ENHANCING URBAN CENTRE RESILIENCE UNDER CLIMATE-INDUCED DISASTERS USING DATA ANALYTICS AND MACHINE LEARNING TECHNIQUES

ENHANCING URBAN CENTRE RESILIENCE UNDER CLIMATE-INDUCED DISASTERS USING DATA ANALYTICS AND MACHINE LEARNING TECHNIQUES

By

May Mahmoud Haggag

BSc., MSc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

McMaster University © Copyright by M. Haggag, June 2021

Doctor of Philosophy (2021)	McMaster University
(Civil Engineering)	Hamilton, Ontario
TITLE:	Enhancing urban centre resilience under climate-induced disasters using data analytics and machine learning techniques
AUTHOR:	May Mahmoud Haggag BSc., MSc. (The American University in Cairo)
SUPERVISORS:	Dr. Wael El-Dakhakhni Dr. Hassini Elkafi
COMMITTEE MEMBERS:	Dr. Paulin Coulibaly Dr. Manish Verma
NUMBER OF PAGES:	xxi, 236

Abstract

According to the Centre for Research on the Epidemiology of Disasters, the global average number of CID has tripled in less than four decades (from approximately 1,300 Climate-Induced Disasters (CID) between 1975 and 1984 to around 3,900 between 2005 and 2014). In addition, around 1 million deaths and $1.7 trillion damage costs were attributed to CID since 2000, with around $210 billion incurred only in 2020. Consequently, the World Economic Forum identified extreme weather as the top ranked global risk in terms of likelihood and among the top five risks in terms of impact in the last 4 years. These risks are not expected to diminish as: *i*) the number of CID is anticipated to double during the next 13 years; *ii*) the annual fatalities due to CID are expected to increase by 250,000 deaths in the next decade; and *iii*) the annual CID damage costs are expected to increase by around 20% in 2040 compared to those realized in 2020. Given the anticipated increase in CID frequency, the intensification of CID impacts, the rapid growth in the world's population, and the fact that two thirds of such population will be officially living in urban areas by 2050, it has recently become extremely crucial to enhance both community and city resilience under CID. Resilience, in that context, refers to the ability of a system to bounce back, recover or adapt in the face of adverse events. This is considered a very farfetched goal given both the extreme unpredictability of the frequency and impacts of CID and the complex behavior of cities that stems from the interconnectivity of their comprising infrastructure systems. With the emergence of data-driven machine learning which assumes that models can be

trained using historical data and accordingly, can efficiently learn to predict different complex features, developing robust models that can predict the frequency and impacts of CID became more conceivable. Through employing data analytics and machine learning techniques, this work aims at enhancing city resilience by predicting both the occurrence and expected impacts of climate-induced disasters on urban areas. The first part of this dissertation presents a critical review of the research work pertaining to resilience of critical infrastructure systems. Meta-research is employed through topic modelling, to *quantitatively* uncover related latent topics in the field. The second part aims at predicting the occurrence of CID by developing a framework that links different climate change indices to historical disaster records. In the third part of this work, a framework is developed for predicting the performance of critical infrastructure systems under CID. Finally, the aim of the fourth part of this dissertation is to develop a systematic data-driven framework for the prediction of CID property damages. This work is expected to aid stakeholders in developing spatio-temporal preparedness plans under CID, which can facilitate mitigating the adverse impacts of CID on infrastructure systems and improve their resilience.

Dedications

To Mahmoud & Azza,

Ahmed & Samar,

Lina,

Little Dodo

Acknowledgments

I would like to express my deep gratitude for the continuous help, support, and motivation of my supervisors, Dr. Wael El-Dakhakhni and Dr. Hassini Elkafi. I really consider myself fortunate to be supervised by such supportive advisors who always provided me with great advice on both the technical and personal levels. Their unceasing support provided me with the opportunity to develop my technical, research, writing, and presentation skills. Throughout my Ph.D. journey, they became not only my supervisors but my role models whom I truly respect and look up to. Special thanks are due to my supervisory committee members, Dr. Paulin Coulibaly and Dr. Manish Verma for their valuable advice and suggestions. I truly appreciate their constructive feedback and helpful discussions during our meetings.

I would also like to express my sincere appreciation for Dr. Mohamed Ezzeldin, Dr. Ahmed Siam, and Dr. Ahmed Yosri for their continuous support, constructive feedback, and valuable advice. Without their support this work would have never been completed. I am also grateful for the support of Sarah Sullivan who provided me with great support during my first couple of years. I am also grateful for the support of my peers and colleagues: Mostafa Naiem, Ahmed Gaith, Ahmed Gondia, Mohamed Salama, Yassin Salaheldin, and Brinda Narayanan. I am especially thankful for my colleagues and sisters: Rasha El Sawy, Mouna Reda, Eman Risk, and Yara Maged who always supported and encouraged me.

I am truly grateful for the endless support of my sisters: Didi Hameed, Yasmine Magdy, and Amira Mostafa who made Canada feel like home, were patience enough to listen to my anxious thoughts over and over again and were always there to celebrate my little accomplishments. Special thanks to Omar and Lina who always filled me with warmth and hope. Special thanks are also due to Heba Teama, Reem Zeiton, Nehal Shaker, Fatima Alamara, Amany Elanany,

Nervana Farok, Nourhan Hassan, Shahy Elkady, Merhan Radwan, and Nayera Waheed. I am extremely grateful for the unceasing support of my best friends in Egypt who never failed to make me feel included even when we are actually thousands of miles apart: Lina Habib, Amira Shalaby, Hanan Nazir, Menna Assal, Reem Gamal, Sarah Halawa, Ola Mohamed, Mohamed Gazouli, Ahmed Assal, Mary Soliman, Nermine Edward, and Cathy Selim.

I am thankful for the research funding support provided through the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Ontario Trillium Scholarship Program. I would also like to acknowledge the fruitful discussions with the research teams of the NSERC-CaNRisk-CREATE program, the INTERFACE.

At last, but by no means the least, I would like to admit that no words are enough to describe my genuine gratitude for my parents, Mahmoud and Azza, without them by my side I would have never been the person I am today. This work would have never seen the light without their prayers, patience, encouragement, support, and motivation. I would also like to express my sincere gratitude for my brother, Ahmed, who is the kindest and most caring big brother anyone can hope for. I am especially thankful for my sister-in-law, Samar for everything she did, and are still doing for me. I am truly thankful for my nephew, little Mahmoud “Dodo” whose unconditional and unwavering love gave me hope during my long sleepless nights.

Co-Authorship

This dissertation was prepared in accordance with the guidelines set by the school of graduate studies at McMaster University for sandwich theses that contain a compilation of research papers published or prepared for publishing as journal articles. Chapters 2, 3, and 4 are already published as journal articles whereas the research paper presented in Chapter 5 is submitted for review. This dissertation presents the work carried out solely by May Haggag. Advice and guidance were provided for the whole thesis by the academic supervisors Dr. Wael El-Dakhakhni and Dr. Hassini Elkafi. Additional advice, and guidance for the papers presented in Chapters 2 and 3 were provided by Dr. Mohamed Ezzeldin, and Dr. Ahmed Siam, respectively. Moreover, Dr. Ahmed Yosri provided advice and guidance for the papers presented in Chapters 4 and 5. Information from outside sources, which has been used towards analysis or discussion, has been cited where appropriate. This thesis consists of the following manuscripts in the following chapters:

Chapter 2

Haggag, M., Ezzeldin, M., El-Dakhakhni, W., & Hassini, E. (2020). **Resilient cities critical infrastructure interdependence: a meta-research**. *Sustainable and Resilient Infrastructure*, 1-22.

Chapter 3

Haggag, M., Siam, A. S., El-Dakhakhni, W., Coulibaly, P., & Hassini, E. (2021). **A deep learning model for predicting climate-induced disasters**. *Natural Hazards*, 1-26.

Chapter 4

Haggag, M., Yorsi, A., El-Dakhakhni, W., & Hassini, E. (2021). **Infrastructure performance prediction under Climate-Induced Disasters**

using data analytics. *International Journal of Disaster Risk Reduction*, *56*, 102121.

Chapter 5

Haggag, M., Yorsi, A., El-Dakhakhni, W., & Hassini, E. (2021). **A Data-Driven Model for Climate-Induced Disaster Damage Prediction**. *International Journal of Disaster Risk Reduction*, submitted for publication in June 2021

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

Hazards are nature- or human-induced events that can cause potential damage (Zimmerman, Zhu, De Leon, & Guo, 2017). Natural hazards are typically impossible to control and include those climate-induced such as tornados, floods, hurricanes, and wildfires, and those that are not related to climate such as volcanos and earthquakes. Anthropogenic (human-induced) hazards are deliberately or unintentionally caused by human activities such as wars, terroristic attacks, industrial accidents and computer viruses. Those hazards can be identified as disasters if they induce adverse impacts on people or properties. These negative impacts can be quantified through three parameters (Zimmerman et al., 2017): (1) the number of people affected by the disaster referred to as System Average Interruption Frequency Index, (2) the total duration of the aftermath of the disaster referred to as Average Interruption Duration, and (3) the total cost for bouncing back to a normal condition after the disaster referred to as Performance Loss Costs.

The frequency and impacts of Climate-Induced Disasters (CID) have been increasing drastically over the past few of decades (Thomas, Albert, & Hepburn, 2014). These CID are driven by the changing climate (i.e., temperature and precipitation) and can be divided into climatological (i.e., heat waves and droughts), meteorological (i.e., wind and winter storms) and hydrological disasters (i.e., floods and flash floods) ("Climate change | EU Science Hub," 2018; Shaftel, 2018a, 2018b). According to the Centre for Research on the Epidemiology of Disasters, the global average number of CID has tripled

in less than four decades (from approximately 1,300 Climate-Induced Disasters (CID) between 1975 and 1984 to around 3,900 between 2005 and 2014) (Thomas & López, 2015). In addition, around 1 million deaths and $1.7 trillion damage costs were attributed to CID since 2000 (Guha-sapir, Hoyois, & Below, 2015; Thomas & López, 2015), with around $210 billion incurred only in 2020 (Newburger, 2021). Consequently, the World Economic Forum identified extreme weather as the top ranked global risk in terms of likelihood and among the top five risks in terms of impact in the last 4 years (World Economic Forum, 2020). Furthermore, almost quarter of the world's population is officially threatened by storm surges and tsunamis ("Climate Change - Oxfam Canada," n.d.). Moreover, since mid 2017, floods have affected about 41 million people with 150 million people living in areas that will be officially under sea level by the end of this century ("Five Ways Climate Change Is Already Affecting Canada," n.d.; *Impacts of climate change - Canada*, n.d.). These risks are not expected to diminish as: *i*) the number of CID is anticipated to double during the next 13 years (Lopez, Thomas, & Troncoso, 2020); *ii*) the annual fatalities due to CID are expected to increase by 250,000 deaths in the next decade (World Health Organization, 2018); and *iii*) the annual CID damage costs are expected to increase by around 20% in 2040 compared to those realized in 2020 ("Natural Disasters Could Cost 20 Percent More By 2040 Due to Climate Change - Yale E360," 2020; "New approaches to help businesses tackle climate change | University of Cambridge," 2020).

Given that 70% of the world population will be living in urban centres by 2050 (Haggag, Ezzeldin, El-Dakhakhni, & Hassini, 2020) together with the increasing complexity of modern cities, the effects of CID on both the human- and community- levels is expected to intensify in the coming decades. In this context, cities can be visualized as

complex systems comprised of interdependent infrastructure systems that provide the necessary services to carry out basic operational functions. Among such infrastructure systems, some are extremely vital for cities to preserve their adaptive nature. As per the Executive Order 13010 "*certain national infrastructures are so vital that their incapacity or destruction would have a debilitating impact on the defense or economic security of the United States*" (Moteff & Parfomak, 2004). These systems include, but are not limited to, electric power systems, telecommunication system, gas and oil storage and transportation systems, water supply systems, and transportation systems. The interdependence among the critical infrastructure systems is what increases the complexity and vulnerability of cities under CID. To minimize the negative impacts of interdependence, complex systems within cities must be resilient enough to absorb any disturbance and retain their basic functions during and following any extreme event. Within this context, Thomas Frieden, director of the US Centers for Disease Control, noted that *"resilient systems are everyday systems that can be scaled up. Managing in an emergency is like managing normally, except more so"* ("Toronto - 100 Resilient Cities," n.d.).

1.1.1. SYSTEM RESILIENCE: DEFINITIONS AND METRICS

Resilience is the ability of a system to adjust and adapt to internal and external changes (Pickett, Cadenasso, & Grove, 2004). Another definition of resilience is the ability of a system to be restored to its original balance (Holling, 1973). Resilience determines *"the persistence of relationships within a system and is a measure of the ability of these systems to absorb changes and still persist"* (Holling, 1973). Furthermore, resilience is a function of time or how long it takes a system to bounce back after a shock (Spaans & Waterhout, 2017). Resilience is also about the amount of disturbance a system can take and yet remain

within its "*critical thresholds*" (Davoudi et al., 2012). Moreover, resilience pertains to planning and evaluation to decrease the vulnerability of the underlying system (Desouza & Flanery, 2013). Finally, resilience refers to the act of *bouncing back* (Vale, 2014).

In 2013 and as part of the Rockefeller Foundation's efforts to promote the well-being of humanity and improve the quality of life of humans, the foundation adopted a new program that focuses on urban resilience. The program, under the name of 100 Resilient Cities, is currently helping cities worldwide to fight shocks and stresses pertaining to the physical, social and economic aspects of daily life. The program defines urban resilience as the capacity of individuals, communities, institutions, businesses and systems within a city to survive, adapt and grow no matter what kind of chronic stresses or acute shocks happen ("About Us | 100 Resilient Cities," n.d.; Spaans & Waterhout, 2017). The City Resilience framework, developed by Arup in conjunction with the Rockefeller Foundation, proposed four basic city elements, twelve performance indicators and seven qualities that are measures of a city's ability to adapt to extreme events and their consequences (Rockefeller Foundation, 2014).

System resilience quantification has proved to be a highly complicated issue given the diverse definitions and metrics that are currently being used to measure resilience. Resilience for environmental, physical and social systems was defined as a term that consists of four dimensions (Bruneau et al., 2003), each has to be optimized in order to exploit system resilience. These four dimensions are: (1) robustness: the ability of a system to maintain its functionality under adverse events, (2) rapidity: the time it takes the system to bounce back to a pre-existing state, (3) redundancy: the availibi8lity of replacement components in the system, (4) resourcefulness: the availability of resources that aid the

system at the time of disaster including forecasting services and preparedness plans. The first two dimensions (i.e., robustness and rapidity) are referred to as resilience *goals* which can be achieved through maximizing the last two dimensions of resilience (i.e., redundancy and resourcefulness) which are referred to as resilience *means* (Rose & Krausmann, 2013).

Through analyzing resilience research trends three actions were proposed to improve the quality of resilience research which are (Opdyke, Javernick-Will, & Koschmann, 2017): (1) conducting resilience-related studies in developing countries; (2) using mixed-methods (qualitative and quantitative) in order to evaluate the resilience of the underlying system; and (3) analyzing critical infrastructure systems closely by including both social (i.e., *"the capacity of social ties and networks in limiting negative impacts from hazards"*) (Aldrich & Meyer, 2015) and environmental (i.e., "*the perturbation of hazard impacts through ecological systems"*) (Prior & Eriksen, 2013) dimensions of resilience which relate to individuals/groups and ecological systems, respectively. As such, the first part of the thesis focuses on analyzing both qualitatively and quantitively previous research work pertaining to interdependence of critical infrastructure systems and city resilience. This analysis aims to: (1) assess in a quantitative manner the current status of systems resilience research, (2) evaluate the different resilience metrics used in literature, (3) categorize the different types of systems interdependencies, (4) compare the mostly used system modelling techniques and identify their drawbacks, and (5) uncover key research gaps and vulnerable critical infrastructure systems.

Given that resourcefulness is key for maximizing the resilience of a system, it recently became extremely crucial to develop effective and efficient models that could predict both the frequency and impacts of CID which in return would ultimately increase

the preparedness of communities to such disasters. This is considered a challenging goal given both the extreme unpredictability of the frequency and impacts of CID and the complex behavior of cities that stems from the interconnectivity of their comprising infrastructure systems. Fortunately, the emergence of data-driven modelling and machine learning techniques which assume that models can be trained using historical data and accordingly, can efficiently learn to predict different complex features, developing robust models that can predict the frequency and impacts of CID became more conceivable. Accordingly, recent research studies have started to employ data analytics and machine learning in disaster prediction and mitigation.

1.1.2. DATA ANALYTICS FOR CID RISK MITIGATION

Data Analytics which is a relatively new field that has been developing rapidly since the 1960's (Keith Foote, 2018), pertains to analyzing raw data to derive meaningful information in an attempt to assist decision makers answer key questions regarding a certain problem or phenomena (Liberty, 2019). It is divided into three main categories: (1) descriptive analytics; (2) predictive analytics; and (3) predictive analytics ("The Power of Analytics," n.d.). Descriptive analytics aims at analyzing historical data to answer key questions concerning complex processes. On the other hand, predictive analytics is more concerned with future rather than historical conditions, thus, in predictive analytics the aim is to predict the future behavior of entities. Finally, prescriptive analytics is concerned with making the best decision after analyzing outcomes of the first two data analytics classes.

The three categories of data analytics were extensively employed in literature to answer key question about CID, derive meaningful relationships between the different

variables related to these disasters, and reach important conclusions that aid the decision-making processes related to disaster impact reduction and risk mitigation. In that context, historical data of the number of fatalities and incurred damages due to natural disasters was used to make inferences on the relationships between those disasters and social and economic aspects. Regression models were used to derive such relationships and the results showed that the impacts of natural disasters are inversely proportional to the increase in social and/or economic conditions (Toya & Skidmore, 2007). The role of social work was identified through qualitatively analyzing interview data in South Africa, where it was found that social work intervention is essential for disaster impact mitigation (Shokane, 2019). Furthermore, the impact of natural disasters on economic growth was studied by qualitatively analyzing case studies following large disasters. It was shown that only very high impact disasters that essentially lead to rebellions have an impact on economic growth (Cavallo, Galiani, Noy, & Pantano, 2013). Instead, empirical analysis for historical disaster data showed that climate related disasters have an effect on economic growth which was attributed to the increased productivity due to adopting new technologies for future disaster mitigation. (Skidmore & Toya, 2002). On another perspective, natural disasters were linked to instability and conflicts using a multivariate model. Historical data from 1990 to 1999 was used to derive relationships between natural disasters and social conflicts, and it was found that both are directly related (Bhavnani, 2006). In addition, natural disasters data from 1971 to 2000 was used to identify and assess risks related to climate variability. Among the identified risks are the number of fatalities and people affected by climate related disasters. The study showed that developing countries are the ones with the highest risk pertaining to climate related disasters (Brooks & Adger, 2003). The link between

natural disasters and population mobility was also established using historical household and natural disasters data. The multivariate analysis performed showed that crop failure due to climate disasters is strongly related to long-term population mobility (Gray & Mueller, 2012). Alternatively, the cost of damage related to a tropical cyclone landfall in China was used to estimate historical damage functions, which were found to be different for each damage dataset analyzed. This difference lead to divergence in future predictions, which calls for standardizing the damage reporting of international disaster (Bakkensen, Shi, & Zurita, 2018). Focusing on disaster direct loss estimation, billion-dollar disasters in the United States from 1980 to 2011 were used to detail the methodology used to convert from insured costs to direct disaster losses. The analysis showed that for the direct losses estimate to be more accurate, the spatial and temporal differences in insurance rates have to be considered in the analysis (Smith & Katz, 2013). Descriptive analytics was also employed to draw meaningful conclusions from a developed climate disaster dataset in Greece for the period from 2001 to 2011. The analysis was able to highlight the most frequent disasters, the months when disasters occur the most, the most costly disasters, and the spatial distribution of disasters in Greece (Papagiannaki, Lagouvardos, & Kotroni, 2013). Turning to Asia-Pacific, historical data was used to assess factors related to climate disasters. It was shown that the changes in temperature are highly associated with climatological disasters, whereas the high fluctuations in precipitation together with the increased population exposure are associated with hydrological disasters (Thomas et al., 2014).

1.1.3. CID PREDICTION USING MACHINE LEARNING APPLICATIONS

Machine learning, a class of artificial intelligence, assumes that models can be trained using data and thus, can efficiently learn to predict different complex phenomena. The two classes of machine learning are: (1) supervised learning, and (2) unsupervised learning. The first class uses labelled data to train and test the model, whereas the second class used unlabelled data for building the model. The two classes have experienced rapid advancements in natural phenomena simulation and prediction recently. For example, flood damage was linked to some household predictors including house structure, flood awareness, literacy and other factors using three types of models: linear regression, random forests and artificial neural networks (Ganguly, Nahar, & Hossain, 2019). Neural network models were also proposed to predict number of hurricanes per season in a specific place, with a prediction accuracy of 73% (Kahira, Gomez, & Badia Sala, 2018). Machine learning was also recently used for wildfire event prediction, specifically anthropogenic wildfire events were predicted using random forests, boosting and support vector machines (Rodrigues & De la Riva, 2014). Additionally, a spatial prediction of wildfire probabilities was proposed by combining different machine learning models with optimization algorithms (i.e., genetic algorithms), and the results showed that optimization algorithms were able to refine the developed machine learning model (Jaafari, Zenner, Panahi, & Shahabi, 2019). Neural networks and logistic regression were used to predict binary and four-class wind damage (Hanewinkel, Zhou, & Schill, 2004), whereas wind gust occurrence was predicted as binary outcome using the same two techniques together with decision trees (Sallis, Claster, & Herna, 2011). Decision trees together with bagging, random forest and boosting were also used to predict heavy rain damage as binary outcome

(Choi et al., 2018), while both neural networks and random forest were employed to classify flood severity into three classes (Khalaf et al., 2018) and predict flood household damage (Ganguly et al., 2019). Moreover, neural networks were used to predict property damage caused by tornado disasters (Diaz & Joseph, 2019). Furthermore, unsupervised machine learning was used to quantify community flood resilience across different US counties (Abdel-Mooty, Yosri, El-Dakhakhni, & Coulibaly, 2021).

While the majority of the previous studies focus on predicting the frequency and direct impacts of CID, these studies clearly lack several key considerations which include: (1) establishing the link between climate change and CID occurrences, (2) predicting the specific impacts of CID on critical infrastructure systems rather than their lumped monetary impacts, (3) integrating different types of data (i.e., hazard, socio-economic, and climate, etc.) for the prediction of CID impacts, (4) developing systematic and standardized approaches for the prediction of both CID occurrences and impacts. Consequently, the second, third and fourth phases of this work aim to address the above-mentioned gaps in an attempt to enhance the preparedness, and thus, the overall resilience of urban communities.

1.2. RESEARCH OBJECTIVES AND PHASES

The aim of this work is to enhance urban centre preparedness and resilience to CID by employing data analytics techniques. To achieve this goal the following specific objectives were outlined:

(1) Conducting, both qualitatively and quantitively, a comprehensive review of the previous research work pertaining to resilience of cities and their comprising critical infrastructure systems through employing text analytics.

(2) Developing a deep learning modelling approach that predicts CID occurrences through linking different climate change indices to historical disaster records.

(3) Developing a standardized data-driven framework for predicting the performance of critical infrastructure systems under CID by employing text mining and data imputation techniques.

(4) Developing a standardized data-driven framework for predicting the lumped impacts of CID on urban communities through integrating different types of data and employing feature selection techniques.

To attain the above-mentioned objectives, this dissertation is divided into four main phases as shown in Figure 1-1. In Phase 1 a comprehensive field exploration is conducted through employing meta research. As such, previous publications in the field of infrastructure systems resilience and interdependence are analyzed quantitatively using topic modelling, and qualitatively using a detailed critical review process. Upon the completion of this exploration, it was shown that a key gap in systems resilience research pertains to hazard considerations including the availability of resources that aid in predicting the occurrences and impacts of disasters on both the system- and community-levels. Moreover, it was shown that the most critical infrastructure system that all other systems depend on for their daily operations is the power system.

To increase the preparedness of urban communities and thus, foster their resilience, Phases 2, 3, and 4 of this dissertation focuses on developing systematic approaches for the

prediction of various CID-related aspects. These aspects include the occurrences of CID, their impacts on specific critical systems, and their related damages. As such, Phase 2 focuses on predicting the occurrences of CID through establishing a link between these disasters and climate change, which is considered their key drive. Phase 3, on the other hand, aims at predicting the specific impacts of CID on critical infrastructure systems. In that context, a standardized data-driven framework that utilizes data analytics in the form of text mining, data imputation, and predictive modelling is developed. Finally, in Phase 4 of the current dissertation the monetary cost of CID is predicted using a systematic data-driven approach that integrates different data types and applies several machine learning techniques to select the optimal prediction model.

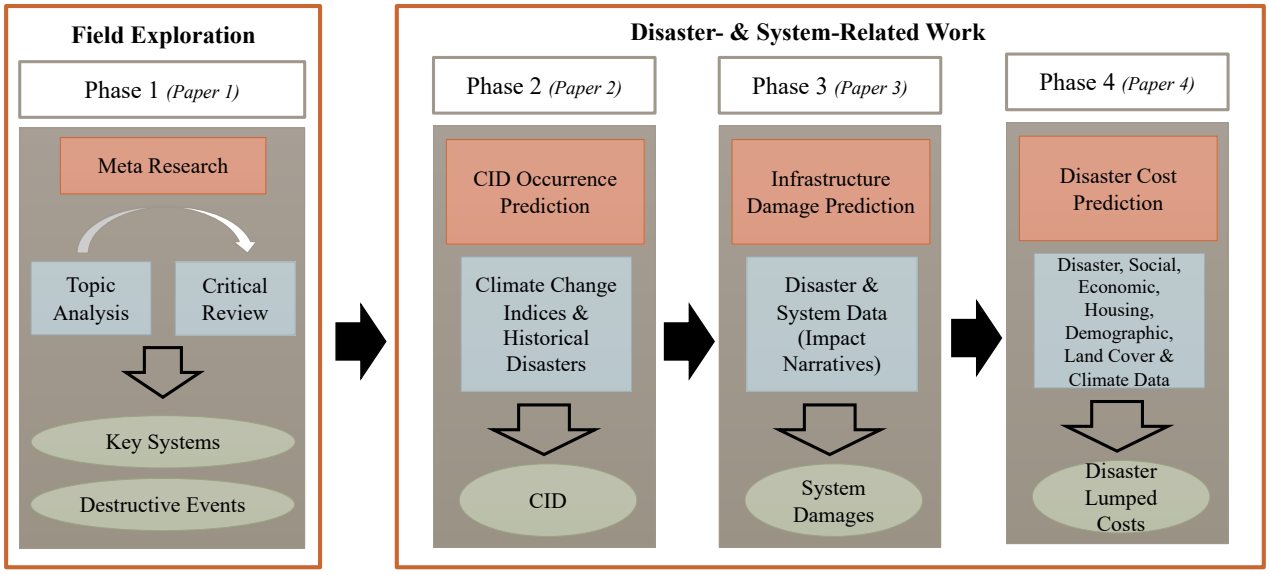


Figure 1-1: Research Phases

1.3. ORGANIZATION OF THE DISSERTATION

This section outlines the contents of the different chapters that comprise this dissertation:

- **Chapter 1** provides the background and motivation for the research conducted herein, together with an outline for the specific research objectives, a description of the employed research phases and a summary of the contents of the dissertation.
- **Chapter 2** presents a critical review of the research work pertaining to resilience of critical infrastructure systems. First, a meta-research approach is employed in the form of topic modelling to *quantitatively* uncover latent topics and their statistical distributions in pertinent literature. Subsequently, the identified topics are *qualitatively* analyzed in terms of established resilience definitions and quantification metrics as well as currently adopted simulation approaches for city infrastructure systems interdependence. Moreover, the contribution of the identified topics and the research gaps pertaining to systems resilience research are uncovered in an attempt to identify possible blue ocean research opportunities.
- **Chapter 3** presents a deep learning modelling approach that aims to predict CID occurrences by linking historical disaster records to different climate change indices. The chapter is divided into two main parts, the first part details the general methodology which can be employed to predict any class of CID in any location. The four stages that comprise this methodology are outlined which include the model architecture analysis, input variable analysis, model selection and prediction, and model validation. The second part of the chapter demonstrates the applicability of the developed model using the province of Ontario's disaster rerecords and relevant climate change indices data.
- **Chapter 4** discusses the developed systematic framework for predicting infrastructure system damages under CID. The chapter is divided into two main

sections, the first section details the damage prediction framework's stages, whereas the second section demonstrates the framework's applicability through a case study. As detailed in the first section, the internal processes of the framework consist of four main stages: linking CID to infrastructure systems, investigating and exploiting the influencing attributes, employing data imputation, and developing and testing the machine learning model. To demonstrate its applicability and viability, the second section of this chapter presents the application of the framework to the historical disaster data collected by the US National Weather Services between 1996 and 2019.

- **Chapter 5** presents a systematic framework that aims to predict CID direct impacts through employing data-driven approaches. This chapter is divided into two main parts: the first part explains in detail the systematic stages pertaining to the CID direct impact prediction framework, and the second part includes a demonstration case study that is used to assess the viability of the proposed framework. In the first part, a detailed description of the framework's four main phases is provided. These stages are: data collection and compilation, feature selection, model development, and result analysis and interpretation. The second part of this chapter presents the case study which is used to assess the applicability of the proposed framework by linking wind disaster data collected by the National Weather Services to climate, land cover, social, housing, demographic, and economic data in the state of New York from 2010 to 2018.
- Chapter 6 provides a summary for the dissertation and its main conclusions and findings. Recommendations for future work are also provided in this chapter.

1.4. REFERENCES

Abdel-Mooty, M. N., Yosri, A., El-Dakhakhni, W., & Coulibaly, P. (2021). Community Flood Resilience Categorization Framework. *International Journal of Disaster Risk Reduction*, 102349. https://doi.org/10.1016/j.ijdrr.2021.102349

About Us | 100 Resilient Cities. (n.d.). Retrieved November 10, 2017, from http://www.100resilientcities.org/about-us#/-_/

Aldrich, D. P., & Meyer, M. A. (2015). Social capital and community resilience. *American Behavioral Scientist, 2*(59), 254–269.

Bakkensen, L., Shi, X., & Zurita, B. (2018). The Impact of Disaster Data on Estimating Damage Determinants and Climate Costs. *Economics of Disasters and Climate Change*, *2*(1), 49–71. https://doi.org/10.1007/s41885-017-0018-x

Bhavnani, R. (2006). Natural Disaster Conflicts. Harrvard University.

Brooks, N., & Adger, W. N. (2003). Country level risk measures of climate-related natural disasters and implications for adaptation to climate change. In *Change* (Vol. 26). Retrieved from http://www.uea.ac.uk/env/people/adgerwn/wp26.pdf

Bruneau, M., Chang, S., Eguchi, R., Lee, G., O'Rourke, T., Reinhorn, A., & Von Winterfeldt, D. (2003). A framework to quantitatively assess and enhance the seismic resilience of communities. *Earthquake Spectra*, *19*(4), 733–752.

Cavallo, E., Galiani, S., Noy, I., & Pantano, J. (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, *95*(5), 1549–1561. https://doi.org/10.1162/REST_a_00413

Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y., & Kim, H. S. (2018). Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data. *Advances in Meteorology*, *2018*. https://doi.org/10.1155/2018/5024930

Climate Change - Oxfam Canada. (n.d.). Retrieved November 10, 2018, from https://www.oxfam.ca/themes/water/

Climate change | EU Science Hub. (2018). Retrieved November 15, 2018, from https://ec.europa.eu/jrc/en/research-topic/climate-change

Davoudi, S., Shaw, K., Haider, L., Quinlan, A., Peterson, G., Wilkinson, C., & Davoudi, S. (2012). Resilience: a bridging concept or a dead end? *Planning Theory & Practice*, *13*(2), 299–333. https://doi.org/https://doi.org/10.1080/14649357.2012.677124

Desouza, K., & Flanery, T. (2013). Designing, planning, and managing resilient cities: A conceptual framework. *Cities*, *35*, 85–99. https://doi.org/https://doi.org/10.1016/j.cities.2013.06.003

Diaz, J., & Joseph, M. B. (2019). Predicting property damage from tornadoes with zero-in fl ated neural networks. *Weather and Climate Extremes*, *25*(July 2018), 100216. https://doi.org/10.1016/j.wace.2019.100216

Five Ways Climate Change Is Already Affecting Canada. (n.d.). Retrieved November 10, 2019, from https://www.c-bc.ca/natureofthings/blog/how-climate-change-is-already-affecting-canada

Ganguly, K., Nahar, N., & Hossain, M. (2019). A machine learning-based prediction and

analysis of flood affected households: A case study of floods in Bangladesh. *International Journal of Disaster Risk Reduction*, *34*, 283–294. https://doi.org/10.1016/j.ijdrr.2018.12.002

Gray, C. L., & Mueller, V. (2012). Natural disasters and population mobility in Bangladesh. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6000–6005. https://doi.org/10.1073/pnas.1115944109

Guha-sapir, D., Hoyois, P., & Below, R. (2015). Annual Disaster Statistical Review 2014: The numbers and trends. In *Ciaco Imprimerie, Louvain-la-Neuve (Belgium)*. Retrieved from http://www.cred.be/sites/default/files/ADSR_2010.pdf

Haggag, M., Ezzeldin, M., El-Dakhakhni, W., & Hassini, E. (2020). Resilient cities critical infrastructure interdependence: a meta-research. *Sustainable and Resilient Infrastructure*, 1–22. https://doi.org/10.1080/23789689.2020.1795571

Hanewinkel, M., Zhou, W., & Schill, C. (2004). A neural network approach to identify forest stands susceptible to wind damage. *Forest Ecology and Management*, *196*(2–3), 227–243. https://doi.org/10.1016/j.foreco.2004.02.056

Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, *4*(1), 1–23.

Impacts of climate change - Canada. (n.d.).

Jaafari, A., Zenner, E., Panahi, M., & Shahabi, H. (2019). Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. *Agricultural and Forest Meteorology*,

266–267, 198–207. https://doi.org/10.1016/j.agrformet.2018.12.015

Kahira, A., Gomez, B., & Badia Sala, R. (2018). A Machine Learning Workflow for Hurricane Prediction. *Book of Abstracts. Barcelona Supercomputing Center*, 72–73.

Keith Foote. (2018). A Brief History of Analytics. Retrieved from DATAVERSITY website: https://www.dataversity.net/brief-history-analytics/

Khalaf, M., Hussain, A. J., Al-jumeily, D., Baker, T., Keight, R., Lisboa, P., … Kafri, S. Al. (2018). A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction. *2018 IEEE Congress on Evolutionary Computation (CEC)*, 1–8.

Liberty, D. (2019). Data Science vs. Big Data vs. Data Analytics. Retrieved from https://www.sisense.com/blog/data-science-vs-data-analytics/

Lopez, R., Thomas, V., & Troncoso, P. (2020). Impacts of Carbon Dioxide Emissions on Global Intense Hydrometeorological Disasters. *Climate, Disaster and Development Journal*, *4*(1), 30–50. https://doi.org/10.18783/cddj.v004.i01.a03

Moteff, J., & Parfomak, P. (2004). Critical infrastructure and key assets: definition and identification. In *Library of Congress Washington DC Congressional Research Service*.

Natural Disasters Could Cost 20 Percent More By 2040 Due to Climate Change - Yale E360. (2020). Retrieved January 2, 2021, from https://e360.yale.edu/digest/natural-disasters-could-cost-20-percent-more-by-2040-due-to-climate-change

New approaches to help businesses tackle climate change | University of Cambridge.

(2020). Retrieved March 22, 2021, from https://www.cam.ac.uk/research/news/new-approaches-to-help-businesses-tackle-climate-change

Newburger, E. (2021). Disasters cause $210 billion in damage in 2020. Retrieved March 22, 2021, from https://www.cnbc.com/2021/01/07/climate-change-disasters-cause-210-billion-in-damage-in-2020.html

Opdyke, A., Javernick-Will, A., & Koschmann, M. (2017). Infrastructure hazard resilience trends: an analysis of 25 years of research. *Natural Hazards*, *87*(2), 773–789. https://doi.org/10.1007/s11069-017-2792-8

Papagiannaki, K., Lagouvardos, K., & Kotroni, V. (2013). A database of high-impact weather events in Greece: A descriptive impact analysis for the period 2001-2011. *Natural Hazards and Earth System Science*, *13*(3), 727–736. https://doi.org/10.5194/nhess-13-727-2013

Pickett, S., Cadenasso, M., & Grove, J. (2004). Resilient cities: meaning, models, and metaphor for integrating the ecological, socio-economic, and planning realms. *Landscape and Urban Planning*, *69*(4), 369–384. https://doi.org/https://doi.org/10.1016/j.landurbplan.2003.10.035

Prior, T., & Eriksen, C. (2013). Wildfire preparedness, community cohesion and social–ecological systems. *Global Environmental Change*, *6*(23), 1575–1586. https://doi.org/https://doi.org/10.1016/j.gloenvcha.2013.09.016

Rockefeller Foundation. (2014). City Resilience Framework. In *Arup*. Retrieved from http://www.seachangecop.org/files/documents/URF_Booklet_Final_for_Bellagio.pdf%5Cnhttp://www.rockefellerfoundation.org/uploads/files/0bb537c0-d872-467f-

9470-b20f57c32488.pdf%5Cnhttp://resilient-cities.iclei.org/fileadmin/sites/resilient-cities/files/Image

Rodrigues, M., & De la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling and Software*, *57*, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003

Rose, A., & Krausmann, E. (2013). An economic framework for the development of a resilience index for business recovery. *International Journal of Disaster Risk Reduction*, *5*, 73–83. https://doi.org/10.1016/j.ijdrr.2013.08.003

Sallis, P. J., Claster, W., & Herna, S. (2011). A machine-learning algorithm for wind gust prediction. *Computers & Geosciences*, *37*, 1337–1344. https://doi.org/10.1016/j.cageo.2011.03.004

Shaftel, H. (2018a). Causes | Facts – Climate Change: Vital Signs of the Planet. Retrieved March 15, 2019, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/causes/

Shaftel, H. (2018b). Evidence | Facts – Climate Change: Vital Signs of the Planet. Retrieved November 15, 2018, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/evidence/

Shokane, A. L. (2019). Social work assessment of climate change: Case of disasters in greater Tzaneen municipality. *Jàmbá Journal of Disaster Risk Studies*, *11*(3), 1–7. https://doi.org/10.4102/jamba.v11i3.710

Skidmore, M., & Toya, H. (2002). Do natural disasters promote long-run growth?

Economic Inquiry, *40*(4), 664–687. https://doi.org/10.1093/ei/40.4.664

Smith, A. B., & Katz, R. W. (2013). US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases. *Natural Hazards*, *67*(2), 387–410. https://doi.org/10.1007/s11069-013-0566-5

Spaans, M., & Waterhout, B. (2017). Building up resilience in cities worldwide–Rotterdam as participant in the 100 Resilient Cities Programme. Cities. *Cities*, *61*, 109–116. https://doi.org/https://doi.org/10.1016/j.cities.2016.05.011

The Power of Analytics. (n.d.). Retrieved October 28, 2019, from GUROBI Optimization website: https://www.gurobi.com/company/about-gurobi/prescriptive-analytics/

Thomas, V., Albert, J. R. G., & Hepburn, C. (2014). Contributors to the frequency of intense climate disasters in Asia-Pacific countries. *Climatic Change*, *126*(3–4), 381–398. https://doi.org/10.1007/s10584-014-1232-y

Thomas, V., & López, R. (2015). Global Increase In ClImate-Related DIsasters. In *ADB economics working paper series*. https://doi.org/10.4449/aib.v121i2.3187

Toronto - 100 Resilient Cities. (n.d.). Retrieved December 12, 2018, from https://www.100resilientcities.org/cities/toronto/

Toya, H., & Skidmore, M. (2007). Economic development and the impacts of natural disasters. *Economics Letters*, *94*(1), 20–25. https://doi.org/10.1016/j.econlet.2006.06.020

Vale, L. (2014). The politics of resilient cities : whose resilience and whose city ? The politics of resilient cities : whose resilience and whose city ? *Building Research &*

Information, *42*(2), 191–201. https://doi.org/10.1080/09613218.2014.850602

World Economic Forum. (2020). *The Global Risks Report*. *15*, 1–114. Retrieved from http://wef.ch/risks2019

World Health Organization. (2018). Climate change and health. Retrieved June 6, 2019, from https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

Zimmerman, R., Zhu, Q., De Leon, F., & Guo, Z. (2017). Conceptual modeling framework to integrate resilient and interdependent infrastructure in extreme weather. *Journal of Infrastructure Systems*, *23*(4). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000394

Chapter 2

RESILIENT CITIES CRITICAL INFRASTRUCTURE INTERDEPENDENCE: A META-RESEARCH

ABSTRACT

Given the unforeseen events that take place worldwide, most cities are experiencing rapid transformations. To maintain their basic functions, cities have to be resilient– possess the ability to bounce back to their original state following extreme events. Unfortunately, the behavior of cities is complex because of the high degree of interdependence among their comprising infrastructure systems which stems from their systems-of-systems architecture. The current work presents a critical review of the research work pertaining to resilience of cities' critical infrastructure systems. To conduct such critical review, meta-research is employed through text analytics, in the form of topic modelling, to *quantitatively* uncover related latent topics and their statistical distributions in pertinent literature. Subsequently, the identified topics are *qualitatively* analyzed in terms of established resilience definitions and quantification metrics as well as currently adopted simulation approaches for city infrastructure systems interdependence. Based on the text analytics conducted, nine common topics and five major research gaps are identified. Through both its quantitative and qualitative analyses, this meta-research study is a steppingstone towards better understanding of city infrastructure systems interdependence simulation and their resilience quantification.

Keywords: Cities; Complex Systems; Meta-Research, Resilience; Topic Modelling

2.1. INTRODUCTION

Cities are conceived as complex adaptive systems with inherent abilities to reorganize their comprising components to reach an equilibrium state under unforeseen circumstances. Cities are also spatial and temporal *systems-of-systems* as per Figure 2-1, comprised of different infrastructure systems that facilitate performing their basic functions, with some of these systems being key to preserve cities' adaptive nature (Moteff and Parfomak 2004). Such systems include electric power, telecommunication, oil and gas, water, and transportation infrastructure. In addition to their own *intra*-complexity, predicting the behavior of these systems is further complicated due to their *inter*dependence-induced vulnerability (Zimmerman, Zhu, and Dimitri 2016).

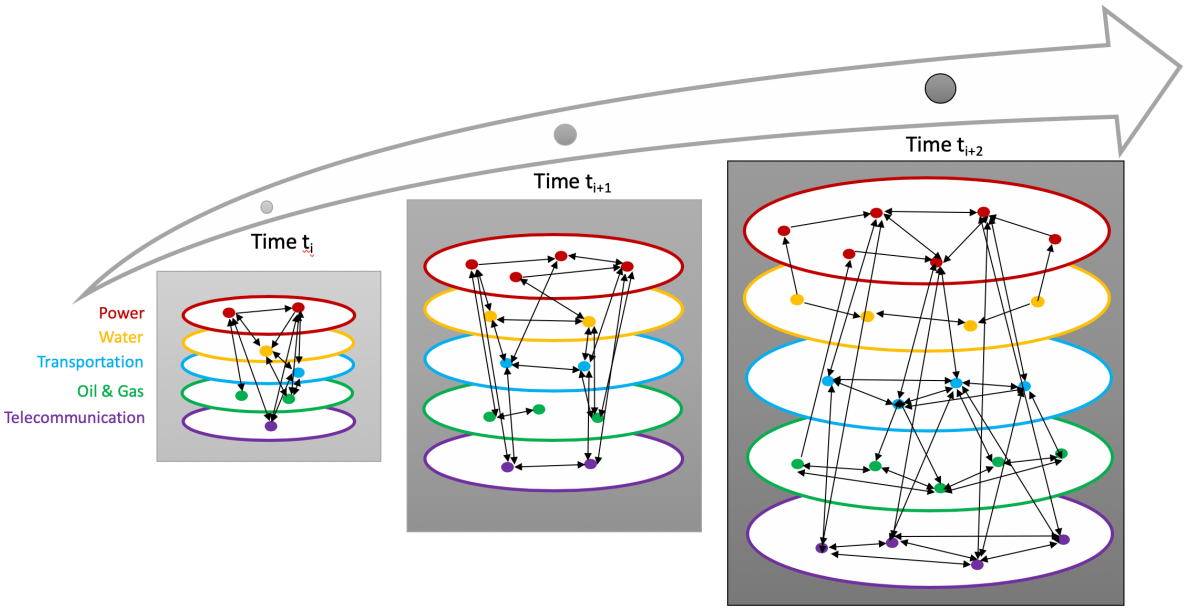


Figure 2-1: Spatio-temporal Evolution of Cities as Systems-of-Systems

The devastating consequences of such interdependence were vividly displayed during the North-eastern blackout which crippled parts of Canada and the United States in 2003. The

time taken to restore different power-dependent infrastructure systems to their initial states varied significantly (Zimmerman and Restrepo 2006). For example, the power infrastructure took a full 72 hours to restore, whereas the New York rail transit, New York traffic signals and Detroit water supply systems took about 94, 188 and 216 hours, respectively. Such a major blackout, among other disruptive events, the most recent of which is the currently evolving COVID-19 epidemic, demonstrates the interdependence-induced cascade vulnerability of our societies' critical systems. Accordingly, there exists an urgent need to evaluate the interdependence of infrastructure systems and, more importantly, to minimize the negative impacts of such interdependence through adopting a *resilient-by-design* philosophy.

Resilience is a multidisciplinary concept that has roots and is perceived differently across different fields including material science (Davoudi et al. 2012; Lu and Stead 2013), ecology (Holling 1996; Standish et al. 2014), urban planning (Spaans and Waterhout 2017), organizational management (Vale 2014), engineering (Vale 2014) and many others. In applied science, resilience was first used to describe the stability of materials under shocks (Davoudi et al. 2012; Lu and Stead 2013). In ecology, resilience indicates how much disturbance an ecosystem can absorb before switching to another state (Holling 1996; Standish et al. 2014). Urban resilience was also advocated for by the Rockefeller Foundation as the capacity of city systems to survive, adapt and grow in spite of *"chronic stresses or acute shocks"* (Spaans and Waterhout 2017). In another example, decisionmakers define resilience as the ability of their organization/facility/system to recover from a certain disruption and return to its original operations (Vale 2014) . In engineering, resilience is the ability of a system to bounce back to a pre-existing or a more

desirable state (Vale 2014). As such, a city that is resilient-by-design have to meet some contradictory objectives to improve inhabitants' everyday life, where it should include redundant (alternative) components/systems yet remain efficient, be diverse (i.e., with some aspects of independence) but benefit from interdependence, operate autonomously but behave collaboratively, and be structured but remain adaptable.

Reviewing relevant literature showed that topics addressing resilience of city infrastructure systems are very broad due to both the multiplicity of quantification metrics and the diversity of city system types and components. As such, a meta-research (conducting research on research) approach, in which text analytics - a field of machine learning which extracts meaningful information from text data through topic modelling (Blei, Andrew, and Micheal 2003), was adopted. In this study, topic modelling is employed to quantitatively identify common topics presented in relevant research publications and subsequently gain corresponding insights (Gatti, Brooks, and Nurre 2015). In this respect, topic modelling has been used to identify research trends in different fields such as transportation (L. Sun and Yin 2017), operational research (Gatti, Brooks, and Nurre 2015), and structural engineering (Ezzeldin M and El-Dakhakhni W. 2019).

The main objective of the current work is to present a critical review of the research work pertaining to the resilience of city interdependent critical infrastructure systems. To conduct such critical review, latent topics are uncovered using topic modelling in Stage 1, as shown in Figure 2-2. Subsequently, the identified topics are then qualitatively analyzed in Stage 2 to review established definitions and quantifiable metrics for resilience of city infrastructure systems, and also to categorize their simulation approaches. Such definitions and approaches are presented to provide a basis for establishing metrics for quantifying

system resilience. It is worth noting that although quantifying the *immediate* impact (risk) of adverse events in terms of inhabitants' lives and injuries is important, the current work mainly focuses on literature pertaining to quantifying the *time-dependant* recovery (resilience), following risk realization, of city infrastructure systems.

