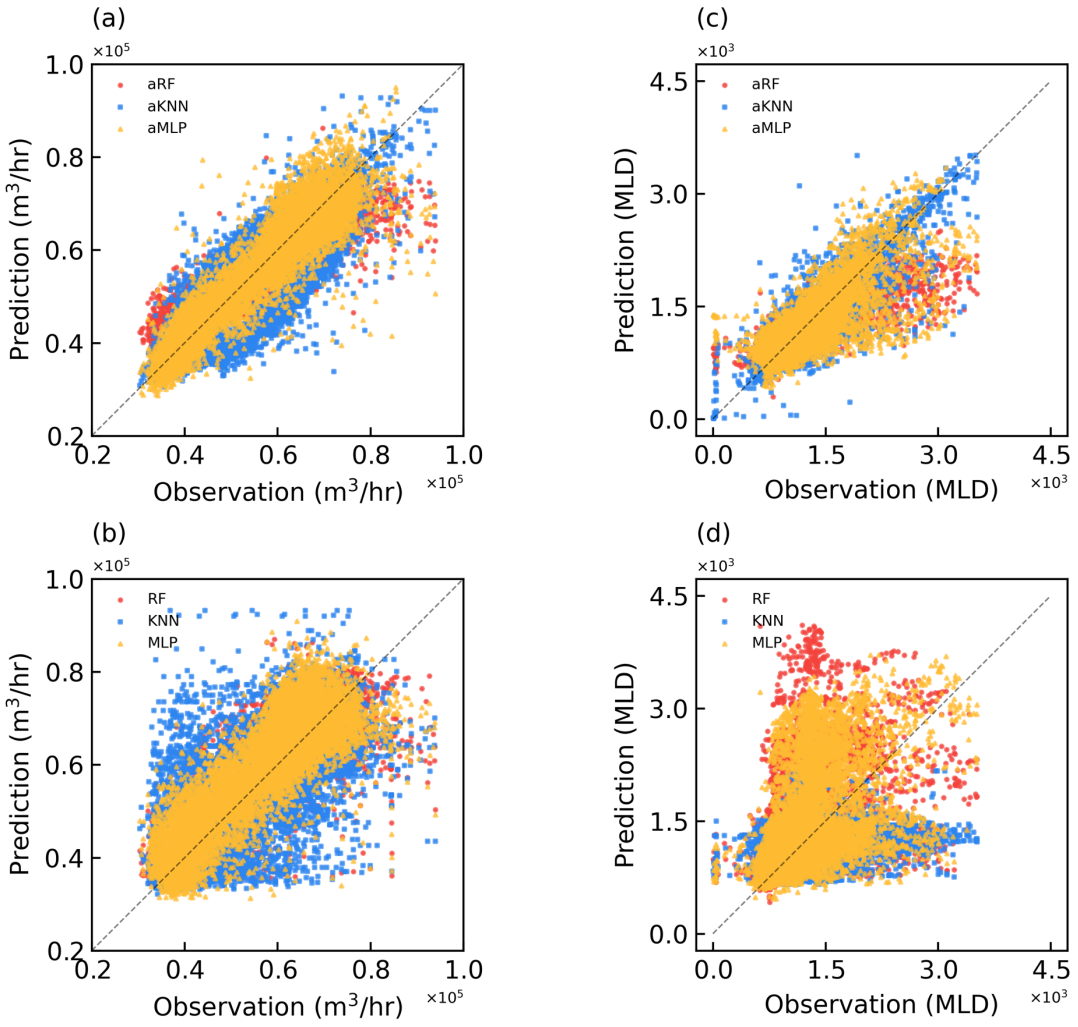

Fig. 5.2 Scatter plots for 24-hour ahead predictions from each model on testing dataset: (a) online learning models for Plant A, (b) batch learning models for Plant A, (c) online learning models for Plant B, and (d) batch learning models for Plant B

5.4.2 Online learning compared to batch learning

Online learning models that achieved the highest R^2, the lowest MAPE, and the lowest RMSE according to Table 5.1, performed superiorly compared to conventional batch learning models for both scenarios across both plants. The scatterplot in Fig. 5.2a illustrates the results of 24-hour ahead online learning predictions for Plant A, with points distributed around the diagonal.

In contrast, the scatterplot of batch learning methods for Plant A in Fig. 5.2b displays a less compact distribution of points. It suggests that the online learning methods had a smaller overall prediction error for Plant A. This phenomenon is even more significant for Plant B. The scatterplot in Fig. 5.2c illustrates the results of 24-hour ahead online learning predictions for Plant B. The points are distributed around and slightly towards the downside of the diagonal, indicating a slight underestimation. While the scatterplot of batch learning methods for Plant B in Fig. 5.2d shows a much more dispersed distribution. This may imply that the changes of influent flow rate at Plant B during the COVID-19 pandemic were more significant, and the mapping functions captured by conventional batching learning models may not be valid and need to be updated. It also emphasizes the importance of adopting new online learning methods to effectively capture and respond to the dynamic changes in influent flow rate during the COVID-19 pandemic.

