5-1-2024

IMPACT OF DATA RELIABILITY ON RESILIENCE-BASED DECISION MAKING IN A WATER DISTRIBUTION SYSTEM

Amrit Babu Ghimire
Southern Illinois University Carbondale, amrit.ghimire@siu.edu

IMPACT OF DATA RELIABILITY ON RESILIENCE-BASED DECISION MAKING IN A WATER DISTRIBUTION SYSTEM

by

Amrit Ghimire

B.E., Tribhuvan University, Nepal, 2018

A Thesis
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

School of Civil, Environmental, and Infrastructure Engineering
in Graduate School
Southern Illinois University Carbondale
May 2024

THESIS APPROVAL

IMPACT OF DATA RELIABILITY ON RESILIENCE-BASED DECISION MAKING IN A WATER DISTRIBUTION SYSTEM

by

Amrit Ghimire

A Thesis Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Civil Engineering

Approved by:

Dr. Sangmin Shin

Dr. Ajay Kalra

Dr. Jia Liu

Graduate School
Southern Illinois University Carbondale
December 20, 2023

AN ABSTRACT OF THE THESIS OF

Amrit Ghimire, for the Master of Science degree in Civil Engineering, presented on December 20, 2023, at Southern Illinois University Carbondale.

TITLE: IMPACT OF DATA RELIABILITY ON RESILIENCE-BASED DECISION MAKING IN A WATER DISTRIBUTION SYSTEM

MAJOR PROFESSOR: Dr. Sangmin Shin

This thesis explores the increasing necessity for resilience in Water Distribution Systems (WDS) facing challenges like leakage, missing data, and cyber-physical attacks. Resilience-based strategies enhance WDS sustainability by minimizing losses and ensuring quick recovery. Integrating smart systems, utilizing real-time data, boosts infrastructure resilience by improving efficiency and responsiveness. Data precision is essential for practical system analysis, development of resilience strategy, and making real-time decisions. The study also investigates into the potential of decentralization, combined with smart systems, to enhance WDS resilience, considering diverse water resources and hybrid systems. It seeks to answer critical questions about resilience in various failure scenarios and the impact of deviating pressure values at demand nodes.

Acknowledging vulnerabilities introduced by smart systems, especially in cyber-physical attacks, the study emphasizes the critical role of data reliability during such threats. Data imputation techniques emerge as a promising solution for challenges like manipulated and missing data, ensuring a more complete dataset for resilience-based decision-making. The study investigates how different degrees of data reliability influence the decision-making process and the evaluation of WDS resilience, specifically focusing on assessing existing data imputation models. The thesis outlines a comprehensive approach, utilizing laboratory-scale experiments

and the C-town benchmark WDS model, to enhance understanding of the significance of data reliability in WDS resilience.

ACKNOWLEDGEMENTS

I sincerely appreciate Dr. Sangmin Shin, who served as my thesis committee chair and advisor, for generously sharing his time, engaging in insightful discussions, and providing invaluable supervision throughout my research, ultimately contributing to the success of this project. I am equally grateful to Dr. Ajay Kalra and Dr. Jia Liu, esteemed members of my thesis committee, for their knowledge, unwavering support, and valuable advice, which significantly contributed to the advancement of my research.

I also want to thank the Department of Civil, Environmental, and Infrastructure Engineering for offering me a conducive research platform. I am grateful to Jennifer, the office manager, for her assistance in coordinating the necessary logistics. I would also like to express my gratitude to my coworkers, including Mr. Utsav Parajuli, Mr. Amrit Bhusal, Mr. Anjan Parajuli, Mr. Abhiru Aryal, Ms. Albira Aacharya, Mr. Mandip Banjara, Mr. Bishal Poudel, Mr. Binod Ale, and Ms. Kriti Acharya, for their support and collaboration.

I sincerely thank my friends Babin Dangal, Bipin KC, and Bibek Karki for their unwavering support throughout my graduate studies. My deepest gratitude goes to my father, mother, and family for their unconditional love and support during my educational and personal journey. Additionally, I want to thank all my colleagues, professors, and staff at SIUC's School of Civil, Environmental, and Infrastructure Engineering for their collaborative efforts and support while pursuing a graduate degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1 Research Background

The increasing demand for resilience in WDS is evident due to emerging challenges, including aging infrastructure, population growth, leakage, natural disasters (such as flooding and droughts (Joshi et al., 2020(a); Parajuli et al., 2017; Pokhrel et al., 2020)), cyber-physical attacks, climate change, and concerns about energy security (Bhandari et al., 2018; Ghimire et al., 2023a; Shrestha et al., 2020a). These diverse threats pose a significant risk to the reliability and functionality of WDSs, necessitating proactive measures to ensure continued operation even in uncertain conditions. The importance of resilience strategies in WDS is evident, focusing on minimizing water supply losses and facilitating swift recovery to normal operating conditions (Babu Ghimire et al., 2023). The unpredictability of the challenges, ranging from climate-related incidents to potential cyber threats, underscores the necessity for adaptive and robust approaches. A commonly suggested resilience strategy involves decentralization to enhance the system's capacity to absorb shocks and disturbances (Shin et al., 2018). As an example of resilience strategies, introducing decentralization to vital elements in WDS infrastructure improves adaptability and responsiveness, minimizing interruptions and securing a robust water supply network. In line with this idea, Bhusal et al. (2023) recommended integrating decentralized approaches into decision-making for detention systems to enhance resilience against severe flooding in urban watersheds.