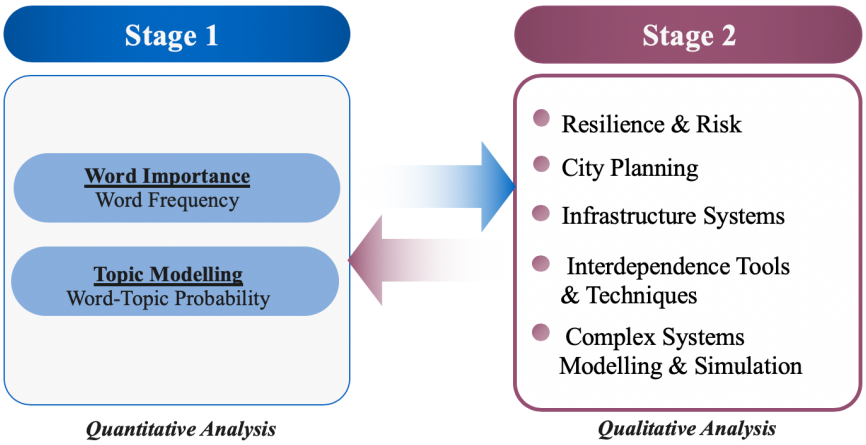


Figure 2-2: The Main Stages of City/Complex Systems Resilience Study

The articles considered in this study were selected from over 600 originally identified publications by considering the latter's relevance to the current field (i.e., city/complex systems resilience). The exploration process started by searching through the Web of Science (https://www.webofknowledge.com) from 1990 to 2017 using the following keywords: City, Resilience, Infrastructure, Interdependence. This search was first performed by topic which resulted in enormous number of publications with the majority of them not related to the current area of study. Thus, the search was further guided by title which resulted in over 600 publications. These publications were subsequently filtered after exploring the abstract of each publication and evaluating each article's relevance to the field. Following this abstract-based screening, the chosen publications were further filtered which yielded the 124 publications that were selected based on their relevance to the field.

2.2. CITY/ SYSTEMS RESILIENCE LITERATURE TEXT ANALYTICS

Instead of subjectively selecting the topics which are relevant in the field and subsequently critically reviewing them, text analytics was used as an objective tool to identify and classify (based on topic modeling) relevant topics prior to reviewing them. Accordingly, the current section is divided into Subsection 2.1 which focuses on highlighting the methodology adopted along with the keywords used in the field; and Subsection 2.2 which focuses on analyzing the results of the topic models generated and selecting the most appropriate accordingly.

2.2.1. METHODOLOGY

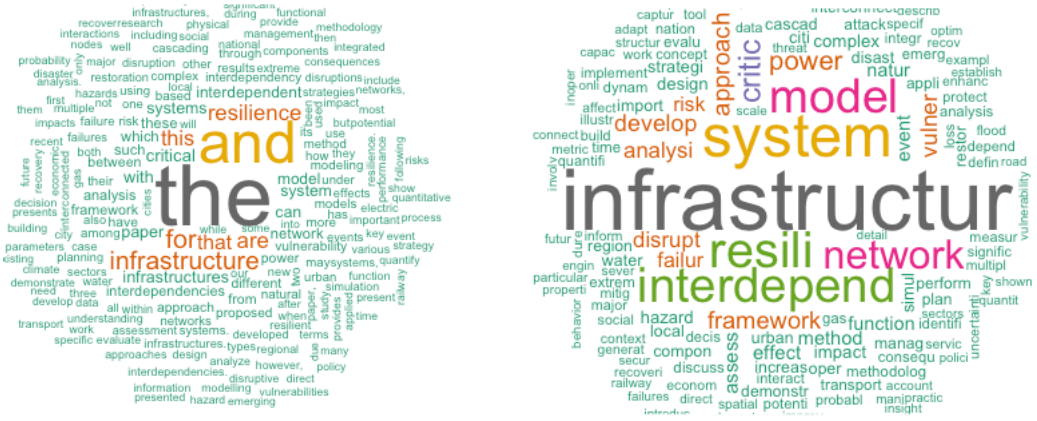
PRE-PROCESSING

Following data collection, the raw abstract dataset was processed using a package that applies different transformation features on text (i.e., tm_map) and is available in R (tm_map function | R Documentation 2019). The abstracts were preprocessed through four steps as in (Miner et al. 2012): (1) transformation, which was used to change all words to be in a lower case format; (2) tokenization, where unstructured text was converted into words in preparation for analysis; (3) treatment, where a standard filter "stop" list was used to remove common words; (4) stemming, where all the affixes were removed.

WORD IMPORTANCE

A word importance can be evaluated through the frequency of this word in the abstract dataset. Figure 2-3 (a and b) show the word cloud before and after pre-processing the abstract dataset, respectively, where large size words have a higher probability of

occurrence compared to that of small size words. The raw abstract dataset has numerous nontechnical common words (i.e., the, this, and) with high frequency, whereas the preprocessed dataset contains technical words, related to city/infrastructure resilience, with high frequency which shows the importance of pre-processing data in text mining.



(a) (b)

Figure 2-3: Word Cloud for the Terms with the Highest Frequencies (a) Before Pre-processing; (b) After Pre-processing

LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that focuses on identifying key topics from a collection of textual documents (Blei, Andrew, and Micheal 2003). The following steps describe how the LDA algorithm adopted herein is used to identify common topics:

1. The user selects the documents to be analyzed (*M*)
2. The user selects the number of Topics (*K*)
3. The algorithm starts by randomly assigning a topic to each word (*w*) in a document (*m*)

   a. Initial per document per topic representation (*Gamma*)
   b. Initial per topic per word representation (*Beta*)

4. To improve on *Gamma* and *Beta*, for each document (*m*), the algorithm checks every word (*w*) and computes two metrics
   a. P (*topic I | document m*): percentage of words in document (*m*) that are assigned to a certain topic (*i*)
   b. P (*word w | topic i*): percentage of assignments to topic i across all documents (*M*) for this word *w*
5. The algorithm assigns the word (*w*) a revised topic based on the above two metrics which together represent the probability that topic (*i*) generated word (*w*)
6. The algorithm repeats Steps 3-5 for all words across all documents (*M*)

Figure 2-4 shows the methodology used to determine the optimum number of topics—one of the main challenges in topic modelling. A threshold of 30 topics was chosen as the maximum number of topics to be considered which is a relatively large number compared to the number of publications under consideration. This high threshold was chosen to visualize how evaluation metrics change as the number of topics is varied. The first metric is the perplexity (Labs n.d.), which is a statistical metric to evaluate the model's ability to predict the sample through calculating the relative degree of uncertainty among models. Generally, lower perplexity values indicate better fitted models; the second metric is the Griffiths' measure (Griffiths et al. 2004), which is based on the Gibbs sampling algorithm. The best number of topics is represented as the measure is maximized.

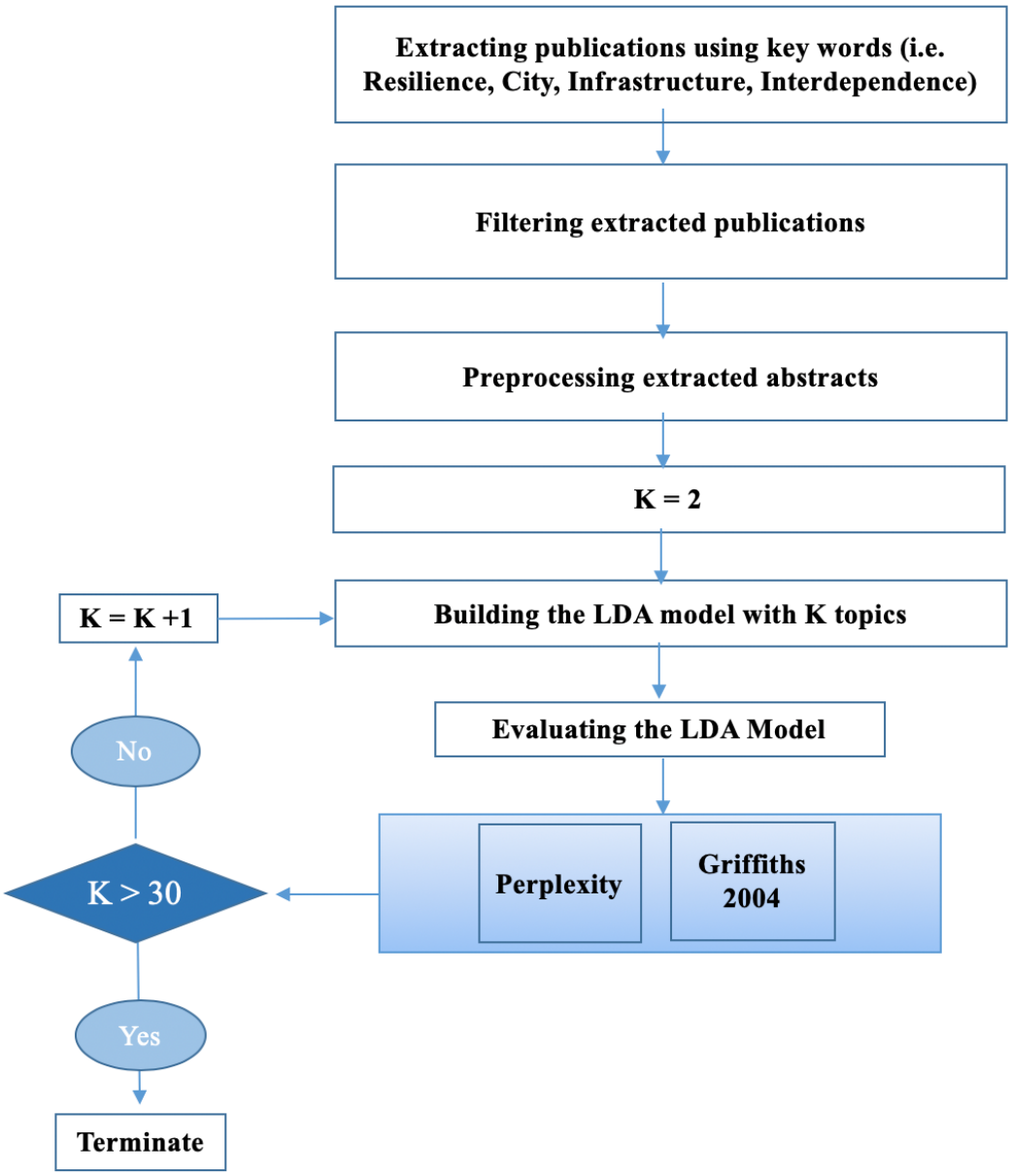


Figure 2-4: Topic Modelling Methodology

2.2.2. RESULTS

MODEL SELECTION

Figure 2-5 shows the sensitivity of the evaluation measures to the number of topics (*k*). Considering perplexity, the best number of topics should be attained at the least value of perplexity calculated, which is 30 topics. This number of topics is quite large considering the number of publications considered. In addition, perplexity decreases with a high and an almost constant rate as the number of topics is increased. The maximum value of Griffiths' measure was attained at 14 topics. However, as can be seen from Figure 5b, the measure experiences a sharp increase at 8 topics, and between 9 and 12 topics, the measure is maximum and is close to its maximum value at 14 topics. Thus, the range of optimal topics is between 9 and 14 topics as proposed by the perplexity and Griffiths' measures. For the purpose of this analysis and due to the number of papers considered, nine topics were chosen to conduct the analysis herein.

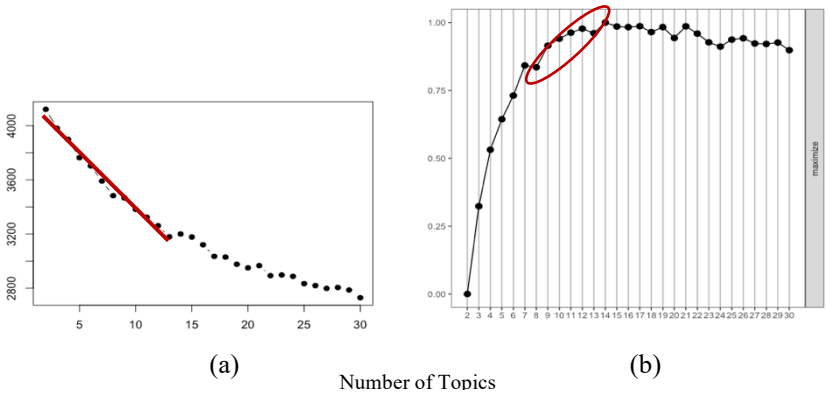


Figure 2-5: (a) Perplexity, vs. the number of topics, and (b) Griffiths vs. the number of topics

TOPIC ANALYSIS

The nine identified topics are shown in figures 6 and 7. **Figure 2-6** shows per-topic-per-word probabilities (*Beta*) for the most ten frequent words within each topic, while **Figure 2-7** shows the word cloud for each of the extracted topics. In the former figure, *Beta* can

be interpreted as the probability that a certain term is generated from a specific topic. It can be observed that, within some of the extracted topics, the most frequent words have very high *Beta* values compared to other extracted topics. For example, comparing the *Beta* values for the most probable word in *Topics A and B*, *Beta* is calculated as 0.169 for the term "resili" in *Topic A* and 0.026 for the term "citi" in *Topic B*. This comparison shows that the probability that the term "resili" is generated from *Topic A* is more than six times the probability that the term "citi" is generated from *Topic* B. In **Figure 2-7**, words with higher probabilities are shown in larger font sizes compared to those of lower probability words.

Starting with *Topic A*, the words "resilience, analysis, framework, etc." are typically related to (system resilience analysis and related frameworks), whereas the words "city, urban, plan, assess, etc." in *Topic B* are typically related to (city assessment and urban planning), and the words "infrastructure, critical, effect, etc." in *Topic* C are typically related to (critical infrastructure systems). Furthermore, the words "interdependence, vulnerability, infrastructure, etc." in *Topic D* are typically related to (infrastructure systems interdependence and vulnerability). Moreover, the words "risk, disrupt, hazard, method, etc." in *Topic E* are mostly related to (risk and disruption due to hazards), whereas the words "model, interdependence, system, etc." in *Topic F* are typically related to (modelling infrastructure systems interdependence). On the other hand, the words "network, complex, etc." in *Topic G* are typically connected to (complex network theory), whereas the words "system, power, water, restore, gas, etc." in *Topic H* are mostly related to (performance and restoration of the three systems: power, gas and water). The LDA model also identified a cross-cutting topic (i.e., *Topic I*) characterized by the words "approach, disrupt, event,

disaster, etc." This topic relates the probability of disasters and their adverse impacts on the disruption of infrastructure systems which is the reason behind the study of system interdependence, vulnerability, risk and resilience. Given the fact that Topic I can be considered as the natural trigger behind all other identified topics, it will not be discussed separately herein. Consequently, research in the field is mainly dominated by concepts pertinent to resilience, risk, and vulnerability. Moreover, the analysis shows that power, water and gas systems are perceived as key to maintaining functionality of cities.

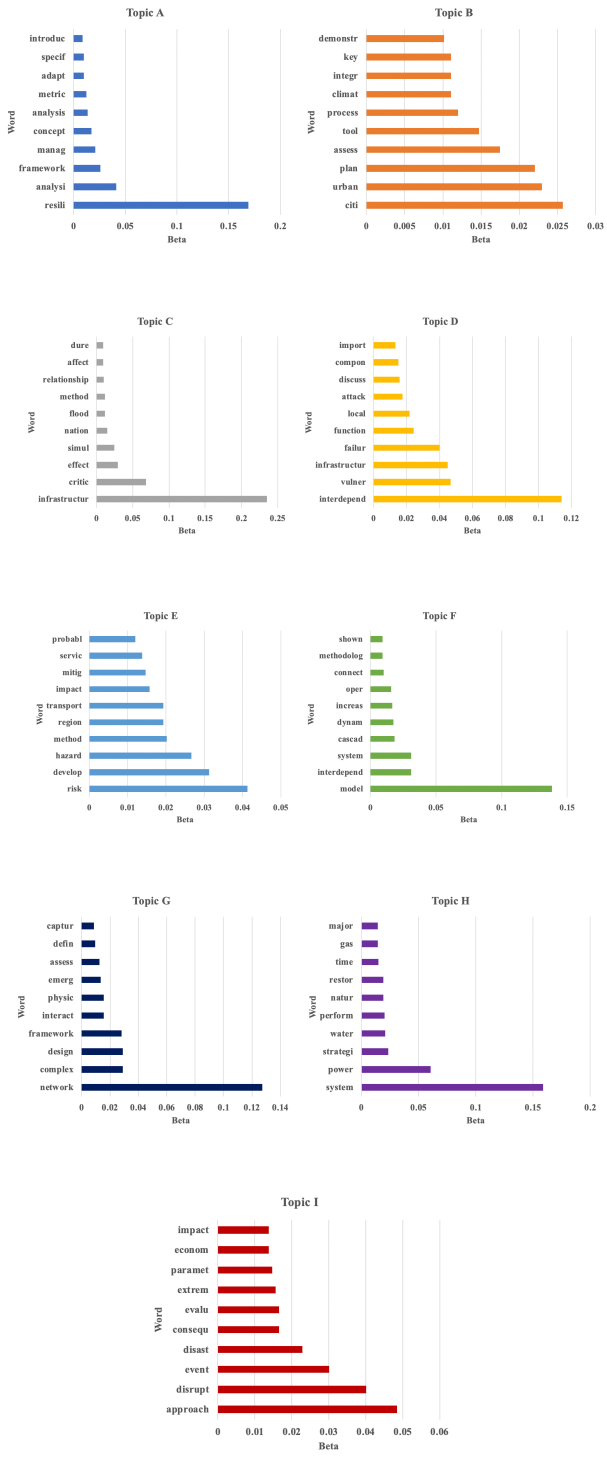


Figure 2-6: *Beta* Distribution for Extracted Topics

Topic A

Topic B

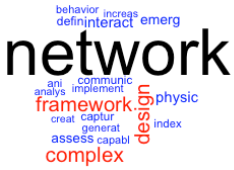
Topic C

Topic D

Topic E

Topic F

Topic G



Topic H

Topic I

Figure 2-7: Word Clouds for Extracted Topics

2.3. CITY/COMPLEX SYSTEMS RESILIENCE PREVIOUS RESEARCH

The current section reviews some of the research studies that have been conducted together with a critical analysis of each of the identified topics. The chosen publications (among the 124 modelled ones) are the ones that significantly add to the analysis of the identified topics. It is worth mentioning that the identified topics are titled hereafter as per **Figure 2-8**.

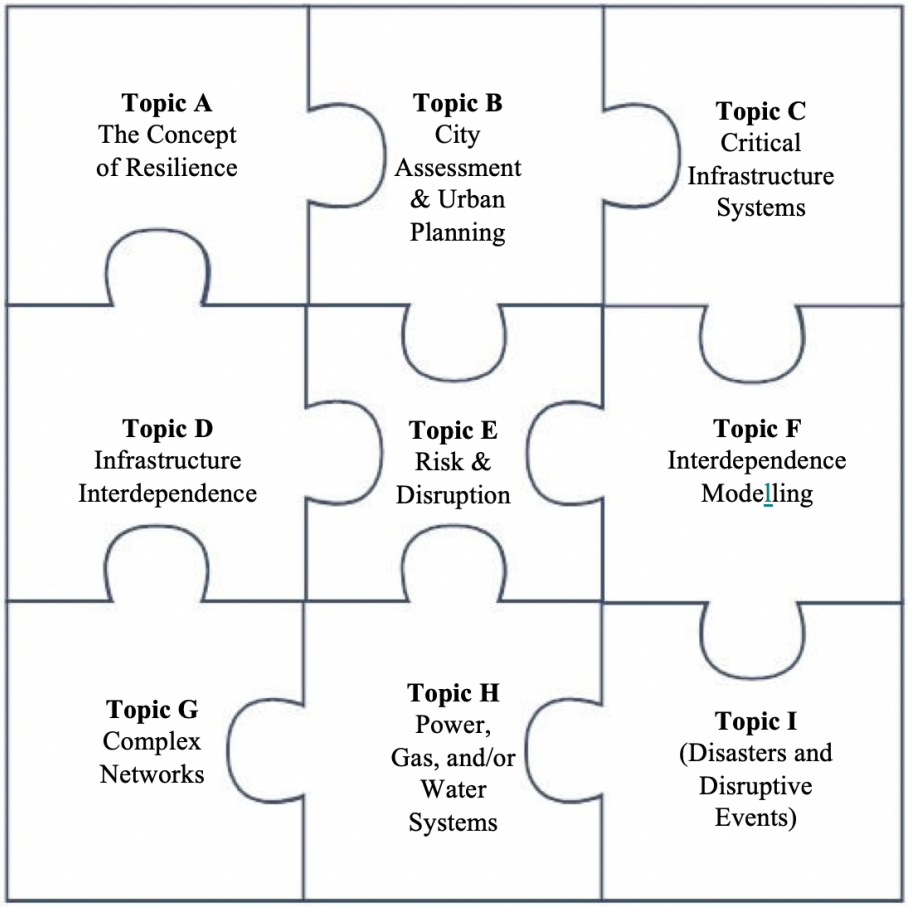


Figure 2-8: Extracted Topics within City/Complex Systems Resilience Field

2.3.1. TOPIC A: THE CONCEPT OF RESILIENCE

Resilience is related to both the time it takes the system to bounce back after a shock and the magnitude of disturbance a system can absorb and yet remain within its "*critical thresholds*" (Davoudi et al. 2012). Reduction of recovery time (Bruneau et al. 2003) was proposed as a resilience assessment measure which was further divided into three distinct processes which are recovery planning, execution and closure (Sharma, Tabandeh, and Gardoni 2017). The main goal of having a resilient city/system is not to prevent the hazard realization per se, but to enhance the performance of the city/system when such hazard materializes (About Us | 100 Resilient Cities n.d.).

RESILIENCE METRICS

Quantifying system resilience is a challenging and controversial subject due to the availability of many measures and metrics which claim to capture the full definition of system resilience. Resilience metrics fall under one of three approaches including those related to: (1) system properties (i.e., the four dimensions of resilience) (Bruneau et al. 2003); (2) system behaviour (i.e., the three resilience capabilities) (Ouyang 2017); and (3) other system requirements/characteristics (Hamida, Amine, and Mostafa 2016; Martin and Ludek 2013; Slivkova et al. 2017). **Figure 2-9** shows the three approaches together with the interaction between the different metrics.

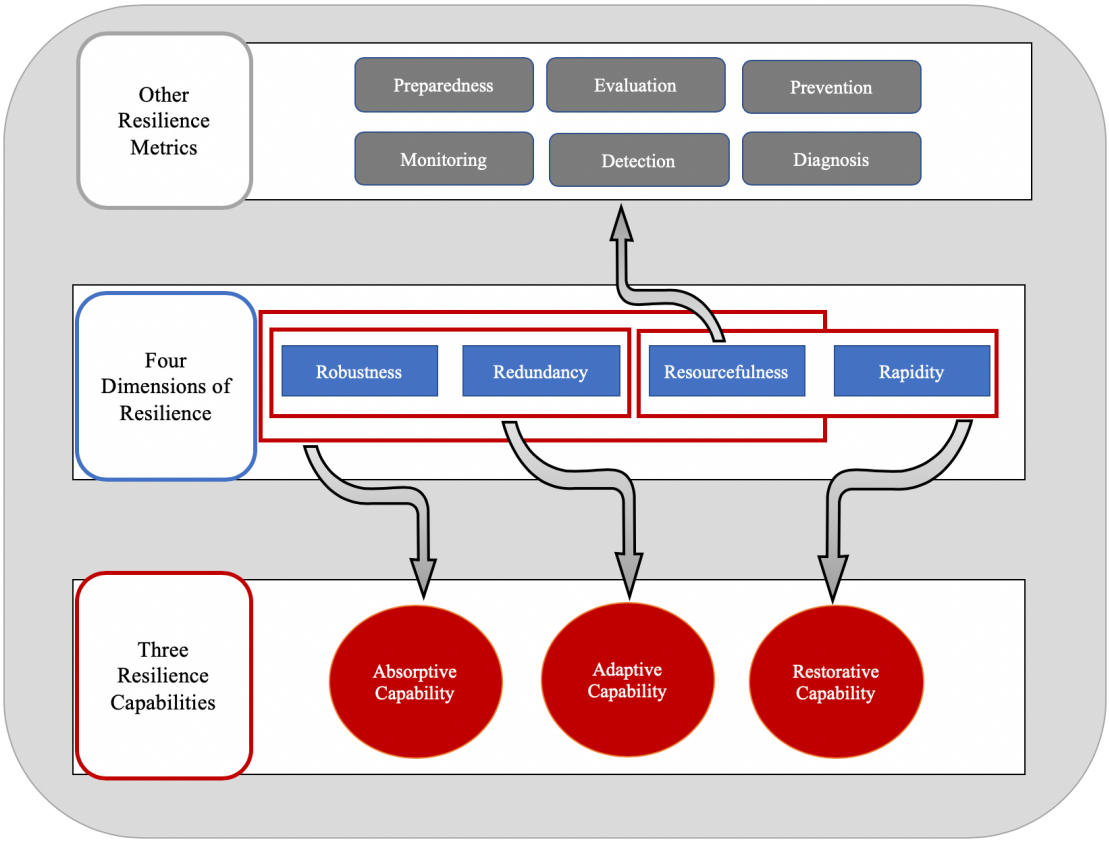


Figure 2-9: The Different Metrics of Resilience (Bruneau et al. 2003; Hamida, Amine, and Mostafa 2016; Martin and Ludek 2013; Ouyang 2017; Slivkova et al. 2017)

Approach I: The Four Dimensions of Resilience

Several studies have been conducted to assess resilience of different systems in terms of four dimensions, each is related to a certain system property. The four dimensions (Bruneau et al. 2003) are: (1) Robustness: the ability of a system to remain operational after an extreme event; (2) Redundancy: the availability of replacement components within a system; (3) Resourcefulness: the availability of resources that can help in detecting, diagnosing and surviving an extreme event; and (4) Rapidity: the time it takes a system to bounce back to an initial, or other predefined, state following an extreme event (i.e., recovery time). Robustness and rapidity can both be considered as resilience *goals* which are achieved through improving the two resilience *means*: redundancy and resourcefulness

(Rose and Krausmann 2013). As such, increasing the number of alternative (redundant) components as well as ensuring resourcefulness is key to alleviate the impact of adverse events by enhancing the system's immediate response (robustness) and functionality recovery (rapidity) following these events. Both goals can be measured by assessing system functionality at time of crisis when it comes to robustness and at full operation after crisis when it comes to rapidity. The complexity of quantifying resilience in terms of the two means resides in the fact that they are hard to directly relate to the loss of functionality of the system.

As shown in **Figure 2-10**, following a hazard realization at time t_h, the system recovers its functionality from just before t_h to t_{nr} and the functionality loss area (i.e., reflected by the red and blue areas together) is used to represent the system's loss of functionality which in turn illustrates how non-resilient or vulnerable the system is. As this area is minimized (e.g., to the red area only), the system is considered to be more resilient as it is able to retain almost 70% of its functionality (compared to the previous 30%) and is also able to recover at an earlier time t_r instead of t_{nr}. The concept of a system being *resilient-by-design* aims at minimizing the loss of functionality area which can be easily achieved by maximizing the system robustness and/or minimizing the rapidity.

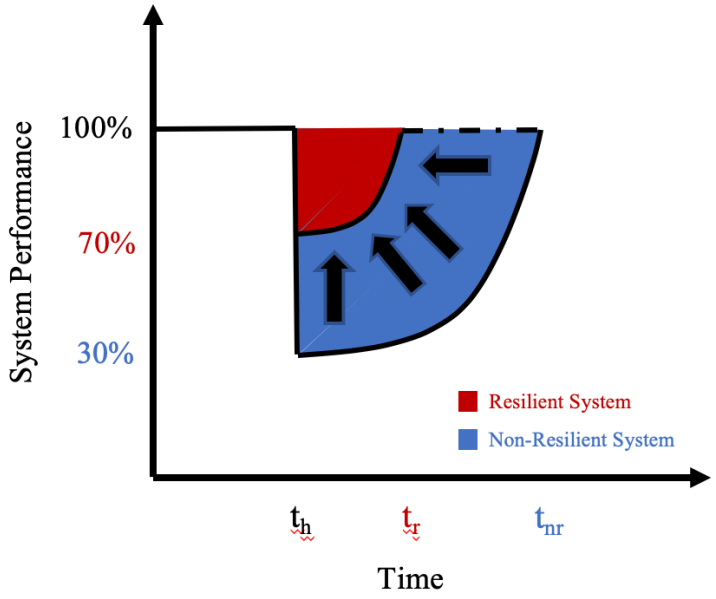


Figure 2-10: Definition of System Resilience (Shen and Tang 2015)

Consequently, resilience of environmental, physical and social systems was qualitatively defined through the aforementioned four dimensions based on the reduction of three system related measures: (1) adverse consequences following a hazard; (2) failure probability when the hazard materializes; and (3) required recovery time after a hazard, reduction (Bruneau et al. 2003). Moreover, resilience of critical infrastructure systems was quantified in terms of recovery time (i.e., rapidity) (Shen and Tang 2015), where it was measured as the ratio between the area under the actual system performance curve and that under the targeted performance curve. Furthermore, two functionality metrics were proposed for measuring the robustness of transportation systems which are topological and traffic related metrics (W. Sun, Bocchini, and Davison 2018), whereas the importance of robustness to resilience evaluation was emphasized (Martin and Ludek 2013) where component redundancy together with other system related metrics including the latter's (i.e., topology,

complexity, technologies, flexibility, and geography) were proposed to evaluate infrastructure systems robustness. Assuming that robustness can be solely used for resilience quantification, the resilience of infrastructure systems under spatially localized attacks was assessed in terms of system performance level following a hazard and before any restoration is carried out (Ouyang 2017). Furthermore, the distinction between resilience, robustness and rapidity concepts was presented by the "*resilience triangle*" concept in (Chang et al. 2014). Moreover, it was highlighted that system maintenance, design, capacity, planning initiative, redundancy and learning abilities are directly related to enhancing both robustness and rapidity (McDaniels et al. 2008). A framework was also developed to include the two dimensions of resilience goals (i.e., robustness and rapidity) (McDaniels et al. 2008) in which flow diagrams were used to evaluate potential decisions that can strengthen the two dimensions within infrastructure systems.

APPROACH II: THE THREE RESILIENCE CAPABILITIES

The second resilience metrics quantification approach relates system resilience to its behaviour, whereby three capabilities were suggested (Ouyang 2017) to measure system resilience as shown in Figure 2-9 as: (1) Absorptive Capability: the ability of a system to minimize the consequences of an extreme event, which can be linked to the robustness of the underlying system and can be enhanced by increasing resilience means (i.e. redundancy or resourcefulness); (2) Adaptive Capability: the ability of a system to adapt (i.e., reorganize) itself after an extreme event to minimize the corresponding consequences, which can also be linked to the two resilience means; and (3) Restorative Capability: the ability of a system to be repaired after an extreme event, which can be directly related both resourcefulness as a resilience mean and rapidity as a resilience goal (Rose and Krausmann

2013). As can be observed from the definitions of the three resilience capabilities, they can be related to the system behavior rather than its inherent properties. Nevertheless, these capabilities can be related to the previously established dimensions of resilience (i.e., goals and means). Based on these three capabilities, different frameworks for quantifying and evaluating system resilience were proposed (Francis and Bekera 2014; Ouyang 2017; Zhao, Liu, and Zhuo 2017).

APPROACH III: OTHER RESILIENCE METRICS

In addition to the previously discussed approaches for resilience metrics/quantification, frameworks were presented to assess the resilience of critical infrastructure systems based on preparedness and responsiveness (Slivkova et al. 2017) and protection, monitoring, maintenance and evaluation (Martin and Ludek 2013). Furthermore, a two-stages strategy for resilience evaluation was outlined (Hamida, Amine, and Mostafa 2016), where Stage I includes defence, detection, remediation and recovery, while Stage II includes cause diagnosis and future refinement.

2.3.2. TOPIC B: CITY ASSESSMENT AND URBAN PLANNING

Several frameworks were developed to define, analyze and assess resilience of cities. The Resilient Cities Framework (Rockefeller Foundation; Arup 2014) highlighted four city elements, twelve performance indicators and seven qualities for resilient cities as per **Figure 2-11**. The seven qualities that were demonstrated to be consistent with all resilient cities are: (1) Reflectiveness: the ability of individuals, communities and institution to build on past experiences (i.e., shared history) that can affect their current and future decisions; (2) Resourcefulness; (3) Robustness; (4) Redundancy; (5) Flexibility: the ability of a

system to adapt in the face of an extreme event; (6) Inclusiveness: the ability of a system to engage different experiences and visions to build or plan for city resilience; and (7) Integration: the ability of a system to bring together resources and share them across various systems. Actions to build resilience within cities were also identified (Bruneau et al. 2003; Chang et al. 2014; Desouza and Flanery 2013; Martin and Ludek 2013; Shen and Tang 2015) as per Figure 2-11. Acknowledging the importance of physical infrastructure systems for city resilience, a framework was proposed which specifies core systems for ensuring city resilience as: green infrastructure, transportation, energy, communication, water, sanitation and buildings (Reiner and McElvaney 2017). Another framework (Yang et al. 2018) provided a way for multisystem asset management to enhance city resilience.

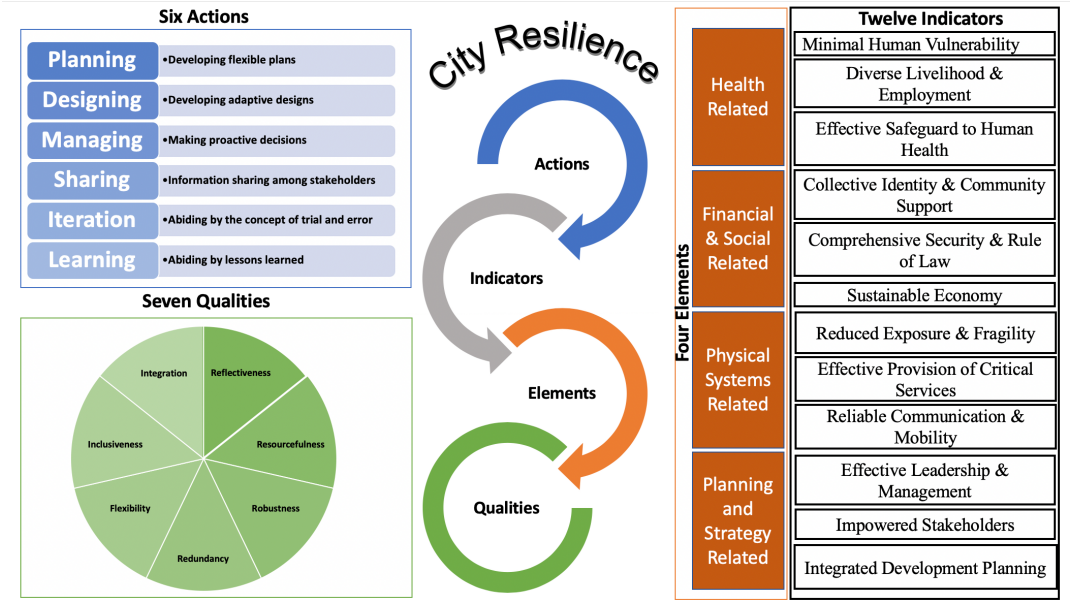


Figure 2-11: Actions for City Resilience (Bruneau et al. 2003; Chang et al. 2014; Desouza and Flanery 2013; Martin and Ludek 2013; Shen and Tang 2015)

2.3.3. TOPIC C: CRITICAL INFRASTRUCTURE SYSTEMS

The importance of designing a resilient critical infrastructure system was comprehensively discussed (Hudson, S., Cormie 2012), whereas resilient infrastructure rating systems and design tools were presented (Pitilakis et al. 2016) and a cycle for infrastructure resilience management was proposed (Yang et al. 2019). An approach was further introduced (Ouyang and Dueñas-Osorio 2012) to assess the time-dependent resilience of infrastructure systems where three values of resilience were proposed at different times to account for the fact that infrastructure systems are continuously evolving. Furthermore, (Timashev 2015) defined infrastructure resilience in probabilistic terms to represent the random parameters of resilience. The use of the three branches of data analytics (i.e., descriptive, predictive and prescriptive) in infrastructure systems resilience enhancement was also investigated (Kash Barker et al. 2017).

The influence of climate-induced risks on infrastructure systems was also explored (Giordano 2012). More specifically, infrastructure planning processes were reviewed to ensure that these systems can adapt to climate-induced hazards (Giordano 2012). Moving from system resilience to system protection, a methodology was presented to evaluate critical infrastructure systems in which protection of critical infrastructures was considered to depend on acceptance, anticipation, and planning (Robert et al. 2015). Furthermore, different modelling tools for infrastructure systems protection were evaluated (Santella, Steinberg, and Parks 2009).

2.3.4. TOPIC D: INFRASTRUCTURE INTERDEPENDENCE

To address interdependencies of infrastructure systems, researchers outlined several interdependence types. Four types of interdependence, among infrastructure systems, were specified (Gillette et al. 2002; Rinaldi, Peerenboom, and Kelly 2001):

(1) Physical: when interdependence is due to reliance on material flow between two or more systems;

(2) Cyber: when interdependence is due to reliance on information transfer between two or more systems;

(3) Geographic: when interdependence is due to proximity of two or more systems; and

(4) Logical: when interdependence is due to other factors that do not fall into the three above categories.

Whereas, other five types of infrastructure interdependence were presented (Lee, Mitchell, and Wallace 2007; Wallace et al. 2003):

(1) Input Dependence: when an infrastructure requires input from another infrastructure;

(2) Mutual Dependence: when at least one activity in an infrastructure is dependent upon another activity from another infrastructure, while at least one activity in the later infrastructure is dependent upon another activity from the former infrastructure;

(3) Shared Dependence: when physical components or activities are shared between infrastructures;

(4) Exclusive/Or Dependence: when an infrastructure is unable to operate once another infrastructure is operating; and

(5) Collocated Dependence: when components of two or more infrastructures are in the same location.

Relationships can be established between the four and five types of interdependence discussed above. The physical and cyber dependences which entitle system dependence of material or non-material components can be closely related to input and mutual dependence, whereas the geographic and collocated dependence are both related to proximity of system components locations. Finally, both shared and exclusive are logical types of systems dependence.

2.3.5. TOPIC E: RISK AND DISRUPTION

Most studies, among the considered 124 publications, do not relate the concepts of risk and resilience. In the context of infrastructure systems, risk assessment depends on the threat level of the hazard (i.e., probability) realization, the consequences of such a hazard realization on the system (i.e., impacts) and the vulnerability of the exposed systems/components to a specific hazard (Linkov et al. 2014), whereas resilience focuses on achieving system adaptation and recovery goals through working on both redundancy and resourcefulness as means. One way of looking at the resilience's relationship to risk is through considering that the former builds on the latter. The reason is that beyond quantifying system functionality loss or robustness through traditional risk analysis, resilience considers the temporal dimension of restoring system's functionality (rapidity), whereas risk focuses on the immediate system functionality loss through assessing

disruptive event threat level, related consequences and system vulnerabilities (Salem et al. n.d.).

Considering the previously established definition of risk, a framework was proposed to present a risk informed decision-making approach for the lifecycle performance of infrastructure systems (Lounis and McAllister 2016). Focusing on the consequence component of risk, a method was proposed (Bristow and Hay 2017) to estimate probabilities after a shock to derive a model that can assess risk consequences and treatment options. Focusing on the vulnerability component of risk, vulnerability and risk assessment frameworks for transportation systems were presented (Blockley, Agarwal, and Godfrey 2012; Pitilakis et al. 2016). Furthermore, it was emphasized that when considering the resilience of critical infrastructure systems, interdependence actually signifies the highest risk (Risk and Critical Infrastructure System Protection 2917). This was illustrated by highlighting the effect of the breakdown of a natural gas transmission line on power system generation plants in a southern state in USA. To study these risks and in aim of having a resilient infrastructure system, an algorithm was developed to simulate the effect of the worst-case failure scenario on the power grid network.

2.3.6. TOPIC F: INTERDEPENDENCE MODELLING

Topic F presents research conducted on simulating interdependence among infrastructure systems. Existing approaches for simulating complex infrastructure systems were divided into four main categories: multi agent-based, system dynamics, economic theory, and complex network theory. Three of these approaches (i.e., multi agent-based, system dynamics, and economic theory) are outlined in the current subsection. The fourth approach (i.e., complex network theory) is thoroughly detailed in the next subsection (i.e., *Topic G*).

Table 2-1 lists the techniques together with their applications and reference for *Topics F, G, and H.*

Table 2-1: Complex Systems Resilience and Interdependence Modelling Techniques

Application	Reference
Multi-Agent Simulation	
Conceptual simulation of infrastructure interdependence	(Dudenhoeffer, Permann, and Manic 2006)
Modelling interdependence among power, water and wastewater systems	(Pereyra, He, and Mostafavi 2016)
System Dynamics	
Employing mitigation strategies to increase transportation system resilience	(Croope and McNe 2011)
Building transportation, energy and telecommunication system maps from defining system relations	(Cavallini et al. 2014)
Input-Output Inoperability Model	
Characterizing interdependence for power and telecommunication systems	(Reed, Kapur, and Christie 2009)
Probabilistic assessment for the resilience of interdependent infrastructure	(Xu et al. 2013)
Uncertainty recovery analysis for two hypothetical interdependent systems	(Xu et al. 2015)
Uncertainty recovery analysis for transportation, utilities, construction, manufacturing and mining	(K Barker and Haimes 2009)
Complex Network Theory	
Resilience assessment for high-pressure natural gas system	(Golara and Esmaeily 2017)
Resilience and reliability assessment of transportation system	(Lam and Tai 2012)
Analyses of vulnerabilities and resilience of railway system	(Chopra et al. 2016)

Evaluate performance of energy, water and wastewater systems	(Holden et al. 2013)
Assessment of the effect of interdependence on resilience of power, water and telecommunication systems	(Mao and Li 2017)
Simulation of power distribution, gas pipeline and telephony transport	(Svendsen and Wolthusen 2007)
Simulation of the vulnerability and resilience of power and water systems	(Zhang, Yang, and Lall 2016)
Modelling edge attack strategies on power and gas systems	(Wang et al. 2013)
Studying the behaviour of power and water systems at the time of flood	(Val, Holden, and Nodwell 2014)
Characterizing infrastructure interdependence under the power outages	(McDaniels et al. 2008)
Other Techniques	
Ranking the IEEE 30 bus system components	(Fang, Pedroni, and Zio 2016)
Evaluating restoration and planning of high-voltage power transmission lines	(Fang and Sansavini 2017)
Simulating interdependence within power systems	(Rahman et al. 2008)
Assessment of interdependence in Taiwan's Northern region power systems	(Chou, Tseng, and Ho 2009)
Maximizing the global connectivity of two interdependent power systems	(Chen and Zhu 2016)
Resilience evaluation of power and gas systems	(Liu, Ferrario, and Zio 2017)
Power and gas interdependence enhancement using microgrids	(Yodo and Arfin 2020)
Quantifying the performance of power and water systems after an earthquake	(Omidvar, Malekshah, and Omidvar 2014)
Resilience assessment for power and water systems	(Ulieru 2007)

Assessing the effect of spatially localized attacks on power and water systems	(Ouyang 2017)
Assessing the effect of spatially localized attacks on power and gas systems	(Ouyang 2016)
Analyzing joint restoration processes for power and gas systems resilience assessment	(Ouyang and Wang 2015)
Resilience evaluation for power, water and gas systems using restoration measures	(Cimellaro, Solari, and Bruneau 2014)
Power, water and gas systems interdependence modelling using Bayesian networks	(Johansen and Tien 2018)

MULTI-AGENT-BASED SIMULATION

Multi-agent-based simulation approach is associated with a relatively recent style of programming called "*object oriented programming*", in which programming languages are *"encapsulated"* in objects that can control their own behavior, and interact with other objects (Rouse n.d.). The *agent-level* represents the basic elements or the physical components that make up the entire system. The next level is the *system-level*, which is made up of aggregating the *agent-level*. Finally, the *system-of-system-level* integrates all systems to represent the highest and more sophisticated systemic level (Pereyra, He, and Mostafavi 2016). Using this approach, a framework was proposed for the simulation of infrastructure components (Dudenhoeffer, Permann, and Manic 2006) and to simulate the components of power, water and wastewater systems (Pereyra, He, and Mostafavi 2016).

SYSTEM DYNAMICS

Control theory is an interdisciplinary branch of engineering and mathematics that deals with the control of continuously operating dynamic processes and machine systems, and how their behaviour is modified by feedback loops (Teknomo 2004). System dynamics

simulation approach extends the toolbox of the control theory from machines to systems (How does system dynamics relate to control theory? - Quora n.d.). More specifically, system dynamics—originally developed in the 1950s, is based on the fact that any system can rely on circular, interlocking and time-delayed relationships among its comprising components (Croope and McNe 2011). This approach was adopted to develop a decision support system that aims to reduce the vulnerability of transportation (Croope and McNe 2011) and energy/telecommunication (Cavallini et al. 2014) systems.

ECONOMIC THEORY (INPUT-OUTPUT INOPERABILITY MODEL)

The input-output model was first proposed by Wassily Leiontief in 1973 to describe the equilibrium between economic sectors through modelling the interconnections among these sectors (Haimes et al. 2005). The model was applied to assess the behaviour of power and telecommunication systems under natural hazards (Reed, Kapur, and Christie 2009). The static input output inoperability model is presented in Equation 2.1 (Reed, Kapur, and Christie 2009).

$$x_i = Ax_j + C \qquad \textbf{Equation 2-1}$$

where x_i and x_j are vectors of systems' "*i*" and "*j*" inoperability, respectively, A is the interdependence matrix between the different subsystems and C is a disturbance or perturbation vector. The values of matrix A coefficients range from 0 to 1, where a_{ij} represents the probability of inoperability that system "*j*" contributes to system "*i*". When the failure of system "*j*" leads to 100% failure in system "*i*", a_{ij} is equal to 1. Conversely, a_{ij} is equal to 0 if the failure of system "*j*" has no effect on system "*i*". Finally, C represents the direct reduction of functionality resulting from exposure to the hazard.

The concept of the static evaluation (i.e., only at a given time) was further developed to propose a dynamic input-output inoperability model, where a continuous evaluation (i.e., overtime) is permitted (Miller and Blair 1985). The dynamic input out[put inoperability model is presented in Equation 2.2 (Miller and Blair 1985).

$$q(t) = K[A^* q(t) + c^*(t) - q(t)] \qquad \textbf{Equation 2-2}$$

Where, $q(t)$, A^*, and $c^*(t)$ are defined as *X*, *A* and *C* in the static input-output inoperability model, respectively. The difference between the static and dynamic inoperability models is that the variables x_i, x_j and C in the static model are modelled as $x(t)_i, x(t)_j$ and $C(t)$ to account for the time dimension. On the other hand, K is a resilience coefficient that represents the system's ability to recover following a disruption where the higher the value of K, the better the system is with respect to its response. Based on the dynamic inoperability input-output model, frameworks for resilience (Xu et al. 2013) and recovery (K Barker and Haimes 2009; Xu et al. 2015) assessment of interdependent infrastructure systems were developed.

2.3.7. TOPIC G: COMPLEX NETWORK THEORY

Thriving as of the first decade of the 21st century, Complex Network Theory (CNT) builds upon the use of the mathematical graph theory (Barabasi 2016). In CNT, a system is typically simulated by nodes and links; each has unique attributes based on the underlying application. Most CNT based models fall under one of two main types as per **Figure 2-12**. The first type is the physical network-based model, in which the relationships between different network components is based on a system architecture property (i.e., their geographical proximity). The second type is the functional network-based model, in which

the rate of flow of system commodities (i.e., water, power, etc.) derives the relationships between network components. In the functional network-based model, the capacity of each component in the network is of key importance.