Fig. 5.3 presents a further performance comparison of online learning models and batch learning models, where dark colors represent the online learning models, and light colors represent the batch learning models. Histograms in Figs. 5.3a and 5.3c were used to show the distribution of prediction errors of the online learning and batch learning models at Plant A and Plant B, respectively. The prediction errors were calculated by subtracting the actual values from the predicted values. The y-axis represents the error range, and the z-axis represents the density of the errors. It is noticed that the histogram of the aMLP model at Plant A, as well as the histograms of the aRF and aKNN models for Plant B, have apparently taller bars near the center close to 0. This suggests that the models have fewer errors. To quantitatively compare the prediction errors, boxplots shown in Figs. 5.3b and 5.3d were utilized. The notch in the box represents the median, and the lower and upper of the box are the first quartile ($Q1$) and third quartile ($Q3$), respectively. The lower whisker extends to the first datum greater than $Q1 - 1.5 \cdot (Q3 - Q1)$, while the upper

whisker extends to the last datum less than $Q3 + 1.5 \cdot (Q3 - Q1)$. All presented boxplots adhere to the same guidelines. The aMLP model for Plant A, as well as the aRF and aKNN models for Plant B dramatically reduced the prediction errors. Although the reductions were not dramatical, the aKNN model for Plant A and aMLP model for Plant B also showed lower prediction errors compared to their respective batching learning models. The aRF model for Plant A displayed similar prediction errors as the RF model, which could be attributed to the RF model already achieving reasonable performance and the limitations in improvement potential for tree-structured algorithms. Notably, at Plant B, online learning methods had significantly decreased prediction errors. This may indicate that the influent flow rate pattern at Plant B during the pandemic has undergone a more significant change. Overall, in response to the dynamic changes in influent flow rate during the COVID-19 pandemic, online learning methods showed improved performance when compared to batch learning methods.

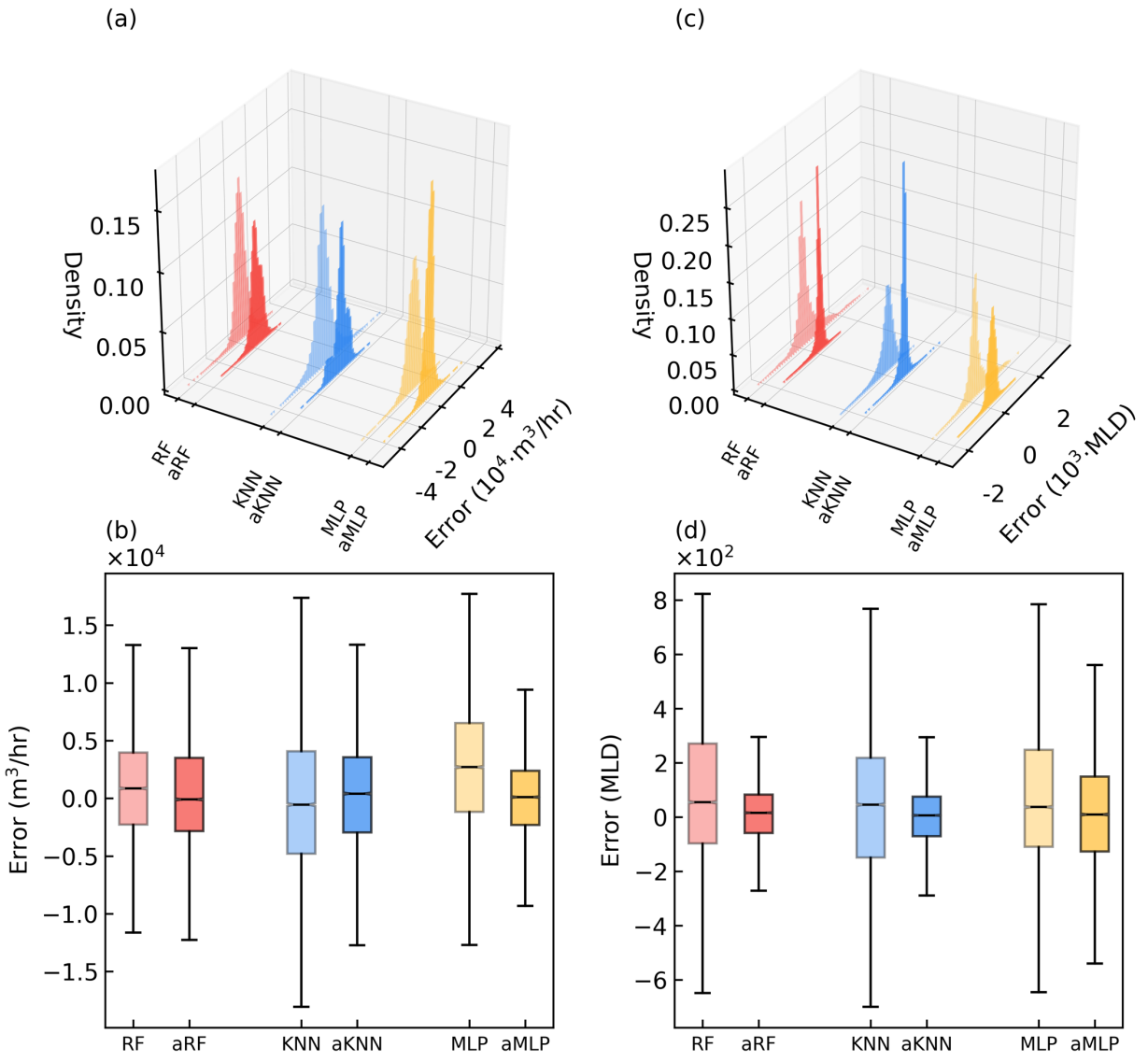

Fig. 5.3 Performance comparison of online learning methods and batch learning methods: (a) histograms of prediction errors at Plant A, (b) boxplots of prediction errors at Plant A, (c) histograms of prediction errors at Plant B, and (d) boxplots of prediction errors at Plant B

5.4.3 Comparison of online learning methods

While online learning methods demonstrated improved performance compared to batch learning methods, further examination is necessary to discern the differences among the three

online learning methods. A comparison of the three online learning methods, along with observations, is presented in Fig. 5.4. All three methods were able to capture the observed flow rate distribution (in grey) with reasonable variations at both plants. It appears that the aKNN models were most successful in capturing the distributions with the least variations at both plants. This is likely due to the fact that aKNN models only store data from the previous four weeks, making the prediction range more stable when real-world variations within the four weeks are limited. Interestingly, the influent flow rate at Plant A exhibited a bi-modal distribution, while that at Plant B exhibited a Poisson distribution. The possible reason for the bi-modal distribution at Plant A is that the data could be classified into two typical seasons, one with a noticeably higher influent flow rate than the other. For aRF predictions, the R^2 and MAPE metrics (Table 5.1) range from 0.75 to 0.90 and 7.35% to 14.82%, respectively. Comparatively, for aKNN predictions, the R^2 and MAPE metrics range from 0.73 to 0.78 and 8.67% to 11.48%, respectively. For aMLP prediction, the R^2 and MAPE metrics range from 0.56 to 0.88 and 5.17% to 22.83%, respectively. Referring to the metrics presented in Table 5.1, the aMLP algorithm is considered as the most suitable among the three algorithms for Plant A, and the aKNN algorithm is deemed the most appropriate for Plant B. Additionally, aRF is considered as the most stable as it achieves overall highest R^2 in all scenarios.