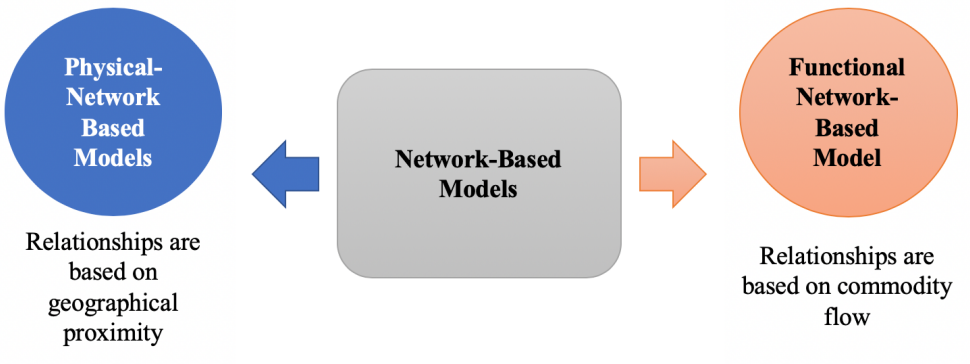


Figure 2-12: Types of Network-based Models

The resilience of transportation systems was evaluated using CNT (Lam and Tai 2012) where the network's ability to maintain basic functions under an extreme event was determined using *"the weighted sum of the resilience of all the reliable independent connection paths between all pairs of nodes."* A framework was also developed to investigate the topology, spatial organization and passenger flow of metro networks (Chopra et al. 2016). Similarly, electrified railway systems were simulated to consider structural, functional and geographical system properties (Johansson and Hassel 2010). In addition the resilience of high pressure natural gas networks was evaluated using CNT (Golara and Esmaeily 2017).

Recently, CNT was used to simulate interdependence among infrastructure systems. Physical and functional network-based models were adopted and used to simulate the interdependence of power, water and telecommunication systems where it was shown that interdependence between the three systems minimize their resilience (Mao and Li

2017). Likewise, CNT flow models were proposed (Holden et al. 2013) to evaluate the performance of energy, water and wastewater systems at times of flood.

2.3.8. TOPIC H: POWER, WATER AND/OR GAS SYSTEMS

The reliance of infrastructure systems on power was clearly illustrated in the consequences of many disasters including the 2003 North-eastern blackout mentioned earlier. As such, among the keywords in *Topic H*, power systems have the highest probability of occurrence which is about three times that of water or gas systems. Thus, focusing on power systems, empirical approaches were used to develop a framework to evaluate how power outages, due to extreme events, can lead to failures in other interdependent infrastructure systems (McDaniels et al. 2007). Furthermore, ranking among power system components was performed based on the optimal repair time by considering the IEEE bus 30 components (Fang, Pedroni, and Zio 2016). Focusing on transmission lines, a credibility-based fuzzy mixed integer programming approach was proposed to evaluate the restoration and planning of high-voltage power transmission lines (Fang and Sansavini 2017). Adding interdependence to the picture, a matrix-based technique (Rahman et al. 2008), a knowledge discovery process (Chou, Tseng, and Ho 2009), and game theoretic approach (Chen and Zhu 2016) were used to simulate interdependence among power systems.

As per the topic analysis results, the three critical infrastructure systems that were considerably tackled together in the literature are the power, water and gas systems. This linkage may indicate the importance of interdependence between these systems. In that sense, the vulnerability of power and water systems was evaluated (Zhang, Yang, and Lall 2016). In addition, CNT was used to simulate power, gas and telephony transport (Svendsen and Wolthusen 2007), power and gas (Wang et al. 2013), and power and water

systems (Val, Holden, and Nodwell 2014). The resilience of power and gas systems was also assessed using differential equations (Liu, Ferrario, and Zio 2017) and the effect of installing microgrids on the interdependence of both systems was investigated using complex network theory (Yodo and Arfin 2020). Similarly, the performance of power and water systems following an earthquake was quantified using a Markov chain approach (Omidvar, Malekshah, and Omidvar 2014), whereas a probabilistic procedure was proposed to assess their resilience (Ulieru 2007). Additionally, the effect of spatially localized attacks which were found to have the ability to cause *"direct damage or interruption of system components that exist within some localized area while those outside this area remain operating"* (Ouyang 2017) on power and water systems, and power and gas systems was presented in (Ouyang 2017) and (Ouyang 2016), respectively. By focusing on joint restoration processes a framework was used to assess the resilience of power and gas systems (Ouyang and Wang 2015). Furthermore, resilience of power, water and gas systems was evaluated by quantifying restoration measures under natural disasters (Cimellaro, Solari, and Bruneau 2014), whereas the interdependence among the three systems was modeled using a Bayesian approach (Johansen and Tien 2018).

2.4. Discussion and Research Opportunities

This section aims to explore research gaps pertaining to the resilience of complex systems in an attempt to create blue ocean research opportunities (i.e., opportunities within unexplored research areas) (Ezzeldin M and El-Dakhakhni W. 2019). Prior to uncovering these gaps, the per-document-per-topic probability (gamma) values are evaluated to assess the contribution of each topic to the content of the publications used in the proposed LDA topic model. The gamma probabilities are summed up for each topic over the analyzed documents and are shown in **Figure 2-13**. *Topic C* (Critical Infrastructure Systems), *Topic D* (Infrastructure Interdependence), and *Topic H* (Power, and/or Water, and/or Gas Systems) have the highest sum of gamma for the publications considered which indicates that these topics contribute more with respect to the overall content of the considered publications.

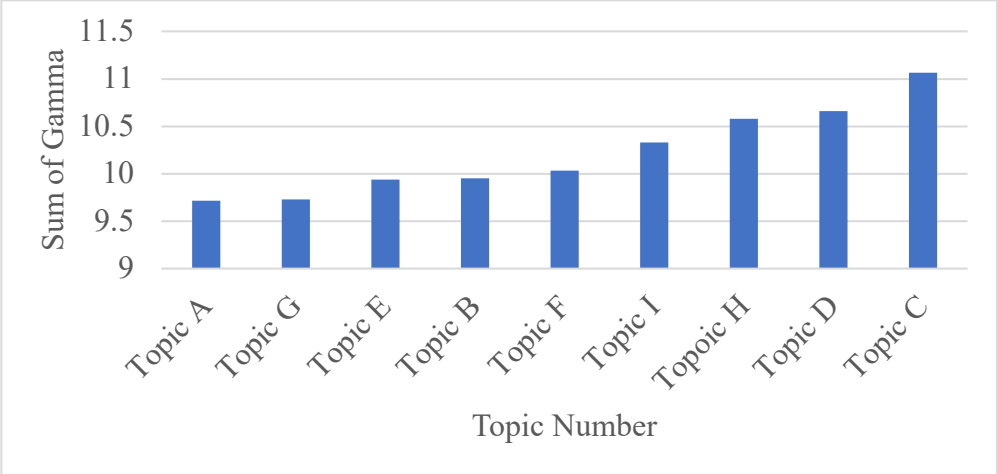


Figure 2-13: Summation of the Per-Document-Per-Topic Contribution for all Topics

Figure 2-14 summarizes the five gaps uncovered based on the current analysis. These gaps were identified based on a qualitative assessment for the identified topics (i.e., shown in

Figure 2-13) that constitute the current state-of-the-art. It is worth mentioning that although some topics have (quantitatively) high contributions to the considered documents, gaps still remain (qualitatively) in terms of the research breadth/depth within these topics. Therefore, the current discussion of the knowledge/research gaps is a key complement to the topic contribution analysis discussed earlier.

The *first* research gap is related to resilience quantification in complex systems which pertains to *Topic A*. Several resilience metrics were previously proposed including the four dimensions of resilience and the three resilience capabilities, as well as others. Nevertheless, previous research studies either failed to quantify some of these metrics or were not able to reach a consensus on how to quantify them. Specifically, resilience was either quantified based on repair time, repair cost or loss of system functionality. However, to have a comprehensive resilience metric, in addition to repair time (i.e., rapidity), repair cost, or loss of functionality (i.e., robustness), other identified resilience metrics have to be quantified. These measures include, to name but a few examples, the availability of redundant components in the system (i.e., redundancy), the availability of resources in the system that can help diagnose and control system failure (i.e., resourcefulness), the ability of a system to be structured yet adaptable under extreme events (i.e., system adaptability), the ability of the system to activate its comprising components to restore its functionality (i.e., responsiveness), the number of people affected by system's lack of resilience.

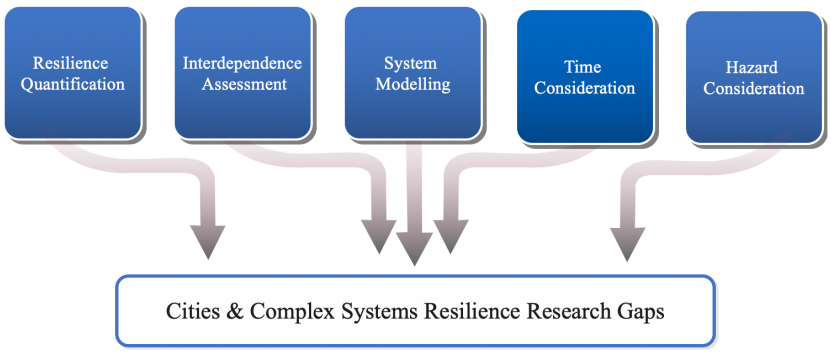


Figure 2-14: City/Complex Systems Resilience Research Gaps

The *second* research gap exists in defining interdependence between different systems which relates to *Topic D*. Given the complexity of infrastructure systems and their interrelated components, there exists a gap in quantifying the defined interdependence types. Further research is critically needed to quantify all types of interdependence between modelled systems to have a comprehensive system-of-systems model that can evaluate the behaviour of complex systems as close to real life as possible (e.g., city's digital twin). This comprehensive system-of-systems model would face three key challenges, (1) data acquisition; (2) computational ability for modelling; and (3) availability of accurate metrics for quantification.

The *third* research gap is related to modelling of real-life systems which is associated with *Topic F*. Most previous research studies simulated either small parts of existing systems without being able to capture the behaviour of the entire system or hypothetical systems (i.e., IEEE power systems). As mentioned before, the unavailability of data and the limited computation ability hinder the future development required to model entire real-life systems. Nevertheless, given the current availability of data worldwide and given the high computation ability reached, it is expected that future research would have the

sufficient resources to tackle such systems. The *fourth* research gap pertains also to *Topics F* and *G* and is related to considering the effect of time (i.e., dynamic) when modelling complex systems. Future research should focus more on the dynamic behaviour of systems to mimic the behaviour of real-life systems.

The *fifth* research gap exists in linking hazards which are the main cause behind system disruption to the performance of infrastructure systems. This gap can be related to *Topic I* which is the cross-cutting topic related to disasters and subsequent disruption of infrastructure systems. Given the fact that ensuring system's resilience is key to protect the system against damages and adverse consequences of future disasters, there is a pressing need to predict the occurrence of these events, and thus, prepare the system accordingly which in return can optimize relevant system's resources and significantly enhance its resilience capabilities. This is especially important given the fact that almost quarter of the world's population is officially threatened by storm surges and tsunamis (Climate Change - Oxfam Canada n.d.), and that, since the 1990s, natural disasters have affected 217 million people annually with damages of more than $1.2 trillion from 2001 to 2010 compared to $528 billion from 1981 to 1990 (Are Natural Disasters Increasing? n.d.).

2.5. CONCLUSIONS

It is estimated that more than 50% of the world population will be living in cities by the year 2050. These cities are currently experiencing rapid transformations due to the unforeseen extreme events that disturb their basic functions. However, only a few studies focused on comprehensively simulating the infrastructure systems of these cities or

assessing their corresponding resilience. The current study utilized a meta-research approach which aims at quantitatively and qualitatively assessing previous work pertaining to the resilience of critical infrastructure systems. Using topic modelling, nine common topics were identified from 124 research publications. Such topics are: the concept of resilience, city assessment and urban planning, critical infrastructure systems, infrastructure interdependence, risk and disruption, complex systems modelling, complex network theory, power, gas and/or water systems, and disasters and system disruption which is the trigger behind studying complex systems and designing for their resilience. Following topic extraction, several complex systems resilience metrics along with simulation approaches were reviewed from previous studies. This paper also outlined existing research/knowledge gaps including those pertaining to quantifying resilience, quantifying interdependence, modelling of full/real life systems, incorporating the dynamic behaviour of complex systems, and linking adverse/extreme events to system performance. Finally, the current study is expected to guide future research thrusts by highlighting knowledge gaps and research opportunities to make a real impact on the existing field of resilient cities and their interdependent infrastructure systems.

2.6. ACKNOWLEDGEMENT

The authors are grateful to the financial support of the Ontario Trillium Scholarship Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors would also like to acknowledge the fruitful discussions with the research teams of the NSERC-CaNRisk-CREATE program, the INTERFACE Institute and the INViSiONLab, both of McMaster University.

2.7. REFERENCES

"About Us | 100 Resilient Cities." http://www.100resilientcities.org/about-us#/-_/ (November 10, 2017).

"Are Natural Disasters Increasing?" *The Borgen Project*. https://borgenproject.org/natural-disasters-increasing/ (October 28, 2019).

Barabasi, Albert-László. 2016. *Network Science*. http://networksciencebook.com/.

Barker, K, and Y Haimes. 2009. "Uncertainty Analysis of Interdependencies in Dynamic Infrastructure Recovery: Applications in Risk-Based Decision Making." *Journal of Infrastructure Systems* 15(4): 394.

Barker, Kash et al. 2017. "Defining Resilience Analytics for Interdependent Cyber-Physical-Social Networks." *Sustainable and Resilient Infrastructure* 2(2): 59–67. http://dx.doi.org/10.1080/23789689.2017.1294859.

Blei, D., Y. Andrew, and J. Micheal. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan): 993–1022.

Blockley, David, Jitendra Agarwal, and Patrick Godfrey. 2012. "Infrastructure Resilience for High-Impact Low-Chance Risks." *Proceedings of the Institution of Civil Engineers - Civil Engineering* 165(6): 13–19. http://www.icevirtuallibrary.com/doi/10.1680/cien.11.00046.

Bristow, David, and Alexander Hay. 2017. "Graph Model for Probabilistic Resilience and Recovery Planning of Multi-Infrastructure Systems." *Journal of Infrastructure Systems* 23(3): 04016039. http://ascelibrary.org/doi/10.1061/%28ASCE%29IS.1943-

555X.0000338.

Bruneau, M. et al. 2003. “A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities.” *Earthquake spectra* 19(4): 733–52.

Cavallini, Simona et al. 2014. “A System Dynamics Framework for Modeling Critical Infrastructure Resilience.” In *International Conference on Critical Infrastructure Protection*, , 141–54. http://link.springer.com/10.1007/978-3-662-45355-1_10.

Chang, S. et al. 2014. “Toward Disaster-resilient Cities: Characterizing Resilience of Infrastructure Systems with Expert Judgments.” *Risk Analysis* 34(3): 416–34.

Chen, Juntao, and Quanyan Zhu. 2016. “Interdependent Network Formation Games with an Application to Critical Infrastructures.” In *2016 American Control Conference (ACC)*, , 2870–75. http://arxiv.org/abs/1602.07745.

Chopra, S, T Dillon, M Bilec, and V Khanna. 2016. “A Network-Based Framework for Assessing Infrastructure Resilience: A Case Study of the London Metro System.” *Journal of the Royal Society Interface* 13(118). http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L610669898%0Ahttp://dx.doi.org/10.1098/rsif.2016.0113.

Chou, C, S Tseng, and T Ho. 2009. “Data Collection and Analysis of Critical Infrastructure Interdependency Relationships.” *Computing in Civil Engineering*: 280–89.

Cimellaro, G, D Solari, and M Bruneau. 2014. “Physical Infrastructure Interdependency and Regional Resilience Index after the 2011 Tohoku Earthquake in Japan.” *Journal*

of Earthquake Engineering and Structural Dynamics 43(6): 1763–84.

"Climate Change - Oxfam Canada." https://www.oxfam.ca/themes/water/ (November 10, 2018).

Croope, S, and S McNe. 2011. "Improving Resilience of Critical Infrastructure Systems Postdisaster." *Transportation Research Record* 2234(1): 3–13.

Davoudi, S. et al. 2012. "Resilience: A Bridging Concept or a Dead End?" *Planning theory & practice* 13(2): 299–333.

Desouza, K., and T. Flanery. 2013. "Designing, Planning, and Managing Resilient Cities: A Conceptual Framework." *Cities* 35: 85–99.

Dudenhoeffer, D, M Permann, and M Manic. 2006. "CIMS: A Framework for Infrastructure Interdependency Modeling and Analysis." In *Winter Simulation Conference*, , 478–85.

Ezzeldin M and El-Dakhakhni W. 2019. "Meta-Researching Structural Engineering: Trend Identification and Knowledge Gap Discovery Using Text Mining." *ASCE Journal of Structural Engineering (Accepted)*.

Fang, Y, N Pedroni, and E Zio. 2016. "Resilience-Based Component Importance Measures for Critical Infrastructure Network Systems." *IEEE Transactions on Reliability* 65(2): 502–12.

Fang, Y, and G Sansavini. 2017. "Optimum Post-Disruption Restoration for Enhanced Infrastructure Network Resilience: A Fuzzy Programming Approach." *Risk, Reliability and Safety: Innovating Theory and Practice - Proceedings of the 26th*

European Safety and Reliability Conference, ESREL 2016.

Francis, Royce, and Behailu Bekera. 2014. "A Metric and Frameworks for Resilience Analysis of Engineered and Infrastructure Systems." *Reliability Engineering and System Safety* 121: 90–103. http://dx.doi.org/10.1016/j.ress.2013.07.004.

Gatti, C., J. Brooks, and S. Nurre. 2015. "A Historical Analysis of the Field of OR/MS Using Topic Models." *arXiv preprint arXiv:1510.05154*. http://arxiv.org/abs/1510.05154.

Gillette, Jerry, Ronald Fisher, James Peerenboom, and Ronald Whitfield. 2002. "Analyzing Water/Wastewater Infrastructure Interdependencies." In *Analyzing Water/Wastewater Infrastructure Interdependencies (No. ANL/DIS/CP-107254). Argonne National Lab., IL (US).*, www.ipd.anl.gov/anlpubs/2002/03/42598.pdf.

Giordano, Thierry. 2012. "Adaptive Planning for Climate Resilient Long-Lived Infrastructures." *Utilities Policy* 23: 80–89. http://dx.doi.org/10.1016/j.jup.2012.07.001.

Golara, A., and A. Esmaeily. 2017. "Quantification and Enhancement of the Resilience of Infrastructure Networks." *Journal of Pipeline Systems Engineering and Practice* 8(1). http://ascelibrary.org/doi/10.1061/%28ASCE%29PS.1949-1204.0000250.

Griffiths, Thomas., Mark Steyvers, By Blei, and Jordan Blei. 2004. "Finding Scientific Topics." *Proceedings of the National academy of Sciences* 101(1): 5228–35. www.pnas.orgcgidoi10.1073pnas.0307752101.

Haimes, Y et al. 2005. "Inoperability Input-Output Model for Interdependent

Infrastructure Sectors. I: Theory and Methodology." *Journal of Infrastructure Systems* 11(2): 80–92.

Hamida, Yaou, Baina Amine, and Bellafkih Mostafa. 2016. "Toward Resilience Management in Critical Information Infrastructure." *Proceedings of the 2015 5th World Congress on Information and Communication Technologies, WICT 2015*: 101–6.

Holden, R, D Val, R Burkhard, and S Nodwell. 2013. "A Network Flow Model for Interdependent Infrastructures at the Local Scale." *Safety Science* 53: 51–60. http://dx.doi.org/10.1016/j.ssci.2012.08.013.

Holling, C. S. 1996. *Engineering Resilience versus Ecological Resilience*. Engineering within ecological constraints.

"How Does System Dynamics Relate to Control Theory? - Quora." https://www.quora.com/How-does-system-dynamics-relate-to-control-theory (January 1, 2019).

Hudson, S., Cormie, D. 2012. "Engineering Resilient Infrastructure." In *Institution of Civil Engineers - Civil Engineering*, , 5–12.

Johansen, Chloe, and Iris Tien. 2018. "Probabilistic Multi-Scale Modeling of Interdependencies between Critical Infrastructure Systems for Resilience." *Sustainable and Resilient Infrastructure* 3(1): 1–15. http://doi.org/10.1080/23789689.2017.1345253.

Johansson, Jonas, and Henrik Hassel. 2010. "An Approach for Modelling Interdependent

Infrastructures in the Context of Vulnerability Analysis." *Reliability Engineering and System Safety* 95(12): 1335–44. http://dx.doi.org/10.1016/j.ress.2010.06.010.

Labs, L. "How Does Perplexity Function in Natural Language Processing?" https://www.quora.com/How-does-perplexity-function-in-natural-language-processing (November 15, 2018).

Lam, C., and K. Tai. 2012. "Evaluating the Reliability of Infrastructure Networks by Resilience Analysis." In *IEEE International Conference on Industrial Engineering and Engineering Management*, , 1165–69.

Lee, E., J. Mitchell, and W Wallace. 2007. "Restoration of Services in Interdependent Infrastructure Systems: A Network Flows Approach." *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 37(6): 1303–17.

Linkov, I. et al. 2014. "Changing the Resilience Paradigm." *Nature Climate Change* 4(6): 407–9.

Liu, X., E. Ferrario, and E. Zio. 2017. "Resilience Analysis Framework for Interconnected Critical Infrastructures." *ASCE-ASME J. Risk and Uncert. in Engrg. Sys., Part B: Mech. Engrg.* 3(2): 021001. http://risk.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4035728.

Lounis, Zoubir, and Therese. McAllister. 2016. "Risk-Based Decision Making for Sustainable and Resilient Infrastructure Systems." *Journal of Structural Engineering* 142(9): F4016005. http://ascelibrary.org/doi/10.1061/%28ASCE%29ST.1943-541X.0001545.

Lu, P., and D. Stead. 2013. "Understanding the Notion of Resilience in Spatial Planning: A Case Study of Rotterdam, The Netherlands." *Cities* (35): 200–212.

Mao, Q., and N. Li. 2017. "Resilience Assessment of Interdependent Critical Infrastructure." In *Joint Conference on Computing in Construction*, , 7–10.

Martin, H, and L Ludek. 2013. "The Status and Importance of Robustness in the Process of Critical Infrastructure Resilience Evaluation." *2013 IEEE International Conference on Technologies for Homeland Security, HST 2013* (1): 589–94. http://www.scopus.com/inward/record.url?eid=2-s2.0-84893243320&partnerID=40&md5=f061cd428b9edbdff1cdc033ad1f1b26.

McDaniels, T. et al. 2007. "Empirical Framework for Characterizing Infrastructure Failure Interdependencies." *Journal of Infrastructure System* 13(3): 175–84.

McDaniels, T et al. 2008. "Fostering Resilience to Extreme Events within Infrastructure Systems: Characterizing Decision Contexts for Mitigation and Adaptation." *Global Environmental Change* 18(2): 310–18.

Miller, R, and P Blair. 1985. *Input-Output Analysis : Foundations and Extensions*. Cambridge university press.

Miner, G. et al. 2012. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Academic Press.

Moteff, John, and Pual Parfomak. 2004. Library of Congress Washington DC Congressional Research Service *Critical Infrastructure and Key Assets: Definition and Identification*.

Omidvar, Babak, Mohammad Malekshah, and Hamed Omidvar. 2014. "Failure Risk Assessment of Interdependent Infrastructures against Earthquake, a Petri Net Approach: Case Study-Power and Water Distribution Networks." *Natural Hazards* 71(3): 1971–93.

Ouyang, Min. 2016. "Critical Location Identification and Vulnerability Analysis of Interdependent Infrastructure Systems under Spatially Localized Attacks." *Reliability Engineering and System Safety* 154: 106–16. http://dx.doi.org/10.1016/j.ress.2016.05.007.

———. 2017. "A Mathematical Framework to Optimize Resilience of Interdependent Critical Infrastructure Systems under Spatially Localized Attacks." *European Journal of Operational Research* 262(3): 1072–84.

Ouyang, Min, and Leonardo Dueñas-Osorio. 2012. "Time-Dependent Resilience Assessment and Improvement of Urban Infrastructure Systems." *Chaos* 22(3).

Ouyang, Min, and Zhenghua Wang. 2015. "Resilience Assessment of Interdependent Infrastructure Systems: With a Focus on Joint Restoration Modeling and Analysis." *Reliability Engineering and System Safety* 141: 74–82. http://dx.doi.org/10.1016/j.ress.2015.03.011.

Pereyra, Jose, Xudong He, and Ali Mostafavi. 2016. "Multi-Agent Framework for the Complex Adaptive Modeling of Interdependent Critical Infrastructure Systems." *Construction Research Congress 2016*: 1556–66. http://ascelibrary.org/doi/10.1061/9780784479827.156.

Pitilakis, K., S. Argyroudis, K. Kakderi, and J. Selva. 2016. "Systemic Vulnerability and

Risk Assessment of Transportation Systems under Natural Hazards Towards More Resilient and Robust Infrastructures." *Transportation Research Procedia* 14: 1335–44. http://dx.doi.org/10.1016/j.trpro.2016.05.206.

Rahman, H, M Armstrong, D Mao, and J Martí. 2008. "I2Sim: A Matrix-Partition Based Framework for Critical Infrastructure Interdependencies Simulation." *2008 IEEE Electrical Power and Energy Conference - Energy Innovation*: 1–8.

Reed, D, K Kapur, and R Christie. 2009. "Methodology for Assessing the Resilience of Networked Infrastructure." *IEEE Systems Journal* 3(2): 174–80.

Reiner, Mark, and Lisa McElvaney. 2017. "Foundational Infrastructure Framework for City Resilience." *Sustainable and Resilient Infrastructure* 2(1): 1–7. http://dx.doi.org/10.1080/23789689.2017.1278994.

Rinaldi, S., J. Peerenboom, and T. Kelly. 2001. "Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies." *IEEE Control Systems Magazine* 21(6): 11–25.

"Risk and Critical Infrastructure System Protection." 2917. *Argonne National Laboratory*. https://www.anl.gov/article/better-infrastructure-risk-and-resilience (February 28, 2020).

Robert, Benoît, Luciano Morabito, Irène Cloutier, and Yannick Hémond. 2015. "Interdependent Critical Infrastructures Resilience: Methodology and Case Study." *Disaster Prevention and Management* 24(1): 70–79.

Rockefeller Foundation; Arup. 2014. Arup *City Resilience Framework*.

http://www.seachangecop.org/files/documents/URF_Booklet_Final_for_Bellagio.pdf%5Cnhttp://www.rockefellerfoundation.org/uploads/files/0bb537c0-d872-467f-9470-b20f57c32488.pdf%5Cnhttp://resilient-cities.iclei.org/fileadmin/sites/resilient-cities/files/Image.

Rose, Adam, and Elisabeth Krausmann. 2013. "An Economic Framework for the Development of a Resilience Index for Business Recovery." *International Journal of Disaster Risk Reduction* 5: 73–83. http://dx.doi.org/10.1016/j.ijdrr.2013.08.003.

Rouse, M. "What Is Object-Oriented Programming (OOP)? - Definition from WhatIs." https://searchapparchitecture.techtarget.com/definition/object-oriented-programming-OOP (January 2, 2019).

Salem, S., A. Siam, W. El-Dakhakhni, and M. Tail. Forthcoming. "Probabilistic Resilience-Guided Infrastructure Risk Management." Journal of Management in Engineering. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000818

Santella, N., L. Steinberg, and K. Parks. 2009. "Decision Making for Extreme Events: Modeling Critical Infrastructure Interdependencies to Aid Mitigation and Response Planning." *Review of Policy Research* 26(4): 409–22.

Sharma, Neetesh, Armin Tabandeh, and Paolo Gardoni. 2017. "Resilience Analysis: A Mathematical Formulation to Model Resilience of Engineering Systems." *Sustainable and Resilient Infrastructure* 9689: 1–19. http://doi.org/10.1080/23789689.2017.1345257.

Shen, Lijuan, and Loonching Tang. 2015. "A Resilience Assessment Framework for Critical Infrastructure Systems." *Proceedings of 2015 the 1st International*

Conference on Reliability Systems Engineering, ICRSE 2015.

Slivkova, Simona, David Rehak, Veronika Nesporova, and Michaela Dopaterova. 2017. "Correlation of Core Areas Determining the Resilience of Critical Infrastructure." *Procedia Engineering* 192: 812–17. http://dx.doi.org/10.1016/j.proeng.2017.06.140.

Spaans, M., and B. Waterhout. 2017. "Building up Resilience in Cities Worldwide– Rotterdam as Participant in the 100 Resilient Cities Programme. Cities." *Cities* 61: 109–16.

Standish, R.J. et al. 2014. "Resilience in Ecology: Abstraction, Distraction, or Where the Action Is?" *Biological Conservation* 177: 43–51. http://dx.doi.org/10.1016/j.biocon.2014.06.008.

Sun, Lijun, and Yafeng Yin. 2017. "Discovering Themes and Trends in Transportation Research Using Topic Modeling." *Transportation Research Part C: Emerging Technologies* 77: 49–66. http://dx.doi.org/10.1016/j.trc.2017.01.013.

Sun, Wenjuan, Paolo Bocchini, and Brian D. Davison. 2018. "Resilience Metrics and Measurement Methods for Transportation Infrastructure: The State of the Art." *Sustainable and Resilient Infrastructure* 9689: 1–32. https://doi.org/10.1080/23789689.2018.1448663.

Svendsen, N, and S Wolthusen. 2007. "Graph Models of Critical Infrastructure Interdependencies." *Physical Review Letters* 4543: 208–11. http://dx.doi.org/10.1007/978-3-540-72986-0_27.

Teknomo, By Kardi. 2004. "System Dynamics Tutorial © 1992-2004." *Cycle*.

https://people.revoledu.com/kardi/tutorial/SystemDynamic/ (January 1, 2019).

Timashev, S A. 2015. "Infrastructure Resilience: Definition, Calculation, Application." In *2015 International Conference on Interactive Collaborative Learning (ICL)*, , 1075. http://queens.ezp1.qub.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=111593805&site=eds-live&scope=site.

"Tm_map Function | R Documentation." 2019. https://www.rdocumentation.org/packages/tm/versions/0.7-7/topics/tm_map (December 12, 2019).

Ulieru, Mihaela. 2007. "Design for Resilience of Networked Critical Infrastructures." *Proceedings of the 2007 Inaugural IEEE-IES Digital EcoSystems and Technologies Conference, DEST 2007*: 540–45.

Val, D, R Holden, and S Nodwell. 2014. "Probabilistic Analysis of Interdependent Infrastructures Subjected to Weather-Related Hazards." *Civil Engineering and Environmental Systems* 31(2): 140–52. https://doi.org/10.1080/10286608.2014.913032.

Vale, L. 2014. "The Politics of Resilient Cities : Whose Resilience and Whose City ? The Politics of Resilient Cities : Whose Resilience and Whose City ?" *Building Research & Information* 42(2): 191–201. http://dx.doi.org/10.1080/09613218.2014.850602.

Wallace, W. et al. 2003. "Managing Disruptions to Critical Interdependent Infrastructures in the Context of the 2001 World Trade Center Attack." *Impacts of and Human Response to the September 11, 2001 Disasters: What Research Tells Us*: 166–98.

Wang, Shuliang et al. 2013. "Vulnerability Analysis of Interdependent Infrastructure Systems under Edge Attack Strategies." *Safety Science* 51(1): 328–37. http://dx.doi.org/10.1016/j.ssci.2012.07.003.

Xu, W et al. 2013. "The Resilience Framework for Interdependent Infrastructure Systems Using the Dynamic Inoperability Input-Output Model." *Intelligence Computation and Evolutionary Computation*: 355–361.

———. 2015. "The Uncertainty Recovery Analysis for Interdependent Infrastructure Systems Using the Dynamic Inoperability Input–Output Model." *International Journal of Systems Science* 46(7): 1299–1306.

Yang, Yifan et al. 2019. "A Physics-Based Framework for Analyzing the Resilience of Interdependent Civil Infrastructure Systems: A Climatic Extreme Event Case in Hong Kong." *Sustainable Cities and Society* 47(February): 101485. https://doi.org/10.1016/j.scs.2019.101485.

Yang, Yifan, S. Thomas Ng, Frank J. Xu, and Martin Skitmore. 2018. "Towards Sustainable and Resilient High Density Cities through Better Integration of Infrastructure Networks." *Sustainable Cities and Society* 42(January): 407–22. https://doi.org/10.1016/j.scs.2018.07.013.

Yodo, Nita, and Tanzina Arfin. 2020. "A Resilience Assessment of an Interdependent Multi-Energy System with Microgrids." *Sustainable and Resilient Infrastructure* 00(00): 1–14. https://doi.org/10.1080/23789689.2019.1710074.

Zhang, Yanlu, Naiding Yang, and Upmanu Lall. 2016. "Modeling and Simulation of the Vulnerability of Interdependent Power-Water Infrastructure Networks to Cascading

Failures." *Journal of Systems Science and Systems Engineering* 25(1): 102–18.

Zhao, S., X. Liu, and Y. Zhuo. 2017. "Hybrid Hidden Markov Models for Resilience Metrics in a Dynamic Infrastructure System." *Reliability Engineering and System Safety* 164: 84–97. http://dx.doi.org/10.1016/j.ress.2017.02.009.

Zimmerman, R., and C. E. Restrepo. 2006. "The next Step: Quantifying Infrastructure Interdependencies to Improve Security." *International Journal of Critical Infrastructures* 2(2/3): 215–30.

Zimmerman, R., Q. Zhu, and C. Dimitri. 2016. "Promoting Resilience for Food, Energy, and Water Interdependencies." *Journal of Environmental Studies and Sciences* 6(1): 50–61.

Chapter 3

A DEEP LEARNING NEURAL NETWORK MODEL FOR PREDICTING CLIMATE-INDUCED DISASTERS

ABSTRACT

The increased severity and frequency of Climate-Induced Disasters (CID) including those attributed to hydrological-, meteorological-, and climatological effects are testing the resilience of cities worldwide. The World Economic Forum highlights - in its 2019 Global Risk Report - that from 2017 to 2019 the top five risks with respect to likelihood and impact are all climate related with the highest ranked risk being extreme weather events. To alleviate the adverse impacts of CID on cities, this paper aims at predicting the occurrence of CID by linking different climate change indices to historical disaster records. In this respect, a deep learning (neural network) model was developed for spatial-temporal disaster occurrence prediction. To demonstrate its application, flood disaster data from the Canadian Disaster Database was linked to climate change indices data in Ontario in order to train, test and validate the developed model. The results of the case study showed that the model was able to predict flood disasters with an accuracy of around 96%. In addition to its association with precipitation indices, the study results affirm that flood disasters are closely linked to temperature-related features including the daily temperature gradient, and the number of days with minimum temperature below zero. This work introduces a new perspective in CID prediction, based on historical disaster data, global climate models, and climate change metrics, in an attempt to enhance urban resilience and mitigate CID risks on cities worldwide.

Keywords: Climate Change, Natural Hazards, Natural Disasters, Prediction, Machine Learning, Artificial Neural Networks.

3.1. INTRODUCTION

Climate change has been linked to increased loss of snow cover, accelerated sea level rise, more frequent heat waves and droughts, more intense hurricanes, and more importantly, a continuous and rapid rise in global temperatures (Callery, 2018; "Climate change | EU Science Hub," 2018; Shaftel, 2018a, 2018b). Since 1960s, the number of climate-induced hazards worldwide has tripled (World Health Organization, 2018), and comparing the total number of climate-induced hazards in 2011 and 2012, a total of 183 and 905 hazards were recorded worldwide, respectively, ranging between storms, tornados, hurricanes, floods, and wildfires (Nestler & Jackman, 2014; The Brookings Institution - London School of Economics project on Internal Displacement, 2012). The World Economic Forum - in its *2019 Global Risk Report* - highlights that from 2017 to 2019 the top three risks with respect to likelihood are all climate related with the highest ranked risk being extreme events. The three climate related risks have also been among the top five highest impact risks for the last three years (*Global Risks Report*, 2019).

Generally, climate-induced hazards can be classified into three main categories (The International Federation of Red Cross and Red Crescent Society, 2019): (1) hydrological hazards, governed by hydrological processes and include floods, droughts, and avalanches; (2) climatological hazards, concerned with extreme temperature related hazards including heat waves, cold waves and wildfires; and (3) meteorological hazards, representing storms of all types including snow storms, thunderstorms, hurricanes and tornados. Such hazards, on their own, do not necessarily signify risks unless they influence elements-at-risk (i.e., located within an urban setting) that are both exposed (i.e., unprotected) and vulnerable (i.e., susceptible to damage) under such hazard levels, thus leading to possible disasters as

per **Figure 3-1**. A disaster, thus can be defined as a "devastating impact of a hazard that negatively affects life, health, property or the environment on a scale sufficient to require outside assistance" (Babu, 2017).

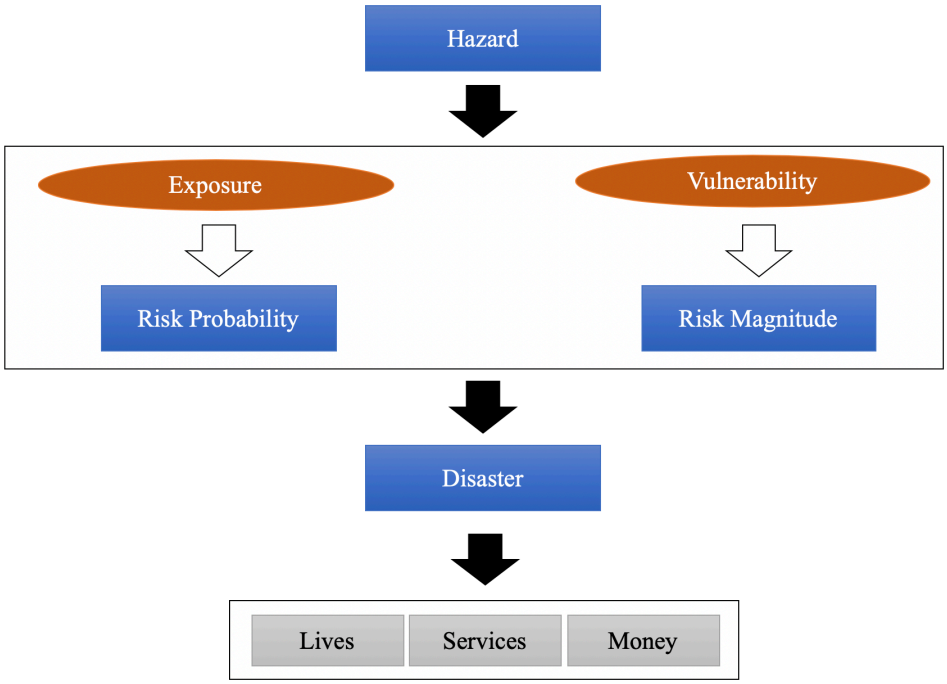


Figure 3-1: Relationships between Hazard, Risk, and Disaster

Between 2000 and 2012, a total of $1.7 trillion were reported as disaster-induced damages with more than 2.9 billion people affected by these disasters globally (Nestler & Jackman, 2014). Furthermore, about 60,000 people die annually due to such disasters, which are expected to cause 250,000 additional annual deaths between 2020 and 2030 (World Health Organization, 2018). By 2050, about 570 cities and 800 million people around the world will be threatened by rising sea levels and storm surges (Muggah, 2019). Over the last decade, more than 90 coastal cities in the United States alone are experiencing chronic flooding as a result of sea level rise (Muggah, 2019). In addition to the expected fatalities, adverse health impacts are expected to reach $2 to $4 billion annually by year 2030 (World

Health Organization, 2018). Moreover, the organization for economic cooperation and development has been showing in its annual reports that floods are causing more than $40 billion worth of damage annually (CHRISTINA, 2019). Currently, the annual liabilities of the Disaster Financial Assistance Arrangement program in Canada have increased from *"$10 million in 1970-1995 to $110 million in 1996-2010 to $360 million in 2011-2016"* (Public Safety Canada, 2017). In 2018 alone, the world has incurred a total of $160 billion cost of natural disasters, whereas the United States recorded a total natural disaster cost of about $91 billion with the camp wildfire of California and Hurricane Michael having a total cost of $16.5 and $16 billion, respectively (Chappell, 2019; Wright, 2019).

To minimize the impacts of Climate-Induced Disasters (CID), cities must be resilient—able to absorb disturbance and retain their basic functions, during and following such disturbances (Nan & Sansavini, 2017). One way of addressing the preparedness aspect of resilience is through disaster forecasting which would enable adequate planning and proactive risk management. In this respect, machine learning can be employed because of its ability to as an efficient modelling tool for the prediction of different processes including extreme weather driven hazard realizations (Zanchetta & Coulibaly, 2020). Originating from artificial intelligence, machine learning assumes that machines (i.e., computational models) can be trained using data records and subsequently, can learn to efficiently predict different complex phenomena. Machine learning was recently used for wildfire event prediction, specifically anthropogenic wildfire events were predicted using random forests, boosting and support vector machines (Rodrigues & De la Riva, 2014). Additionally, a spatial prediction of wildfire probabilities was proposed by combining different machine learning models with optimization algorithms (i.e., genetic algorithms),

and the results showed that optimization algorithms were able to refine the developed machine learning model (Jaafari, Zenner, Panahi, & Shahabi, 2019). Furthermore, different machine learning techniques were used to develop binary predictions related to CID. Crop losses were predicted as binary outcome on village level due to droughts using random forests (Mann, Warner, & Malik, 2019). Moreover, heavy rain damage was predicted as a binary outcome using decision trees, bagging, random forests and boosting (Choi et al., 2018).

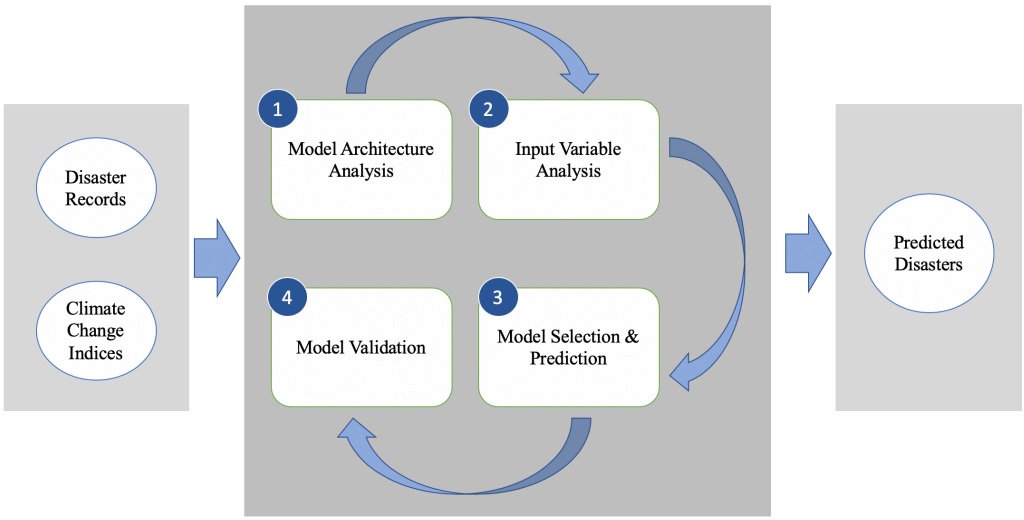
Initially formulated in an attempt to simulate the behavior of the human brain, artificial neural networks were first proposed by Warren McCulloch and Walter Pitts in 1943 (Abiodun et al., 2019; Jaspreet, 2016). Generally, an artificial neural network consists of an input layer, an output layer and a hidden layer(s). The input layer consists of the input attributes, whereas each hidden layer extracts different features from the input data to finally allow the network to formulate the desired outcome which is represented by the output layer (Dormehl, 2019; Kurt Hornik, 1991). The fact that hidden layers work as a black box is considered the main limitation of artificial neural networks as such ambiguity fails to provide the user with information behind the reasoning of the network's outputs. Another disadvantage of artificial neural networks is the fact that their performance depends on many parameters including the number of hidden layers and the number of neurons per layer. To develop a robust artificial neural network model the user needs to first optimize such parameters to meet the intended modeling objectives with the least possible computational effort (Donges, 2019; Hagan, Demuth, & Beale, 1997). Nevertheless, the main advantage of artificial neural networks lies in its strong ability to learn composite behaviors and thus, predict complex phenomena from large datasets.

Within each hidden layer, relevant information is drawn from the input data which explains the typically enhanced performance facilitated by using multilayer (i.e., deep) networks. As such, over the last few decades, applications and modeling approaches adopting artificial neural networks in natural phenomena simulation and prediction have been experiencing rapid advancement. For example, rainfall runoff processes were simulated using artificial neural networks which proved to be more efficient than most used physical models for rainfall runoff simulation (Hu et al., 2018). In addition, simplified deep learning-based extreme learning machine was used in rainfall prediction (Cholissodin & Sutrisno, 2018). On the other hand, deep learning artificial neural networks were used in flash flood mapping prediction and showed excellent results compared to support vector machine models (Bui et al., 2020). Flood damage was also linked to some household predictors including house structure, flood awareness, literacy and other factors using three types of models: linear regression, random forests and artificial neural networks (Ganguly, Nahar, & Hossain, 2019). Neural network models were also proposed to predict the number of hurricanes per season in vulnerable areas, with a prediction accuracy of 73% (Kahira, Gomez, & Badia Sala, 2018). In addition, hurricane path was predicted using a neural network model which proved to be of comparable efficiency of traditional prediction models (Alemany, Beltran, Perez, & Ganzfried, 2018; Giffard-roisin et al., 2018). Artificial neural networks and support vector machines were also used to predict the binary occurrence of wildfires using satellite images (Sayad, Mousannif, & Al Moatassime, 2019). In addition, both logistic regression and artificial neural networks were used to classify wind damage to forests as a binary outcome (Hanewinkel, Zhou, & Schill, 2004).

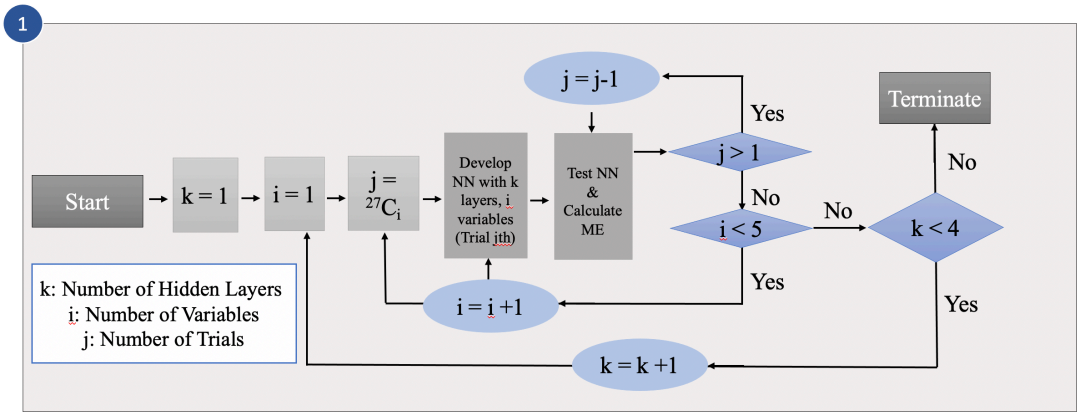
Subsequently, in this paper a deep learning (neural network) model for CID prediction is developed by linking historical disaster records to different climate change indices. The paper is divided into two main parts, the first part involves the general model structure that is generic enough to be employed to predict any class of CID in any location given the availability of the influencing spatiotemporal climate data. The second part demonstrates the applicability of the developed model using Ontario's disaster rerecords and relevant climate change indices data. This work is considered the first step in CID prediction, based on historical disaster data, global climate models, and climate change metrics, in an attempt to maximize urban resilience and mitigate CID impacts on cities worldwide.