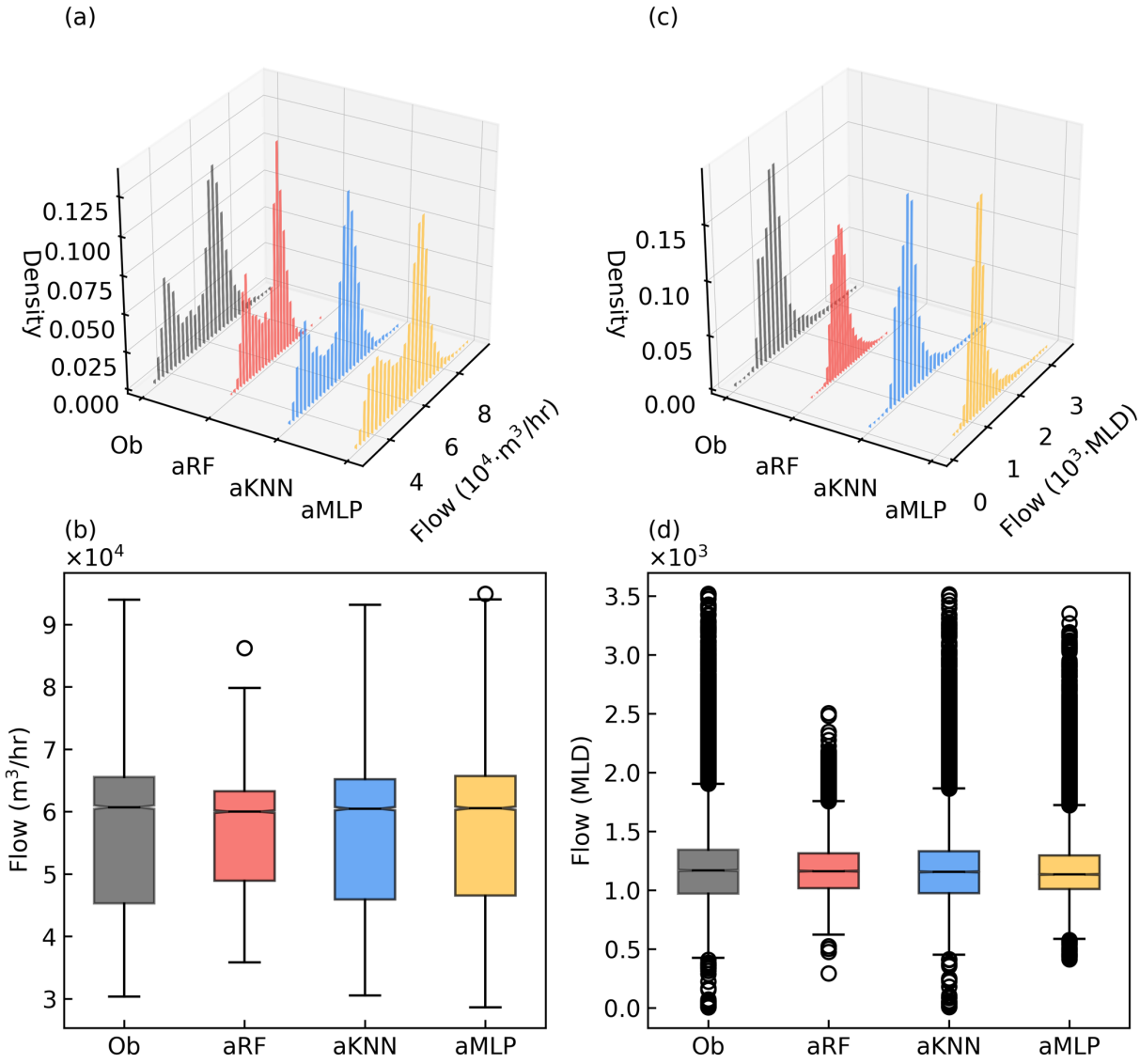

Fig. 5.4 Performance comparison of online learning methods: (a) histograms of predictions at Plant A, (b) boxplots of predictions at Plant A, (c) histograms of predictions at Plant B, and (d) boxplots of predictions at Plant B

5.4.4 Ensemble of online learning methods

The use of an ensemble of different online learning methods, such as taking the average of predictions, can leverage the strengths of each method. Fig. 5.5 shows the scatterplots of averaged online learning predictions versus observations at the two plants. For Plant A, all the scatters were densely distributed around the diagnose and no apparent offset was observed. When compared to the single-algorithm model, the averaged online learning predictions achieved a tied highest R^2 (0.90, same with aRF), the second lowest MAPE (5.65%, slightly higher than aMLP), and the lowest RMSE (4023.58 m^3/hr). For Plant B, while most scatters were distributed around the diagnose, there was an overestimation of observations near 0 and slight underestimation of extreme high observations. Practically, attention should be paid to these underestimated high observations as they may lead to more severe issues such as biomass washout. The averaged online learning predictions for plant B achieved the highest R^2 (0.79), the second lowest MAPE (14.25%, higher than aKNN), and the lowest RMSE (179.31 MLD) when compared to the single-algorithm models. Overall, the averaged online learning predictions performed superior compared to each single-algorithm model.

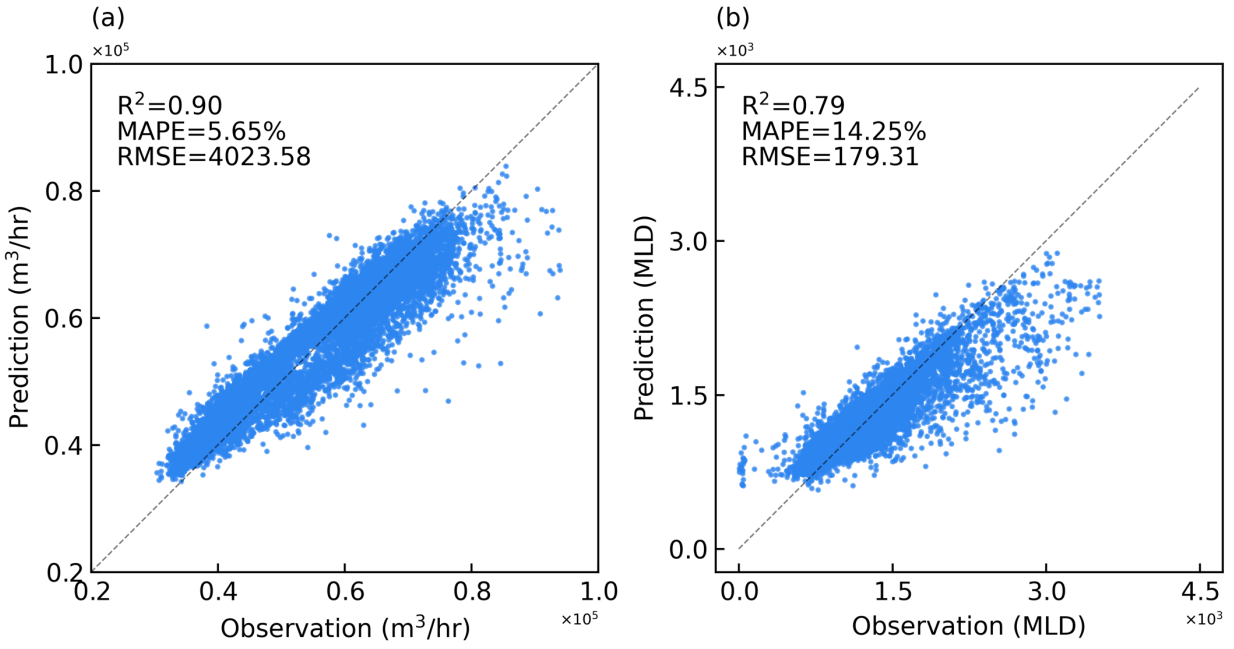

Fig. 5.5 Scatterplots of averaged online learning predictions versus observations at: (a) Plant A, and (b) Plant B

An interval prediction that incorporates the predictions from all three online learning models can be generated to provide more information about the ensemble online learning predictions. The upper bound of the interval is the maximum prediction from the three models, while the lower bound is the minimum prediction. As shown in Fig. 6, the cumulative density functions (CDFs) of the maximum, average and minimum predictions as well as the observations were compared. The CDF of a random variable X, evaluated at x, is the probability that X will take a value less than or equal to x. When comparing the minimum, maximum, and observed CDFs, it is noticeable that the overall shape of them is similar at both plants and the observed CDF is in-between. The average CDF and the observation CDF show overlap in the middle at both plants, while the greatest deviation occurs at the segments: for low flow rates, the observation CDF is higher than the average CDF; for high flow rates, the observation CDF is lower than the average

CDF. This suggests that the observation CDF has a greater likelihood of extremely small and extremely large values for the influent flow rate, compared to the average CDF. To improve the prediction of influent flow rate, it is suggested to use the average CDF for moderate flow rates, the minimum CDF for low flow rates, and the maximum CDF for high flow rates.

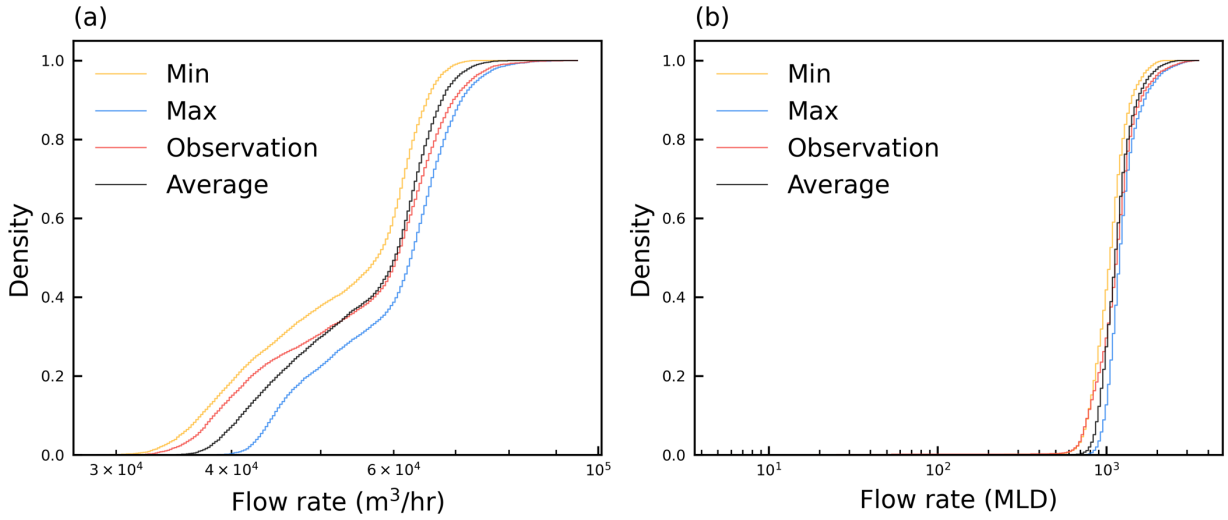

Fig. 5.6 Comparison of cumulative density functions at: (a) Plant A, (b) Plant B

5.5 Conclusion

In this study, a series of online learning models (aRF, aKNN, and aMLP) were developed for predicting the changing influent flow rate under the impact of COVID-19. Online learning models that can adapt to changing influent flow rate patterns show distinct advantages in comparison with traditional batch learning models. These models were developed based on 3-4 years of hourly influent flow rate data and meteorological data, collected from two Canadian wastewater treatment plants. The developed online learning models were compared to their respective conventional batch learning models (RF, KNN, and MLP) for influent flow rate prediction at two Canadian wastewater treatment plants. Two scenarios were considered, including

24-hour ahead prediction and no lead time prediction. The online learning models produced predictions with good accuracy. Additionally, an optimal prediction strategy for influent flow rate was found through the uncertainty analysis of each model. The online learning models were found superior to batch learning models. They can not only easily adapt to a dynamic input-output relationship but are also friendly to extremely large data streams. The proposed models can provide more robust decision support to wastewater operators or managers for coping with changing influent patterns due to emergencies such as COVID-19.