3.2. CLIMATE-INDUCED DISASTER PREDICTION MODEL

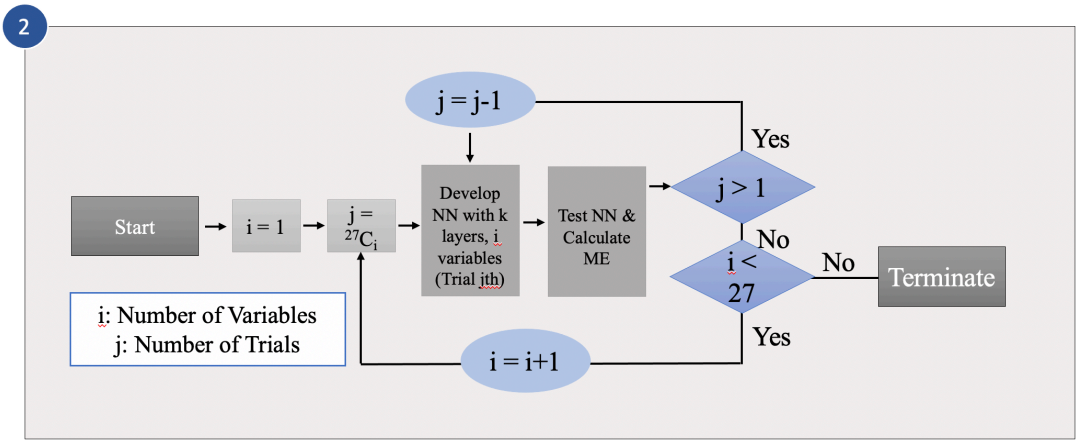
Through employing deep learning, the CID prediction model developed herein utilizes previous disaster records and specific climate change indices as inputs and returns whether or not (Yes/No) a given type of disaster would occur in a given place and time as a binary output. Following four main stages, the model takes in the inputs and generates the required outputs as per **Figure 3-2**.


Disaster Records
Climate Change Indices
1
Model Architecture Analysis
2
Input Variable Analysis
4
Model Validation
3
Model Selection & Prediction
Predicted Disasters

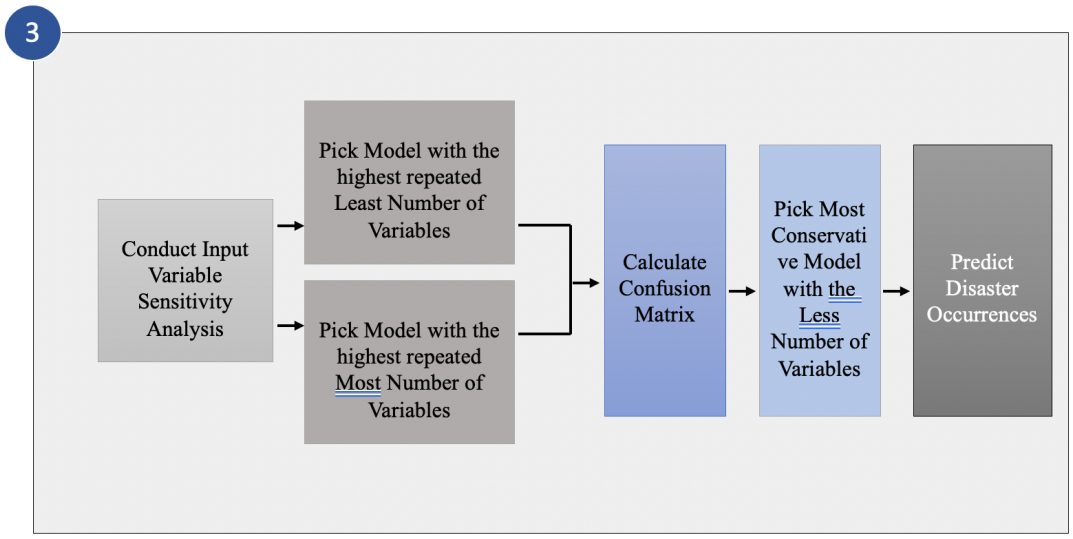
(a)


1
Start
k = 1
i = 1
j = $^{27}C_i$
Develop NN with k layers, i variables (Trial jth)
Test NN & Calculate ME
j = j-1
j > 1
Yes
No
i < 5
No
Yes
i = i +1
k < 4
No
Terminate
Yes
k = k +1
k: Number of Hidden Layers
i: Number of Variables
j: Number of Trials

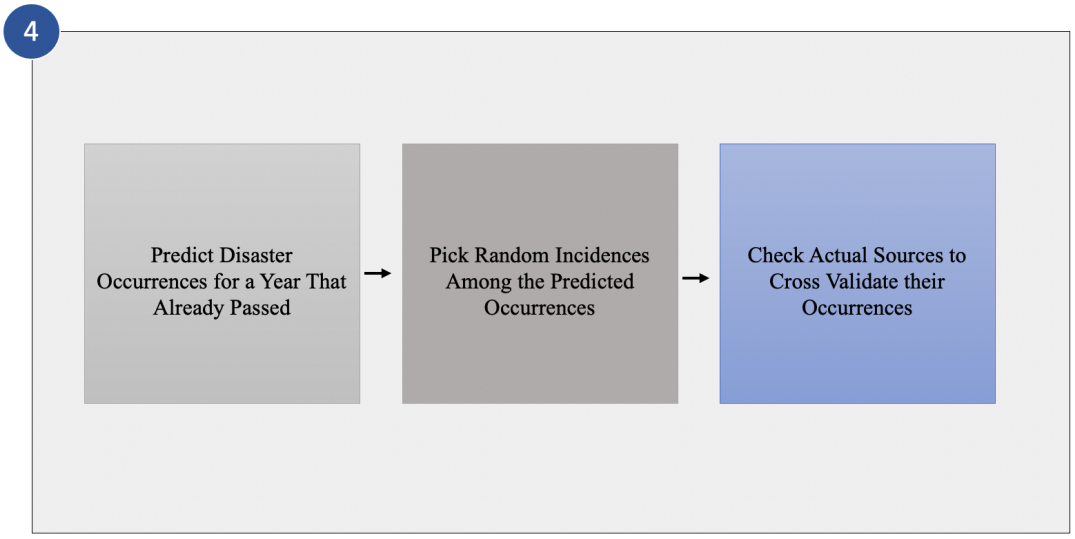
(b)


2
Start
i = 1
j = $^{27}C_i$
Develop NN with k layers, i variables (Trial jth)
Test NN & Calculate ME
j = j-1
j > 1
Yes
No
i < 27
No
Terminate
Yes
i = i+1
i: Number of Variables
j: Number of Trials

(c)



(d)



(e)

Figure 3-2: Deep Learning Model(a) Concept; (b) Stages 1; (c) Stages 2; (d) Stages 3; and (e) Stages 4

3.2.1. Model Inputs: Disaster Data & Climate Change Indices

As a prerequisite for any machine learning model, relevant data availability is key for training, testing and validating the model. Within the context of the current study, the

variables considered for CID prediction are: (1) type of disaster (i.e., hydrological, climatological or meteorological), (2) date of occurrence; and (3) specific location.

The *Expert Team on Climate Change Detection and Indices* (Expert Team on Climate Change Detection and Indices, 2009) and the *Working Group on Climate Change Detection* (Peterson et al., 2001) developed a set of standardized indices to be used by stakeholders around the world for monitoring climate change (Karl, Nicholls, & Ghazi, 1999; Peterson, 2005; Peterson et al., 2001). In total, 40 indices were approved, 27 of which were considered to be *core* indices. These 27 climate change indices are divided into two distinct subsets. The first subset of indices represents different measures that captures the variations in temperature as per **Table 3-1**, whereas the second subset is mainly related to detecting the change in precipitation as per **Table 3-2** (Expert Team on Climate Change Detection and Indices, 2009). These two subsets of indices will be used together with disaster data to predict CID occurrence.

Table 3-1: The 16 Climate Change Temperature Indices

Index	Definition
FD- Number of Frost Days	Annual count of days when TN (i.e., daily minimum temperature) < 0oC
IC - Number of Icing Days	Annual count of days when TX (i.e., daily maximum temperature) < 0°C
SU - Number of Summer Days	Annual count of days when TX > 25°C
TR - Number of Tropical Nights	Annual count of days when TN > 20°C
GSL - Growing Season Length	Annual (1st Jan to 31st Dec in Northern Hemisphere, 1st July to 30th June in Southern Hemisphere) count between first span of at least 6 days with TG (i.e., daily mean temperature) > 5°C and first span after July 1st (Jan 1st in Southern Hemisphere) of 6 days with TG < 5°C

DTR - Daily Temperature Range	Monthly mean difference between TX and TN
TN10p	Percentage of days when TN < 10th percentile
TX10p	Percentage of days when TX < 10th percentile
TN90p	Percentage of days when TN > 90th percentile
TX90p	Percentage of days when TX > 90th percentile
TX_X	Monthly maximum value of daily maximum temperature
TN_X	Monthly maximum value of daily minimum temperature
TX_N	Monthly minimum value of daily maximum temperature
TN_N	Monthly minimum value of daily minimum temperature
WSDI - Warm Spell Duration Index	Annual count of days with at least 6 consecutive days where TX > 90th percentile
CSDI - Cold Spell Duration Index	Annual count of days with at least 6 consecutive days where TN < 10th percentile

Table 3-2: The 11 Climate Change Precipitation Indices

Index	Definition
CDD - Maximum Length of Dry Spell	Maximum number of consecutive days with RR (i.e., daily precipitation) < 1mm
CWD - Maximum Length of Wet Spell	Maximum number of consecutive days with RR ≥ 1mm
R10	Annual count of days when RR ≥ 10
R20	Annual count of days when RR ≥ 20mm
PRCPTOT	Annual total precipitation in wet days (i.e., days with precipitation over 1mm)
R95	Annual total precipitation when RR > 95th percentile
R99	Annual total precipitation when RR > 99th percentile
Rx1day	Monthly maximum 1-day precipitation
Rx5day	Monthly maximum consecutive 5-day precipitation
SDII - Simple Precipitation Intensity Index	The sum of precipitation in wet days during the year divided by the number of wet days in the year
Rnn	Annual count of days when PRCP ≥ nn, where nn is a user defined threshold

3.2.2. OUTPUT: FORECASTED CLIMATE CHANGE INDUCED DISASTERS

The developed model aims to predict CID occurrence in a specific location on an annual basis. Thus, the output of the model would be either a "Yes" if a disaster is predicted or a "No" if no disaster is predicted in a certain location.

3.2.3. STAGE 1: MODEL ARCHITECTURE ANALYSIS

Stage 1, as per Figure 3-2(b), involves determining the model architecture which, in the case of artificial neural networks, involves determining which model class to use (i.e., multiple- or single- layer model). This stage also entitles selecting the most suitable number of hidden layers and neurons to use in the prediction model. Multiple layer models are developed for different input variables trials. The trials are comprised of 27 single variable models corresponding to the 27 climate change indices defined earlier as well as any two, three, four or five index combinations of the 27 indices resulting in a total of 101,583 model trials. Model architecture analysis is based on the combination of one to five input variables instead of the combination of the total number of variables (i.e., 1 to 27 input variables) in an attempt to minimize both the modelling time and computational effort. For each set of inputs, the model is trained and tested, and the misclassification error is recorded. Furthermore, for each number of layers, the number of neurons is specified. The average of all misclassification errors for the models with a specific number of layers together with the number of models with the least misclassification error are compared to help in selecting the best performing model and thus, comply with this specific number of layers and neurons for the rest of the analysis.

3.2.4. STAGE 2: INPUT VARIABLES ANALYSIS

Stage 2, as per Figure 3-2(c) aims to select the significant input variables among the 27 available variables to be used for disaster predictions. A sensitivity analysis involving all input variable combinations is conducted, resulting in a total of more than 16.6 million models with different combination of input variables ranging from 1 to 10 variables as shown in **Figure 3-3**. For instance, Combination 4 refers to combining any four variables out of the 27 input variables. In each model trial, the model inputs are changed, the model is trained, tested, and the misclassification error is calculated. The models with the least misclassification error are then selected as the best models among all model trials.

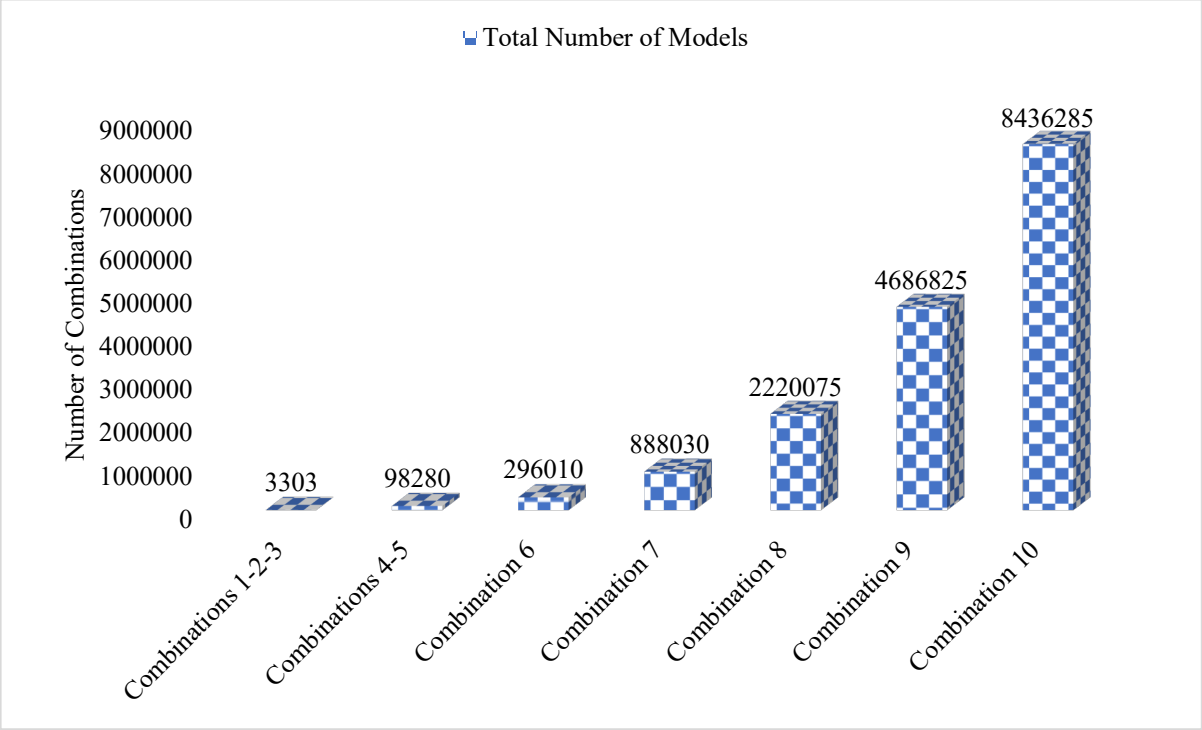


Figure 3-3: Different Number of Input Variables Combinations

3.2.5. STAGE 3: MODEL SELECTION & PREDICTION

Stage 3 involves model selection and prediction as per Figure 3-2 (d). Based on the sensitivity analysis performed in Stage 2, the best performing models is selected (i.e., among the models with the lowest misclassification error). Furthermore, the confusion matrix of the selected models will be compared to select the best performing model accordingly. Following model selection, disaster prediction analysis can be conducted.

3.2.6. STAGE 4: MODEL VALIDATION

Finally, as per Figure 3-2 (e), in Stage 4 model validation is employed by comparing the model prediction results to actual disasters records that are not included in the model input data.

3.3. APPLICATION: PREDICTING FLOOD DISASTERS IN ONTARIO, CANADA

To demonstrate its applicability, the developed model was applied to predict disasters in Ontario which is Canada's most populated province and the one with the highest number of recorded CID between 1900 and 2016 as per **Figure 3-4**. The model is specifically employed to predict flood disasters in Ontario as they account for 42% of the total number of CID from 1900 to 2016 as per **Figure 3-5**, which have at least four times the frequency of occurrence of any other CID in Ontario for this period of analysis.

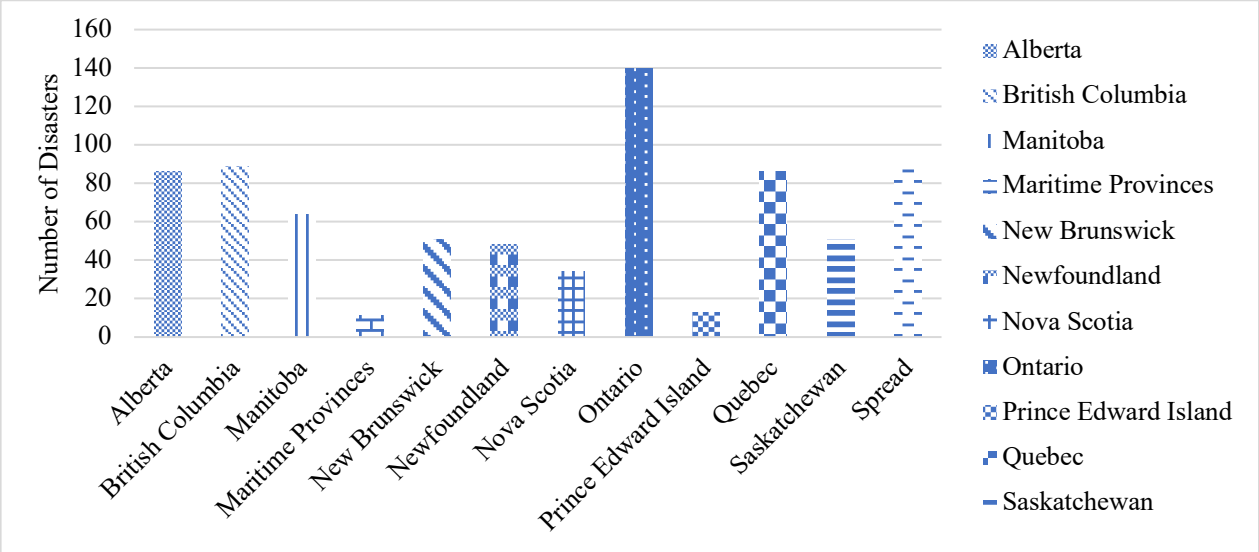


Figure 3-4: CID per province from 1900 to 2016

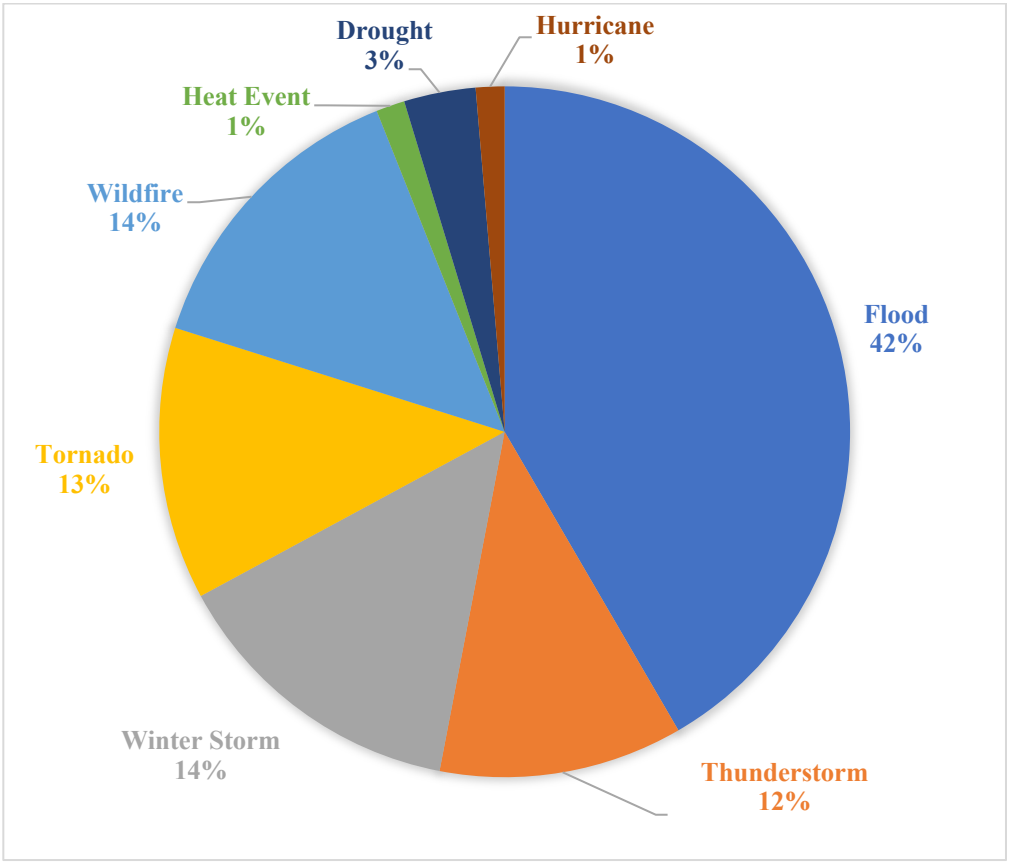


Figure 3-5: CID Distribution in Ontario

3.3.1. DATA PREPARATION

The Canadian Disaster Database used in this paper was created by Public Safety Canada (Public Safety Canada, 2019). The database includes approximately 1,000 significant disaster events (i.e., natural, technological and conflict events) with one or more of the following characteristics: (1) 10 or more people killed, (2) 100 or more people affected/injured/infected/evacuated or homeless, (3) an appeal for national/international assistance, (4) historical significance, and/or (5) significant damage/interruption of normal processes. As this case study focuses on flood prediction, the database was filtered to include only flood designated disasters in Ontario between 1900 and 2016 as shown on the map of **Figure 3-6**. The "Yes" to "No" ratio in the flood disaster data considered herein is around 7:3 which, while slightly imbalanced, is not expected to significantly affect the final model accuracy since data-driven models are considered to be biased when data imbalance reaches a ratio of 100:1 or more (He & Shen, 2007; Kubat, Holte, & Matwin, 1998; Ramyachitra & Manikandan, 2014; Viola, Emonet, Habrard, Metzler, & Sebban, 2020).

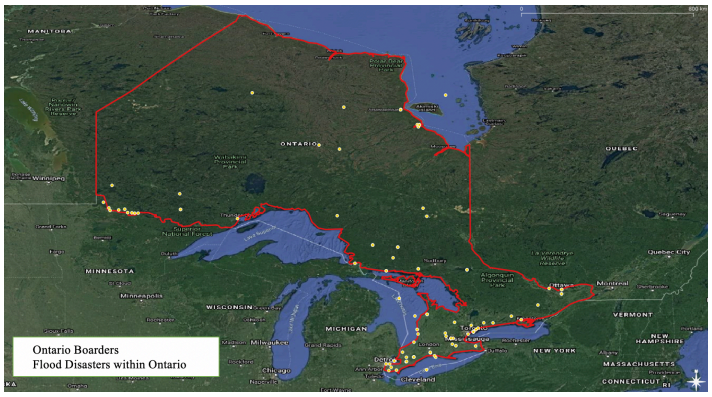


Figure 3-6: Flood Disasters Spatial Distribution in Ontario

Furthermore, the 27 climate change indices were calculated at 24 different watersheds in Ontario as shown in **Figure 3-7**. The indices were calculated based on historical and future

climate data from 1950 to 2016 and from 2017 to 2100, respectively (Wazneh, Arain, & Coulibaly, 2019).

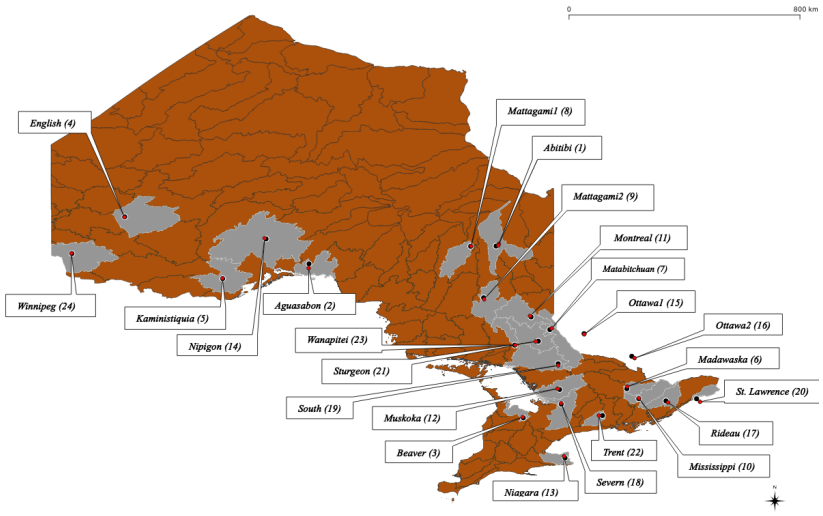


Figure 3-7: The 24 Ontario Watershed Locations

To initiate global climate model simulations, IPCC proposed different "benchmark emission scenarios" widely known as Representative Pathway Concentrations (RCP). These scenarios represent greenhouse gas concentrations based on different proposed volume of greenhouse gas emissions in the future. There are currently four RCP values that are approved by the IPCC and widely adopted in climate model simulations. The first is RCP 2.6 which assumes that a reduction in greenhouse gas emission will occur over the next few decades and is referred to as the peak scenario. Specifically, RCP 2 .6 assumes that carbon dioxide emission will not only start declining over the next couple of years but will reach zero by 2100. RCP 8.5, on the other hand, is based on the assumption that greenhouse gas emissions will increase rapidly over time as such it is considered the worst-case scenario among all proposed RCPs. The two other pathways, RCP 6 and RCP 4.5, are both stabilization scenarios (i.e., intermediate scenarios) that assume greenhouse gas

emissions to reach a peak and then decline after 2040 and 2080, respectively. Both pathways assume that different decisions, techniques and laws will be established by governments around the world which will contribute to stabilizing greenhouse gas emissions on the long run (Kolp & Riahi, 2009; Meehl et al., 2013; Wayne, 2013). The indices used herein were calculated based on the 12 global climate models (Bi et al., 2013; Block & Mauritsen, 2013; Chylek, Li, Dubey, Wang, & Lesins, 2011; Collier et al., 2011; Flato et al., 2013; Gent et al., 2011; Griffies et al., 2010; Meehl et al., 2013; Ongoma, Chen, & Gao, 2019; Tongwen et al., 2014; Voldoire et al., 2013; Volodin, Dianskii, & Gusev, 2010) given in **Table 3-3** and are based on RCP 4.5 and 8.5 (Wazneh et al., 2019). The models simulated will employ the indices calculated based on RCP 4.5, the reason for this resides in the fact that RCP 8.5 is characterized by a steady and large increase in the greenhouse gas emissions over time which represents the worst-case scenario with respect to dealing with these emissions. Given the fact that governments are working on controlling greenhouse gas emissions (Prairie Climate Centre, 2018), in this paper the model is based on the more conservative assumption for forecasting of greenhouse gas emissions rather than the worst-case scenario (Collins et al., 2013).

Table 3-3: The 12 Global Climate Models

Model	Institution
ACCESS1	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia
BCC-CSM1	Beijing Climate Center, China Meteorological Administration
CanESM2	Canadian Centre for Climate Modelling and Analysis
CCSM4	National Center for Atmospheric Research (NCAR)

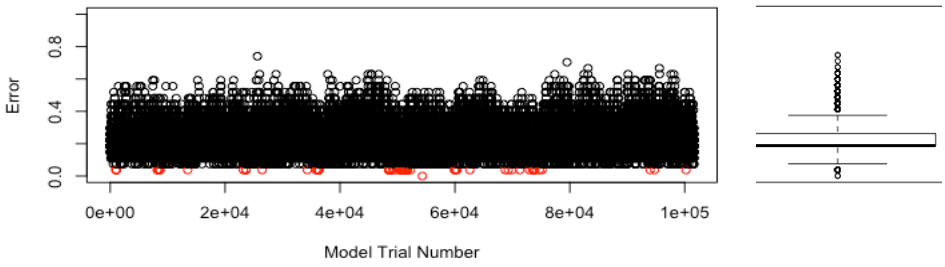
CESM1-BGC	Community Earth System Model Contributors
CNRM-CM5	Centre National de Recherches Météorologiques/ Centre Européen de Recherche et Formation Avancée en Calcul Scientifique
CSIRO-MK3-6-0	Commonwealth Scientific and Industrial Research Organization, Queensland Climate Change Centre of Excellence
GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory
Inmcm4	Institute for Numerical Mathematics
IPSL-CM5A-LR	Institut Pierre-Simon Laplace
MICROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology
MPI-ESM-LR	Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)

3.3.2. MODEL ARCHITECTURE ANALYSIS

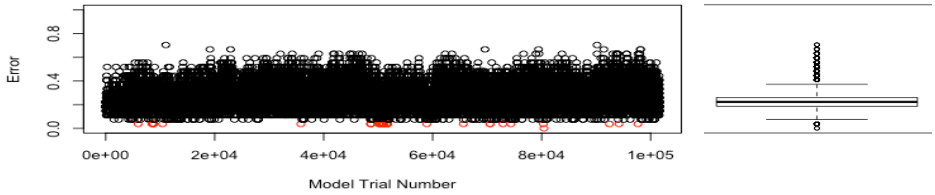
The model architecture, including the number of neurons in hidden layers, is formulated based on guidance of previous artificial neural networks research (Hagan et al., 1997; Heaton, 2017) and including the least number of input variables, out of the 27 studied indices, that would yield accurate predictions. As such, the size of the hidden layer was selected based on the following conditions: 1) be less than the sizes of the input and the output variables; 2) be at most 2/3 the size of both the input and output layers; and 3) be less than two times the input layer size. Increasing the number of hidden neurons has two widely acknowledged drawbacks (Cybenko, 1989; Hinton, Osindero, & Teh, 2006; Kurt Hornik, 1991). The first is related to overfitting, when the training information available is not enough to train the considered number of neurons, and as such the model fails in terms of generalizability as it memorizes the training data. The second shortcoming of having a

wide network is the excessive computational effort required to train the model. As such, the study commenced with a relatively narrow network configuration and its performance (i.e., accuracy) was checked through testing and validation. Furthermore, as the number of neurons in the input layer is much larger than that in the output layer, adopting a narrowing neural network (i.e., inverted pyramid network architecture) leads to data noise removal since, as each layer is further narrowed, the model is forced to drop irrelevant information which explains why such models were shown to yield higher accuracy at lower computational cost (Czanner et al., 2015; Srivastava, 2019).

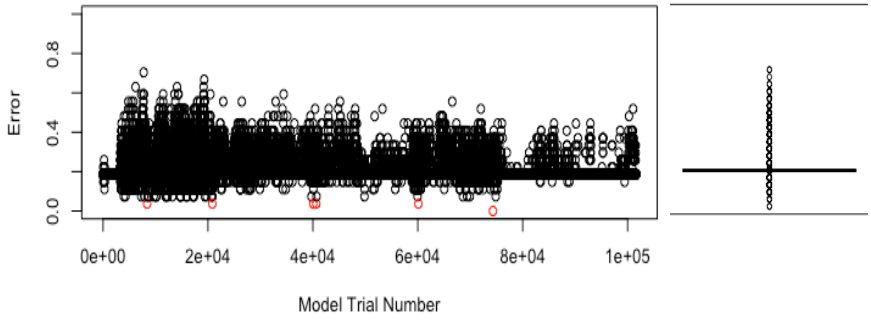
As such, to select the most suitable number of hidden layers and neurons for the artificial neural network model, the model was simulated with 1, 2, 3 and 4 hidden layers. The number of neurons was specified for each model i.e., for the single layer models 3 neurons were specified, as for the two layers model, 5 neurons were selected for layer one and 3 for layer two. For the three layers models, 5, 3, and 2 neurons were selected for layers 1, 2 and 3, respectively. Finally, for the four layers models, 5, 3, 2, and 2 neurons were selected for Layers 1, 2, 3 and 4, respectively. To choose the model with the optimum number of layers the average of misclassification errors for all models with a specific number of layers was calculated. In addition, the minimum error was compared across the models with different number of hidden layers. **Figure 3-8** shows the misclassification error plotted against the 101,583 models; the red dots represent the models' which yielded the minimum error reached which is 3.8%. As can be observed, the number of models with minimum error is decreasing rapidly as the number of hidden layers increases. Fig. 8 also shows the box plots for the misclassification error of the four simulated layers which includes the average and the 25^{th} and 75^{th} percentiles.



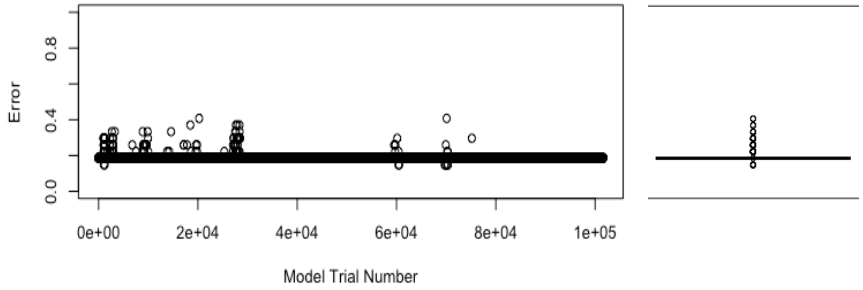
(a)



(b)



(c)



(d)

Figure 3-8: Hidden Layers Misclassification Error Analysis, (a), Single Hidden Layer, (b) Two Hidden Layers, (c) Three Hidden Layers, and (d) Four Hidden Layers

Table 3-4 shows the average misclassification error together with the number of models with the least error (i.e., 3.8%) for the four model types simulated. The average

misclassification error for the models with two hidden layers is the lowest among the number of layers simulated. Excluding the models with two layers, the average of the misclassification errors decreases as the number of layers increases.

Table 3-4: Hidden Layers Sensitivity Analysis

	One Hidden Layer	Two Hidden Layers	Three Hidden Layers	Four Hidden Layers
Average Error	0.21	0.24	0.20	0.19
Minimum Error Models	114	37	5	0

As per (D. Liu, Zhang, Polycarpou, Alippi, & He, 2011), increasing the number of hidden layers in a neural network may sometimes cause the network to overfit (i.e., influenced by the training dataset) which results in lower prediction ability when it comes to the testing dataset. This is the reason why the single layer trials resulted in better performing models compared to the multiple layer models. The universal approximation theorem also confirms the usefulness of single-layer networks (Dong & Li, 2012; Kumar, 2019; Sanger, 1989; Stathakis, 2009). Consequently, the model with one hidden layer and three neurons was proved to be the best performing model for flood disaster prediction in Ontario.

3.3.3. INPUT VARIABLES ANALYSIS

After selecting the number of hidden layers, input variables sensitivity analysis was carried out and the models with up to 10 variables were evaluated. The best performing models were selected based on the least obtained misclassification error which was found to be 3.8% corresponding to 20,093 models out of more than 16.6 million models (i.e., only 0.12% of the models had such a low error) as per **Figure 3-9**.

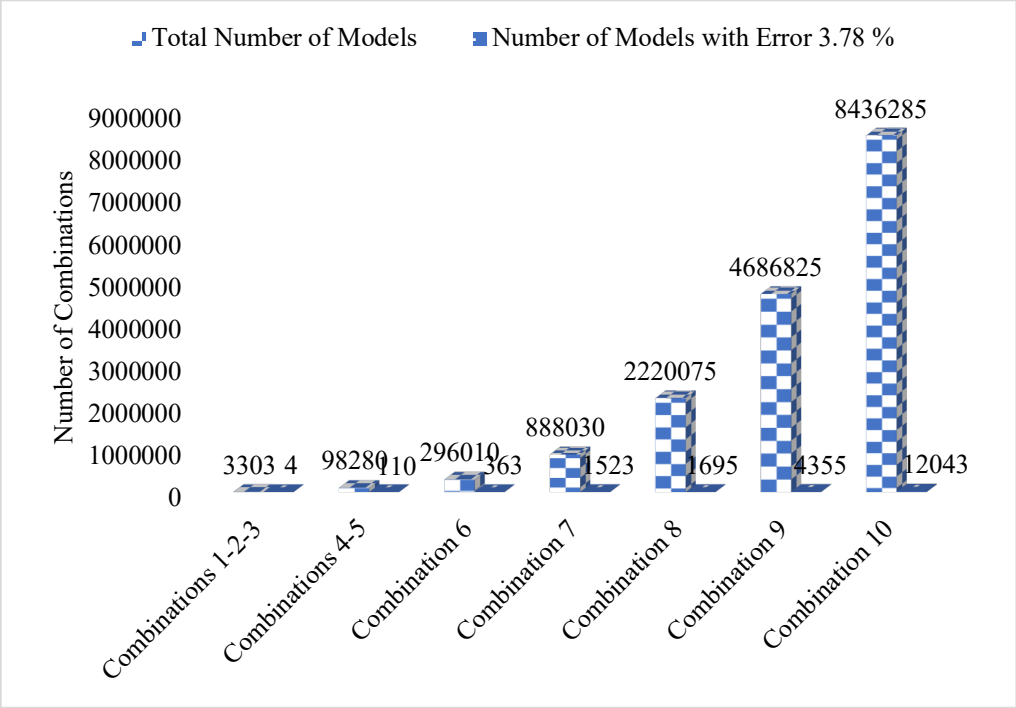


Figure 3-9: Total Number of Models versus Number of Models with Least Misclassification Error

The input variables together with the number of times each variable was repeated in the 20,093 models are shown in Fig. 10 which reflects the importance of each input variable in flood disaster prediction in Ontario. As per **Figure 3-10**, the four most reoccurring climate change indices in the 20,093 models are:

(1) DTR (daily temperature range) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh, Arain, & Coulibaly, 2017) represented by equation 1, where TX_{ij} and TN_{ij} are the daily maximum and minimum temperature respectively on day *i* in period *j*. If *I* represents the number of days in period *j*, then:

$$DTR_j = \frac{\sum_{i=1}^{I}(Tx_{ij} - Tn_{ij})}{I} \quad (1)$$

(2) CWD (maximum length of wet spell) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which is calculated by counting the largest number of consecutive days where $RR_{ij} \geq 1$mm, where RR_{ij} is the daily precipitation amount on day *i* in period *j*;

(3) TN10P (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) refers to the percentage of days when $TN_{ij} < TN_{in}10$, where TN_{ij} is the daily minimum temperature on day *i* in period *j* and $TN_{in}10$ is the calendar day 10^{th} percentile centred on a 5-day window for the base period 1961-1990; and

(4) FD (number of frost days) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which refers to the annual count of days when TN < 0°C. To calculate FD, let TN_{ij} be the daily minimum temperature on day *i* in year *j*. Count the number of days where $TN_{ij} < 0$°C.

On the other hand, the least four repeated indices are:

(1) CSDI (cold spell duration index) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which refers to the annual count of days with at least 6 consecutive days when daily minimum temperature < 10^{th} percentile;

(2) R99 (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which refers to the sum of annual total precipitation in days with precipitation more than 1 mm when RR > 99^{th} percentile;

(3) R95 (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which refers to the sum of annual total precipitation in days with precipitation more than 1 mm when RR > 95^{th} percentile; and

(4) CDD (maximum length of dry spell) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al., 2017) which refers to the maximum number of consecutive days with $RR_{ij} < 1$mm, where RR_{ij} is the daily precipitation amount on day *i* in period *j*. It is calculated by counting the largest number of consecutive days where $RR_{ij} < 1$mm.

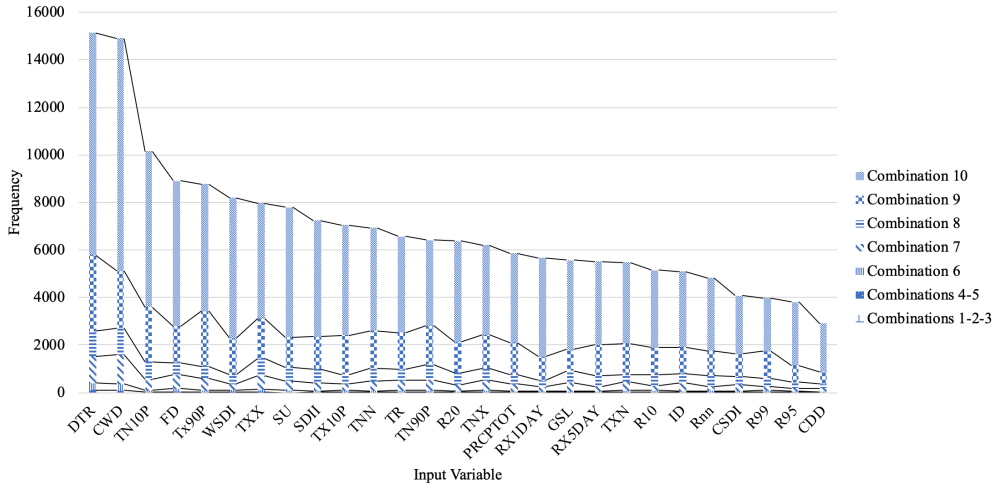


Figure 3-10: Input Variable Frequency Analysis

The variable frequency analysis shown in Fig. 10 shows that three of the most frequent variables are actually temperature related which supports the direct and strong relationship that exists between temperature changes and flooding. As air gets warmer, its ability to hold water vapour increases (e.g., air contains around 7% more moisture for each 1°C temperature increase), and this moisture turns into rain when condensation occurs due to temperature drop (American Association for the Advancement of Science, 2011). Moreover, 1- to 5-day-long intense precipitation episodes are intensified with continued warming (Min, Zhang, Zwiers, & Hegerl, 2011). The interdependence and feedbacks between temperature and precipitation are well documented in the literature (Cong & Brady, 2012; Mishra, Wallace, & Lettenmaier, 2012; Wazneh, Arain, Coulibaly, & Gachon, 2020). Furthermore, different studies (Liu, Cheng, & Su, 2014; Mirza, 2011;

Wasko & Sharma, 2017) demonstrated strong correlation between air temperature changes and flooding. As such, the input variable sensitivity analysis which indicates that three of the four most frequently introduced variables are temperature-related, confirms the ability of the developed data-driven model to establish physical relationship between temperature changes and flooding.

To further validate the number of hidden layers and neurons selected in Stage 1, **Figure 3-11** shows the results of the misclassification error for the model with the four most frequent input variables as the number of layers is varied from 1 to 4 and for each hidden layer 3, 10, 15, 20, 25, and 30 neurons are modelled. It can be observed that the least misclassification error is reached when the single hidden layer model is comprised of 3 neurons. It can also be observed that generally as the number of hidden layers increase, the misclassification error increases.

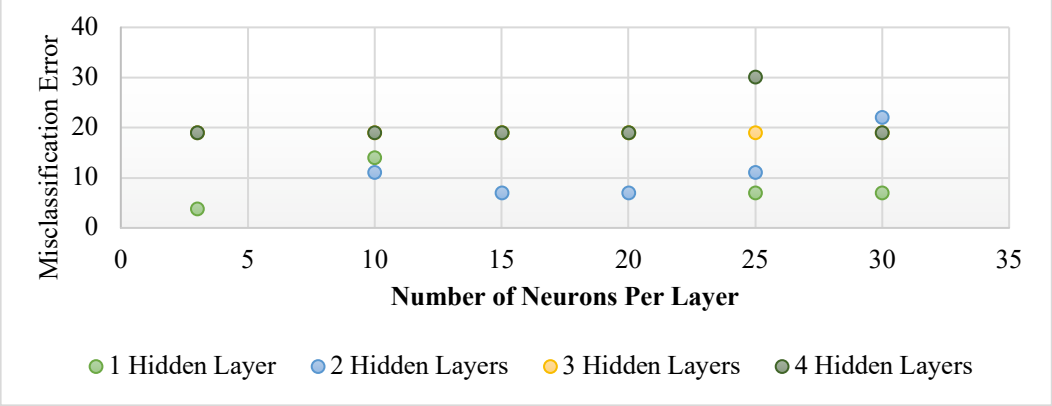


Figure 3-11: Misclassification Error versus Number of Hidden Layers and Neurons

3.3.4. MODEL SELECTION

To select the best performing model among the 20,093 models, two models are compared, the first model includes the most frequent four input variables (Model 1), while the second

model includes ten input variables among these variables are nine of the most repeated input variables (Model 2) as shown in **Figure 3-12**. The training of both models is shown in **Figure 3-13**. The developed artificial neural networks together with the weights to and from the single hidden layer are shown. The models with four and ten input variables needed 174 and 52 steps to converge with sum of squares errors 6 and 5.58, respectively.

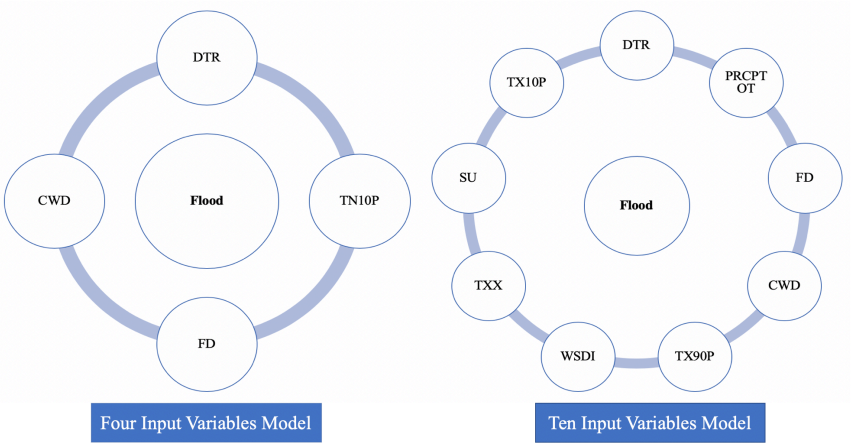


Figure 3-12: Flood Disaster Prediction Models Input Variables

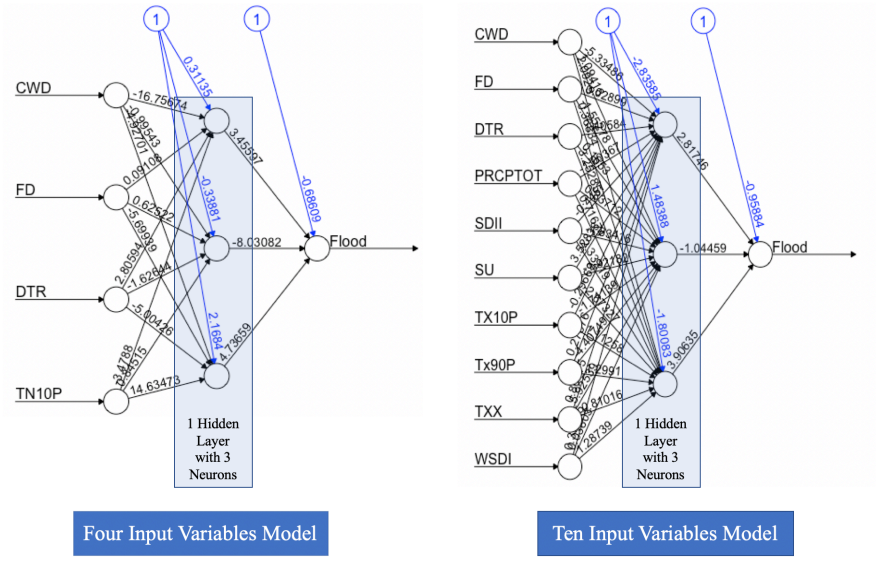


Figure 3-13: Neural Network Model Training

To select the best performing model among Model 1 and Model 2, the confusion matrixes for both models were compared and were found to be identical as shown in **Table 3-5**. The matrix indicates that 22 flood disasters which actually occurred were correctly predicted, among the five years in which no flood disasters happened four of which were correctly predicted while one was misclassified. Additionally, the actual flood disasters as per the Canadian Disaster Database are compared to model predictions in

Table 3-6. The results for the testing of the two models are also shown on the map in **Figure 3-14**. Locations where flood disasters were correctly predicted are shaded in blue, while the red shaded areas refer to locations where misclassifications were found. The yellow shades represent areas where no predictions were materialized.