Online learning models that achieved the highest R^2, the lowest MAPE, and the lowest RMSE performed better compared to conventional batch learning models for both scenarios across both plants. The R^2 values on testing dataset for 24-hour ahead prediction of the aRF, aKNN, and aMLP at Plant A were 0.90, 0.73, and 0.87, respectively; these values at Plant B were 0.75, 0.78, and 0.56, respectively. The averaged online learning predictions performed superior compared to each single-algorithm model. The averaged online learning predictions achieved a tied highest R^2 (0.90, the same with aRF), the second lowest MAPE (5.65%, slightly higher than aMLP), and the lowest RMSE (4023.58 m^3/hr) at Plant A and the highest R^2 (0.79), the second lowest MAPE (14.25%, higher than aKNN), and the lowest RMSE (179.31 MLD) at Plant B. To improve the prediction of influent flow rate, cumulative density functions (CDF) corresponding upper bound, lower bound, and average of the ensemble online learning predictions were generated. It is suggested to use the average CDF for moderate flow rate predictions, the minimum CDF for low flow rates, and the maximum CDF for high flow rates. In future studies, these proposed online learning models can be utilized for predicting not only influent flow rate but also other wastewater characteristic at other WWTPs. The online learning models presented provide promising results;

however, this study is limited to two case studies. Future studies should include more case studies and consider more prediction scenarios to further validate the developed models.

References

Abu-Bakar, H., Williams, L., Hallett, S.H., 2021. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. npj Clean Water 2021 4:1 4, 1–9. https://doi.org/10.1038/s41545-021-00103-8

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. Environmental Modelling & Software 21, 430–446.

Ahmed, N.K., Atiya, A.F., Gayar, N. el, El-Shishiny, H., 2010. An empirical comparison of machine learning models for time series forecasting. Econom Rev 29, 594–621.

Alfano, V., Ercolano, S., 2020. The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel Analysis. Applied Health Economics and Health Policy 2020 18:4 18, 509–517. https://doi.org/10.1007/S40258-020-00596-3

Andreides, M., Dolejš, P., Bartáček, J., 2022. The prediction of WWTP influent characteristics: Good practices and challenges. Journal of Water Process Engineering 49, 103009. https://doi.org/10.1016/J.JWPE.2022.103009

Ansari, M., Othman, F., Abunama, T., El-Shafie, A., 2018. Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant, Malaysia. Environmental Science and Pollution Research 25, 12139–12149. https://doi.org/10.1007/S11356-018-1438-Z

Bechmann, H., Nielsen, M.K., Madsen, H., Kjølstad Poulsen, N., 1999. Grey-box modelling of pollutant loads from a sewer system. Urban Water 1, 71–78. https://doi.org/10.1016/S1462-0758(99)00007-2

Bifet, A., Gavalda, R., 2007. Learning from time-changing data with adaptive windowing, in: Proceedings of the 2007 SIAM International Conference on Data Mining. SIAM, pp. 443–448.

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q., Zhou, P., 2019. Influent forecasting for wastewater treatment plants in North America. Sustainability (Switzerland) 11. https://doi.org/10.3390/su11061764

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32.

Bzdok, D., Krzywinski, M., Altman, N., 2018. Machine learning: supervised methods. Nat Methods 15, 5. https://doi.org/10.1038/NMETH.4551

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. ACM International Conference Proceeding Series 148, 161–168. https://doi.org/10.1145/1143844.1143865

Domingos, P., Hulten, G., 2000. Mining high-speed data streams, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 71–80.

Fontenla-Romero, Ó., Guijarro-Berdiñas, B., Martinez-Rego, D., Pérez-Sánchez, B., Peteiro-Barral, D., 2013. Online machine learning, in: Efficiency and Scalability Methods for Computational Intellect. IGI global, pp. 27–54.

Gautam, S., Hens, L., 2020. COVID-19: impact by and on the environment, health and economy. Environment, Development and Sustainability 2020 22:6 22, 4953–4954. https://doi.org/10.1007/S10668-020-00818-7

Gomes, H.M., Barddal, J.P., Ferreira, L.E.B., Bifet, A., 2018. Adaptive random forests for data stream regression., in: ESANN.

Gomes, H.M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G., Abdessalem, T., 2017. Adaptive random forests for evolving data stream classification. Mach Learn 106, 1469–1495.

Hillary, L.S., Farkas, K., Maher, K.H., Lucaci, A., Thorpe, J., Distaso, M.A., Gaze, W.H., Paterson, S., Burke, T., Connor, T.R., McDonald, J.E., Malham, S.K., Jones, D.L., 2021. Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. Water Res 200, 117214. https://doi.org/10.1016/J.WATRES.2021.117214

Hoi, S.C.H., Sahoo, D., Lu, J., Zhao, P., 2021. Online learning: A comprehensive survey. Neurocomputing 459, 249–289.

Hoi, S.C.H., Wang, J., Zhao, P., 2014. Libol: A library for online learning algorithms. Journal of Machine Learning Research 15, 495.