Table 3-5: Model Confusion Matrix

		Actual
Predicted	**0**	**1**
0	4	0
1	1	22

Table 3-6: Actual versus Predicted Results

Year	**Watershed**	**Actual**	**Model 1**	**Model 2**
1954	Niagara	Yes	Yes	Yes
1956	Niagara	Yes	Yes	Yes
1968	Beaver	Yes	Yes	Yes
1968	Niagara	Yes	Yes	Yes
1970	Abitibi	Yes	Yes	Yes
1979	Beaver	Yes	Yes	Yes
1979	South	Yes	Yes	Yes
1979	Wanapitei	Yes	Yes	Yes
1980	Trent	Yes	Yes	Yes
1985	Niagara	Yes	Yes	Yes
1989	Abitibi	Yes	Yes	Yes

1989	Nipigon	Yes	Yes	Yes
1992	Beaver	Yes	Yes	Yes
1996	Mattagami2	Yes	Yes	Yes
1996	Niagara	Yes	Yes	Yes
1996	Rideau	Yes	Yes	Yes
2002	Mattagami1	Yes	Yes	Yes
2002	Winnipeg	Yes	Yes	Yes
2004	Mattagami1	Yes	Yes	Yes
2008	Nipigon	Yes	Yes	Yes
2013	Niagara	Yes	Yes	Yes
2016	Winnipeg	Yes	Yes	Yes

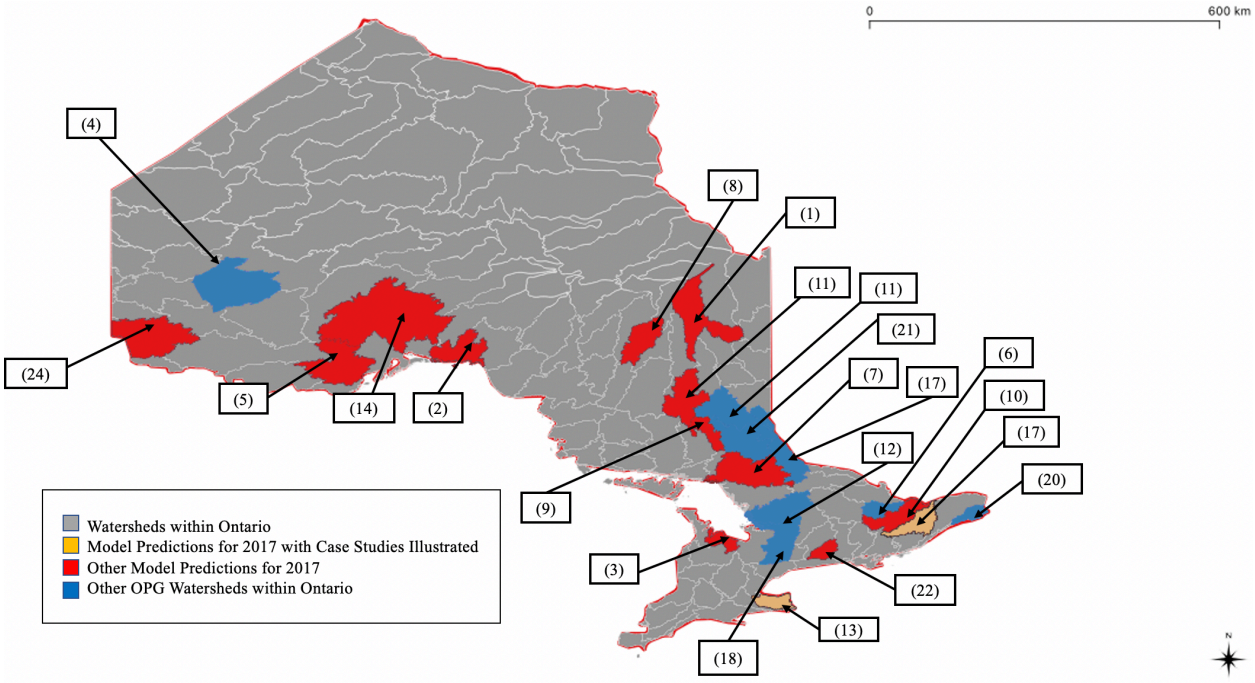


Figure 3-14: Actual versus Predicted Model Results

3.3.5. DISASTER PREDICTION

After comparing the testing results of both models, it is apparent that both models yielded the same accuracy for predicting flood disasters. Thus, Model 1 with the four input variables is selected as it resulted in the same prediction accuracy with a smaller number of inputs which yields a simpler and more applicable model. Thus, Model 1's flood disaster prediction from 2020 to 2030 is shown in **Figure 3-15**. The figure shows whether or not a flood disaster is expected to take place in a certain location during a specific year. For example, the model predicts yearly flood disasters in Kiministiquia over the next decade, whereas only a single flood disaster in Aguasabon in 2030. Furthermore, the model predicts no flood disasters in Niagara over the next decade.

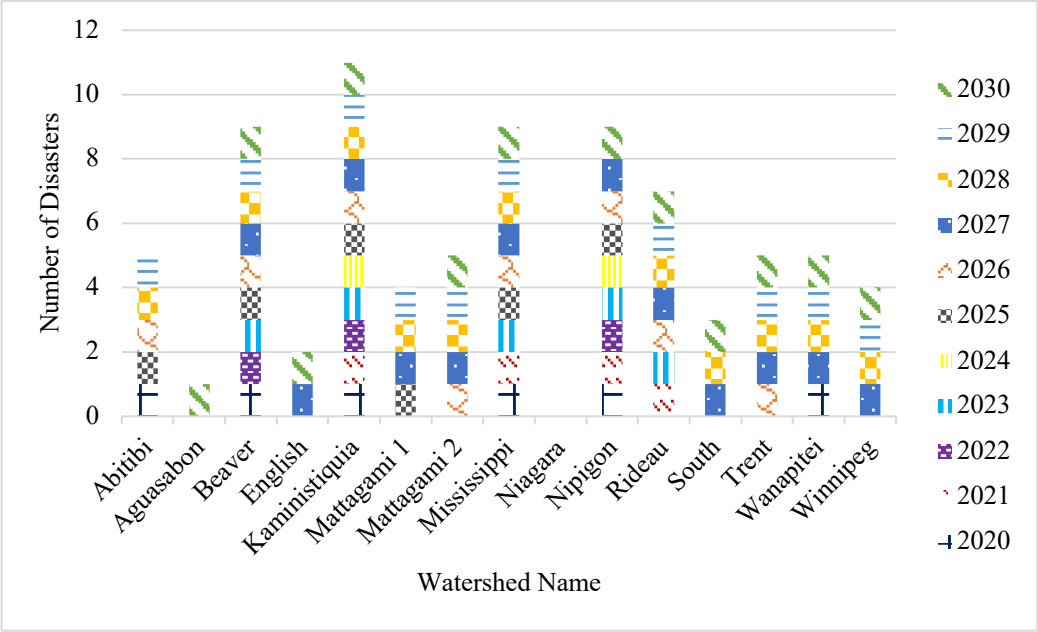


Figure 3-15: Flood Disaster Prediction until 2030

3.3.6. MODEL VALIDATION

The Canadian Disaster Database used herein tracks flood disasters until 2016 thus, some events that actually took place in 2017 and that were predicted by the model are illustrated herein to validate the developed model. **Figure 3-16** shows the model predictions for 2017 which are linked to two of the actual flood disaster occurrences that were reported in 2017.

An extraordinary warm winter resulted in a much higher water level in Lake Ontario which were raised 1.0 m beyond its normal levels, thus leading to flooding Toronto Island in May 2017 (Longley, 2017). These floods caused more than 40% of the islands to be overrun by water which resulted in a rehabilitation cost of about 7.38 million dollars (Roberts, 2017). These floods are predicted by the model as model predictions indicated that flood disaster will take place in proximity of Niagara watershed which is the closest to where Toronto Islands is located. Furthermore, Ottawa experienced devastating floods in May 2017. These floods turned fields and roads into lakes, destroyed homes, and led to the evacuation of about 500 homes which led the provincial government of Quebec to turn to the army to help mitigate the impacts of this disaster (Lynch & Meagher, 2017). The model was able to predict these floods by indicating that a flood disaster will take place in proximity of Rideau watershed.

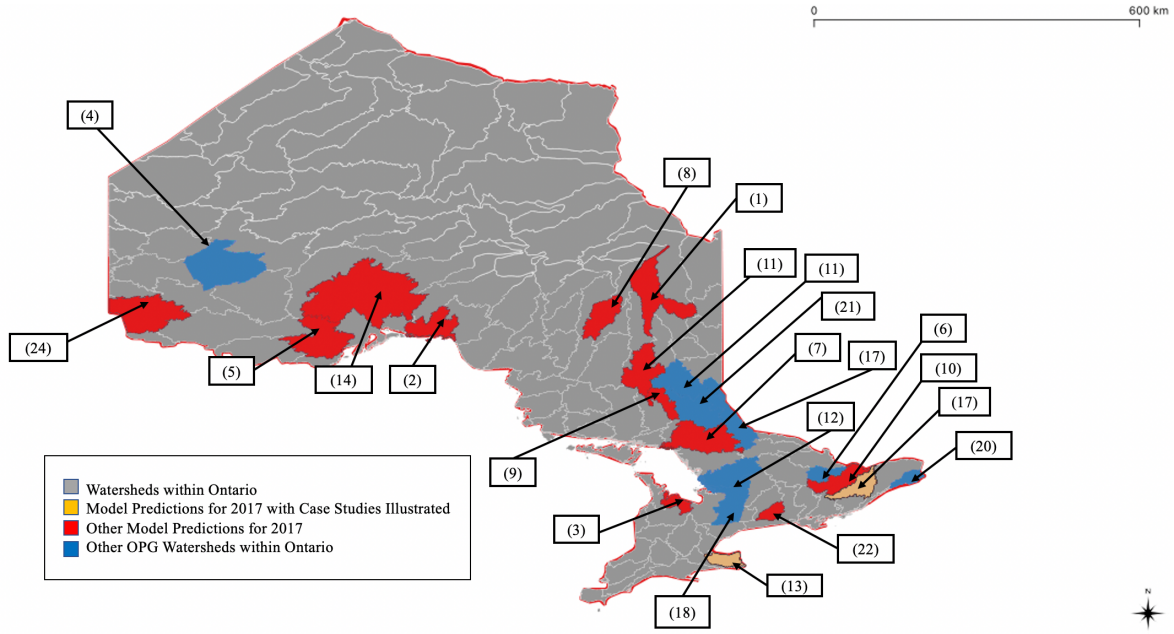


Figure 3-16: Model Validation in 2017

The occurrence of both the 2017 Toronto Islands and the 2017 Ottawa River flood disasters further validates the model. The ability of the model to predict disasters supports the evidence that a link actually exists between climate change and natural disasters. The importance of this link resides in the fact that accurate disaster predictions can be reached given the availability of climate variability data. In this paper the case study was applied using annual climate change indices as climate variability data and the Canadian Disaster Database as historical disaster records to train and test the proposed model. However, the use of higher resolution indices would certainly improve the model's accuracy and overall utility. Yet, currently available climate change indices are calculated in each location on a yearly basis as "the compilation, provision, and update of a globally complete and readily available full resolution daily dataset is a very difficult task" (Expert Team on Climate Change Detection and Indices, 2009). Nevertheless, the Expert Team on Climate Change Detection and Indices is currently working on higher resolution data which can enhance

the utility of the developed model as such higher resolution data become available in the future.

3.4. CONCLUSIONS

The aim of this paper is the prediction of CID occurrence in an attempt to enhance urban centers resilience under such disasters. The modelling approach proposed herein employs machine learning techniques to develop a spatio-temporal model for disaster prediction. The developed model links previous disaster records with climate change indices which represent the change in temperature and precipitation on a yearly basis. The spatial-temporal model aims to predict both whether or not a disaster will occur and where it would take place on a yearly basis. The developed model is divided into four different stages which are: (1) Model Architecture Analysis; (2) Input Variables Analysis; (3) Model Selection and prediction, and (4) Model Simulation. To test the applicability of the proposed model, a case study was presented which focuses on flood disaster prediction in Ontario. In the case study, a machine learning model was developed using disaster data from the Canadian Disaster Database together with calculated historical and future climate change indices data (Wazneh et al., 2019). Upon testing, the model was able to predict flood disasters occurrence in Ontario with an average error of 4%. This work is considered the first step in CID prediction, based on historical disaster data, global climate models, and climate change metrics, in an attempt to maximize urban resilience and mitigate CID impacts on cities worldwide.

3.5. ACKNOWLEDGEMENTS

The authors are grateful to the financial support of the Ontario Trillium Scholarship Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors would also like to acknowledge the fruitful discussions with the research teams of the NSERC-CaNRisk-CREATE program and the INViSiONLab. In addition, the authors would like to acknowledge the support of Dr. Hussein Wazneh and the Canadian Strategic Research Network on Floods (FloodNet).

3.6. DECLARATIONS

Funding The authors are grateful to the financial support of the Ontario Trillium Scholarship Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflicts of interest/Competing interests Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Authors' contributions Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication The authors confirm that the consent for publication is granted.

3.7. REFERENCES

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., … Kiru, M. U. (2019). Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. *IEEE Access*, *7*(February 2017), 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545

Alemany, S., Beltran, J., Perez, A., & Ganzfried, S. (2018). Predicting Hurricane Trajectories using a Recurrent Neural Network. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 468–475.

American Association for the Advancement of Science. (2011). Rising Temperatures Bringing Bigger Floods. *Science*, *331*(6020), 994–994. https://doi.org/10.1126/science.331.6020.994-a

Babu, S. (2017). Hazard vs Disaster: The principle behind disaster management. Retrieved May 29, 2019, from https://eco-intelligent.com/2017/01/21/hazard-vs-disaster-the-principle-behind-disaster-management/

Bi, D., Dix, M., Marsland, S. J., O'Farrell, S., Rashid, H. A., Uotila, P., … Puri, K. (2013). The ACCESS coupled model: Description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, *63*(1), 41–64. https://doi.org/10.22499/2.6301.004

Block, K., & Mauritsen, T. (2013). Forcing and feedback in the MPI-ESM-LR coupled model under abruptly quadrupled CO2. *Journal of Advances in Modeling Earth Systems*, *5*(4), 676–691. https://doi.org/10.1002/jame.20041

Bui, D. T., Hoang, N. D., Martínez-Álvarez, F., Ngo, P. T. T., Hoa, P. V., Pham, T. D., & Costache, R. (2020). A novel Deep Learning Neural Network approach for predicting flash flood susceptibility: A Case study at a high frequency tropical storm area. *Science of The Total Environment*, *701*(134413). https://doi.org/10.1016/j.neubiorev.2019.07.019

Callery, S. (2018). Effects | Facts – Climate Change: Vital Signs of the Planet. Retrieved December 5, 2018, from https://climate.nasa.gov/effects/

Chappell, C. (2019). Natural disasters cost $91 billion in 2018, according to federal report. Retrieved October 1, 2019, from CNBC website: https://www.cnbc.com/2019/02/06/natural-disasters-cost-91-billion-in-2018-federal-report.html

Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y., & Kim, H. S. (2018). Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data. *Advances in Meteorology*, *2018*. https://doi.org/10.1155/2018/5024930

Cholissodin, I., & Sutrisno, S. (2018). Prediction of Rainfall using Simplified Deep Learning based Extreme Learning Machines. *Journal of Information Technology and Computer Science*, *3*(2), 120. https://doi.org/10.25126/jitecs.20183258

CHRISTINA, N. (2019). Floods—facts and information. Retrieved June 6, 2019, from https://www.nationalgeographic.com/environment/natural-disasters/floods/

Chylek, P., Li, J., Dubey, M. K., Wang, M., & Lesins, G. (2011). Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmospheric Chemistry and Physics Discussions*, *11*(8), 22893–

22907. https://doi.org/10.5194/acpd-11-22893-2011

Climate change | EU Science Hub. (2018). Retrieved November 15, 2018, from https://ec.europa.eu/jrc/en/research-topic/climate-change

Collier, M., Jeffrey, S., Rotstayn, L., Wong, K., Dravitzki, S., Moeseneder, C., … Atif, M. (2011). The CSIRO-Mk3.6.0 Atmosphere-Ocean GCM: Participation in CMIP5 and data publication. *MODSIM 2011 - 19th International Congress on Modelling and Simulation - Sustaining Our Future: Understanding and Living with Uncertainty*, (December), 2691–2697.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J. L., Fichefet, T., Friedlingstein, P., & Shongwe, M. (2013). Long-term climate change: Projections, commitments and irreversibility. In *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1029–1136). https://doi.org/10.1017/CBO9781107415324.024

Cong, R. G., & Brady, M. (2012). The interdependence between rainfall and temperature: Copula analyses. *The Scientific World Journal*, *2012*. https://doi.org/10.1100/2012/405675

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *4*(2), 303–314. https://doi.org/10.1007/BF02836480

Czanner, G., Sarma, S. V., Ba, D., Eden, U. T., Wu, W., Eskandar, E., … Brown, E. N. (2015). Measuring the signal-to-noise ratio of a neuron. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7141–7146.

https://doi.org/10.1073/pnas.1505545112

Dong, Y., & Li, D. (2012). Efficient and effective algorithms for training single-hidden-layer neural networks. *Pattern Recognition Letters*, *33*(5), 554–558. https://doi.org/10.1016/j.patrec.2011.12.002

Donges, N. (2019). 4 Disadvantages Of Neural Networks | Built In. Retrieved November 30, 2010, from https://builtin.com/data-science/disadvantages-neural-networks

Dormehl, L. (2019). What is an artificial neural network? Here's everything you need to know | Digital Trends. Retrieved November 30, 2020, from https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

Expert Team on Climate Change Detection and Indices. (2009). Climate Change Indices - Definitions of the 27 core indices. Retrieved February 1, 2019, from http://etccdi.pacificclimate.org/indices_def.shtml

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., … Rummukainen, M. (2013). IPCC AR5. WG1. Chap. 9. Evaluation of Climate Models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 741–866. https://doi.org/10.1017/CBO9781107415324

Ganguly, K., Nahar, N., & Hossain, M. (2019). A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh. *International Journal of Disaster Risk Reduction*, *34*(December 2018), 283–294. https://doi.org/10.1016/j.ijdrr.2018.12.002

Gent, P., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., … Zhang, M. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Giffard-roisin, S., Yang, M., Charpiat, G., Kégl, B., Giffard-roisin, S., Yang, M., … Giffard-roisin, S. (2018). *Deep Learning for Hurricane Track Forecasting from Aligned Spatio-temporal Climate Datasets*.

Global Risks Report. (2019). Retrieved from http://www3.weforum.org/docs/WEF_Global_Risks_Report_2019.pdf

Griffies, S., Winton, M., Donner, L., Horowitz, L., Downes, S., Farneti, R., & Palter, J. (2010). GFDL's CM3 Coupled Climate Model : Characteristics of the Ocean and Sea Ice Simulations. *Journal of Climate*, *24*(13), 3520–3544. https://doi.org/https://doi.org/10.1175/2011JCLI3964.1

Hagan, M. T., Demuth, H. B., & Beale, M. (1997). *Neural network design*. https://doi.org/10.1007/1-84628-303-5

Hanewinkel, M., Zhou, W., & Schill, C. (2004). A neural network approach to identify forest stands susceptible to wind damage. *Forest Ecology and Management*, *196*(2–3), 227–243. https://doi.org/10.1016/j.foreco.2004.02.056

He, H., & Shen, X. (2007). A ranked subspace learning method for gene expression data classification. *Proceedings of the 2007 International Conference on Artificial Intelligence, ICAI 2007*, 358–364.

Heaton, J. (2017). Heaton Research The Number of Hidden Layers. Retrieved from

https://www.heatonresearch.com/2017/06/01/hidden-layers.html

Hinton, G. E. ., Osindero, S. ., & Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1109/TNN.2006.880582

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water (Switzerland)*, *10*(11), 1–16. https://doi.org/10.3390/w10111543

Jaafari, A., Zenner, E., Panahi, M., & Shahabi, H. (2019). Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. *Agricultural and Forest Meteorology*, *266–267*(2018), 198–207. https://doi.org/10.1016/j.agrformet.2018.12.015

Jaspreet. (2016). A Concise History of Neural Networks - Towards Data Science. Retrieved November 30, 2020, from https://towardsdatascience.com/a-concise-history-of-neural-networks-2070655d3fec

Kahira, A., Gomez, B., & Badia Sala, R. (2018). A Machine Learning Workflow for Hurricane Prediction. *Book of Abstracts. Barcelona Supercomputing Center*, 72–73.

Karl, T., Nicholls, N., & Ghazi, A. (1999). CLIVAR/GCOS/WMO Workshop on Indices and Indicators for Climate Extremes - Workshop summary. *Climatic Change*, *42*(1), 3–7. https://doi.org/10.1023/A:1005491526870

Kolp, P., & Riahi, K. (2009). RCP Database. Retrieved May 10, 2019, from http://www.iiasa.ac.at/web-apps/tnt/RcpDb/dsd?Action=htmlpage&page=welcome

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*(2–3), 195–215. https://doi.org/10.1023/a:1007452223027

Kumar, N. (2019). *Illustrative Proof of Universal Approximation Theorem*. Retrieved from https://hackernoon.com/illustrative-proof-of-universal-approximation-theorem-5845c02822f6

Kurt Hornik. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks, 4*(2), 251–257. https://doi.org/10.1109/72.363453

Liu, D., Zhang, H., Polycarpou, M., Alippi, C., & He, H. (2011). Advances in Neural Networks. *8th International Symposium on Neural Networks*, *9*. https://doi.org/10.1016/0020-7101(78)90038-7

Liu, J. J., Cheng, Z. L., & Su, P. C. (2014). The relationship between air temperature fluctuation and Glacial Lake Outburst Floods in Tibet, China. *Quaternary International*, *321*, 78–87. https://doi.org/10.1016/j.quaint.2013.11.023

Longley, R. (2017). After the flood: can Toronto Islands be saved from the next disaster? - NOW Magazine. Retrieved July 1, 2019, from https://nowtoronto.com/news/after-the-flood-toronto-island-preservation-climate-change/

Lynch, C., & Meagher, P. (2017). EASTERN ONTARIO: Flooding rocks urban areas, soaks down farmland and floods fields | FarmersForum. Retrieved July 1, 2019, from https://farmersforum.com/eastern-ontario-flooding-rocks-urban-areas-soaks-down-farmland-and-floods-fields/

Mann, M. L., Warner, J. M., & Malik, A. S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia. *Climatic Change*, *154*(1–2), 211–227. https://doi.org/10.1007/s10584-019-02432-7

Meehl, G., Washington, W. M., Arblaster, J. M., Hu, A., Teng, H., Kay, J. E., … Strand, W. G. (2013). Climate change projections in CESM1(CAM5) compared to CCSM4. *Journal of Climate*, *26*(17), 6287–6308. https://doi.org/10.1175/JCLI-D-12-00572.1

Min, S. K., Zhang, X., Zwiers, F. W., & Hegerl, G. C. (2011). Human contribution to more-intense precipitation extremes. *Nature*, *470*(7334), 378–381. https://doi.org/10.1038/nature09763

Mirza, M. M. Q. (2011). Climate change, flooding in South Asia and implications. *Regional Environmental Change*, *11*(SUPPL. 1), 95–107. https://doi.org/10.1007/s10113-010-0184-7

Mishra, V., Wallace, J. M., & Lettenmaier, D. P. (2012). Relationship between hourly extreme precipitation and local air temperature in the United States. *Geophysical Research Letters*, *39*(16), 1–7. https://doi.org/10.1029/2012GL052790

Muggah, R. (2019). The world's coastal cities are going under. Retrieved August 5, 2019, from https://www.weforum.org/agenda/2019/01/the-world-s-coastal-cities-are-going-under-here-is-how-some-are-fighting-back/

Nan, C., & Sansavini, G. (2017). A quantitative method for assessing resilience of interdependent infrastructures. *Reliability Engineering and System Safety*, *157*, 35–53. https://doi.org/10.1016/j.ress.2016.08.013

Nestler, G., & Jackman, A. (2014). 21st Century Emergency Management. Retrieved October 1, 2019, from IBM White Paper: Smarter Cities Thought Leadership website: https://www.ibm.com/downloads/cas/QVPL4PJK

Ongoma, V., Chen, H., & Gao, C. (2019). Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial East Africa. *Theoretical and Applied Climatology*, *135*(3–4), 893–910. https://doi.org/10.1007/s00704-018-2392-x

Peterson, T. (2005). Climate Change Indices. In *World Meteorological Organization Bulletin* (Vol. 54). https://doi.org/WMO, Rep. WCDMP-47,WMO-TD 1071

Peterson, T., Folland, C., Gruza, G., Hogg, W., Mokssit, A., & Plummer, N. (2001). Report on the activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998–2001. In *Rep. WCDMP-47, WMO-TD 1071*. https://doi.org/WMO, Rep. WCDMP-47,WMO-TD 1071

Prairie Climate Centre. (2018). How does Canada plan to reduce its Greenhouse Gas Footprint? Retrieved June 5, 2019, from http://prairieclimatecentre.ca/2018/05/how-does-canada-plan-to-reduce-its-greenhouse-gas-footprint/

Public Safety Canada. (2017). *2016-2017 Evaluation of the Disaster Financial Assistance Arrangements*.

Public Safety Canada. (2019). The Canadian Disaster Database. Retrieved January 10, 2019, from https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research*

(IJCBR), *5*(4).

Roberts, E. (2017). Environmental Impact of 2017 — Flooding at Toronto Islands. Retrieved July 1, 2019, from https://torontostoreys.com/environmental-flooding-toronto-islands/

Rodrigues, M., & De la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling and Software*, *57*, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003

Sanger, T. (1989). Optimal unsupervised learning in a single-layer network. *Neural Networks*, *2*, 459–473.

Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*, *104*(September 2018), 130–146. https://doi.org/10.1016/j.firesaf.2019.01.006

Shaftel, H. (2018a). Causes | Facts – Climate Change: Vital Signs of the Planet. Retrieved March 15, 2019, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/causes/

Shaftel, H. (2018b). Evidence | Facts – Climate Change: Vital Signs of the Planet. Retrieved November 15, 2018, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/evidence/

Srivastava, S. (2019). *On Foveation of Deep Neural Networks Sanjana Srivastava*. Massachusetts Institute of Technology.

Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of*

Remote Sensing, *30*(8), 2133–2147. https://doi.org/10.1080/01431160802549278

The Brookings Institution - London School of Economics project on Internal Displacement. (2012). The year that shook the rich: A review of natural disasters in 2011. Retrieved October 1, 2019, from https://www.brookings.edu/multi-chapter-report/the-year-that-shook-the-rich-a-review-of-natural-disasters-in-2011/

The International Federation of Red Cross and Red Crescent Society. (2019). Types of disasters: Definition of hazard. Retrieved March 8, 2019, from https://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/

Tongwen, W., Lianchun, S., Weiping, L., Zaizhi, W., Hua, Z., Xiaoge, X. I. N., … Mingyu, Z. (2014). An Overview of BCC Climate System Model Development and. *Journal of Meteorlogical Research*, *28*(1), 34–56. https://doi.org/10.1007/s13351-014-3041-7.Supported

Viola, R., Emonet, R., Habrard, A., Metzler, G., & Sebban, M. (2020). Learning from Few Positives: a Provably Accurate Metric Learning Algorithm to Deal with Imbalanced Data. *The 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, 2155–2161. https://doi.org/10.24963/ijcai.2020/298

Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Sénési, S., … Chauvin, F. (2013). The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics*, *40*(9–10), 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

Volodin, E., Dianskii, N., & Gusev, A. (2010). Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya - Atmospheric and Ocean Physics*, *46*(4), 414–431. https://doi.org/10.1134/S000143381004002X

Wasko, C., & Sharma, A. (2017). Global assessment of flood and storm extremes with increased temperatures. *Scientific Reports*, *7*(1), 1–8. https://doi.org/10.1038/s41598-017-08481-1

Wayne, G. (2013). The Beginner's Guide to Representative Concentration Pathways. Retrieved March 1, 2019, from https://skepticalscience.com/docs/RCP_Guide.pdf

Wazneh, H., Arain, A., & Coulibaly, P. (2017). Historical spatial and temporal climate trends in Southern Ontario, Canada. *Journal of Applied Meteorology and Climatology*, *56*(10), 2767–2787. https://doi.org/10.1175/JAMC-D-16-0290.1

Wazneh, H., Arain, M. A., & Coulibaly, P. (2019). Climate indices to characterize climatic changes across southern Canada. *Meteorological Applications*, *27*(1), 1–19. https://doi.org/10.1002/met.1861

Wazneh, H., Arain, M. A., Coulibaly, P., & Gachon, P. (2020). Evaluating the Dependence between Temperature and Precipitation to Better Estimate the Risks of Concurrent Extreme Weather Events. *Advances in Meteorology*, *2020*, 1–16. https://doi.org/10.1155/2020/8763631

World Health Organization. (2018). Climate change and health. Retrieved June 6, 2019, from https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

Wright, P. (2019). 2018 Global Disasters Cost $160 Billion; Climate Change a Factor, Report Says. Retrieved October 2, 2019, from The Weather Channel website: https://weather.com/science/environment/news/2019-01-09-disasters-cost-damage-climate-change

Zanchetta, A. D. L., & Coulibaly, P. (2020). Recent Advances in Real - Time Pluvial Flash Flood Forecasting. *Water*, *12*(2). https://doi.org/10.3390/w12020570

Chapter 4

INFRASTRUCTURE PERFORMANCE PREDICTION UNDER CLIMATE-INDUCED DISASTERS USING DATA ANALYTICS

ABSTRACT

The frequency of Climate-induced Disasters (CID) has tripled in the last three decades, driving the World Economic Forum to identify them as the most likely and most impactful risks worldwide. With more than 70% of the world population expected to be living in cities by 2050, ensuring the resilience of urban infrastructure systems under CID is crucial. The present work employs data analytics and machine learning techniques to develop a performance prediction framework for infrastructure systems under CID. The framework encompasses four stages related to: extracting meaningful information about the impact of CID on infrastructure systems and identifying the latter's performance; investigating the relationship between different CID attributes and previously identified system performance; employing data imputation using unsupervised machine learning techniques; and developing and testing a supervised machine learning model based on the different influencing CID attributes. To demonstrate its application, the developed framework is applied to disaster data compiled by the National Weather Services between 1996 and 2019 in the state of New York. The analysis results showed that: *i*) power systems in New York are the most vulnerable infrastructure to CID, and particularly to wind-related hazards; *ii*) power system performance level depends on hazard-system interactions rather than solely hazard characteristics; and *iii*) a 4-predictors random forest-based model can effectively predict power system performance with an accuracy of 89%. This work is expected to aid

stakeholders in developing spatio-temporal preparedness plans under CID, which can facilitate mitigating the adverse impacts of CID on infrastructure systems and improve their resilience.

Keywords: *climate-induced disasters, urban centres, infrastructure system, resilience, machine learning, data analytics*

4.1. INTRODUCTION

Climatological, meteorological, and hydrological hazards have been increasing in magnitude and frequency due to the changing climate (i.e., temperature, precipitation, and humidity) [1]. The risks due to such hazards to urban areas can hinder daily activities, incur costly damages, and contribute to life losses, which is the reason why, when such risks are realized, they are often referred to as *disasters* [1]. More specifically, a disaster is manifested only when: *i*) a hazard is realized; *ii*) a vulnerable system is exposed to that hazard; and *iii*) severe negative consequences strike this exposed system. The frequency and magnitude of Climate-Induced Disasters (CID) have increased dramatically over the past three decades [1] leading to identifying CID to be the top risk in terms of both likelihood and impact in 2020 [2]. Globally over the last decade, CID resulted in a mortality rate of approximately 60,000 people per year [3]. In addition, around 25% of the world's population live in coastal areas threatened by CID such as storm surges and tsunamis [4]. In the United States alone, the average annual number of CID causing economic losses of more than $1 billion has increased from 2.9 CID from 1980 to 1989 to 11.9 CID from 2010 to 2019 (cost adjusted using Consumer Price Index) [5]. It should also be noted that the increase in both population and "material wealth" contributes to such increasing CID-induced losses [5]. Moreover, a total of $1.75 trillion were reported as CID-related damages, with windstorms being the most impactful meteorological events that caused more than $1 trillion economic losses and 5,000 fatalities over the past four decades [6–8]. Among the different types of windstorms that affected the United States, thunderstorms and tornado winds alone caused around $26 billion as economic losses between 2005 and 2015 [8].

With more than 70% of the world's population are expected to live in cities by 2050 [9], and given the adverse impacts of CID on urban areas and their infrastructure systems, developing tools for predicting the impacts of different CID on critical infrastructure systems is critically urgent [2]. As such, several studies were conducted to predict the frequency of CID and their social and economic impacts using different data analytics and machine learning techniques [10,11,12–19,20–22].

Data analytics aims at uncovering hidden information that cannot be explored through classic mathematical and statistical tools, and is generally divided into descriptive, predictive, and prescriptive analytics [23]. Descriptive analytics is concerned with analyzing historical data to understand the processes being studied, answering key questions about these processes, and subsequently drawing valuable conclusions. Building on such conclusions, predictive analytics aims at predicting the future behavior of systems and entities. Finally, prescriptive analytics focuses on finding the best future decision(s) supported by the outcomes of descriptive and predictive analytics. The three fields of data analytics were extensively employed to investigate CID consequences [24,25], derive meaningful relationships between the different attributes controlling CID [1,18,20,21,26], and develop effective risk reduction and mitigation strategies [17–19,22,27]. It is noteworthy that data analytics techniques can be applied using structured/unstructured data (i.e., text), quantitative/qualitative data, or a combination of different data types.

Machine learning is a branch of artificial intelligence is built on the premise that a computer model can *learn* through being exposed to data and information representing real world interactions. Using different algorithms, machine learning techniques can automatically find solutions to complex problems by identifying patterns and relationships

within datasets [28]. In general, machine learning can be classified as either supervised or unsupervised learning, in which the former uses labelled data to train and test a model, whereas the latter employs unlabelled data for the model development and testing. Both machine learning classes have experienced rapid advances in natural phenomena simulation and prediction, and have been recently adopted for identifying the different factors controlling flood damage and severity [10,29], estimating the number of hurricanes per season [16], calculating wildfire risk [15,30], predicting wind risk [12,14], estimating heavy rain impacts [13], and predicting tornado-related damages [11].

Although it is essential to predict the damages induced by CID on different critical infrastructure systems, most previous work focused on predicting CID occurrence and lumped impacts (i.e., without segregating CID impacts on the affected infrastructure systems). Therefore, the present work develops a systematic framework that can be used to predict infrastructure system damages under CID. As will be discussed next, the developed damage prediction framework (**DPF**) consists of three main phases: the input phase, the internal processes phase, and the output phase. The internal processes phase is further divided into four stages in order to facilitate the prediction of infrastructure system damages. To demonstrate its applicability and viability, the developed DPF was applied to the historical disaster data collected by the National Weather Services (**NWS**) between 1996 and 2019. This DPF provides a better conceptualization of CID impacts, which can aid the decision makers to develop effective preparedness plans and risk mitigation strategies under future CID risks. This can, in turn, improve the overall urban resilience under CID. The present work is organized as follows: a detailed description of the developed DPF including the methods and techniques to be implemented; a description of

the study used to demonstrate the applicability of the developed DPF; and finally, key decision-making insights and conclusions.

4.2. INFRASTRUCTURE SYSTEMS DAMAGE PREDICTION FRAMEWORK

The developed DPF shown in **Figure 4-1** provides a systematic approach for predicting infrastructure system damage under CID. The input phase comprises collecting the data required to predict infrastructure system damages such as event related narratives (i.e., a description of the hazard and its impacts), geographic characteristics of the affected location (i.e., coordinates), time-related attributes (i.e., start and end time of the hazard), hazard-related attributes (i.e., magnitude, duration, intensity), and any other attributes (i.e., climate-related, social-related and economic-related). It is worth mentioning that, similar to any other data driven framework, the exactitude and certainty of the collected input data (i.e., event narratives) is key for the development of the DPF established herein. The internal processes phase consists of four main stages, where in Stage 1 the link between CID and infrastructure systems is established in two steps: the systems affected by CID are identified; and the distinct damages are subsequently defined. This can be achieved through mining the text data describing the hazards and their impacts provided in the input phase (i.e., event narratives). Upon identifying both the systems affected and their distinctive damages, influencing attributes are investigated and explicated to select those which will be employed in the model development and testing (i.e., Stage 2). In Stage 3, data imputation is performed in an attempt to enhance the predictive capability of the DPF. The predictive model is then developed in Stage 4, where different techniques can be used such as regression or classification trees with- and without ensample techniques (i.e., bagging, boosting, and random forests). The model is subsequently tested, and several evaluation

criteria can be used to assess the model performance such as the mean squared error, misclassification error, and the confusion matrix measures (i.e., precision, recall, f1-score). It is noteworthy that the use of ensemble techniques in Stage 4 can significantly boost the model accuracy, as will be discussed in greater detail next. It is important to note that both filled and unfilled data records can be used for the model development, and their corresponding performances can be compared. Finally, the output phase of the DPF involves predicting the infrastructure system damages which may include continuous monetary damages or damage severity classes as will be further illustrated.

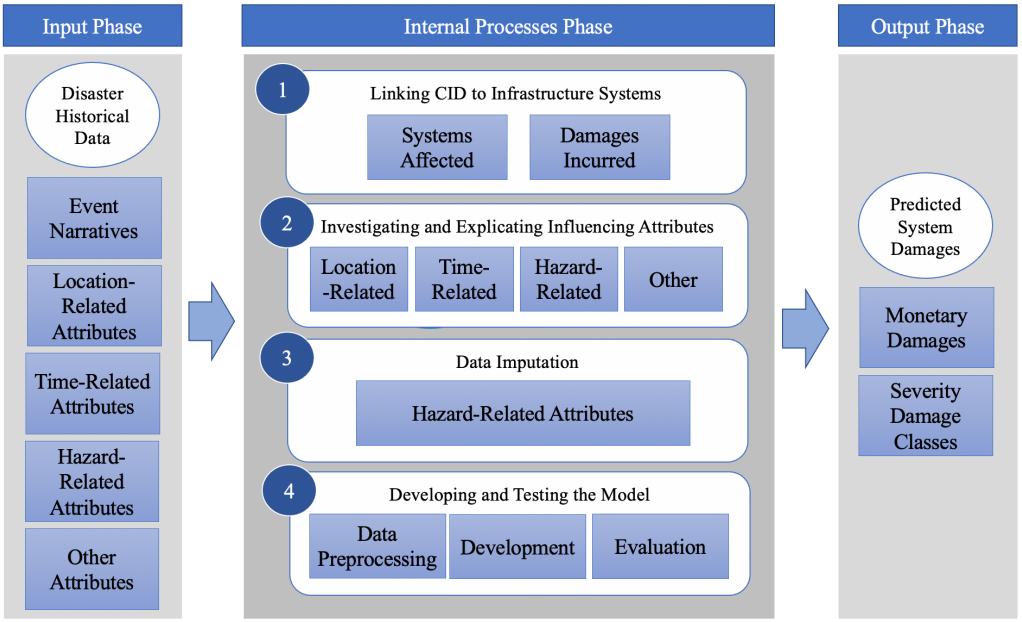


Figure 4-1: A Schematic of the Damage Prediction Framework

4.2.1. Stage 1: Linking CID to Infrastructure Systems

To explore the impacts of CID on infrastructure systems, event narratives describing the CID and their impacts are collected and investigated using text analytics as illustrated in **Figure 4-2**. Text analytics works on converting text data into quantifiable information in

order to extract patterns and draw viable conclusions [31]. Generally, text analytics is used to analyze text data through either *semantic parsing*, pertaining to the word type (i.e., being a positive or negative word); or *bag of words*, in which words are treated as single tokens without considering the word type or order [32,33]. The present DPF employs the bag of words analysis technique to investigate the characteristics and impacts of CID affecting infrastructure systems due to the insignificance of the word type in this type of investigation. This can be implemented using any available commercial or open-source packages such as the *tm_map* package (available in the *R* language) which will be employed in the framework demonstration study. Prior to applying the bag of words analysis, text data are preprocessed to convert it into quantifiable information. The preprocessing steps include [34]: *i*) transformation, where all words are converted into lower case format to avoid having the same word repeated in upper- and lower-cases; *ii*) tokenization, where the unstructured text is converted into words; *iii*) treatment, where a standard filter "stop" list is used to remove common words (i.e., the, to, a, an, and, or); and, *vi*) stemming, where all affixes are removed in order to return words to their roots. The resulting group of words is used to identify the frequent infrastructure systems being affected by CID based on a word frequency analysis. These systems can be subsequently linked to the CID that contribute mostly to their damage through a system-disaster association process.

To investigate the severity of system damages, an N-gram analysis [35] can be employed to estimate the frequency of *N* associated words. Consequently, the frequencies of *N* associated words are evaluated, and the system damage is classified according to the corresponding level of damage severity. For instance, damage severity levels can be

discriminated into the following wide-ranging categories: 1) no damage; 2) damage at the component-level; and 3) damage at the system-level. The N-gram analysis can be conducted using for example, the *N-gram* [36] package available *R*.

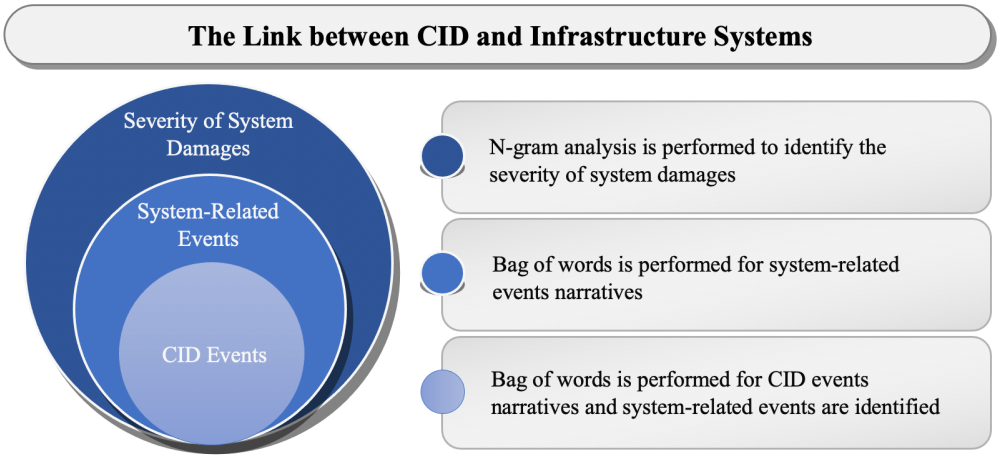


Figure 4-2: Establishing the Link between CID and Infrastructure Systems

4.2.2. Stage 2: Investigating and Explicating Influencing Attributes

At this stage, the interrelationship between the different inputs (i.e., location-, time-, hazard-related and any other attributes) and outputs (i.e., monetary damage or level of damage severity) are thoroughly investigated to determine the inputs that should be included in the predictive model. Another objective of this stage is to explicate the identified influential factors in order to have a better understanding of the CID-system interaction. Such exploration can be conducted through exploratory-, sensitivity-, and correlation analyses; a feature selection/extraction process; or a combination of both. For example, in the demonstration study, the geographic distribution of each damage class can

be investigated to decide whether the spatial attribute of CID need to be included in the prediction stage.

4.2.3. STAGE 3: DATA IMPUTATION

The development of a data-driven model necessitates having a database that is as complete as possible, which is often very challenging. Several alternatives have therefore been proposed for filling missing data instances (i.e., data imputation), including: a complete removal from the database; a replacement with the average instance encountered; or, the use of unsupervised machine learning to divide the database into clusters and generate missing instances accordingly [37–39]. The latter approach is preferred over other former alternatives as the removal of missing instances decreases the model accuracy and the use of an average value to replace missing instances requires the underlying variable to be normally distributed [38,40,41] and ignores correlation [42] unless used with large datasets [43]. Clustering is an unsupervised learning technique that aims at grouping instances based on their degree of similarity [44]. Several clustering algorithms have been developed over the past decades such as K-Means Clustering (**KMC**) and Model-Based Clustering (**MBC**). KMC aims at grouping observations through minimizing their distance to the cluster center. KMC is the most widely applied clustering algorithm due to its simplicity and ability to partition data into clusters with a spectrum of shapes and sizes. However, the application of KMC requires predefining the number of clusters (K). Therefore, K-Means clustering is typically applied through changing K between 2 and X, where X is the maximum number selected by the user. The within-cluster-sum-of-squares (**WCSS**) is subsequently employed to determine the optimum number of clusters, where the highest drop in the value of wcss corresponds to the optimum K value. MBC is, instead, used to

discretize observations into clusters based on an appropriate finite mixture model [45], where each of the resulting clusters is defined as a unimodal component within that model [45]. A *Gaussian* mixture model is typically employed, with parameters estimated using an Expectation Maximization algorithm. Bayesian Information Criterion (**BIC**) is subsequently used to estimate the optimum number of clusters (i.e., *K*), where the optimum number of clusters corresponds to the maximum BIC value. It is important to note that MBC is most often preferred over KMC as it does not require a prior definition of *K*. Both KMC and MBC, together with many other unsupervised machine learning techniques [37–39], can be used within the DPF for filling the missing data records; however, the framework demonstration study uses the KMC and MBC only. The application procedures of both approaches are summarized in **Figure 4-3**. After identifying the optimum number of clusters, the CID records are assigned to clusters and missing data records (i.e., hazard-related attributes) are replaced by the corresponding cluster average.

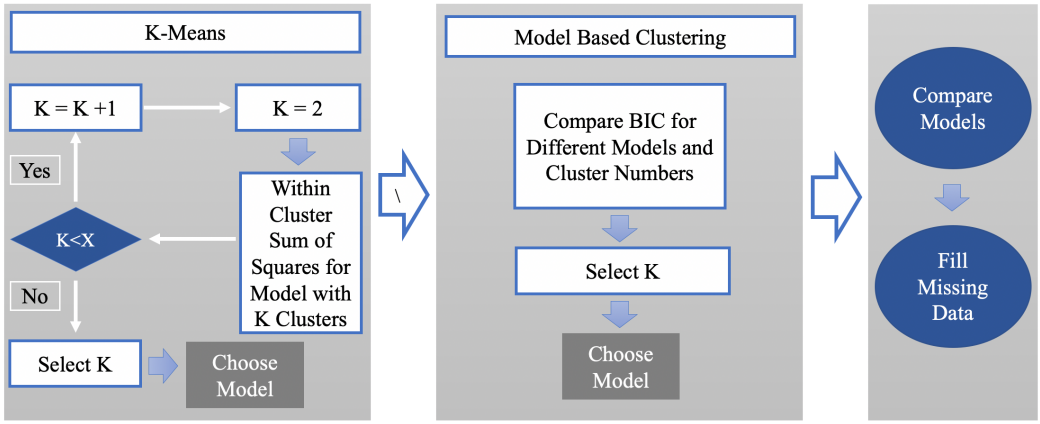


Figure 4-3: The Application Procedures of KMC and MBC for Data Imputation

4.2.4. Stage 4: Model Development and Testing

After filling the missing data records, a supervised machine learning model is developed to predict infrastructure system damages under CID. Unlike cluster analysis, supervised machine learning techniques rely on training and testing the model using specific input-output pairs. A supervised machine learning model can be developed to predict continuous outputs (i.e., regression trees) or distinct classes (i.e., classification trees). Unlike regression trees which predict numerical outputs, classification trees discriminate instances and allocate them to different classes [51]. In both cases, the dataset is divided into training and testing subsets. The training subset is used for model development, whereas the testing subset is used to assess the model generalizability to the whole dataset. Several evaluation criteria are thus used to evaluate the model performance in both the development and testing stages. The mean squared error is commonly utilized to evaluate regression trees performance, whereas misclassification error and confusion matrix measure are typically used for evaluating the performance of classification trees.

To improve the model accuracy, ensemble techniques such as bagging, random forest, and boosting [46] can be used. Bagging is an ensemble technique in which M bootstrap trees are generated from the training subset, and the resulting predictions are combined using the majority vote [13,47]. Random forest is a modification of the bagging technique in which M trees are distinguished from one another using a random sample of $\mathcal{M}$ predictors at each split (i.e., sampling with replacement) rather than using all predictors [13,47]. This facilitates obtaining uncorrelated predictions, and thus enhance the model accuracy [13,47]. The value of $\mathcal{M}$ for trees with P predictors should be equal to $\sqrt{P}$ [48,49]. Boosting is another ensemble technique that depends on sequentially developing n trees (classifiers) from the training subset, each with d splits and a shrinkage parameter (λ)

[48,49]. To further enhance the model accuracy, classifiers are combined in each iteration and the misclassified data points are assigned higher weights so that they can be classified correctly during the following iterations [13].

4.3. FRAMEWORK DEMONSTRATION STUDY

4.3.1. DATA DESCRIPTION

The disaster database provided by the NWS, a sub-agency under the National Oceanic and Atmospheric Administration (**NOAA**), is exploited to demonstrate the applicability of the DPF developed in the present study. The NWS database outlines different types of CID that affected the United States between 1950 and 2019 [50], and includes: *i*) storms and other weather phenomena that caused loss of life, injuries, significant property damage, and/or disruption to commerce; *ii*) rare, unusual, weather phenomena that generate media attention (i.e., snow flurries in South Florida or San Diego coastal areas); and, *iii*) severe meteorological events (i.e., maximum/minimum temperature, precipitation coupled with other events). From 1950 to 1995, only tornados, thunderstorm wind, and hail events were recorded by the NWS. As of 1996, more than 45 event types were added to the NWS database resulting in a total of 1,355,969 records from 1996 to 2019. Event types included in the NWS fall under the three types of CID discussed earlier (i.e., meteorological, hydrological, and climatological). Each recorded event is characterized by several variables categorized into location-, time-, and hazard-related. The DPF was applied using the NWS database to enable the prediction of infrastructure system damage severity within the state of New York following a classification approach.

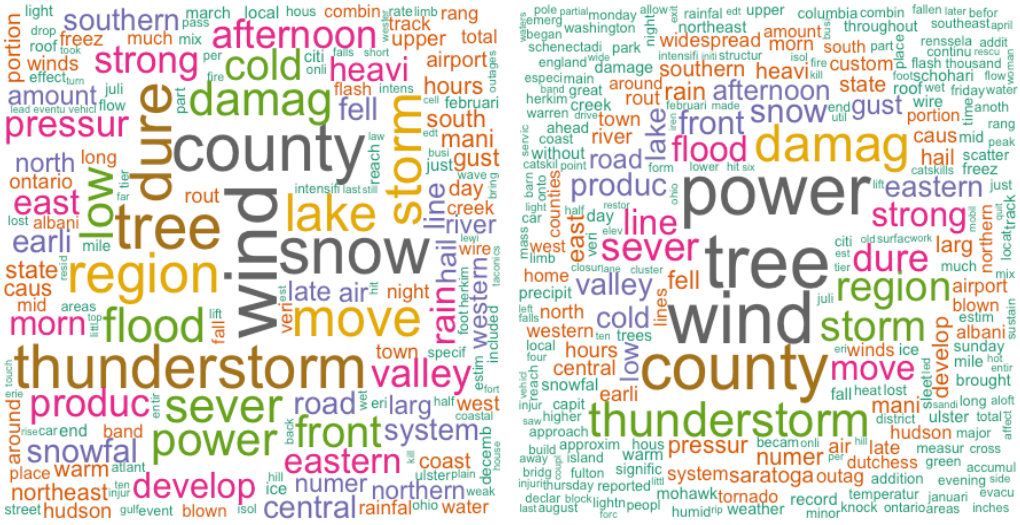
4.3.2. STAGE 1: LINKING CID TO INFRASTRUCTURE SYSTEMS

A bag of words analysis is used to estimate the word frequency based on episode and event narratives of CID records in New York State between 1996 and 2019, as shown in **Figure 4-4**(a). The word *"wind"* is the most frequent among all words mentioned in the CID narratives, followed by *"thunderstorm"*, *"snow"*, *"storm"*, and *"damage"*. The word *"tree"* is also among the highly mentioned words. For the affected infrastructure systems, the two highly mentioned words were *"power"* followed by *"road"*. The results from the bag of words analysis support that interruptions of power and transportation systems across New York state were mainly due to wind hazards, and also highlights that such interruptions may be related to fallen trees (an indirect effect of wind).

To explore the dominant CID affecting power infrastructure (i.e., the most affected system based on the bag of words analysis), **Figure 4-4**(b) shows the results of the bag of words analysis based on episode and event narratives of power-related CID only. The results of such analysis support the strong relationship between *power* and *wind* as the word *"wind"* was also most frequent among the narratives of the power-related CID affecting New York. The word *"line"* also appeared frequently in these narratives, which indicates that the power system damage may be at the component level. It is worth mentioning that the power system in New York State may be also vulnerable to thunderstorms as the frequency of the word *"thunderstorm"* was relatively high. However, the vulnerability of New York's power system to wind-related hazards only is considered in the current case study.

A bi-gram analysis (i.e., N-gram analysis with N=2) was further conducted to uncover the common power system damage scenarios due to CID. The frequency of two-associated words highlighted that the power system damage can be in a form of: a power

line damage, a power pole damage, and an overall system failure. These scenarios represent component- (i.e., power line or pole damage) or system- (i.e., power outage) failures. The wind-related hazards affecting New York are thus divided into those causing no power damage (Class 1), a damage at the component level (Class 2), and a damage that led to a system-wide power failure (Class 3). The number and percentage of hazards allocated to Classes 1, 2, and 3 are 8,547 and 58%, 2,348 and 16%, and 3,932 and 26%, respectively. This supports the higher vulnerability of the power system in New York to wind-related hazards as around 42% of such hazards (Classes 2 and 3) caused adverse impacts to the



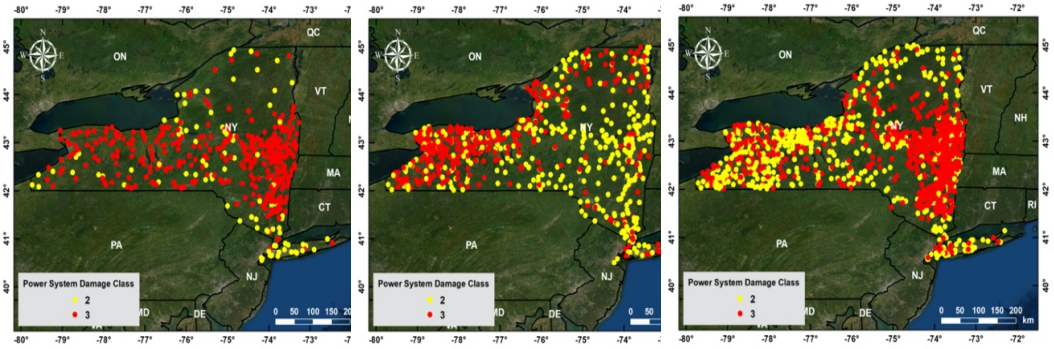
system.

Figure 4-4: Bag of Words Based on: (a) Episode and Event Narratives of all CID Records, and (b) Episode and Event Narratives of Power-Related CID Records only

4.3.3. STAGE 2: INVESTIGATING AND EXPLICATING INFLUENCING ATTRIBUTES

The relationship between each of the location-related, time-related, and hazard-related attributes and the damage class is investigated to determine which of these attributes can

significantly influence the power system damage severity. **Figure 4-5** shows the spatio-temporal distribution of each damage class for the periods of 1996-2003, 2004-2010, and 2011-2019. Prior to 2004, hazards causing a system failure (i.e., Class 3) overnumbered those causing a component damage (i.e., Class 2) and impacted the whole state, particularly the eastern part. Between 2004 and 2010, hazards causing a component damage (i.e., Class 2) were more common and were scattered across the state. However, hazards causing a system failure (i.e., Class 3) occurred significantly more frequently at the western part of the state. After 2010, the western part of New York experienced more frequent Class 2 damages, the eastern part experienced a very high frequency of Class 3 with relatively a low frequency of Class 2, and the central part experienced a low frequency of both classes.



(a) (b) (c)

Figure 4-5: Distribution of Wind-Related Hazards Affecting New York (a) between 1996-2003, (b) 2004-2010, and (c) 2011-2019

Figure 4-6 shows boxplots for the magnitude and duration of the wind-related hazard allocated to the three damage classes. It can be observed that the median and the 25th percentile of the wind magnitude are nearly similar for all classes (**Figure 4-6**(a)). This indicates that around 50% of hazards allocated to the three classes are of the same intensity. In addition, the maximum, minimum, and 75th percentile of the wind magnitudes for the

hazards allocated to Classes 1 and 2 (i.e., causing no damage and component-level damage, respectively) are identical, which supports the observation that the statistical distributions of wind magnitude are nearly the same in these two classes (as the mean value is slightly different between the two classes). On the other hand, for Class 3 (i.e., power outage), wind magnitudes have a wider range compared to the other two classes. Furthermore, the 75th percentile of the wind magnitude is much higher for Class 3. This indicates that the magnitude of wind-related hazards causing power outages is expected to be higher than those causing either no damage or a damage at the component level. The mean wind magnitude of Class 3 was also found to be higher than that of Classes 1 and 2, indicating that, on average, the power damage gets more severe with the increasing wind magnitudes. For the duration of wind-related hazards, it is apparent from **Figure 4-6**(b) that the wind duration is not directly related to the damage severity, as the mean, median, maximum, and 75th percentiles are all smaller for Class 2 followed by Classes 3 then 1. Overall, the analysis asserts that the severity of the power system damage depends on the complex hazard-system interactions rather than the hazard characteristics only supporting the need for a machine learning-based model to tackle the complexity associated with predicting the power system damage under CID.

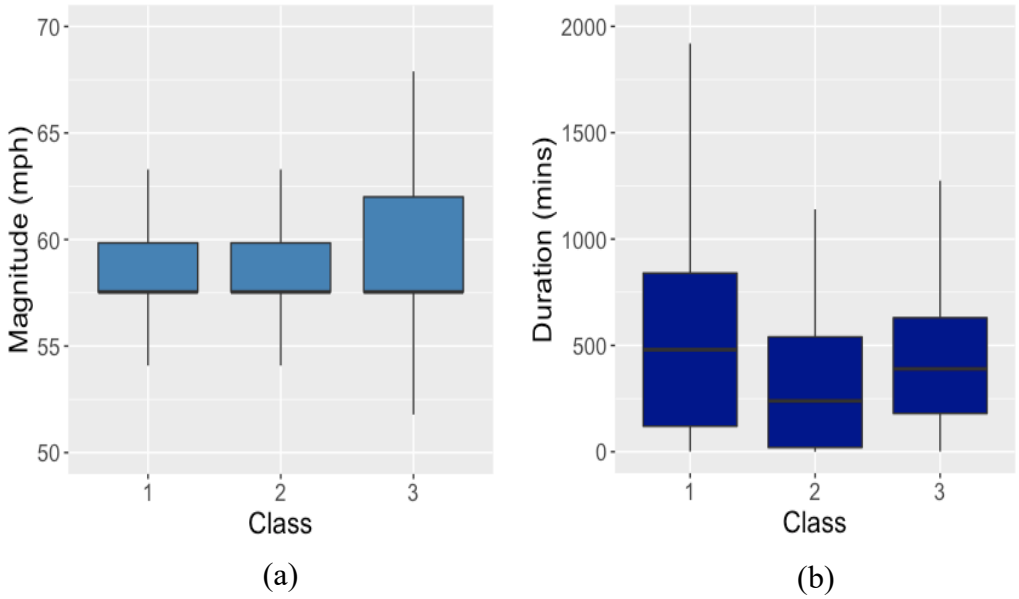
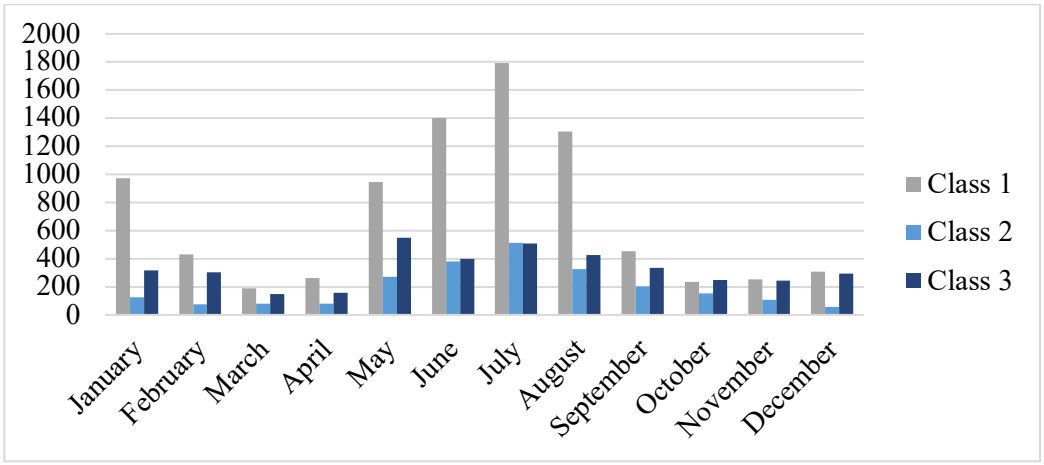
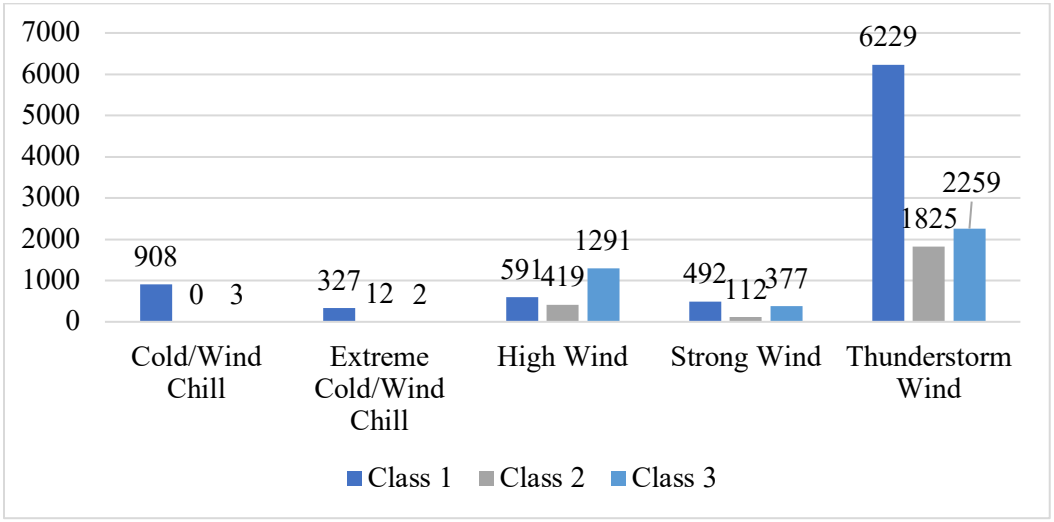


Figure 4-6: Boxplots for the (a) Magnitude and (b) Duration of the Wind-Related Hazards that Affected New York verses the Three Power Damage Classes

Finally, the monthly frequency of the wind-related hazards allocated to Classes 1, 2, and 3 indicate that most of the power system damages (i.e., Classes 2 and 3) took place over May through August (**Figure 4-7**(a)). In addition, the distribution of the different types of wind-related hazards among Classes 1, 2, and 3 show that most of the power system failures and component damages are due to thunderstorms and high winds (**Figure 4-7**(b)). Accordingly, both the occurrence month and type of the wind-related hazard should be considered in the model to account for the influence of both variables on the severity of the power system damage.



(a)



(b)

Figure 4-7: The Distribution of Power system Damage Classes over (a) Months and (b) Wind Types

4.3.4. STAGE 3: DATA. IMPUTATION

Filling the missing magnitude and duration records of the wind-related hazards affecting New York is key to develop a model capable of predicting the three power system damage classes discussed earlier. Therefore, KMC and MBC are used to cluster all of the wind-related hazards affecting New York based on the hazard type, latitude, longitude, month, year, duration, and magnitude. The best clustering algorithm is subsequently selected based

on the performance of the resulting model and is used for filling the missing records. **Figure 4-8** and **Figure 4-9** show K-WCSS and K-BIC relationships for wind-related hazards allocated to the three classes of the power system damage, respectively. For Class 1 (i.e., no damage), the use of nine clusters resulted in the minimum WCSS and the maximum BIC values. Accordingly, hazards allocated to Class 1 were divided into nine clusters and the missing magnitudes and durations were replaced by the corresponding cluster average. Of the 8,545 hazards allocated to Class 1, a total of 1,299 magnitude and 5,850 duration records were missing. All of these missing records were filled except for 51 magnitude records as they were in the same cluster. For Class 2, it can be observed that the elbow of the K-WCSS relationship is at six clusters whereas the maximum BIC value is achieved through five clusters only. Hence, hazards allocated to Class 2 (2,348 records) were discriminated into six clusters and the missing magnitude and duration records were filled similar to those of Class 1. Out of 339 missing magnitude records and 1,791 missing duration records, a total of 297 magnitude records were discarded as they were in the same cluster. For Class 3, the optimum number of clusters was found to be seven according to both KMC and MBC. Therefore, hazards allocated to Class 3 (3,932 records) were categorized into seven clusters and all of the missing magnitudes (303 records) and durations (2,364 records) were filled similar to those in Classes 1 and 2.

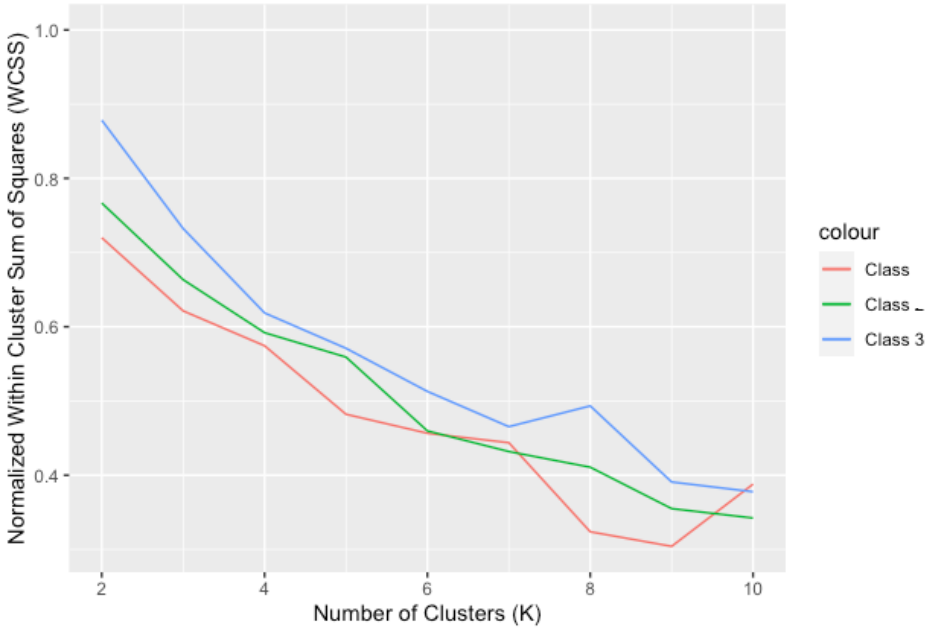


Figure 4-8: WCSS for Class 1 (a), Class 2 (b), and Class 3 (c)

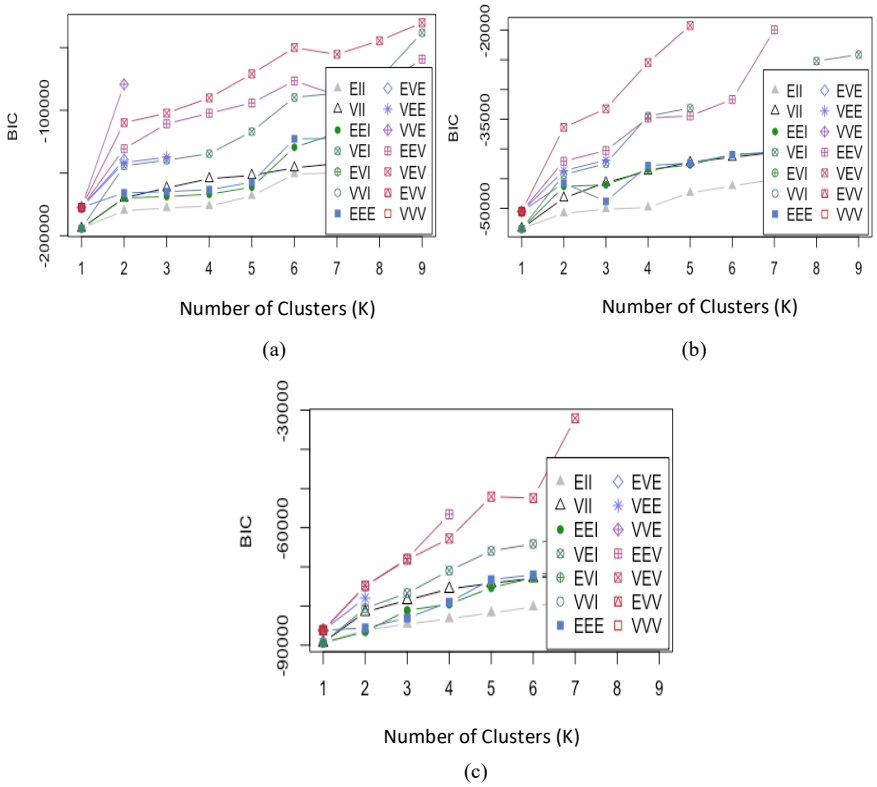


Figure 4-9: BIC for Class 1 (a), Class 2 (b), and Class 3 (c)

It should be emphasized that data imputation can introduce bias in the developed model if the percentage of missing records is more than 10% of the total number of records available [51,52]. Accordingly, since 13% and 67% of magnitude and duration records, respectively, are missing, both filled and unfilled data will be used for the development of the predictive model to assess the effect of the data imputation process.

4.3.5. STAGE 4: MODEL DEVELOPMENT AND TESTING

The CID dataset is divided into training and testing subsets, where the training subset includes 70% of the CID and the testing subset includes the remaining 30% of the CID. The model inputs include the location-related, time-related and hazard-related attributes investigated in Stage 2, whereas the model output is the damage class of the infrastructure system. Two classification models are accordingly developed, where: Model 1 employs the filled data whereas, Model 2 uses the actual dataset (before filling). Both models are trained using 70% of the dataset (filled or unfilled records) and are subsequently tested using the remaining 30%. For each of the two models, classic classification tree together with three ensemble techniques (bagging with 1,000 trees, random forest with 2, 3, or 4 predictors at each split, and boosting with 5,000 trees, 4 splits, and a shrinkage parameter of 0.01) are employed. After training, the performance of each model (and associated ensemble technique) is assessed based on its ability to replicate the testing subset. This is achieved through evaluating the misclassification error (i.e., accuracy) and other confusion matrix measures (i.e., precision, recall, and f1-score), as summarized in Table 4-1. The overall model accuracy is calculated as the ratio between the number of true predictions and the total number of data instances in the testing subset. The other confusion matrix evaluation criteria are class-related, and include: the precision, recall (i.e., sensitivity), and f1-score

[29]. The precision of the model for predicting a certain class can be conceptualized as the model exactness and is calculated as the relative number of true predictions within that class. The recall is a measure of the model completeness and is concerned with the total number of actual records within each class. The recall is calculated as the ratio between the total number of true predictions within a certain class and the total number of actual records within that class. Assessing the model performance can be more effective when the precision and recall are integrated, rather than using each measure separately. Accordingly, the f1-score can be used as it combines the model precision and recall into a single informative measure. The f1-score is related to the precision and recall through: $\text{f1-score} = (2\times\text{precition}\times\text{recall})/(\text{precision}+\text{recall})$.

To compare the accuracy of the models for both training and testing subsets, the overall model accuracy is also calculated for the training subset. The performance measures shown in **Table 4-1** support that the performance of Model 1 outweighs that of Model 2. This boosted performance might be attributed to the fact that Model 1 includes filled data, whereas Model 2 uses only non-missing-data records. To assess whether data imputation has introduced bias in Model 1, the classification trees for both models are compared. It is observed that the first two splits in Model 2 are wind event type and magnitude attributes, whereas, in Model 1 the first split is the duration attribute which had about 68% of its records missing as per the results of Stage 3. This implies that the enhanced performance of Model 1 may be attributed to the bias introduced by the data imputation, accordingly, Model 2 is considered more robust. Comparing the performance of the different ensemble techniques for Model 2, it is apparent that the 4-predictors random forest technique yields the best performance.

Table 4-1: Performance Measures of Different Classification Models

Model ID	Ensemble technique	Training Subset Accuracy (%)	Testing Subset Accuracy (%)	Precision (%)			Recall (%)			f1-score (%)		
				Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Model 1	**Classification Tree**	88.2/11.8	88.2/11.8	91.9	75.3	87.3	90.9	97.6	79.6	91.4	95	83.3
	Bagging	99.5/0.5	95.9/4.1	97.5	87.8	96.8	97.8	93.0	93.6	97.6	90.3	95.2
	Random Forest with 2 Predictors	99.5/0.5	96.9/3.1	98.9	90.4	96.2	97.9	95.2	95.8	98.4	92.8	96.0
	Random Forest with 3 Predictors	99.3/0.7	96.5/3.5	98.6	90.4	95.3	97.4	93.8	95.8	98.0	92.0	95.5
	Random Forest with 4 Predictors	99.4/0.6	96.4/3.6	98.6	91.0	94.9	97.3	93.6	96.1	97.9	92.3	95.5
	Boosting	92.6/7.4	93.1/6.9	97.8	79.7	90.6	92.7	97.9	92.0	95.2	87.9	91.3
Model 2	**Classification Tree**	64.3/35.7	62.1/37.9	83.6	4	51.5	62.1	55.6	62.3	71.3	7.4	56.4
	Bagging	97.8/2.2	86.5/13.5	93.3	46.8	89.8	87.4	73.8	87.6	90.3	57.3	88.7
	Random Forest with 2 Predictors	99.1/0.9	88.7/11.3	94.6	50	93.1	90.4	82.9	87.6	92.4	62.4	90.3
	Random Forest with 3 Predictors	99.3/0.7	88.6/11.4	93.8	53.2	92.9	90.6	80.7	87.6	92.2	64.1	90.1

Random Forest with 4 Predictors	99.4/0.6	89.0/11.0	93.5	54.8	93.9	91.4	77.5	88.3	92.4	64.2	91
Boosting	88.1/11.9	81.1/18.9	90.1	35.7	83.2	83	67.2	80.6	86.4	46.6	81.9

4.4. DECISION-MAKING INSIGHTS

The current work can support risk mitigation and resilience enhancement decision-making in many ways including identifying specific at-risk systems under different types of CID spatially and temporally, thus facilitating effective and efficient planning to optimize the resilience goals of these systems through available means. In addition to identifying systems at risk, the methodology conducted herein for categorizing system damages enables classifying CID according to the severity of the corresponding system damage they induce which is key for infrastructure systems spatio-temporal resilience planning.

Furthermore, in light of the specific analysis conducted in Stage 2 of the case study presented herein, a number of key decision-making insights can be drawn. For instance, the three maps shown in Figure 5 shows that, throughout the last decade, the eastern part of New York was highly susceptible to wind hazards causing power system failure. This should support power system decision makers to closely study the performance of eastern New York state power system and possibly introduce enough redundancy (i.e., either redundant overhead or underground cables) and/or resources to achieve higher overall power system robustness and minimal disruptions under wind-related disasters. In addition, for the considered study space boundaries (NY State) it was found that power system damages (i.e., Classes 2 and 3) were more frequent in the summertime rather than the winter months (i.e., May through August), this indicates that system

performance evaluation and asset management need to be initiated well before the start of the summer season in order to alleviate component or system damages.

Notwithstanding the value of the developed machine learning model in predicting the level of system performance based on various input attributes, further insights can be gained from the developed predictive model. For example, the model input attributes can be sorted according to the corresponding Mean Decrease in Gini (MDG) and Mean Decrease in Accuracy (MDA) [53]. The MDG depends on the Gini impurity of the model, which refers to the probability of incorrectly classifying a new record at a certain tree node. Accordingly, a higher MDG value indicates that the corresponding variable is more important for classifying the data. On the other hand, the MDA represents the decrease in the model accuracy due to the exclusion of a specific variable. Therefore, variables with higher MDA values are more important for predicating the severity of system damages. **Figure 4-10** shows the MDA and MDG values for the attributes used in Model 2 random forest ensemble with 4-predictors. The MDG values support that the magnitude is the most important variable for prediction, and therefore its accurate measurement is key for the effective prediction of power damage severity under wind-related hazards. Moreover, as per the MDG values, year, duration and latitude can also be considered key attributes for predicting power system damage severity. Similarly, the MDA values support that the three most important attributes for damage severity prediction are year, magnitude and latitude.

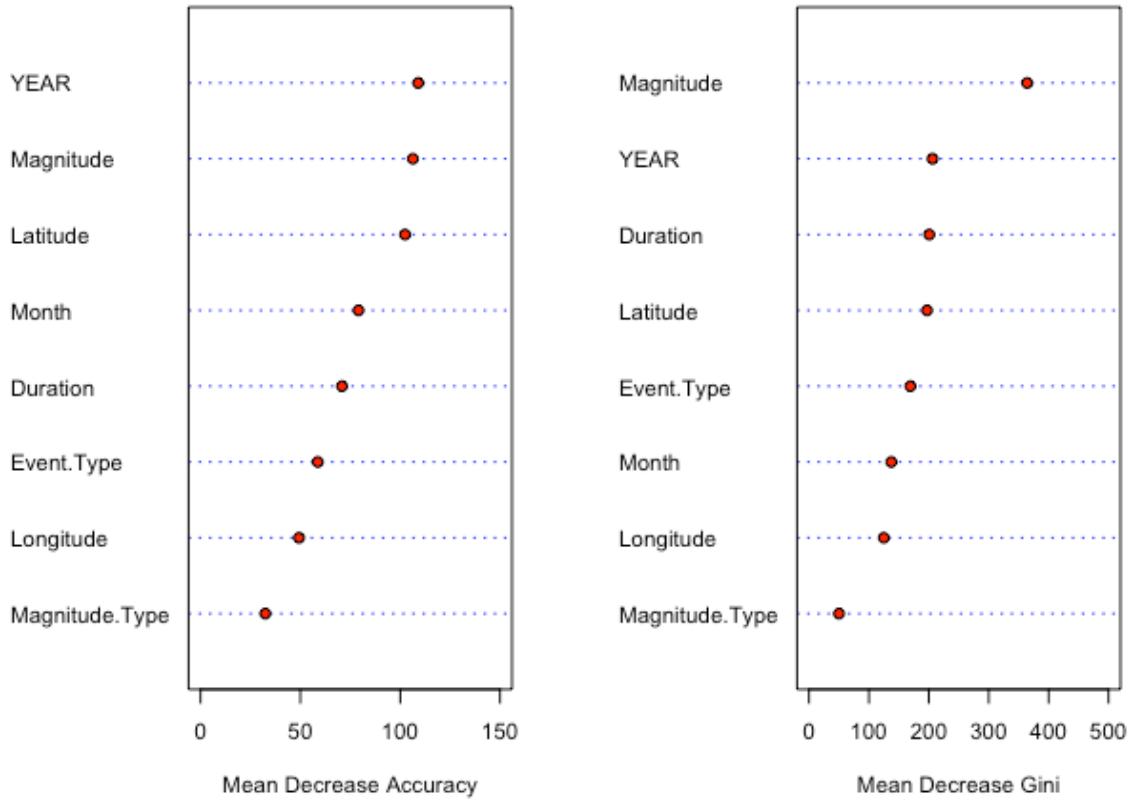


Figure 4-10: Variable Importance for the 4-Predictors Random Forest Model

The high importance of both spatial-related attributes (i.e., latitude) and time-related attributes (i.e., year) show that the power system damage severity does not only depend on hazard-related characteristics (i.e., magnitude and duration), but it is also a function of the inherent system properties and the hazard-system interaction. In addition, the higher accuracy achieved by the random forest ensemble technique employed herein supports the potential of employing nonlinear methods in solving complex problems such as disaster and impact prediction as suggested by recent studies [29,54,55].

4.5. CONCLUSIONS

Considering the significant adverse impacts of CID on infrastructure systems, the current work aims at: 1) identifying specific at-risk systems under different types of CID both spatially and temporally; 2) categorizing CID according to the severity of the resulting system damage; 3) developing a machine learning model which can be used to predict the level of system damage as a function of the CID attributes; and 4) identifying the primary parameters governing the severity of system damage. As such, an infrastructure Damage Prediction Framework (DPF) under Climate Induced Disaster (CID) is developed using textual data analytics and machine learning techniques. The DPF consists of input, internal processes, and output phases. The input phase comprises data collection processes. The internal processes phase incorporates four main stages: 1) establishing the link between CID and infrastructure systems affected; 2) investigating and explicating the influencing attributes; 3) employing data imputation to fill missing records; and 4) developing and testing a prediction model. The output phase of the DPF involves predicting infrastructure system damages using either regression or a classification technique. To demonstrate the applicability and viability of the developed DPF, it was applied to the historical disaster data collected by the National Weather Services (NWS) in New York State between 1996 and 2019.

The results of the first stage of the DPF showed that the power system in New York is the most vulnerable infrastructure system to CID, especially to wind-related hazards. The adverse impacts to the power system were found to range from a damage at the component level (i.e., power line or pole) to an overall system

failure (i.e., power outage). As such, wind-related hazards were classified depending on the corresponding severity of the power system damage into no damage (Class 1); component level damage (Class 2); and system-wide failure (*Class 3).* The spatio-temporal analysis conducted showed that, in the last decade, the eastern part of New York was highly susceptible to wind hazards causing power system failure compared to the western and central parts. Special attention should therefore be given to the power system in such region through improving the redundancy and resourcefulness aspects in order to boost the system's resilience. In addition, power system damages were more frequent in May through August, which highlight the need to conduct system assessment and implement asset management programs prior to the start of the summer season in order to alleviate component or system damages.

A supervised machine learning model was subsequently developed to predict the severity of the power system damage in New York under future spatio-temporal projections of wind hazard characteristics. To enhance the accuracy of the predictive model: 1) a cluster analysis was used to predict missing magnitude and duration records; and 2) bagging, random forest, and boosting algorithms were utilized to augment the performance of the classical classification tree. Two different models were accordingly trained and tested: Model 1 employed filled data, whereas Model 2 used unfilled data (i.e., actual data). Although Model 1 overperformed Model 2 for all ensemble techniques employed, after comparing the classification trees of both models, such high accuracy was attributed to the bias

introduced in the model through the data imputation performed in Stage 3. As such, Model 2 integrated with 4-predictors random forest, which yielded an overall accuracy of approximately 89% for the testing dataset, was chosen as the best performing model. In addition, hazard magnitude was found to be the most important variable controlling the power system damage severity under CID. This highlights the need for accurately estimating the wind magnitude for the efficient prediction of power system damage severity under CID. On the other hand, hazard duration was not identified as significantly governing the damage severity. Instead, the effects of time (i.e., year) and location were found to be more significant. This supports that the severity of the power system damage under CID depends on the interplay between hazard- and system-related attributes rather than solely the hazard characteristics.

The DPF developed herein is expected to aid state governments and decision-making stakeholders in developing preparedness plans for possible CID. This would in return facilitate mitigating the adverse impacts of CID on infrastructure systems, and therefore improve the overall urban resilience under such disasters. Further research can be implemented to advance the developed DPF through: 1) incorporating detailed system-related data (i.e., type of system components, maintenance information, number of redundant components, number of simultaneous disruptions, overall system recovery time); 2) using different techniques for data imputation (i.e., k-means weighted and inverse weighted distance), 3) integrating the duration of the system disruption to further enhance the

reliability of categorizing the severity of system damages; and, 4) including historical mitigation strategies followed to alleviate the impacts of CID on the affected systems.

4.6. ACKNOWLEDGEMENTS

The authors are grateful to the financial support of the Ontario Trillium Scholarship Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada, NSERC-CaNRisk-CREATE program. The authors would also like to acknowledge the fruitful discussions with the research teams of the INViSiONLab.

4.7. DATA AVAILABILITY

The dataset related to this article is publicly available and is provided by the NWS, a sub-agency under the National Oceanic and Atmospheric Administration (NOAA). It can be found at https://www.ncdc.noaa.gov/stormevents/ftp.jsp.

4.8. REFERENCES

[1] V. Thomas, J.R.G. Albert, C. Hepburn, Contributors to the frequency of intense climate disasters in Asia-Pacific countries, Clim. Change. 126 (2014) 381–398. https://doi.org/10.1007/s10584-014-1232-y.

[2] Global Risks Report, 2020. http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf

[3] H. Ritchie, M. Roser, Natural Disasters, 2012.

[4] Climate Change - Oxfam Canada, (n.d.). https://www.oxfam.ca/themes/water/ (accessed November 10, 2018).

[5] National Oceanic and Atmospheric Administration, USGCRP Indicator Details, (2019). https://www.globalchange.gov/browse/indicators/billion-dollar-disasters (accessed January 4, 2021).

[6] A.B. Smith, 2018's Billion Dollar Disasters in Context, Climate.Gov. (2019) 1. https://www.climate.gov/news-features/blogs/beyond-data/2018s-billion-dollar-disasters-context.

[7] Weather Disasters and Costs, Natl. Ocean. Atmos. Adm. Off. Coast. Manag. (2020) 1–6. https://coast.noaa.gov/states/fast-facts/weather-disasters.html.

[8] Strategic plan for the National Windstorm Impact Reduction Program, 2018. https://doi.org/10.29085/9781783300792.002.

[9] M. Haggag, M. Ezzeldin, W. El-Dakhakhni, E. Hassini, Resilient cities critical infrastructure interdependence: a meta-research, Sustain. Resilient Infrastruct. 00 (2020) 1–22. https://doi.org/10.1080/23789689.2020.1795571.

[10] K. Ganguly, N. Nahar, M. Hossain, A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh, Int. J. Disaster Risk Reduct. 34 (2019) 283–294. https://doi.org/10.1016/j.ijdrr.2018.12.002.

[11] J. Diaz, M.B. Joseph, Predicting property damage from tornadoes with zero-in fl ated neural networks, Weather Clim. Extrem. 25 (2019) 100216. https://doi.org/10.1016/j.wace.2019.100216.

[12] P.J. Sallis, W. Claster, S. Herna, A machine-learning algorithm for wind gust prediction, Comput. Geosci. 37 (2011) 1337–1344. https://doi.org/10.1016/j.cageo.2011.03.004.

[13] C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, H.S. Kim, Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data, Adv. Meteorol. 2018 (2018). https://doi.org/10.1155/2018/5024930.

[14] M. Hanewinkel, W. Zhou, C. Schill, A neural network approach to identify forest stands susceptible to wind damage, For. Ecol. Manage. 196 (2004) 227–243. https://doi.org/10.1016/j.foreco.2004.02.056.

[15] M. Rodrigues, J. De la Riva, An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environ. Model. Softw. 57 (2014) 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003.

[16] A. Kahira, B. Gomez, R. Badia Sala, A Machine Learning Workflow for Hurricane Prediction, in: B. Abstr. Barcelona Supercomput. Cent., 2018: pp. 72–73.

[17] K. Papagiannaki, K. Lagouvardos, V. Kotroni, A database of high-impact weather events in Greece: A descriptive impact analysis for the period 2001-2011, Nat. Hazards Earth Syst. Sci. 13 (2013) 727–736. https://doi.org/10.5194/nhess-13-727-2013.

[18] H. Toya, M. Skidmore, Economic development and the impacts of natural disasters, Econ. Lett. 94 (2007) 20–25. https://doi.org/10.1016/j.econlet.2006.06.020.

[19] A.L. Shokane, Social work assessment of climate change: Case of disasters in greater Tzaneen municipality, Jàmbá J. Disaster Risk Stud. 11 (2019) 1–7. https://doi.org/10.4102/jamba.v11i3.710.

[20] E. Cavallo, S. Galiani, I. Noy, J. Pantano, Catastrophic natural disasters and economic growth, Rev. Econ. Stat. 95 (2013) 1549–1561. https://doi.org/10.1162/REST_a_00413.

[21] M. Skidmore, H. Toya, Do natural disasters promote long-run growth?,

Econ. Inq. 40 (2002) 664–687. https://doi.org/10.1093/ei/40.4.664.

[22] R. Bhavnani, Natural Disaster Conflicts, Harrvard University, 2006.

[23] D. Bertsimas, N. Kallus, From predictive to prescriptive analytics, Manage. Sci. 66 (2020) 1025–1044. https://doi.org/10.1287/mnsc.2018.3253.

[24] L. Bakkensen, X. Shi, B. Zurita, The Impact of Disaster Data on Estimating Damage Determinants and Climate Costs, Econ. Disasters Clim. Chang. 2 (2018) 49–71. https://doi.org/10.1007/s41885-017-0018-x.

[25] A.B. Smith, R.W. Katz, US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases, Nat. Hazards. 67 (2013) 387–410. https://doi.org/10.1007/s11069-013-0566-5.

[26] C.L. Gray, V. Mueller, Natural disasters and population mobility in Bangladesh, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 6000–6005. https://doi.org/10.1073/pnas.1115944109.

[27] N. Brooks, W.N. Adger, Country level risk measures of climate-related natural disasters and implications for adaptation to climate change, 2003. http://www.uea.ac.uk/env/people/adgerwn/wp26.pdf.

[28] E. Alpaydin, Introduction to machine learning, MIT press. Chicago, 2020.

[29] M. Khalaf, A.J. Hussain, D. Al-jumeily, T. Baker, R. Keight, P. Lisboa, P. Fergus, S. Al Kafri, A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction, 2018 IEEE Congr.

Evol. Comput. (2018) 1–8.

[30] A. Jaafari, E. Zenner, M. Panahi, H. Shahabi, Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability, Agric. For. Meteorol. 266–267 (2019) 198–207. https://doi.org/10.1016/j.agrformet.2018.12.015.

[31] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, Word cloud explorer: Text analytics based on word clouds, Proc. Annu. Hawaii Int. Conf. Syst. Sci. (2014) 1833–1842. https://doi.org/10.1109/HICSS.2014.231.

[32] A.B. Alencar, M.C.F. De Oliveira, F. V. Paulovich, Seeing beyond reading: A survey on visual text analytics, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 (2012) 476–492. https://doi.org/10.1002/widm.1071.

[33] A. Endert, P. Fiaux, C. North, Semantic interaction for visual text analytics, Conf. Hum. Factors Comput. Syst. - Proc. (2012) 473–482. https://doi.org/10.1145/2207676.2207741.

[34] G. Miner, I. Elder, A. Fast, T. Hill, R. Nisbet, D. Delen, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[35] T. Banerjee, S., & Pedersen, The design, implementation, and use of the ngram statistics package, in: Int. Conf. Intell. Text Process. Comput.

Linguist. (Pp. 370-381). Springer, Berlin, Heidelberg. Chicago, Springer, Berlin, Heidelberg. Chicago, 2003: pp. 370–381.

[36] ngram package | R Documentation, (n.d.).

[37] K. Yagci Sokat, I.S. Dolinskaya, K. Smilowitz, R. Bank, Incomplete information imputation in limited data environments with application to disaster response, Eur. J. Oper. Res. 269 (2018) 466–485. https://doi.org/10.1016/j.ejor.2018.02.016.

[38] D. Patil, B. M., Joshi, R. C., & Toshniwal, Missing value imputation based on k-mean clustering with weighted distance, in: Int. Conf. Contemp. Comput., Springer, Berlin, Heidelberg., 2010: pp. 600–60.

[39] D. Li, J. Deogun, W. Spaulding, B. Shuart, Dealing with Missing Data: Algorithms Based on Fuzzy Set and Rough Set Theories, Springer, Berlin, 2005.

[40] R.C.T. Lee, J.R. Slagle, C.T. Mong, Application of clustering to estimate missing data and improve data integrity, in: 2nd Int. Conf. Softw. Eng., IEEE Computer Society Press, 1976: pp. 539–544.

[41] D. Li, J. Deogun, W. Spaulding, B. Shuart, Towards missing data imputation: A study of fuzzy K-means clustering method, Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci. 3066 (2004) 573–579. https://doi.org/10.1007/978-3-540-25929-9_70.

[42] S.P. Mandel J, A Comparison of Six Methods for Missing Data Imputation, J. Biom. Biostat. 06 (2015) 1–6. https://doi.org/10.4172/2155-6180.1000224.

[43] Y. Fujikawa, T. Ho, Cluster-based Algorithms for Filling Missing Values, PAKDD '02 Proc. 6th Pacific-Asia Conf. Adv. Knowl. Discov. Data Min. (2002) 549–554.

[44] R. Xu, D. Wunsch, Clustering, John Wiley & Sons, 2008.

[45] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (2002) 611–631. https://doi.org/10.1198/016214502760047131.

[46] A. Nagpal, Decision Tree Ensembles- Bagging and Boosting, (2017). https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (accessed March 1, 2019).

[47] A. Liaw, M. Wiener, Classification and Regression by RandomForest, R News. 3 (2002) 18–22.

[48] S. McNicholas, Introduction to classification, 2018. https://doi.org/10.1017/cbo9780511607493.005.

[49] S. McNicholas, Trees, Bagging, Random Forests and Boosting, 2018. https://ms.mcmaster.ca/~sharonmc/STATS780/.

[50] NCEI, Storm Events Database | National Centers for Environmental

Information, (2016). https://www.ncdc.noaa.gov/stormevents/%5Cnfiles/5576/stormevents.html (accessed October 10, 2019).

[51] Y. Dong, C.Y.J. Peng, Principled missing data methods for researchers, Springerplus. 2 (2013) 1–17. https://doi.org/10.1186/2193-1801-2-222.

[52] J.C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts, BMC Med. Res. Methodol. 17 (2017) 1–10. https://doi.org/10.1186/s12874-017-0442-1.

[53] H. Wang, F. Yang, Z. Luo, An experimental study of the intrinsic stability of random forest variable importance measures, BMC Bioinformatics. 17 (2016) 1–19. https://doi.org/10.1186/s12859-016-0900-5.

[54] S.J. Kim, C.H. Lim, G.S. Kim, J. Lee, T. Geiger, O. Rahmati, O. Son, W.K. Lee, Multi-Temporal Analysis of Forest Fire Probability Using Socio-Economic and Environmental Variables, Remote Sens. 11 (2019).

[55] X. Zhang, X. Li, L. Li, S. Zhang, Q. Qin, Environmental factors influencing snowfall and snowfall prediction in the Tianshan Mountains, Northwest China, J. Arid Land. 11 (2019) 15–28. https://doi.org/10.1007/s40333-018-0110-2.

Chapter 5

A DATA DRIVEN MODEL FOR CLIMATE-INDUCED DISASTER DAMAGE PREDICTION

ABSTRACT

Abstract: The frequency and magnitude of Climate-Induced Disasters (CID) has been increasing consistently over the past few decades and is expected to continue to escalate in the coming years. According to the Emergency Events Database, a three-fold increase in the number of CID was recorded in less than 4 decades, from around 1,300 CID in 1975–1984 to over 3,900 in 2005–2014. As such, alleviating the impacts of such disasters at both the individual and community levels is key. In this respect, the current work proposes a systematic data-driven framework for predicting CID-related damages. The framework encompasses four phases: (1) Data Collection and Integration in which spatial interpolation methods are proposed to facilitate integrating data from multiple sources, (2) Feature Selection, which aims at comparing several methods to select relevant input variables for inclusion in the prediction model, (3) Model Development where supervised machine learning techniques are employed to train and test the prediction model, and (4) Result Analysis and Interpretation in which the Blackbox nature of the machine model is decoded. To demonstrate the utility of the proposed framework, property damages due to wind disasters were linked to event type, magnitude, duration, time and location as well as climate, land cover, social, housing, demographic and

economic data recorded in the state of New York from 2010 to 2018. Features significantly important for the property damage prediction were selected using different feature selection approaches (i.e., filters, wrappers, and embedded methods), and a set of machine learning models were subsequently developed. The best performing model was found to be a random forest-based regression tree and yielded a Root Mean Squared Error of 0.32 and a coefficient of determination of 0.79 between the actual and predicted property damages. Both the feature selection and model interpretation processes showed that, within the considered demonstration application wind-related damages were found to depend on the complete interplay between disaster, climate, socioeconomic, housing, and demographic conditions rather than wind hazard characteristics only. This highlights the need for accurately recording such factors for the effective prediction of wind-related damages. The proposed framework is considered a step forward in enhancing the preparedness of governments for CID, and thus alleviating their adverse impacts and reaching more resilient communities.