Jain, L.C., Seera, M., Lim, C.P., Balasubramaniam, P., 2014. A review of online learning in supervised neural networks. Neural Comput Appl 25, 491–509.

Khan, I., Shah, D., Shah, S.S., 2020. COVID-19 pandemic and its positive impacts on environment: an updated review. International Journal of Environmental Science and Technology 2020 18:2 18, 521–530. https://doi.org/10.1007/S13762-020-03021-3

Kim, M., Kim, Y., Kim, H., Piao, W., Science, C.K.-F. of E., 2016, undefined, 2016. Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. Springer 10, 299–310. https://doi.org/10.1007/s11783-015-0825-7

Kovacs, D.J., Li, Z., Baetz, B.W., Hong, Y., Donnaz, S., Zhao, X., Zhou, P., Ding, H., Dong, Q., 2022. Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study. J Memb Sci 660, 120817.

Ma, S., Zeng, S., Dong, X., Chen, J., Environmental, G.O.-F. of, 2014, undefined, 2014. Short-term prediction of influent flow rate and ammonia concentration in municipal wastewater treatment plants. Springer 8, 128–136. https://doi.org/10.1007/s11783-013-0598-9

Montiel, J., Read, J., Bifet, A., Abdessalem, T., 2018. Scikit-multiflow: A multi-output streaming framework. The Journal of Machine Learning Research 19, 2914–2915.

Nemati, M., Tran, D., 2022. The Impact of COVID-19 on Urban Water Consumption in the United States. Water 2022, Vol. 14, Page 3096 14, 3096. https://doi.org/10.3390/W14193096

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825–2830.

Taunk, K., De, S., Verma, S., Swetapadma, A., 2019a. A brief review of nearest neighbor algorithm for learning and classification, in: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, pp. 1255–1260.

Taunk, K., De, S., Verma, S., Swetapadma, A., 2019b. A brief review of nearest neighbor algorithm for learning and classification. 2019 International Conference on Intelligent

Computing and Control Systems, ICCS 2019 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747

Wei, X., Kusiak, A., 2014. Short-term prediction of influent flow in wastewater treatment plant. Stochastic Environmental Research and Risk Assessment 2014 29:1 29, 241–249. https://doi.org/10.1007/S00477-014-0889-0

Wei, X., Kusiak, A., Sadat, H.R., 2012. Prediction of Influent Flow Rate: Data-Mining Approach. Journal of Energy Engineering 139, 118–123. https://doi.org/10.1061/(ASCE)EY.1943-7897.0000103

Zhang, Q., Li, Z., Snowling, S., Siam, A., El-Dakhakhni, W., 2019. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. Water Science and Technology 80, 243–253. https://doi.org/10.2166/WST.2019.263

Zhou, P., Li, Z., Snowling, S., Baetz, B.W., Na, D., Boyd, G., 2019. A random forest model for inflow prediction at wastewater treatment plants. Stochastic Environmental Research and Risk Assessment 2019 33:10 33, 1781–1792. https://doi.org/10.1007/S00477-019-01732-9

Zhou, P., Li, Z., Snowling, S., Goel, R., Zhang, Q., 2022. Multi-step ahead prediction of hourly influent characteristics for wastewater treatment plants: a case study from North America. Environmental Monitoring and Assessment 2022 194:5 194, 1–14. https://doi.org/10.1007/S10661-022-09957-Y

Zhou, P, Li, Z., Snowling, S., Goel, R., Zhang, Q., Solutions, I., 2019. Short-term wastewater influent prediction based on random forests and multi-layer perceptron. Journal of environmental informatics letters 1, 87–93.

Zhu, J.-J., Anderson, P.R., 2019. Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. Water Science and Technology 80, 695–706. https://doi.org/10.2166/WST.2019.309

Chapter 6 – Conclusions

6.1 Conclusions and contributions

The aim of this dissertation is to develop advanced data-driven approaches for wastewater modeling under uncertainties and thus support the operations and management of wastewater treatment plants (WWTPs). This was achieved through the development of data-driven approaches to assist conventional wastewater process-driven models (PDMs) as well as expanding and elevating the application of data-driven models (DDMs) to tackle more challenging wastewater simulation tasks.

In Chapter 2, an efficient arbitrary polynomial chaos expansion (aPCE) approach based on data-driven techniques was developed for the uncertainty analysis of a conventional process-driven Secondary Settling Tank (SST) models. The use of SST models is crucial for the effective simulation of wastewater treatment systems. By predicting the quality of effluent and underflow, SST models aid in the optimization, management, and design of wastewater treatment systems. SST modeling relies on an empirical settling velocity function, which can be significantly affected by parameter uncertainty. This uncertainty may negatively impact the performance of the SST model, making it important to assess parameter uncertainty. While Monte Carlo simulation (MCS) is a traditional method for assessing uncertainty, it can be computationally expensive and requires explicit knowledge of parameter distribution. To overcome these limitations, a novel approach based on data-driven techniques called arbitrary polynomial chaos expansion has been developed and used for the first time. The well-known Bürger-Diehl SST model is utilized as a subject, and uncertainties originating from five key model parameters are evaluated using both the aPCE and MCS techniques. Both techniques generate probabilistic estimations of the model output sludge blanket height. Comparing the results of aPCE and MCS, it appears that the aPCE approach is as

effective as MCS in quantifying SST model parameter uncertainty while significantly reducing computational cost. This study validates the effectiveness and efficiency of aPCE in quantifying uncertainties of wastewater PDMs and demonstrates that aPCE can be an effective alternative to MCS for uncertainty quantification in the field of wastewater modeling. Additionally, this study shows that utilizing data-driven approaches to assist conventional wastewater PDMs is feasible and can provide more robust support for the design, management, and optimization of wastewater treatment systems.

In Chapter 3, DDMs were developed for tackling the challenging Bisphenol A (BPA) prediction task. BPA is a contaminant of emerging concern that poses a risk to human health and is commonly found in the aquatic environment, with conventional WWTPs being a significant pathway of BPA. Accurately predicting BPA's fate at WWTPs is crucial to controlling and mitigating BPA contamination. Three machine learning models, namely shared layer multi-task neural network, genetic programming, and extra trees, are employed in this study to predict the effluent BPA concentration at twelve municipal WWTPs across Canada. Additionally, network theory is applied to examine the interdependencies among the variables influencing BPA removal. This study validates the abilities of advanced DDMs to accurately predict BPA effluent concentration, with advantages such as alleviating data sparsity and imbalance, improving model interpretability, and measuring predictor importance. The network analysis is shown to be effective in revealing interdependencies among various factors affecting BPA removal. The study demonstrates that BPA removal is unlikely to occur at primary treatment plants, while it can be achieved through secondary or tertiary treatment. More importantly, this study provides an integrated framework for modeling and analyzing emerging contaminants at WWTPs.

In Chapter 4, DDMs were developed to help assess the impact of the COVID-19 pandemic and the subsequent lockdowns on Canadian municipal sewage systems. The focus is on the changes in influent flow rates at two wastewater treatment plants in Ontario, Canada. Weekly patterns and daily average flow rates before and during the lockdowns are compared. Predicted flow rates for a no-lockdown scenario are generated by random forest models and compared with the observed influent flow rates to exclude the meteorological impact. The study shows that influent flow rates exhibited differences in weekly patterns and less variability during the lockdowns compared to pre-lockdowns. Both plants experienced a decrease in influent flow rates during the lockdowns, with a surge after the easing of provincial emergency state. This information is valuable for improving wastewater management strategies and guiding policy decisions during times of crisis in the future.

In Chapter 5, traditional DDMs were modified for capturing the constantly changing influent flow rate patterns during the COVID-19 pandemic. Data-driven models have been proven effective in previous studies for predicting influent flow rates, but most of these studies focused on batch learning, which is insufficient for predicting wastewater patterns during the COVID-19 era because of changing patterns. Online learning has the potential to address this issue due to its distinct advantages of handling stream data, large datasets, and changing data patterns. This study compares the performance of conventional batch learning models (Random Forest, K-Nearest Neighbors, and Multi-Layer Perceptron) with their respective online learning models (Adaptive Random Forest, Adaptive K-Nearest Neighbors, and Adaptive Multi-Layer Perceptron) for predicting influent flow rate at two Canadian WWTPs. The online learning models outperformed the conventional batch learning models in all scenarios. This study proves online learning models are effective in making reliable influent flow rate predictions under changing data patterns and are

efficient in handling continuous and large influent data streams. Additionally, this study proposes new methods for adapting to changing influent data and supports to wastewater treatment management during unprecedented emergencies such as COVID-19 that may alter input-output relationships.

Overall, this dissertation developed a data-driven uncertainty quantification technique to assist conventional process-driven models (PDMs) and filled the gap in the application of data-driven models (DDMs) for challenging wastewater modeling issues, such as predictions with limited data and predictions in emergency scenarios.

6.2 Future research recommendations

(1) While the uncertainty analysis for SST models has been well investigated in this thesis, large uncertainties could also arise from the simulation of other unit processes. For example, uncertainty analysis is crucial for all types of activated sludge models (ASMs), but it is often performed using MCS, which is computationally expensive. Therefore, applying advanced and efficient methods such as aPCE for uncertainty analysis of ASMs, as well as other unit process simulation models, can provide valuable benefits.

(2) Although using aPCE can efficiently achieve uncertainty analysis, stochastic modeling might be another approach to address uncertainties and is worth investigating. For example, stochastic wastewater models can be developed by incorporating intrusive polynomial chaos expansion into wastewater PBMs.

(3) Although new effective DDMs for wastewater modeling have been developed, the discussion on the uncertainty of the developed DDMs is limited. For instance, the hyperparameters of DDMs in wastewater modeling field are usually tuned through grid search or experience.

Assessing the uncertainties associated with hyperparameters can benefit the improvement of DDM performance.