KEYWORDS: *climate-induced disasters, property damages, resilience, regression trees, machine learning, data-driven modelling*

5.1. INTRODUCTION

The frequency and magnitude of Climate-Induced Disasters (CID) has increased tremendously over the past few decades, and their adverse impacts have become more predominant. Such impacts include property and crop damages, evacuations, injuries, and life losses. According to the Emergency Events Database reported by the Centre for Research on the Epidemiology of Disasters, the global average number of CID has tripled in less than four decades (from approximately 1,300 CID between 1975 and 1984 to around 3,900 between 2005 and 2014) [1]. In addition, around 1 million deaths and $1.7 trillion property damages were attributed to CID since the year 2000 [1], [2], with around US $210 billion property damages (i.e., approximately 12.5%) incurred only in 2020 [3]. In the United States, the 2020 CID damage costs reached approximately US $95 billion [4], which is double those reported in 2019 [4], [5]. Consequently, the 2020 Global Risks Report continued to identify extreme weather as the top ranked global risk and among the top five risks in terms of likelihood and impact, respectively, according to the statistics of CID that occurred between 2017 and 2020 [6]. Such rankings are expected to remain unchanged since: (1) the number of CID is anticipated to double during the next 13 years [7]; (2) the annual fatalities due to CID are expected to increase by 250,000 deaths in the next decade [8]; and, (3) the annual CID damage costs are expected to increase by around 20% in 2040 compared to those realized in 2020 [9], [10].

Given the anticipated increase in CID frequency, the intensification of their impacts, the rapid growth of the world's population, and the fact that more than two thirds of such population is expected to be living in urban areas by 2050 [11], it has recently become extremely crucial to enhance both community and city resilience under CID. This necessitates the effective prediction of CID impacts which requires the existence of massive databases that characterize the different drivers of CID (i.e., physical, social, economic, etc.). However, classic mathematical and statistical models are typically unable to describe interrelationships based on large databases. Thus, data-driven research that aims at aiding decision makers in reducing and mitigating CID risks has progressed rapidly. In that context, flood damage was linked to structure architype, flood awareness, and literacy in a regression fashion using different machine learning models [12]. Machine learning techniques were also employed to predict the number of hurricanes per season [13], wildfire frequency [14], spatial probabilities of wildfire [15], wind damage [16], wind gust occurrence [17], heavy rain occurrence [18], flood severity [19], flood-related household damage [12], property damage caused by tornado disasters [20], and also to interpret the spatial distribution of structural damage due to wind events [21].

In addition, the use of data-driven modelling for relating the impacts of CID to the community physical, social, and economic attributes has highlighted that: (1) enhancing the social and/or economic conditions can significantly reduce the impacts of natural disasters [22]; (2) the negatively changing environmental conditions are expected to increase the social vulnerabilities (i.e., sensitivity,

exposure and adaptivity) to natural disasters during the coming decades [23]; (3) adopting new technologies for mitigating CID risks can facilitate the economic growth through increasing productivity [24]; (4) social work intervention is essential for mitigating CID risks [25]; (5) crop failures due to climate disasters are strongly related to long-term population mobility [26]; and, (6) changes in temperature are highly associated with climatological disasters and the high fluctuations in precipitation are associated with hydrological disasters [27].

Consequently, it can be inferred that the prediction of CID-related aspects doesn't depend on a single data type, rather it often depends on the interaction between different feature categories. Hence, it is particularly crucial to both gather and integrate several types of input data to develop an effective data-driven model for predicting CID occurrences and/or their impacts. To build such an integrated database, spatial interpolation methods can be employed for forecasting feature values at disaster locations. Such methods were previously used to estimate environmental [28]–[30], socioeconomic [30]–[32], geotechnical [33], and health-related features [30], [34]. Furthermore, given the correlation between CID-related aspects and different feature categories, selecting the most important features for the prediction of these aspects is key since it contributes to both model performance enhancement and computational efficiency. In this context, several feature selection methods were employed to decrease the dimensionality of socio-economic [35], [36] and environmental data [37], [38].

Despite the emerging efforts conducted to link CID impacts to their deriving factors using machine learning techniques [39], [40], a standardized approach that can be systematically used for the development of such linkage is yet to be developed. In this respect, this paper aims at developing a systematic framework that can be used for predicting CID damages by employing different data preprocessing and machine learning. To demonstrate the utility of the proposed framework, wind disaster data collected by the National Weather Services was linked to climate, land cover, social, housing, demographic, and economic data in the state of New York from 2010 to 2018. The proposed framework is considered a step forward in enhancing the preparedness of governments for CID, to subsequently alleviate their adverse impacts and reaching more resilient communities. This paper is divided into two main parts: the first part explains in detail the systematic phases pertaining to the CID damage prediction framework, and the second part includes a demonstration application to highlight the utility of the proposed framework.

5.2. CLIMATE-INDUCED DISASTER DAMAGE PREDICTION FRAMEWORK

The data-driven damage prediction framework shown in Figure 5-1 presents a systematic approach for predicting CID direct impacts (i.e., human related impacts such as number of injuries, fatalities, or evacuations, and monetary related impacts such as property and crop damages) based on spatio-temporal (i.e., location, time),

community (i.e., social, economic), climate (i.e., temperature, precipitation), and hazard (i.e., magnitude, duration) attributes. The framework is comprised of four main phases: (1) data collection and fusion, (2) feature selection, (3) model development, and (4) result analysis and interpretation.

The first and most important phase for CID impacts prediction is collecting the input data which includes hazard-, community-, and climate-related attributes. As multiple data sources are expected to be used, data collection should be followed by a compilation process to integrate the collected data into a single database that can be used for the development of a CID damage prediction model. The second phase of the framework aims at removing the redundant information present in the database through selecting a set of features that is significantly important for predicting the output of interest. Upon feature selection, the third phase of the framework involves developing multiple machine learning models using different supervised machine learning techniques (i.e., classical and ensemble decision trees, artificial neural networks, support vector regression). Such models are subsequently tested using different sets of input-output pairs and several measures are used for performance evaluation [i.e., Root Mean Squared Error (RMSE), Coefficient of Determination (R^2)]. After selecting the best performing model, the fourth and final phase in the framework involves interpretating the results. In this phase, relationships between the model inputs and the predicted output are uncovered to decode the black-box nature of the machine learning techniques employed.

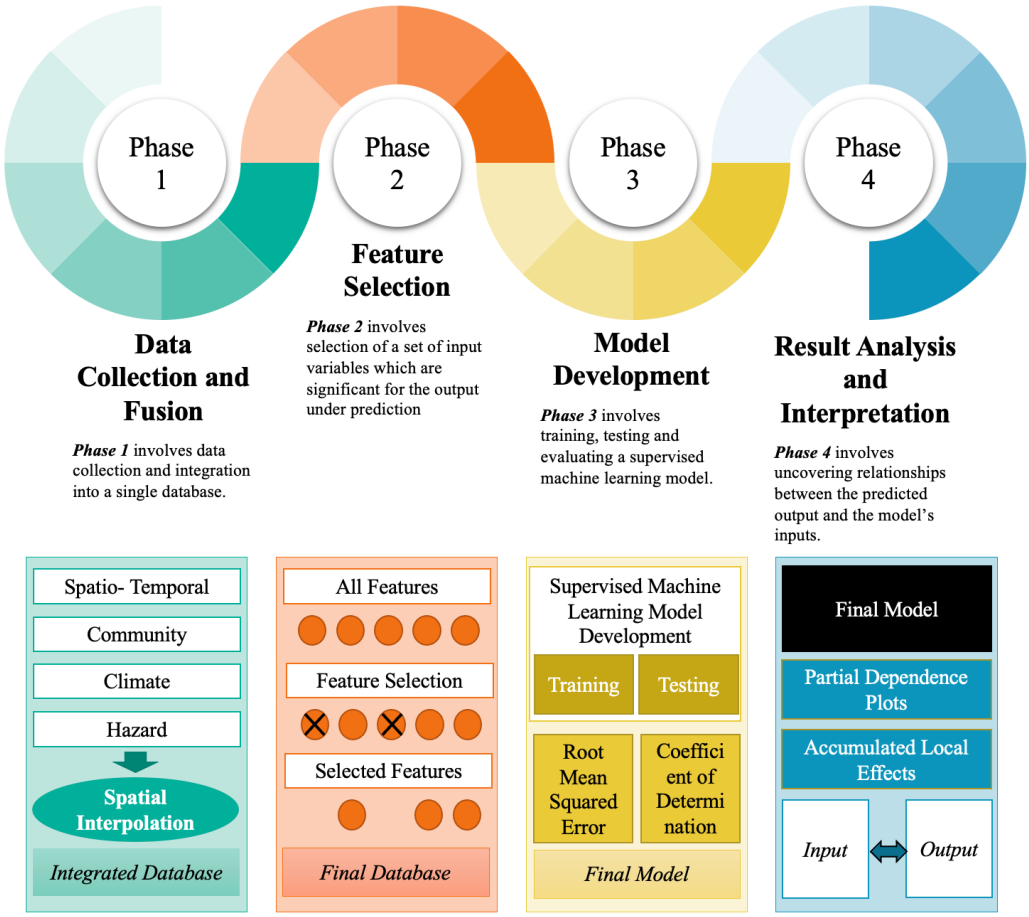


Figure 5-1: CID Impacts Prediction Framework

5.2.1. Phase 1 – Data Collection and Fusion

Data collection and fusion represent the most important phase in the framework as the quality of the data collected can significantly affect both the model predictability and the insights drawn from it. To effectively predict CID impacts, several input features are required and are typically categorized into disaster-, climate-, socioeconomic-, demographic, location-, and physical-related attributes. Disaster-related attributes include the type of the hazard (i.e., flood, tornado,

thunderstorm wind, drought, wildfire), geographical coordinates of the affected location (i.e., latitude and longitude), temporal aspects (i.e., month and year), innate characteristics of the hazard (i.e., scale, magnitude, duration), and induced damages and impacts (i.e., human, and monetary losses). Climate-related attributes include statistics of hourly, daily, or monthly temperature, precipitation, humidity, and air pressure. Socioeconomic-related attributes include nationalities, educational levels, household income and income per capita, whereas demographic-related attributes include the total population, age, and gender distribution in the area of interest. Location-related attributes, including land cover and housing information in the affected area, are also essential for predicting CID property damage. Moreover, information related to physical infrastructure systems (i.e., age and status) at the disaster location are crucial for estimating the system damages induced by CID.

As various categories of input data are required for the accurate prediction of CID impacts, data fusion represents a key step in developing the data-driven model. Such compilation is challenging as each category of the input data may be recorded at different spatial and temporal resolutions. As such, after collecting the different categories of input data, a spatial interpolation is essential to facilitate integrating the different input features into a single database. Through assuming that *"everything is related to everything else, but near things are more related than distant things"* [41], [42], point spatial interpolation uses the known values of a certain feature at several locations (i.e., control points) to estimate unknown values at other locations. Point spatial interpolation methods can be classified according

to the spatial extent (i.e., global verses local methods) and exactness (i.e., exact verses approximate methods) [43]. In global methods (i.e., trend surface analysis) values at all known points are employed to determine those at unknown locations, whereas in local methods (i.e., local polynomial interpolation) only nearby locations are used [43], [44]. On another perspective, the difference between exact and approximate methods is the fact that exact methods reserve the original data point values (i.e., generates a surface that passes through the control points) [43], [44]. Several interpolation methods have been developed to date. Among those methods, inverse distance weighting, kriging, and spline are the most widely used spatial interpolation methods [45]. The inverse distance weighting method enables estimating feature values at unknown points deterministically using a combination of weights from nearby points. Within such method, weights from further points are typically smaller compared to those from nearby points, and thus they have less effects on the value to be estimated [46]. Inverse distance weighting has been used in several applications, including rainfall variability estimation [28], temperature mapping [29], socio-economic feature prediction [30], [31], and disease risk prediction [34]. Kriging interpolation uses known observations (i.e., feature values) to infer the covariance structure of the underlying feature. Weights are subsequently obtained based on a variogram analysis and then used to estimate the feature values at unknown locations [46]. Kriging was previously employed in several applications including environmental [47], [48] and geotechnical [49], [50] feature estimation. Spline interpolation is a deterministic method that fits several

polynomials to different subsets of the feature values rather than using a single polynomial for all values [33]. Spline interpolation was used to estimate several feature types including environmental [51], [52] and demographic [53], [54] features. It should be noted that there is no preference for a specific spatial interpolation method in a certain situation. Therefore, most related studies either use several methods and select the best performing one or employ methods that proved to be viable for similar data types. It is also noteworthy to mention that given the availability of significant missing records in the integrated database, data imputation becomes key for obtaining an effective and accurate predictive model. This imputation can be performed using several unsupervised machine learning techniques including clustering methods.

5.2.2. PHASE 2 – FEATURE SELECTION

Whereas the vast success of machine learning models in predicting CID-related aspects is clear, the data driven nature of such models increases the sensitivity of their predictions to the input features employed. As some input features may be strongly correlated, and thus convey similar information, appropriate selection of input features is crucial for developing an efficient machine learning model. Using the most relevant input features and removing those with redundant information has proved to boost the model performance and decrease the computational cost [55]. Several feature selection methods have been developed in the past few decades, and are generally categorized under filter, wrapper, and embedded methods [56]–[58].

Due to their low computational cost, filter methods are often used for large databases with numerous input features under analysis [59]. These methods use raw data to select features (i.e., inputs) based on their correlation with the dependent variable (i.e., output) and their independence upon other predictors [60]. The value of the correlation coefficient (R) is used for such purposes and typically ranges from -1 to 1, where a negative R value indicates an inverse correlation, zero R values support the independence between variable pairs, and a positive R value indicates a direct relationship. Independent features that are more correlated with the dependent variable, and less correlated with one another, are considered the most significant for prediction.

In wrapper methods, relevant features are selected based on evaluating the performance of the developed data-driven model [61]. The application of wrapper methods start by developing multiple models using different subsets of input features, and subsequently adding new features or removing existing ones based on assessing the model performance [62]. Sequential feature selection is the most widely used wrapper method and can be typically conducted through: (1) backward elimination, in which all available features are initially employed in the model and insignificant features are removed at each iteration; or (2) forward addition, in which significant features are added with each iteration until adding more features cease to have a positive effect on the performance of the model. Recursive feature elimination is another wrapper method which relies on developing numerous models, removing worst performing features, and subsequently ranking the input

features after comparing the performance of such models [62]. Boruta is another wrapper feature selection method that uses a top-down approach for finding the relevant features based on their random forest importance score. Through comparing features with their randomized versions (i.e., shadow features), the Boruta method filters features that have higher importance than the highest randomized feature's importance and marks them as significant. Genetic algorithms (GA) represent another type of wrapper feature selection which starts by randomly creating an initial population of models (each with a different set of inputs), assesses their performance, and reproduces the next generation of models through adding/removing a subset of inputs using genetic operations (i.e., elitism, mutation, and crossover). The reproduction step is repeated until the algorithm reaches a certain termination criterion.

In embedded methods, the feature selection process is inherently included within the developed machine learning model. As such, embedded feature selection methods are basically metrics rooted in the model's training process [61]. Examples of embedded methods include variable importance metrics that are part of the random forest- and boosting-based regression trees. Random Forest variable importance uses either the Mean Squared Error (MSE) and the node impurity to rank input features when the variable to be predicted is continuous, or the mean decrease accuracy and mean decrease gini within categorical variable prediction models. Furthermore, the boosting relative importance algorithm is used to rank the input features based on their relative influence on the performance of the resulting

gradient boosted trees. As in random forest variable importance, the relative influence algorithm ranks features by computing the average increase in prediction error after permuting each input feature.

Prior to model development, it is noteworthy to emphasize the importance of both label encoding and data transformation (i.e., power transforms). Given the availability of different categorical variables, label encoding is crucial before training the machine learning model. As such, different categorical variables are encoded with numerical values (i.e., month names are substituted with month numbers). Moreover, given the vast value ranges for the different variables employed in the model, data transformation (i.e., power transforms) which aims at removing the skewness from the data is key prior to building the model.

5.2.3. PHASE 3 – MODEL DEVELOPMENT

The third phase of the framework involves the development of the machine learning model. As discussed previously, several techniques can be utilized for the prediction of CID-related aspects including artificial neural networks and decision trees. The choice of an appropriate machine learning technique depends on the type of the variable under prediction (i.e., continuous, or categorical), where: (1) artificial neural networks are among many techniques that can be used regardless of the type of the output variable; (2) regression trees are used for continuous outputs; and (3) classification trees are used in case of categorical outputs. In

addition, ensemble techniques (i.e., bagging, random forests, and boosting) can be employed to boost the model performance [63]–[65].

In the framework developed herein, several machine learning and ensemble techniques can be used, and their evaluation criteria should be compared to select the best performing model. In all techniques, the input data is divided into training and testing subsets. Depending on the type of the prediction (i.e., regression or classification), the model performance is assessed using certain evaluation criteria. For instance, the misclassification error can be used when the model output is categorical (i.e., classification), whereas the RMSE can be utilized when the output of the model is continuous (i.e., regression).

5.2.4. PHASE 4 – RESULT ANALYSIS AND INTERPRETATION

The black-box nature of most machine learning models is the primary limitation of such techniques, and this ambiguity hinders the user's ability to interpret the resulting input-output interrelationships. As such the last phase of the framework aims at overcoming this limitation through visualizing the relationships and patterns that the model has already learned. An example of the methods employed for such purpose include partial dependence plots (PDP) in which the input-output relationships are depicted and classified into linear and complex relationships [66]. In PDP, the marginal effect of a single or a group of input features on the predictor is expressed in terms of the average prediction corresponding to all values of other inputs [66]. The key disadvantage of PDP is that these plots are informative when

the input features are not strongly correlated [66], [67]. Other methods can also be used for interpretating the model result such as accumulated local effect plots and sensitivity analysis.

5.3. Demonstrative Application: prediction of Wind-Related Property Damage in New York State

5.3.1. Phase 1 – Data Collection and Fusion

The current case study aims at assessing the applicability and viability of the developed framework through predicting wind-related property damages in the state of New York based on disaster-, landcover-, climate-, social-, economic -, housing-, and demographic- related records collected by different agencies from 2010 to 2018. Disaster data was obtained from the National Weather Services Database (NWS), which aims at providing historical and forecasts of weather and climate data [68]. The collected disaster attributes include location, time, and hazard related properties. Landcover data was represented by the 30-m Landsat obtained from the National Land Cover Database (NLCD) [69]. Climate data was collected from the National Oceanic and Atmospheric Administration (NOAA) online search tool [70], and include statistics of the daily temperature (i.e., minimum, and maximum). Social, economic, housing, and demographic data were obtained from the American Community Survey (ACS) which provides population and housing information in the United States on a yearly basis [71]. More specifically, the social attributes collected include nationality, educational level,

and spoken languages; economic features include mean and median household income and income per capita; housing attributes include total housing units, number of mobile homes, year of household construction, number of vehicles per household, and household value; and demographic attributes include population size, gender, age, and race. It is noteworthy to mention that logarithmic transformation is used to minimize the skewness in the data. Table 5-1 shows the attributes considered in this case study together with their type (i.e., continuous, or categorical), location (disaster, county center, or stations inside counties), and source.

Table 5-1: [i]Case Study Input Features

	Code	*Type*	*Location*	*Source*
Disaster Attributes				
Longitude	DS5	Continuous	Disaster	
Latitude	DS6	Continuous	Disaster	
Year	DS1	Continuous	Disaster	
Month	DS2	Continuous	Disaster	NWS
Event Type	DS4	Continuous	Disaster	
Magnitude	DS3	Continuous	Disaster	
Duration	DS7	Continuous	Disaster	
Property Damage	DS8	Continuous	Disaster	
Land Cover Attributes				
Type of Land Cover	L1	Continuous	Disaster	NLCD
Climate Attributes				
Daily Maximum Temperature	T1	Continuous	Station	NOAA
Daily Minimum Temperature	T2	Continuous	Station	

Social Attributes				
Number of High School Graduates	S1	Continuous	County	
Number of College Graduates	S2	Continuous	County	
Number of Associate Degree Graduates	S3	Continuous	County	
Number of Bachelor's degree Graduates	S4	Continuous	County	
Number of Post Graduate Degree Graduates	S5	Continuous	County	ACS
Number of Foreign-Born US Citizens	S6	Continuous	County	
Number of US Born US Citizens	S7	Continuous	County	
Number of Non-US Citizens	S8	Continuous	County	
Number of People with English Speaking Homes	S9	Continuous	County	
Number of People with Non-English-Speaking Homes	S10	Continuous	County	
Housing Attributes				
Number of Housing Units	H1	Continuous	County	
Number of Mobile Homes	H2	Continuous	County	
Number of Households Built on or after 2000	H3	Continuous	County	
Number of Households Built on or before 1999	H4	Continuous	County	
Number of Households with No Vehicles	H5	Continuous	County	
Number of Households with Vehicles	H6	Continuous	County	
Number of Households with Value of $0 to $99,999	H7	Continuous	County	ACS
Number of Households with Value of $100.000 to $199,999	H8	Continuous	County	
Number of Households with Value of $100.000 to $199,999	H9	Continuous	County	
Number of Households with Value of $500,000 to $999,999	H10	Continuous	County	
Number of Households with Value of $1,000,000 or more	H11	Continuous	County	
Demographic Attributes				
Total Population	D1	Continuous	County	
Number of Males	D2	Continuous	County	ACS
Number of Females	D3	Continuous	County	

Median Age	D4	Continuous	County	
Number of People 65 Years and Over	D5	Continuous	County	
Number of White People	D6	Continuous	County	
Number of Other than White People	D7	Continuous	County	
Economic Attributes				
Number of Households with Income Less than 10K	E1	Continuous	County	
Number of Households with Income 10K to 50K	E2	Continuous	County	
Number of Households with Income 50K to 100K	E3	Continuous	County	
Number of Households with Income 100K to 200K	E4	Continuous	County	ACS
Number of Households with Income 200K or More	E5	Continuous	County	
Median Household Income	E6	Continuous	County	
Mean Household Income	E7	Continuous	County	
Per Capita Income	E9	Continuous	County	

It should be emphasized that wind-related disasters are recorded at scattered locations across the state of New York (as shown in Figure 5-2), whereas climate, social, economic, housing, and demographic data are collected at specific locations within each county. To have a fully integrated database, all input features are predicted at disaster locations using an inverse distance weighting (IDW) spatial interpolation method described earlier.

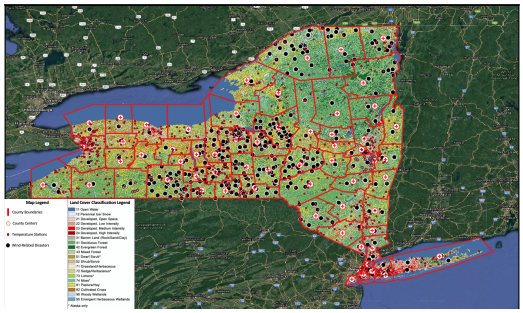


Figure 5-2: Spatial Distribution of Disasters, Counties, and Monitoring Stations

IDW spatial interpolation method estimates feature values at unknown locations using the following equation [29].

$$Z_j = \sum_{i=1}^{n} w_i Z_i \quad \textbf{Equation 5-1}$$

where Z_j is the feature value to be estimated at location *j*, Z_i is feature value at a known location *i*, n is the number of points to be used in the interpolation process, and w_i is the interpolation weight of Z_i which can be calculated as per Equation 5-2:

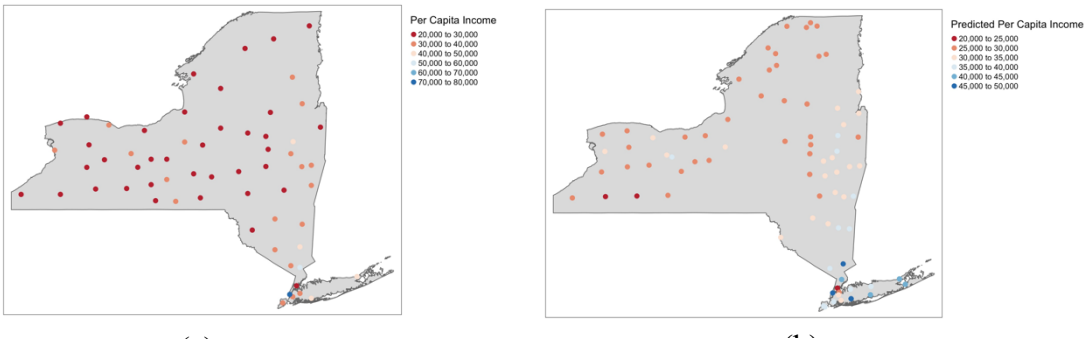
$$w_i = \frac{h_{ij}^{-p}}{\sum_{i=1}^{n} h_{ij}^{-p}} \quad \textbf{Equation 5-2}$$

where h_{ij} is the distance between locations *i* and *j*, and *p* is a power factor that determines the weight strength. When *p* equals 0, the weight becomes independent of the distance and Z_j would turn out to be equal to the mean value of Z_i. As *p* increases, the weights of further points decrease dramatically. Finally, the distance h_i can be quantified using Equation 5-3.

$$h_i = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2} \quad \textbf{Equation 5-3}$$

where x_i and y_i are the coordinates of the location *i*, while x_j and y_j are the coordinates of the location *j*. Figure 5-3 shows an example of applying the IDW method with *p* equals 2 to estimate per capita income at disaster locations in 2018, where each observation of the per capita income is weighted based on the distance from the county center to the disaster location. The same procedure has been

conducted to interpolate the other input features at the disaster locations for the time frame from 2010 to 2018.



(a) (b)

Figure 5-3: Per Capita Income in 2018 (a) Recorded at the County Center and (b) Interpolated at Disaster Locations

5.3.2. PHASE 2 – FEATURE SELECTION

The second phase of the framework aims at optimizing the performance of the machine learning model through selecting the most significant features for property damage prediction (out of the ones listed in Table 1). Several feature selection methods are employed in this application and their results are compared to select the most significant subset of features. These methods are the correlation matrix (a filter approach); the random forest and the relative influence algorithm (embedded approaches); and the Boruta and genetic algorithms (wrapper approaches).

The first feature selection method used herein is filtering based on the correlation between the independent features (i.e., the input variables) and the dependent variable (i.e., the output of interest). Figure 5-4 shows the correlation matrix that include R values between each input and input-output pair for the

features considered in this study. It can be inferred that the social, economic, housing, and demographic features are highly correlated, whereas the disaster-related features are mostly not correlated with each other. Furthermore, the independent variable (i.e., wind-related property damages) has its highest correlation with a few number of independent variables (i.e., year, disaster type, wind magnitude, and median age). The high correlation between the independent features together with the minimal correlations between property damages (i.e., the dependent variable) and other input variables calls for the use of other feature selection methods that could aid in determining he most significant set of features for prediction. As such, several embedded methods that rank features according to their importance in the development of a machine learning model are employed hereafter.

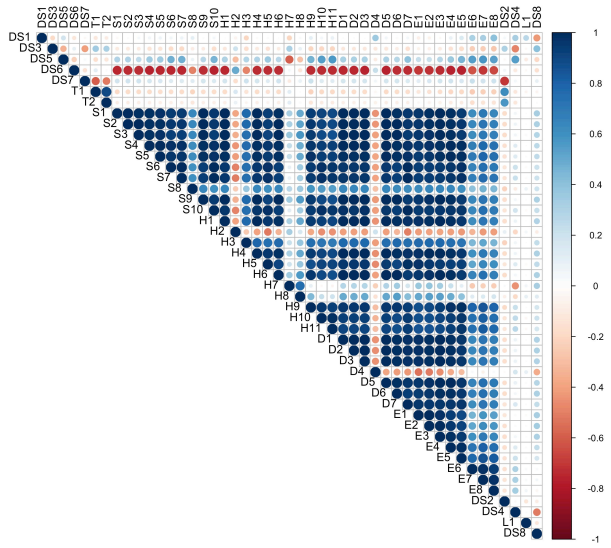


Figure 5-4: Features Correlation Matrix

The first embedded method used herein is the random forest based variable importance algorithm which ranks the independent features according to the Mean Squared Error (MSE) and the node impurity, as shown in Figure 5-5. According to both the MSE and node impurity plots, the most important disaster-related features are the magnitude year, event type, duration, and month. Both measures also indicate that the climate-related features considered in this study (i.e., the maximum and minimum temperatures) are key for property damage prediction, whereas the per capita income was found to be the most vital economic feature. Among the housing attributes, the number of households built after the year 2000 and those with a value less than US $100,000 were found to be the most important housing attributes based on MSE and node impurity values, respectively. Finally, the number of non-US citizens and the number of high school graduates are the most important social variables according to the values of MSE and node impurity, respectively. It should be highlighted that the node impurity metric shows that the location is key for property damage prediction as latitude was found to be the top 5^{th} ranked attribute.

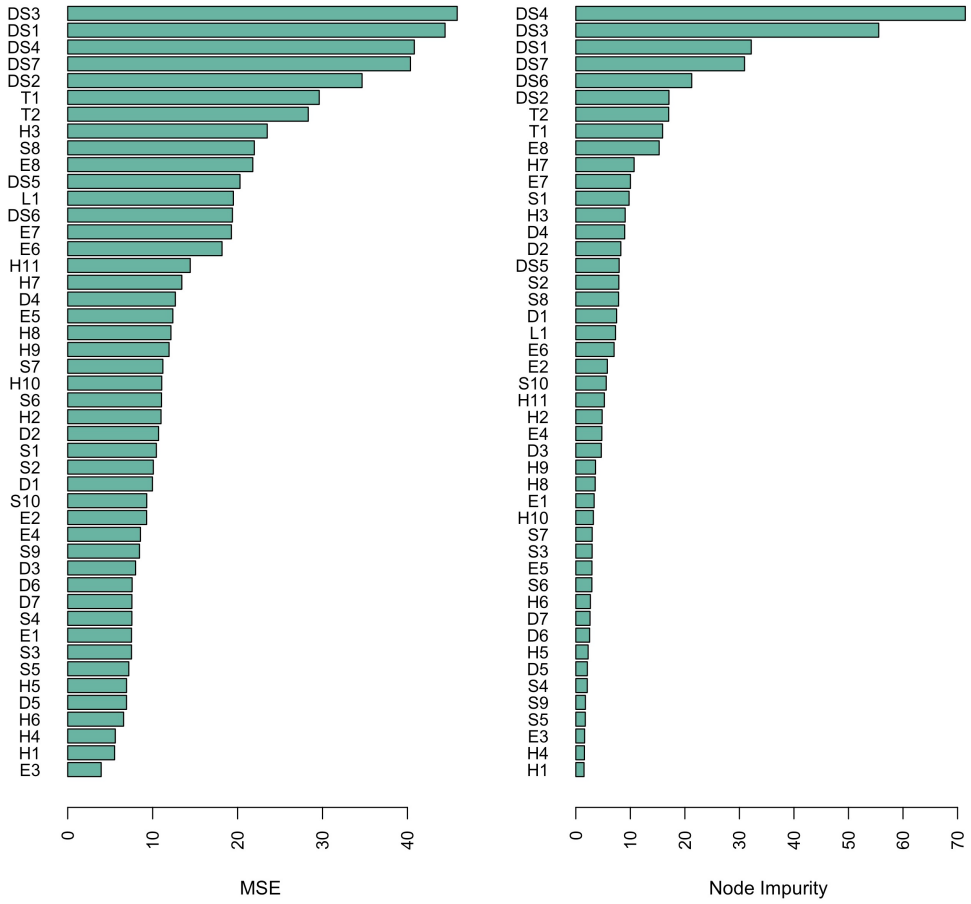


Figure 5-5: Random Forest Variable Importance

The second embedded feature selection method employed in this study relies on the relative influence algorithm, which is rooted within the boosting ensemble, as shown in Figure 5-6. As can be observed, the three most important features for prediction are the magnitude, event type, and duration, which are the same features identified by the variable importance algorithm embedded in the random forest ensemble technique illustrated before. Moreover, as specified by the node impurity, the latitude is identified as 4th most significant feature, whereas the year is identified among the five highest ranked features as per the MSE and node impurity. In

addition, the maximum temperature, the minimum temperature, and the per capita income are identified among the highest ranked 10 features which conforms with the results obtained from the random forest variable importance measures. On the other hand, the median age is identified as the 10th ranked feature according to the boosting relative influence.

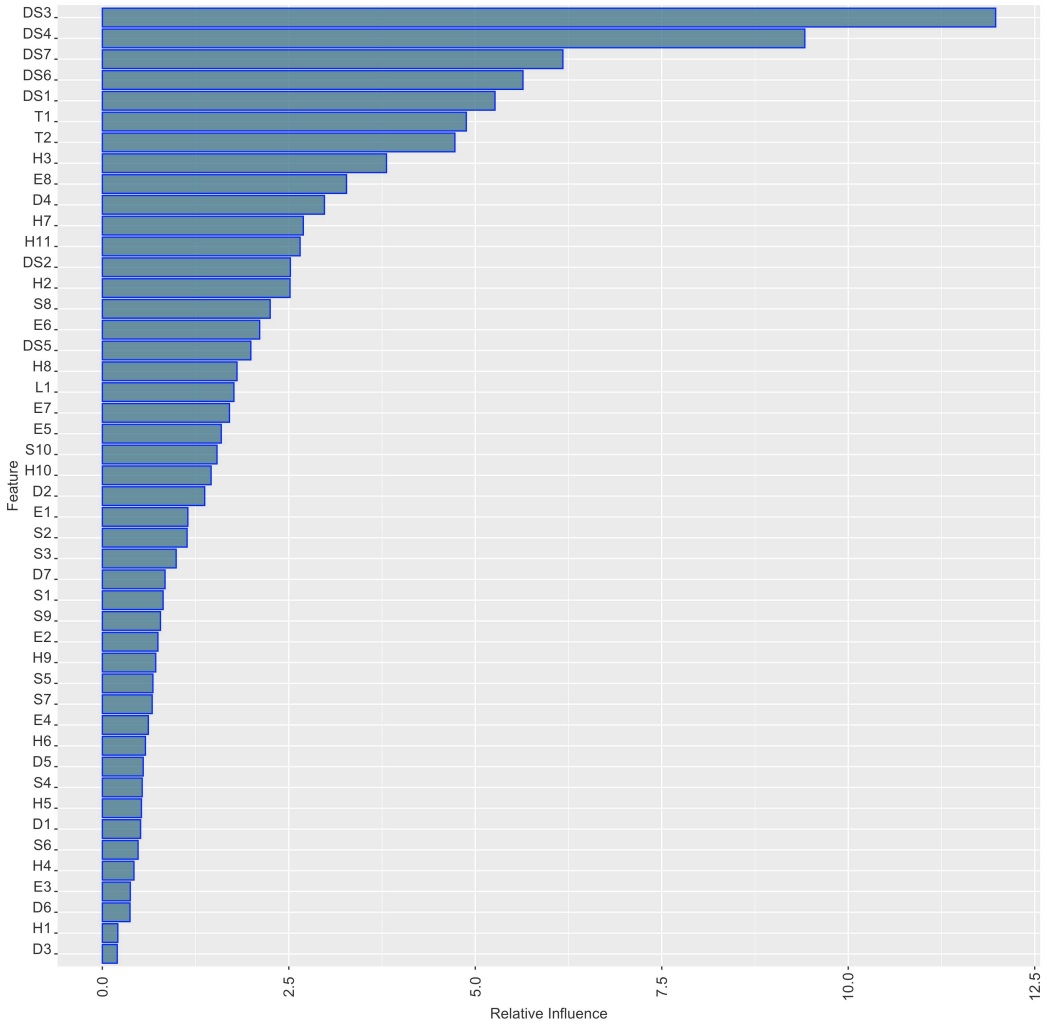


Figure 5-6: Boosting Relative Influence for Features

The first wrapper selection method employed here in is the Boruta algorithm. Figure 5-7 shows the Boruta importance plot for the features listed in Table 5-1. The boxplots displayed in the figure show the distribution of feature's importance over 100 iterations, where the colors of the boxplots are used to distinguish the selected features. It can be observed that all input features are considered significant as per the Boruta algorithm; however, the magnitude, duration, year, event type, month, temperatures, and per capita income are ranked as the seven most important features. This confirms the results of the random forest importance and the boosting relative influence.

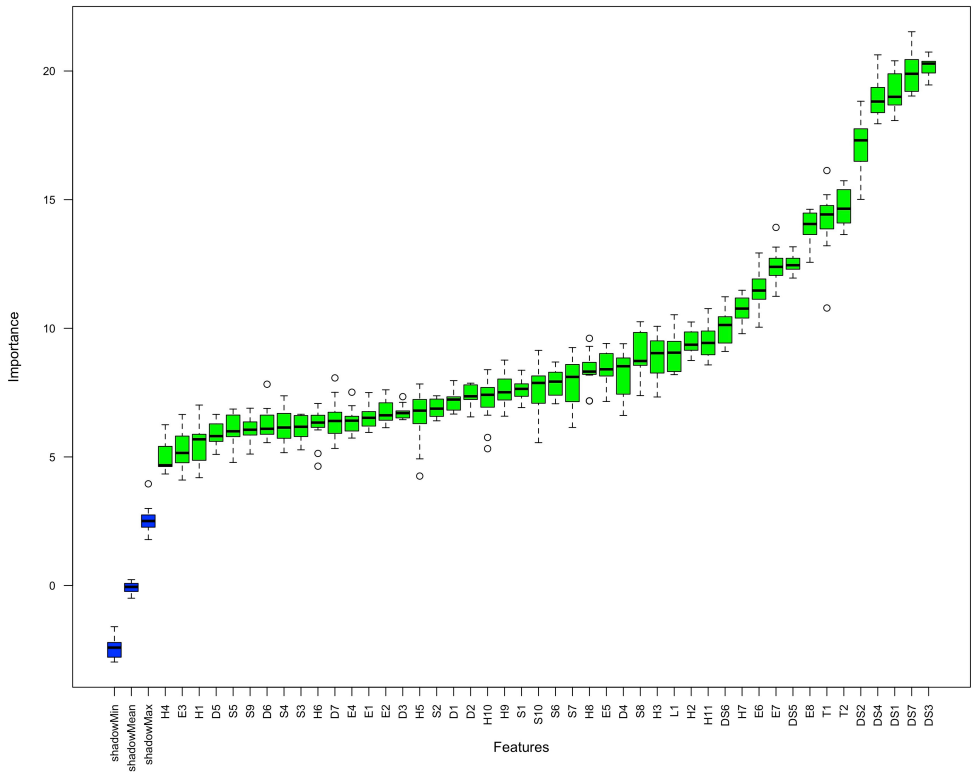


Figure 5-7: Boruta Algorithm Importance Plot

To boost the predictability of the model even further, GA are also employed for feature selection. Being a wrapper feature selection method, GA are integrated herein within a random forest model. After each generation, the out-of-bag RMSE is estimated as shown in Figure 5-8.

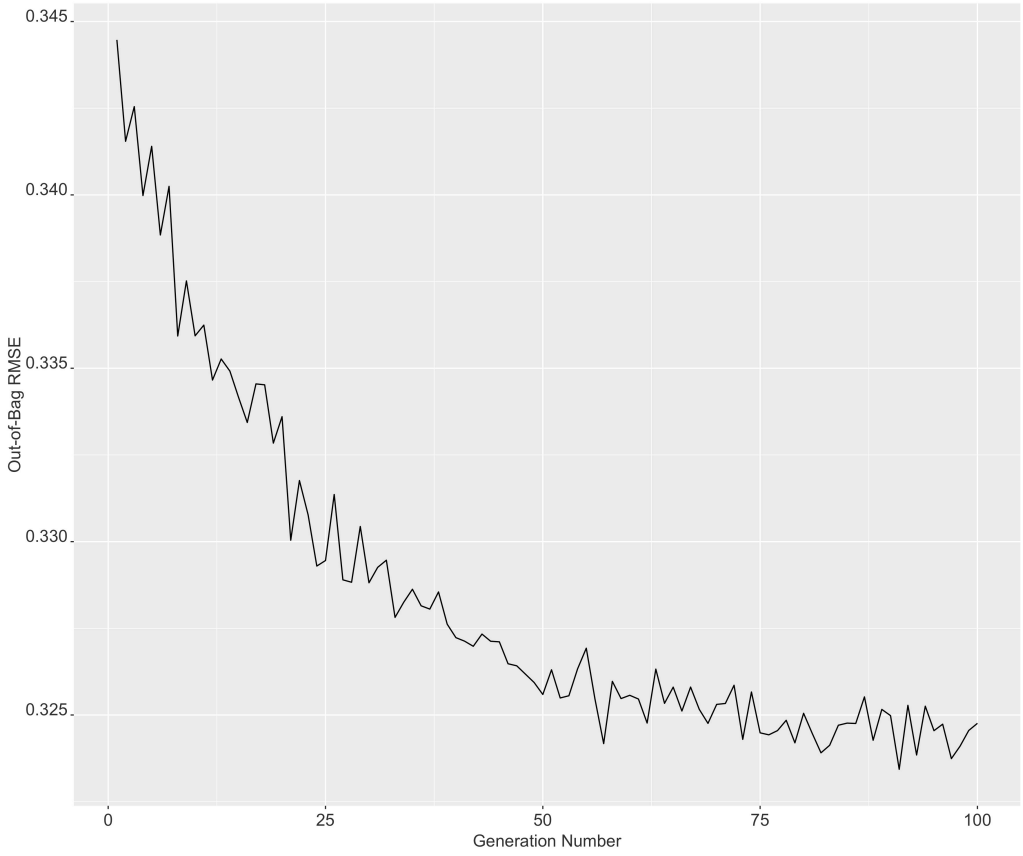


Figure 5-8: Out-of-Bag RMSE of the Integrated GA-Random Forest Model

It should be emphasized that the input features selected by the GA may be sensitive to the initial population when the algorithm converges to a local, rather than a global, optimal solution. To reach a robust model with the optimal set of input features, the integrated GA-random forest model is run for 25 realizations of initial

populations, and the out-of-bag RMSE values are obtained for each generation within each realization. As it can be noticed from Figure 8, the out-of-bag RMSE decreases over the generations, but the rate of decrease of the out-of-bag RMSE can be considered negligible after around 60 generations. As such, for all GA-random forest runs, the criterion is fixed as 60 generations. Table 5-2 shows the out-of-bag RMSE, R^2, generation at which the minimum out-of-bag RMSE was attained (i.e., the Best Generation), and the number of variables selected at the best generations (i.e., Number of Variables). The average out-of-bag RMSE and R^2 are 0.32 and 0.79 respectively, whereas the average number of selected variables is 18.

Table 5-2: GA-Random Forest Model Results for the 25 Initial Population Realizations

Run Number	Out-of-Bag RMSE	R^2	Best Iteration	Number of Variables
1	0.3243	0.7842	45	21
2	0.3211	0.7867	58	9
3	0.3162	0.7945	60	21
4	0.3241	0.7797	59	18
5	0.3201	0.787	60	10
6	0.3257	0.7812	53	18
7	0.3209	0.7897	48	27
8	0.3255	0.7856	50	15
9	0.3184	0.7902	60	18
10	0.3137	0.7945	59	11
11	0.317	0.7924	60	24
12	0.3258	0.7812	52	30

13	0.3244	0.7829	59	18
14	0.3145	0.7926	54	15
15	0.3324	0.7712	54	13
16	0.3256	0.7819	60	18
17	0.3216	0.7856	42	21
18	0.3219	0.7845	59	18
19	0.3148	0.7937	59	15
20	0.321	0.7874	55	21
21	0.3326	0.7718	54	10
22	0.3181	0.7934	60	13
23	0.3235	0.7827	59	15
24	0.3227	0.7858	56	27
25	0.3236	0.7822	46	19

To select the optimum set of input features, a frequency analysis for the features selected at the best iterations is conducted as shown in Figure 5-9. It can be noticed that some features are selected consistently (i.e., DS1, DS2, DS3, DS5, DS6, DS7, L1, T1, and T2), whereas other features are selected only a few times (i.e., H2 and S9), which shows how some features are more significant for predicting property damage compared to others.

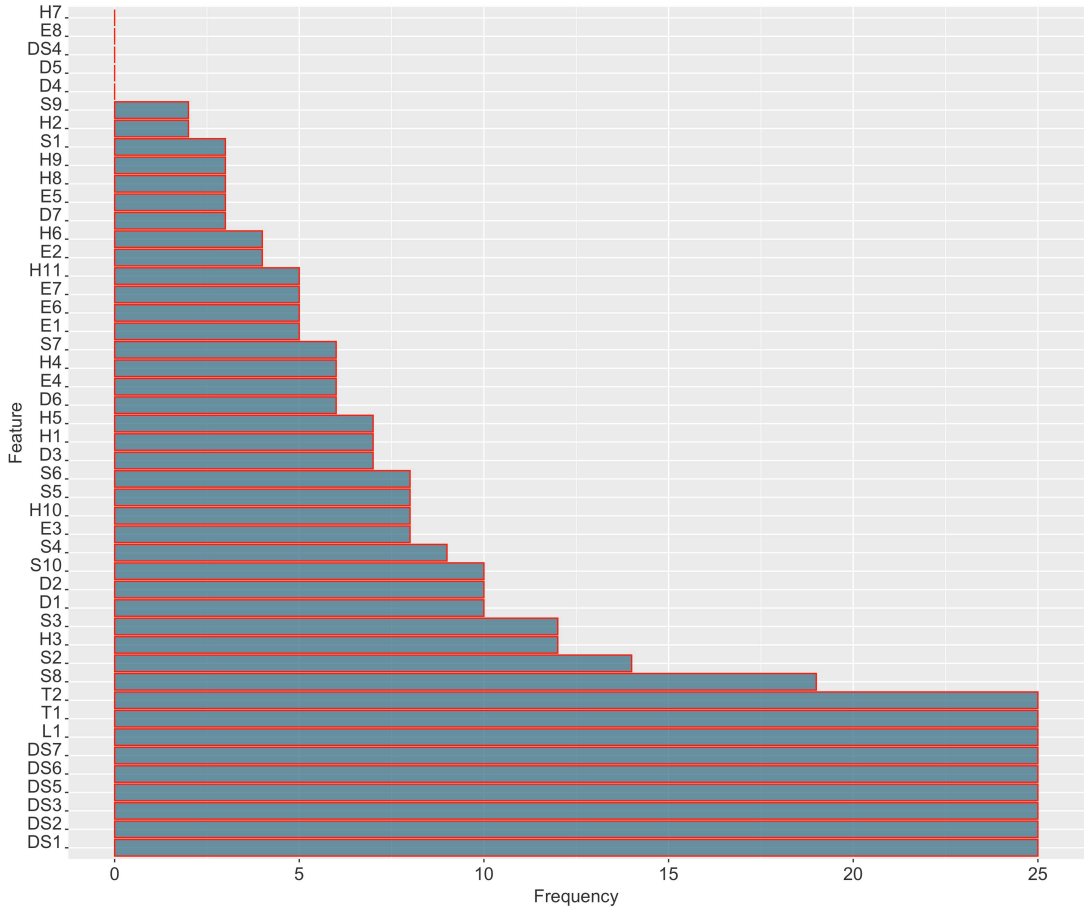


Figure 5-9: The Frequency of Selecting the Input Features within the GA-Random Forest Model

5.3.3. PHASE 3 – MODEL DEVELOPMENT

To select the most efficient model which requires the least number of input features to reach a considerably high level of prediction accuracy, classic regression trees together with regression trees ensemble techniques (i.e., bagging, random forest, and boosting) are trained using four different sets of input features (i.e., four models are developed). For all models, the dataset is divided into a training and a testing subset with a ratio of 70/30. Model 1 is developed using all input features listed in

Table 1, whereas Model 2 is developed based on the feature ranking provided by the random forest feature importance, the Boruta algorithm, and the boosting relative influence algorithm. The ten highest ranked features obtained from each of the three algorithms are compared and those repeated twice, or more are selected as inputs for Model 2. As such, the input features for Model 2 are the magnitude, event type, duration, year, month, latitude, maximum temperature, minimum temperature, per capita income, and number of households built after the year 2000. Model 3 is developed using the nine features that are selected at every realization of the GA-random forest model, whereas Model 4 is developed using the nine features used in Model 3 together with the features that were selected in more than half of the genetic algorithm realizations conducted (i.e., those selected 12 times or more). The performance of all models is evaluated based on RSME and R^2, as shown in Table 5-3.

Table 5-3: Summary of Models Performance

Model Number	Modelling Technique	RMSE	R^2
Model 1	Classic Regression Tree	0.38	0.69
	Bagging	0.36	0.72
	Random Forest	0.34	0.76
	Boosting	0.37	0.71
Model 2	Classic Regression Tree	0.38	0.70
	Bagging	0.36	0.72
	Random Forest	0.32	0.78
	Boosting	0.35	0.74

Model 3	Classic Regression Tree	0.41	0.65
	Bagging	0.38	0.70
	Random Forest	0.31	0.79
	Boosting	0.36	0.73
Model 4	Classic Regression Tree	0.41	0.65
	Bagging	0.38	0.70
	Random Forest	0.33	0.78
	Boosting	0.35	0.75

It can be observed that, on average, Model 2 is the best performing model in terms of both its RMSE and R^2 values. Turning to the best modelling technique, it is apparent that random forests outperform the three other modelling techniques (i.e., classic regression trees, bagging, and boosting) for all of the four models. It can be also concluded that the reduced set of input features selected in Model 2 is able to accurately predict the property damages. As such, Figure 5-10 shows the actual versus predicted damages for all techniques used in Model 2, being the most efficient model for property damage prediction.

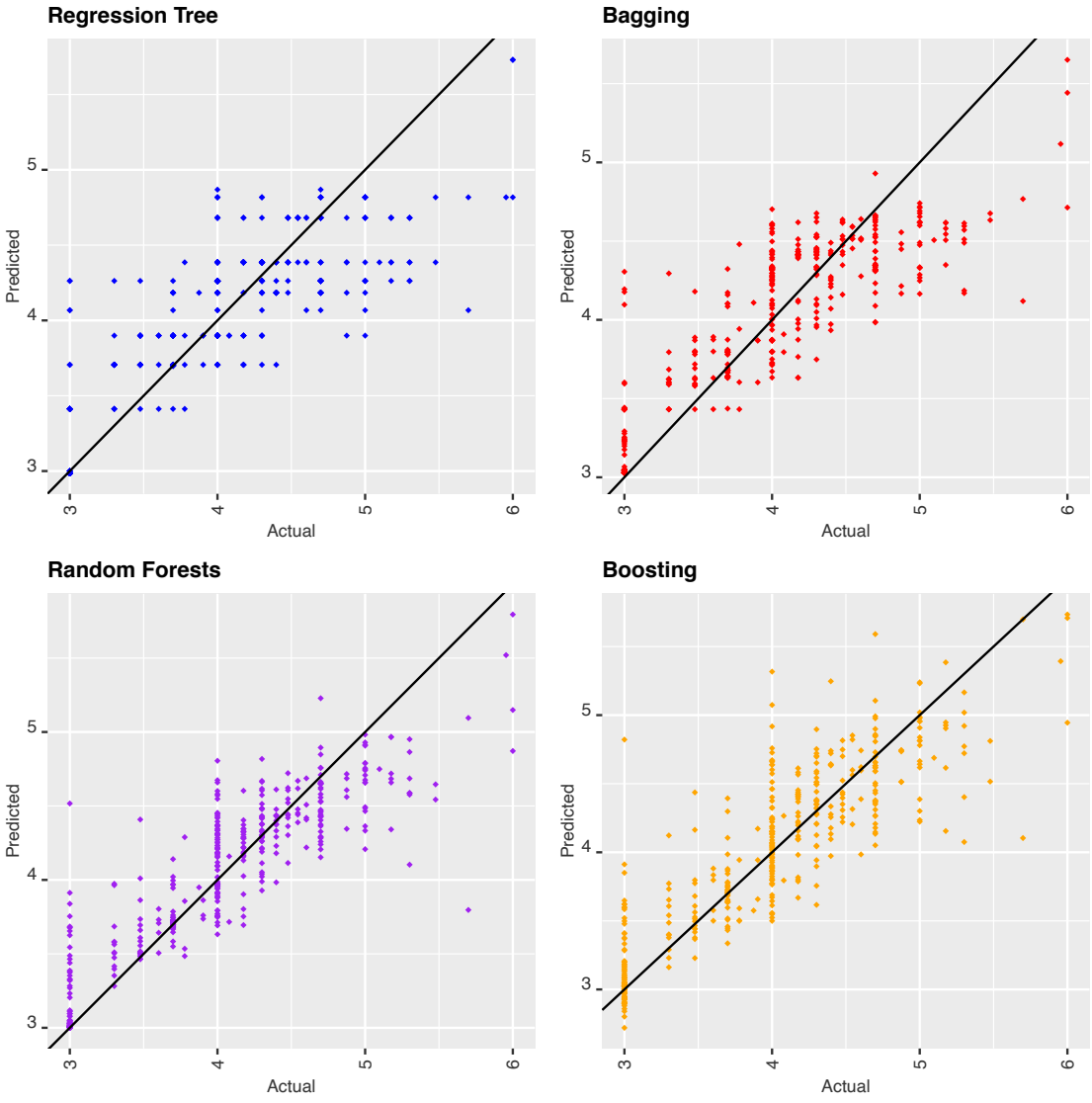


Figure 5-10: Model 2 Predicted verses Actual Property Damages

5.3.4. PHASE 4 – RESULTS ANALYSIS AND INTERPRETATION

PARTIAL DEPENDENCE PLOTS

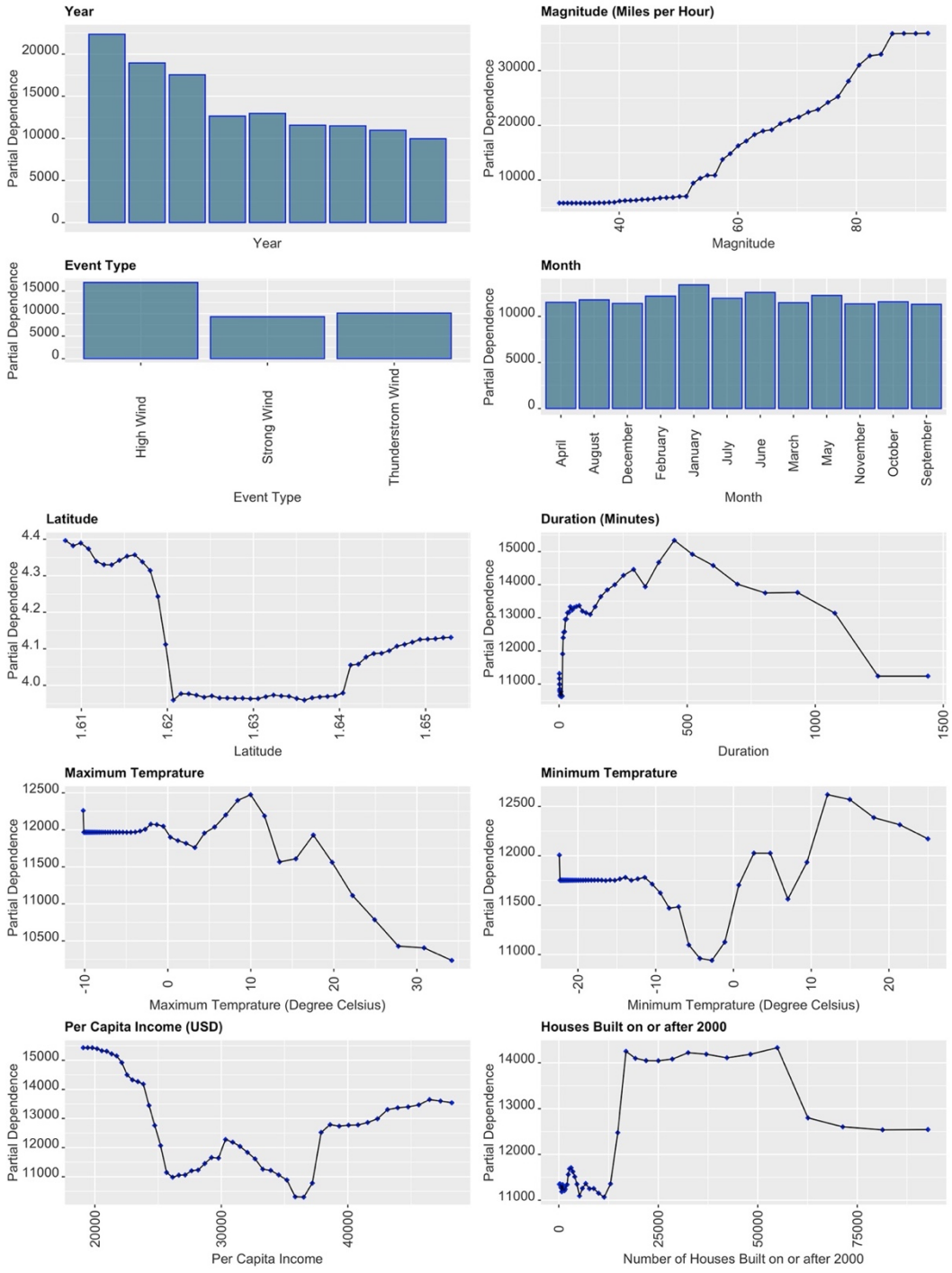
The last phase of the framework aims at understanding the relationships between the model inputs and the predicted property damages based on the selected model

(Model 2). Thus, the effect of changes in the input features on the property damages are depicted using single variable PDP, as shown in Figure 5-11(a). At each value of the input feature under consideration, the dependent variable is quantified for all values of the other input features, and subsequently the arithmetic mean is used as a representative value. The PDP shows that property damages increase significantly as the wind speed increases between 52 mph and 83 mph. The effect of location on property damages can be depicted through the change in the predicted damages with changes in event latitude. On average, higher property damages are expected in the southern part of New York (i.e., latitudes below 42) compared to the mid-state and up-state locations. Turning to the relationship between duration and property damages, it can be observed that for durations less than 500 min, as the event duration increases, the predicted property damages increase, whereas the duration and property damages have an inverse relationship between 500 min and 1250 mins. For the effect of temperature on property damages, it can be observed that more property damages are expected during extremely cold and extremely hot days when the maximum temperature is below 15°C and the minimum temperature is above 10°C, respectively. Economic features can also affect wind disaster property damages as damages generally increase with the increasing income for incomes exceeding US $37,000. The relationship between the number of houses built after the year 2000 and property damages shows that property damages are highest when the number of houses built on or after the year 2000 is between 12,500 and 57,500 houses.

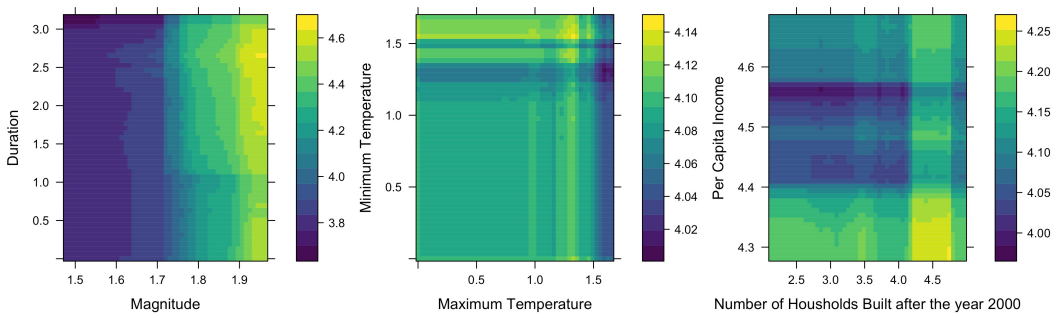
The wind type is also important for the effective prediction of property damage as it can be observed that high wind (i.e., sustained non-convective winds of 40 mph or greater lasting for 1 hour or longer, or gusts of 58 mph or greater for any duration) produces higher property damages than both thunderstorm (i.e., winds arising from thunderstorms) and strong wind (i.e., non-convective winds gusting less than 58 mph, or sustained winds less than 40 mph). Inspecting the relationship between predicted property damages and the month when an event occurred shows that property damages are highest in January compared to all other months.

The effect of the interaction between different pairs of input variables on the predicted property damages can be visualized in Figure 5-11(b) in logarithmic scale. The figure shows the interaction between the magnitude and duration, the maximum and minimum temperatures, and the per capita income and the number of households built after the year 2000. The interaction between wind speed and duration shows that generally property damages increase with the increase in wind speed regardless of the duration of the wind event. Nevertheless, as wind speed exceeds 80 mph, higher property damages are expected as the wind event duration increases. Moreover, the effect of the interaction between the maximum and minimum temperatures on property damages shows that higher property damages are expected with the increase in minimum temperature or the decrease in maximum temperature. However, the highest property damages were observed at very hot days when both the maximum and minimum temperatures were

considerably high. Finally, the interaction between the per capita income and the number of households built after the year 2000 shows that for incomes between US$ 25,000 and US$ 35,000 together with less than 10,000 houses bult after the year 2000, property damages are expected to be considerably lower than those expected in places where the per capita income is lower than US$ 25,000 and the number of houses built after the year 2000 is higher than 10,000.


Year
Partial Dependence
Year
Magnitude (Miles per Hour)
Partial Dependence
Magnitude
Event Type
Partial Dependence
High Wind
Strong Wind
Thunderstrom Wind
Event Type
Month
Partial Dependence
April
August
December
February
January
July
June
March
May
November
October
September
Month
Latitude
Partial Dependence
Latitude
Duration (Minutes)
Partial Dependence
Duration
Maximum Temprature
Partial Dependence
Maximum Temprature (Degree Celsius)
Minimum Temprature
Partial Dependence
Minimum Temprature (Degree Celsius)
Per Capita Income (USD)
Partial Dependence
Per Capita Income
Houses Built on or after 2000
Partial Dependence
Number of Houses Built on or after 2000

(a)



(b)

Figure 5-11: Partial Dependence Plots for Model 2 Random Forest

INSIGHTS FOR DECISION MAKING

Several insights can be drawn from the case study conducted herein. The first and most important insight is related to the complexity of CID-related damages prediction. This complexity is apparent as different categories of input features were proven to be significant for the prediction of such damages, ranging from disaster-related, economic, housing, and climate attributes. This highlights the need for accurately recording such features for the effective prediction of CID-related aspects in the future. The current case study also ensures the effectiveness of data-driven and machine learning techniques for predicting complex phenomena that would otherwise be hard to correlate. The results of applying Phase 4 on the current demonstration application asserts that the only drawback of using machine learning techniques which is the fact that they work as a blackbox can be eliminated through employing different techniques that can interpret and uncover the latent relationships learned by the model, these techniques include but are not limited to

partial dependence plots and accumulated local effects. Furthermore, the use of the feature selection methods proposed in Phase 2 resulted in having a model that both resulted in higher predictive ability and proved to be more efficient as it required considerably less inputs to achieve such performance. Moreover, it was shown in Phase 3 that the performance of the random forest variable importance, Boruta algorithm, and boosting relative influence matched that of genetic algorithms. This asserts that the use of a more resource demanding computational approach (i.e., genetic algorithms) for feature selection does not necessarily result in increased prediction accuracy.

5.4. CONCLUSIONS

The current work aims at developing a systematic data-driven framework for predicting Climate-Induced Disaster (CID) damages. The damage prediction framework (DFP) is comprised of four phases which are: (1) data collection and fusion, (2) feature selection, (3) model development, and (4) result analysis and interpretation. The first phase (i.e., data collection and fusion) comprises collecting the input data which includes hazard-, community, and climate-related attributes. Data collection is followed by a compilation process to integrate the collected data into a single database that can be used for the development of a CID damage prediction model. The second phase of the framework (i.e., feature selection) aims at removing the redundant information present in the database through selecting a set of features that is significantly important for predicting the output of interest.

Upon feature selection, the third phase (i.e., model development) involves developing multiple machine learning models using different supervised machine learning techniques (i.e., classical and ensemble decision trees, artificial neural networks, support vector regression). Such models are subsequently tested using different sets of input-output pairs and several measures are used for performance evaluation [i.e., Root Mean Squared Error (RMSE), Coefficient of Determination (R^2)]. After selecting the best performing model, the fourth phase of the framework (i.e., result analysis and interpretation) involves interpretating the developed machine learning model. In this phase, relationships between the model inputs and the predicted output are uncovered to decode the black-box nature of the machine learning techniques employed.

To demonstrate its utility, the framework was used to estimate the wind-related property damages incurred in the state of New York from 2010 to 2018 based on disaster characteristics, landcover, climatic conditions, socioeconomic attributes, housing information, and demographic statistics. It should be emphasized that wind-related disasters are recorded at scattered locations across the state of New York), whereas climate, social, economic, housing, and demographic data are collected at specific locations within each county. To have a fully integrated database, all input features are predicted at disaster locations using an inverse distance weighting (IDW) spatial interpolation method. This method enables estimating feature values at unknown points deterministically using a combination of weights from nearby points. Within such method, weights from

further points are typically smaller compared to those from nearby points, and thus they have less effects on the value to be estimated. Prior to developing the data-driven model, the most important features for estimating the property damages were selected using a set of filter (i.e., correlation matrix), wrapper (i.e., the Boruta algorithm, and genetic algorithms) and embedded (i.e., random forest variable importance, and boosting relative influence) feature selection methods. Consequently, four different sets of significant input features were identified based on the rankings obtained from the feature selection methods employed. Regression tree models with- and without ensemble techniques (i.e., bagging, random forest, and boosting) were developed based on each of the four input feature sets. For each set, the performance of the corresponding predictive models was compared based on the RMSE and R^2 values. Such comparison revealed that the best performing model (i.e., RMSE = 0.32, and $R^2 = 0.79$) is a random forest-based model with the following input variables: wind magnitude, wind type, wind duration, year, month, latitude, maximum temperature, minimum temperature, per capita income, and the number of households built after the year 2000.

To uncover the relationships between the property damages and the model inputs, single feature partial dependence plots were formulated for the random forest-based model (i.e., the best performing model as described earlier). These plots showed that: (1) property damages increase significantly as wind speed increases between 52 and 83 mph, (2) higher property damages are expected in the southern part of New York, (3) property damages increase as the event duration

increases for durations less than 500 min, (4) higher property damages are expected during extremely cold and extremely hot days when the maximum temperature is below 15°C and the minimum temperature is above 10°C, respectively, (5) property damages increase with the increasing per capita income for incomes exceeding US $37,000, (6) property damages are high when the number of houses built on or after the year 2000 is between 12,500 and 57,500 houses, (7) sustained, non-convective winds of 40 mph or greater lasting for 1 hour or longer, or gusts with a speed larger than or equal 58 mph produce higher property damages compared to those induced by other wind types, and (8) property damages are significantly high in January compared to all other months.

Given the proposed framework's ability to develop an efficient and effective wind-related property damage prediction model, it can be considered a step forward in enhancing the preparedness of governments for CID, and thus alleviating their adverse impacts and reaching more resilient communities. To advance the developed framework, further research can be employed through considering the following key points: (1) implementing several spatial interpolation techniques (i.e., other point and area interpolation methods), (2) collecting infrastructure systems data (i.e., system maintenance information), and integrating it into the database, (3) employing other methods to uncover the relationships between the model variables (i.e., accumulated local effects).

5.5. ACKNOWLEDGEMENTS

The authors are grateful to the financial support of the Ontario Trillium Scholarship Program and the Natural Sciences and Engineering Research Council (NSERC) of Canada, NSERC-CaNRisk-CREATE program. The authors would also like to acknowledge the fruitful discussions with the research teams of the INViSiONLab and the INTERFACE institute.

5.6. DATA AVAILABILITY

The datasets related to this article are publicly available and are provided by the following organizations: (1) NWS, National Weather Services (data available at https://www.ncdc.noaa.gov/stormevents/%5Cnfiles/5576/stormevents.html), (2) NLCD, National Land Cover Database (data available at https://www.usgs.gov/core-science-systems/science-analytics-and-synthesis/gap/science/land-cover-data-download?qt-science_center_objects=0#qt-science_center_objects), (3) NOAA, National Oceanic and Atmospheric Administration online search tool (data available at https://www.ncdc.noaa.gov/cdo-web/search), and (4) ACS, American Community Survey (data available at https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2018/).

5.7. REFERENCES

[1] V. Thomas and R. López, "Global Increase In ClImate-Related DIsasters," 2015.

[2] P. and Guha-sapir, D., Hoyois and R. Below, "Annual Disaster Statistical Review 2014: The numbers and trends," 2015.

[3] E. Newburger, "Disasters cause $210 billion in damage in 2020," 2021. [Online]. Available: https://www.cnbc.com/2021/01/07/climate-change-disasters-cause-210-billion-in-damage-in-2020.html. [Accessed: 22-Mar-2021].

[4] C. Flavelle, "U.S. Disaster Costs Doubled in 2020, Reflecting Costs of Climate Change," 2021. [Online]. Available: https://www.nytimes.com/2021/01/07/climate/2020-disaster-costs.html. [Accessed: 22-Mar-2021].

[5] V. Limaye, "Shattering Records, Climate Disasters Fueled Misery in 2020 | NRDC," 2021. [Online]. Available: https://www.nrdc.org/experts/vijay-limaye/shattering-records-climate-disasters-fueled-misery-2020. [Accessed: 22-Mar-2021].

[6] World Economic Forum, "The Global Risks Report," vol. 15, pp. 1–114, 2020.

[7] R. Lopez, V. Thomas, and P. Troncoso, "Impacts of Carbon Dioxide

Emissions on Global Intense Hydrometeorological Disasters," *Clim. Disaster Dev. J.*, vol. 4, no. 1, pp. 30–50, 2020.

[8] World Health Organization, "Climate change and health," 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health. [Accessed: 06-Jun-2019].

[9] "Natural Disasters Could Cost 20 Percent More By 2040 Due to Climate Change - Yale E360," 2020. [Online]. Available: https://e360.yale.edu/digest/natural-disasters-could-cost-20-percent-more-by-2040-due-to-climate-change. [Accessed: 02-Jan-2021].

[10] "New approaches to help businesses tackle climate change | University of Cambridge," 2020. [Online]. Available: https://www.cam.ac.uk/research/news/new-approaches-to-help-businesses-tackle-climate-change. [Accessed: 22-Mar-2021].

[11] V. Thomas, *Climate change and natural disasters: Transforming economies and policies for a sustainable future*. Routledge, London, 2017.

[12] K. Ganguly, N. Nahar, and M. Hossain, "A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh," *Int. J. Disaster Risk Reduct.*, vol. 34, no. December 2018, pp. 283–294, 2019.

[13] A. Kahira, B. Gomez, and R. Badia Sala, "A Machine Learning Workflow

for Hurricane Prediction," in *Book of abstracts. Barcelona Supercomputing Center*, 2018, pp. 72–73.

[14] M. Rodrigues and J. De la Riva, "An insight into machine-learning algorithms to model human-caused wildfire occurrence," *Environ. Model. Softw.*, vol. 57, pp. 192–201, 2014.

[15] A. Jaafari, E. Zenner, M. Panahi, and H. Shahabi, "Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability," *Agric. For. Meteorol.*, vol. 266–267, no. 2018, pp. 198–207, 2019.

[16] M. Hanewinkel, W. Zhou, and C. Schill, "A neural network approach to identify forest stands susceptible to wind damage," *For. Ecol. Manage.*, vol. 196, no. 2–3, pp. 227–243, 2004.

[17] P. J. Sallis, W. Claster, and S. Herna, "A machine-learning algorithm for wind gust prediction," *Comput. Geosci.*, vol. 37, pp. 1337–1344, 2011.

[18] C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S. Kim, "Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data," *Adv. Meteorol.*, vol. 2018, 2018.

[19] M. Khalaf *et al.*, "A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction," *2018 IEEE Congr. Evol. Comput.*, pp. 1–8, 2018.

[20] J. Diaz and M. B. Joseph, “Predicting property damage from tornadoes with zero-in fl ated neural networks,” *Weather Clim. Extrem.*, vol. 25, no. July 2018, p. 100216, 2019.

[21] S. F. Pilkington and H. N. Mahmoud, “Interpreting the socio-technical interactions within a wind damage-artificial neural network model for community resilience,” *R. Soc. Open Sci.*, vol. 7, no. 11, 2020.

[22] H. Toya and M. Skidmore, “Economic development and the impacts of natural disasters,” *Econ. Lett.*, vol. 94, no. 1, pp. 20–25, 2007.

[23] D. Mafi-Gholami, A. Jaafari, E. K. Zenner, A. Nouri Kamari, and D. Tien Bui, “Vulnerability of coastal communities to climate change: Thirty-year trend analysis and prospective prediction for the coastal regions of the Persian Gulf and Gulf of Oman,” *Sci. Total Environ.*, vol. 741, 2020.

[24] M. Skidmore and H. Toya, “Do natural disasters promote long-run growth?,” *Econ. Inq.*, vol. 40, no. 4, pp. 664–687, 2002.

[25] A. L. Shokane, “Social work assessment of climate change: Case of disasters in greater Tzaneen municipality,” *Jàmbá J. Disaster Risk Stud.*, vol. 11, no. 3, pp. 1–7, 2019.

[26] C. L. Gray and V. Mueller, “Natural disasters and population mobility in Bangladesh,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 16, pp. 6000–6005, 2012.

[27] V. Thomas, J. R. G. Albert, and C. Hepburn, "Contributors to the frequency of intense climate disasters in Asia-Pacific countries," *Clim. Change*, vol. 126, no. 3–4, pp. 381–398, 2014.

[28] K. C. Lam, R. G. Bryant, and J. Wainright, "Application of spatial interpolation method for estimating the spatial variability of rainfall in Semiarid New Mexico, USA," *Mediterr. J. Soc. Sci.*, vol. 6, no. 4S3, pp. 108–116, 2015.

[29] J. P. Musashi, H. Pramoedyo, and R. Fitriani, "Comparison of Inverse Distance Weighted and Natural Neighbor Interpolation Method at Air Temperature Data in Malang Region," *Cauchy*, vol. 5, no. 2, p. 48, 2018.

[30] Y. Meng, M. Cave, and C. Zhang, "Comparison of methods for addressing the point-to-area data transformation to make data suitable for environmental, health and socio-economic studies," *Sci. Total Environ.*, vol. 689, pp. 797–807, 2019.

[31] A. Rastogi, S. Sridhar, and R. Gupta, "Comparison of Different Spatial Interpolation Techniques to Thematic Mapping of Socio-Economic Causes of Crime Against Women," *2020 Syst. Inf. Eng. Des. Symp. SIEDS 2020*, 2020.

[32] M. Farfán, J. François Mas, and L. Osorio, "Interpolating Socioeconomic Data for the Analysis of Deforestation: A Comparison of Methods," *J. Geogr. Inf. Syst.*, vol. 04, no. 04, pp. 358–365, 2012.

[33] G. M. Laslett, A. B. Laslett, P. J. Pahl, and M. F. Hutchinson, "Comparison of several spatial prediction methods for soil ph," *J. Soil Sci.*, vol. 38, no. 2, pp. 325–341, 1987.

[34] Q. M. Ajaj, M. A. Shareef, N. D. Hassan, S. F. Hasan, and A. M. Noori, "GIS based spatial modeling to mapping and estimation relative risk of different diseases using inverse distance weighting (IDW) interpolation algorithm and evidential belief function (EBF) (Case study: Minor Part of Kirkuk City, Iraq)," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 185–191, 2018.

[35] L. Zhao, Z. Chen, Y. Hu, G. Min, and Z. Jiang, "Distributed Feature Selection for Efficient Economic Big Data Analysis," *IEEE Trans. Big Data*, vol. 4, no. 2, pp. 164–176, 2016.

[36] J. Tang and H. Liu, "Feature selection for social media data," *ACM Trans. Knowl. Discov. Data*, vol. 8, no. 4, pp. 1–27, 2014.

[37] A. Gümüşçü, M. E. Tenekeci, and A. V. Bilgili, "Estimation of wheat planting date using machine learning algorithms based on available climate data," *Sustain. Comput. Informatics Syst.*, vol. 28, 2020.

[38] A. Haidar and B. Verma, "A novel approach for optimizing climate features and network parameters in rainfall forecasting," *Soft Comput.*, vol. 22, no. 24, pp. 8119–8130, 2018.

[39] M. Haggag, A. Yorsi, W. El-Dakhakhni, and E. Hassini, "Infrastructure

performance prediction under Climate-Induced Disasters using data analytics," *Int. J. Disaster Risk Reduct.*, vol. 56, no. November 2020, 2021.

[40] M. Haggag, A. S. Siam, W. El-Dakhakhni, P. Coulibaly, and E. Hassini, "A deep learning model for predicting climate-induced disasters," *Nat. Hazards*, vol. 107, no. 1, pp. 1009–1034, 2021.

[41] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, *Geographic Information Systems and Science*, Second. 2005.

[42] W. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," *Econ. Geogr.*, vol. 46, pp. 234–240, 1970.

[43] N. S. N. Lam, "Spatial interpolation methods: A review," *Am. Cartogr.*, vol. 10, no. 2, pp. 129–150, 1983.

[44] GIS Resources, "Classification of Interpolation," *GIS Resources*, 2013. [Online]. Available: https://www.gisresources.com/classification-of-interpolation_2/. [Accessed: 01-Oct-2020].

[45] Y.-H. (Eva) Wu and M.-C. Hung, "Comparison of Spatial Interpolation Techniques Using Visualization and Quantitative Assessment," in *Applications of Spatial Statistics*, 2016.

[46] C. A. Schloeder, N. E. Zimmerman, and M. J. Jacobs, "Comparison of Methods for Interpolating Soil Properties Using Limited Data," *Soil Sci. Soc. Am. J.*, vol. 65, no. 2, pp. 470–479, 2001.

[47] Z. Kebaili Bargaoui and A. Chebbi, “Comparison of two kriging interpolation methods applied to spatiotemporal rainfall,” *J. Hydrol.*, vol. 365, no. 1–2, pp. 56–73, 2009.

[48] M. R. Holdaway, “Spatial modeling and interpolation of monthly temperature using kriging,” *Clim. Res.*, vol. 6, no. 3, pp. 215–225, 1996.

[49] F. Zohra and B. Largueche, “Estimating Soil Contamination with Kriging Interpolation Method,” *Am. J. Appl. Sci.*, vol. 3, no. 6, pp. 1894–1898, 2006.

[50] H. Ha, J. R. Olson, L. Bian, and P. A. Rogerson, “Analysis of heavy metal sources in soil using kriging interpolation on principal components,” *Environ. Sci. Technol.*, vol. 48, no. 9, pp. 4999–5007, 2014.

[51] P. A. Hancock and M. F. Hutchinson, “Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines,” *Environ. Model. Softw.*, vol. 21, no. 12, pp. 1684–1694, 2006.

[52] H. Yan, H. A. Nix, M. F. Hutchinson, and T. H. Booth, “Spatial interpolation of monthly mean climate data for China,” *Int. J. Climatol.*, vol. 25, no. 10, pp. 1369–1379, 2005.

[53] D. R. McNeil, T. J. Trullell, and J. C. Turner, “Spline interpolation of demographic oata,” *Demography*, vol. 14, no. 2, pp. 245–252, 1977.

[54] L. Smith, T. Australian, and S. N. Wood, “Spline Interpolation for

Demographic Variables :," vol. 21, no. 1, pp. 95–98, 2004.

[55] S. Raheel, "Feature Selection Techniques in Machine Learning with Python," *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Accessed: 23-Mar-2021].

[56] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, no. 1997, 2015.

[57] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[58] M. Shardlow, "An Analysis of Feature Selection Techniques," *Univ. Manchester*, vol. 14, no. 1, pp. 1–7, 2016.

[59] N. Sánchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter Methods for Feature Selection – A Comparative Study," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2007, pp. 178–187.

[60] Y. Charfaoui, "Hands-on with Feature Selection Techniques: Embedded Methods | by Younes Charfaoui | Heartbeat," *Heartbeat*, 2020. [Online]. Available: https://heartbeat.fritz.ai/hands-on-with-feature-selection-

techniques-embedded-methods-84747e814dab. [Accessed: 10-Dec-2020].

[61] S. Raschka, “What is the difference between filter, wrapper, and embedded methods for feature selection.” [Online]. Available: https://sebastianraschka.com/faq/docs/feature_sele_categories.html. [Accessed: 01-Jan-2021].

[62] S. KAUSHIK, “Introduction to Feature Selection methods with an example (or how to select the right variables?),” *Analytics Vidhya*, 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/. [Accessed: 01-Oct-2020].

[63] A. Nagpal, “Decision Tree Ensembles- Bagging and Boosting,” 2017. [Online]. Available: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9. [Accessed: 01-Mar-2019].

[64] M. Haggag, A. Yorsi, W. El-Dakhakhni, and E. Hassini, “Infrastructure performance prediction under Climate-Induced Disasters using data analytics,” *Int. J. Disaster Risk Reduct.*, vol. 56, no. February, pp. 1–11, 2021.

[65] M. Haggag, A. S. Siam, W. El-Dakhakhni, P. Coulibaly, and E. Hassini, “A deep learning model for predicting climate-induced disasters,” *Nat. Hazards*, pp. 1–26, 2021.

[66] C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. 2021.

[67] I. K. Kabul, "Interpret model predictions with partial dependence and individual conditional expectation plots - The SAS Data Science Blog," 2018. [Online]. Available: https://blogs.sas.com/content/subconsciousmusings/2018/06/12/interpret-model-predictions-with-partial-dependence-and-individual-conditional-expectation-plots/. [Accessed: 01-Jan-2021].

[68] National Weather Services, "Storm Events Database | National Centers for Environmental Information," 2019. [Online]. Available: https://www.ncdc.noaa.gov/stormevents/%5Cnfiles/5576/stormevents.html. [Accessed: 10-Oct-2019].

[69] U.S Geological Survey, "Land Cover Data Download," 2011. [Online]. Available: https://www.usgs.gov/core-science-systems/science-analytics-and-synthesis/gap/science/land-cover-data-download?qt-science_center_objects=0#qt-science_center_objects. [Accessed: 10-Oct-2020].

[70] National Oceanic and Atmospheric Administration, "Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC)," *National Centers for ENvironmental Information*, 2021. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/search. [Accessed: 01-Oct-2010].

[71] US Census Bureau, "2018 Data Profiles | American Community Survey | US Census Bureau," 2018. [Online]. Available: https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2018/. [Accessed: 01-Sep-2020].

Chapter 6

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

6.1. SUMMARY

The work presented in this dissertation focuses on enhancing urban resilience under Climate-Induced Disasters (CID) through employing different data analytics techniques. To achieve this objective, the first phase of this work aimed at conducting a thorough exploration of the field of infrastructure system resilience. Accordingly, a meta-research approach was employed to quantitatively and qualitatively assess previous relevant work in the field. Consequently, the contribution of the different research topics and the research gaps pertaining to systems resilience research were uncovered. It was shown that to enhance system resilience maximizing the resources that aid in predicting both the occurrences and impacts of disasters is key. As such, Phases 1,2 and 3 of this work focused on using data-driven modelling to predict the occurrences, system-related impacts, and lumped costs of CID. In the second phase, a deep learning modelling approach was proposed to predict CID occurrences by linking historical disaster records to climate change indices. This approach was validated using a case study focused on the province of Ontario, Canada. To relate CID to their impacts on critical infrastructure systems, the third phase of this work focused on linking CID occurrences to their system-related impacts by employing different data analytics techniques including text mining and predictive modelling. A framework was

proposed, and a case study was presented to assess the framework's applicability in the state of New York, USA. In the fourth and final phase, a framework for the prediction of CID costs was developed through integrating several types of input data including disaster-related, socio-economic, land cover, and climate features. The framework also employed diverse feature selection techniques to boost the prediction accuracy. As such, a case study was presented in the state of New York, USA.

6.2. CONCLUSIONS

The results of the research work conducted in this dissertation demonstrate the effectiveness of data analytics in explicating complex phenomena (i.e., CID occurrences and impacts). It can now be inferred that the use of machine learning techniques in predicting CID-related aspects presents an effective and efficient tool compared to the standard methods which include using physical and mathematical models. This efficiency was demonstrated through both the superior prediction accuracy and the reduced computational effort of the models employed herein.

Specifically, the use of text mining in the form of topic modelling was proved to be a very efficient approach in conducting a thorough critical review. Together with its ability to uncover research topics and their respective contribution to the field, the developed models can also be used to identify research gaps. In terms of the infrastructure systems resilience research field exploration, it was shown that

quantifying systems interdependence, especially the relationships between power, water, and gas systems has the most contribution in the research carried out in the last couple of decades. It was also highlighted that quantifying systems resilience represents a key research gap despite the availability of several theoretically proposed metrics. Specifically, resilience was either quantified based on repair time (i.e., rapidity), or loss of system functionality (i.e., robustness). However, previous research failed to quantify the means to reach such a rapid and robust (resilient) system, which include, the availability of redundant components in the system (i.e., redundancy), the availability of resources in the system that can help diagnose and control system failure (i.e., resourcefulness). Moreover, a cross-cutting gap related to disasters and subsequent disruption of infrastructure systems was identified which showed that there is a pressing need to predict the occurrence of these events, and thus, prepare the system accordingly which in return can optimize relevant system's resources and significantly enhance its resilience capabilities. This is especially important given the rising complexity of modern city systems, and the fact that the number of CID is anticipated to double within the next 15 years (Lopez, Thomas, & Troncoso, 2020).

Despite the fact that the increased frequency of heat waves and droughts, more intense hurricanes, tornadoes and snow storms and exaggerated floods was generally attributed to climate change, previous research failed to link the occurrences of these disasters (i.e., CID) to climate change in a quantitative manner (Callery, 2018; "Climate change | EU Science Hub," 2018; Shaftel, 2018a, 2018b).

As such, one of the key contributions of the work conducted in Phase 2 herein is that it successfully linked CID occurrences to climate change using standardized climate change indices that quantify the changes in both temperature and precipitation over the next decades depending on different greenhouse gas emissions scenarios. It was shown that flood disasters in North America were strongly correlated with temperature variations which supports the direct and strong relationship that exists between temperature changes and flooding. These temperature variations include the difference between the daily maximum and minimum temperatures, the number of days with minimum temperature below 10^{th} percentile, and the number of days with minimum temperature below zero degrees Celsius.

Employing deep learning neural networks to model CID occurrences showed that increasing the number of hidden layers in a neural network may sometimes cause the network to overfit which results in lower prediction accuracy as was previously confirmed by the universal approximation theorem (Yu Dong & Li, 2012; Kumar, 2019; Sanger, 1989; Stathakis, 2009). Additionally, it was shown that increasing the number of hidden neurons doesn't necessarily result in an increased model accuracy. This can be attributed to overfitting which is especially common when the training data is not enough to train the considered number of neurons, and as such the model fails in terms of generalizability as it memorizes the training data. Furthermore, it was also shown that when the number of neurons in the input layer is much larger than that in the output layer, adopting a narrowing

neural network (i.e., inverted pyramid network architecture) leads to data noise removal since, as each layer is further narrowed, the model is forced to drop irrelevant information which explains why such models were shown to yield higher accuracy at lower computational cost (Czanner et al., 2015; Srivastava, 2019).

In terms of predicting the impacts of CID on critical infrastructure systems, the analysis showed that the use of text mining in the form of bag of words and n-gram analysis can facilitate linking disasters to their respective infrastructure damages. It was shown that in North America, the most affected infrastructure system by wind-related disasters is the power system followed by the transportation system. It was also asserted that the power damage gets more severe with higher wind speed, whereas the changes in wind duration didn't affect the severity of damage as much. The results of the analysis conducted herein also showed that most of the power system damages took place over May through August and were due to thunderstorms and high winds. It was highlighted that throughout the last decade, the eastern part of New York was highly susceptible to power system failure due to wind-related disasters which calls for introducing either redundant overhead or underground cables and/or resources to achieve higher overall power system robustness. It was also shown that power system damage severity does not only depend on hazard-related characteristics (i.e., magnitude and duration), but it is also a function of the inherent system properties and the hazard-system interaction.

The work conducted in Phase 3 also confirms that ensemble techniques can be effectively used to boost the performance of both regression and decision trees. Among these techniques, random forests proved to be extremely reliable in predicting CID-related aspects (i.e., both system related impacts and lumped impacts). Moreover, after assessing the performance of different models before and after employing data imputation (i.e., filling missing records), it was clear that in some cases data imputation might introduce bias in the model especially when the attributes to be filled have more than 10% of their values missing (Yiran Dong & Peng, 2013; Jakobsen, Gluud, Wetterslev, & Winkel, 2017).

The usefulness of using machine learning models in predicting CID-related aspects was illustrated in Phase 4. It was shown that a diverse range of data categories were significant for the prediction of CID lumped damages, ranging from disaster-related to economic, to housing, and climate attributes. It was also established that the only drawback of using machine learning techniques (i.e., their black box nature) can be overcome through employing different techniques that have the ability to interpret and uncover the latent relationships learned by the model, which include partial dependence plots, and accumulated local effects. In addition, the use of feature selection techniques proved to both increase the prediction accuracy as well as decrease the model's computational time due to the exclusion of redundant input features. Moreover, it was highlighted that the performance of the random forest variable importance, Boruta algorithm, and boosting relative influence matched that of genetic algorithms in feature selection

which asserts that the use of a more resource demanding computational approach (i.e., genetic algorithms) for feature selection does not necessarily result in increased prediction accuracy.

Upon assessing wind disaster property damages in North America, it was shown that these damages are decreasing in the last decade. It was also highlighted that for wind magnitudes between 52 and 83 mph, as wind speed increases the property damages induced by wind disasters increase significantly. Unlike power system damage severity, it was shown that for durations less than 500 min, property damages increase as the event duration increases. It was also shown that instead of increasing in the summer as with power system damage severity, property damages generally are higher in January. The vulnerability of Southern New York State to wind disasters was also emphasized as higher property damages are expected around this part of the state.

6.3. RECOMMENDATIONS FOR FUTURE RESEARCH

The work conduced herein focused on enhancing urban resilience through employing data-driven modelling to predict CID-related aspects. In light of this work, the following issues still require further analysis and investigation:

(1) This work aims to enhance the resilience of urban areas under CID through boosting their resourcefulness. This is accomplished through employing data analytics and machine learning techniques to predict CID occurrences

and impacts both on system and community levels. Since redundancy, along with resourcefulness, is key to enhance system resilience, it is considered extremely crucial to conduct further research to optimize such metric to enhance the resilience of urban systems. Consequently, conducting further research which aims at uncovering critical system components and assessing the economic feasibility of adding such replacement components to the system is essential.

(2) Given the fact that the results of the field exploration conducted in Phase 1 of this dissertation is dependant on the range of publications used in developing the topic model, it is envisioned that, as more relevant research articles are published, the topic model, and the subsequent critical analysis of the extracted topics should be updated. With the massive research being conducted in the field of infrastructure resilience, it is expected that the gaps uncovered herein will be tackled, and other ones will emerge over the next few decades.

(3) The high predictive ability of the modelling approach presented in Phase 2 of this work affirms that accurate disaster predictions can be reached given the availability of climate variability data. However, the use of annually calculated climate change indices is affecting the model's overall utility. Nonetheless, currently available climate change indices are only available on a yearly basis as the compilation of a daily dataset is extremely difficult. Nevertheless, the Expert Team on Climate Change Detection and Indices is

currently working on higher resolution data which can enhance the utility of the developed model as such higher resolution data become available in the future.

(4) The framework proposed in Phase 3 of this work is expected to facilitate mitigating the adverse impacts of CID on infrastructure systems, and therefore improve the overall urban resilience under such disasters. However, further research can be implemented to advance the developed framework through incorporating detailed system-related data which can increase the model's accuracy, using different data imputation methods to eliminate the bias introduced after filling data records in several features, and integrating the duration and monetary cost of system disruption to further enhance the utility of the model.

(5) The key advantage of the framework presented in Phase 4 resides in the fact that it was able to overcome the complexity of predicting CCID lumped impacts through integrating several data types in a single throughout database. Nevertheless, the case study demonstrated in this work applied point interpolation to integrate all data types which is the technique that is successfully employed in similar applications in the literature, however it still might not be the best spatial interpolation method for all features considered in the study. As such, to further advance the presented work, several techniques can be used to interpolate different input features and the best performing technique can be chosen for each feature distinctly.

(6) Although employing Partial Dependence Plots (PDP) to uncover latent relationships between the output and different model inputs facilitates the future implementation of machine learning methods in CID-related aspects prediction, it is noteworthy to mention that PDP is only informative when the input features are not strongly correlated (Kabul, 2018; Molnar, 2021). As such other techniques may be used in case strong correlations exist between features. One of these techniques is Accumulated Local Effects (ALE) which is an unbiased approach that depicts the input-output relationships based on the conditional distribution of the input features. ALE is thus preferred over PDP when input variables are strongly correlated as it depicts the difference in prediction values rather than their averages which aids in eliminating the effect of correlated input features (Molnar, 2021).

6.4. REFERENCES

Callery, S. (2018). Effects | Facts – Climate Change: Vital Signs of the Planet. Retrieved December 5, 2018, from https://climate.nasa.gov/effects/

Climate change | EU Science Hub. (2018). Retrieved November 15, 2018, from https://ec.europa.eu/jrc/en/research-topic/climate-change

Czanner, G., Sarma, S. V., Ba, D., Eden, U. T., Wu, W., Eskandar, E., … Brown, E. N. (2015). Measuring the signal-to-noise ratio of a neuron. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7141–7146. https://doi.org/10.1073/pnas.1505545112

Dong, Yiran, & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1–17. https://doi.org/10.1186/2193-1801-2-222

Dong, Yu, & Li, D. (2012). Efficient and effective algorithms for training single-hidden-layer neural networks. *Pattern Recognition Letters*, *33*(5), 554–558. https://doi.org/10.1016/j.patrec.2011.12.002

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 1–10. https://doi.org/10.1186/s12874-017-0442-1

Kabul, I. K. (2018). Interpret model predictions with partial dependence and

individual conditional expectation plots - The SAS Data Science Blog. Retrieved January 1, 2021, from https://blogs.sas.com/content/subconsciousmusings/2018/06/12/interpret-model-predictions-with-partial-dependence-and-individual-conditional-expectation-plots/

Kumar, N. (2019). *Illustrative Proof of Universal Approximation Theorem*. Retrieved from https://hackernoon.com/illustrative-proof-of-universal-approximation-theorem-5845c02822f6

Lopez, R., Thomas, V., & Troncoso, P. (2020). Impacts of Carbon Dioxide Emissions on Global Intense Hydrometeorological Disasters. *Climate, Disaster and Development Journal*, *4*(1), 30–50. https://doi.org/10.18783/cddj.v004.i01.a03

Molnar, C. (2021). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

Sanger, T. (1989). Optimal unsupervised learning in a single-layer network. *Neural Networks*, *2*, 459–473.

Shaftel, H. (2018a). Causes | Facts – Climate Change: Vital Signs of the Planet. Retrieved March 15, 2019, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/causes/

Shaftel, H. (2018b). Evidence | Facts – Climate Change: Vital Signs of the Planet.

Retrieved November 15, 2018, from NASA's Jet Lab Propulsion Laboratory California Institute of Technology website: https://climate.nasa.gov/evidence/

Srivastava, S. (2019). *On Foveation of Deep Neural Networks* (Massachusetts Institute of Technology). Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/123134/1128816526-MIT.pdf?sequence=1&isAllowed=y

Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing*, *30*(8), 2133–2147. https://doi.org/10.1080/01431160802549278