In this context, smart system approaches have received attention in the strategies decision making on the design and operation of resilient WDSs. These approaches encompass sensing, monitoring, and operational changes, collectively contributing to the real-time optimization of WDS design and operation (Javaid et al., 2021). By employing advanced technologies, smart

systems facilitate continuous data collection, enabling resilience-based decision-making processes vital for the system's adaptability and robust response to unforeseen challenges.

In particular, the smart system provides real-time data crucial for resilience-based decision-making on quickly upgrading, operating, and managing the WDSs during disruptions. The decision-making process in smart systems involves real-time data collection, analysis, formulation, continuous monitoring, and adaptive learning to enhance overall system resilience (Sarker, 2021). By integrating real-time data, smart systems enable adaptive decision-making processes that address immediate challenges and contribute to the system's overall resilience. In this context, securing reliable data for the decision making is important to properly bring the resilience effects from design and operational options that are suggested from the decision-making process. Data reliability includes the correctness and consistency of information, ensuring dependable and trustworthy data for robust analysis and decision-making. Reliability is greatly enhanced by using efficient imputation techniques, particularly when handling missing data. Thus, this study investigates the influence of data reliability on resilience-based decision making, emphasizing proper methods for recovering missing data and accurately analyzing resilience in smart water distribution systems with imputed data.

1.2 Research Motivation

While smart systems undoubtedly provide sensing data to enhance the efficiency and resilience of WDS, it is crucial to acknowledge that their implementation introduces new vulnerabilities, particularly cyber-physical attacks (Shin et al., 2020). Smart systems heavily rely on data, including sensing information, for decision-making processes. This reliance on data makes them susceptible to cyber-attacks, such as malicious manipulation or deletion of sensing data, which can compromise the reliability of the information used in critical decision-making within the WDS.

Ensuring data reliability during cyber-physical threats is critical for the adequate design and efficient operation of WDS. Numerous studies have tackled this challenge across different domains. Javed et al. (2023) not only identified anomalies but also proposed a robust framework to maintain integrity and reliability, explicitly safeguarding Smart Healthcare Cyber-Physical Systems from blackhole and greyhole attacks. Meanwhile, (Cao et al., 2022) investigated the effects of false data injection attacks on microgrid cooperative control. (Cao et al., 2022) also introduces a resilient control method for synchronous mitigation that focuses on local detection to ensure compatibility with reactive power targets. Additionally, Varshini and Latha (2023) explored the repercussions of attacks on WAC applications. Varshini conducted a comparative analysis of model-based and data-driven attack detection methods, employing evaluations and simulations to determine the most effective detection strategy. Collectively, these studies contribute to the ongoing efforts to enhance the resilience of smart systems against evolving threats.

A promising solution to challenges such as cyber-physical attacks, manipulated data, missing data, and sensor reading errors is data imputation. This method involves identifying and

removing problematic data while using statistical and computational techniques to fill in missing or compromised data. This ensures that decision-makers have a completer and more reliable dataset for resilience-based decision-making, even in the face of cyber threats. Although data imputation has been widely studied in various contexts, its impact on precisely calculating resilience in WDSs has yet to be addressed. Previous research has primarily focused on other areas, indicating a need for further investigation to tailor and optimize data imputation techniques. This study is crucial for obtaining accurate real-time data, calculating resilience, and providing system preparedness and recovery recommendations. Bridging this gap can lead to more robust solutions, enhancing data reliability in smart WDSs and contributing to the resilience and security of WDSs.

1.3 Research Overview

This thesis focuses on explaining the critical role of data reliability in assessing resilience within WDS. To illustrate this, a laboratory-scale WDS with sensors is developed. Various scenarios are created, encompassing the conditions involving scenarios like leakage, demand variation, and pump failure. The resilience of the system is subsequently evaluated under each condition, with a specific focus on understanding the system's sensitivity to variations in resilience values and their impact on decision-making processes.

Another part of the research involves the application of a more complicated WDS, where different percentages of missing (or manipulated) data are intentionally introduced with the assumption of the failure conditions from cyber-physical attacks. The study then explores the computation of missing data within WDS, employing various imputation algorithms and assessing their accuracy. In essence, this thesis comprehensively investigates the significance of data reliability concerning the resilience and security aspects of WDS.

1. *Can the degree of data reliability change the decision of resilience-based options in WDS?*

The research used the construction of a lab-scale WDS connecting the sensors to explore resilience-based options within a decision-making framework systematically. The decision-making process, integral to this study, encompasses four key stages: 1) understanding problems, 2) identifying potential solutions, 3) evaluating and analyzing solution performance against management objectives, and 4) selecting the optimal solution aligned with the target objectives. The effectiveness of these activities relies heavily on data reliability, which is crucial for quantifying WDS problems and assessing the performance of resilience-based options. In this regard, I hypothesize that manipulated or falsified data from cyber-attacks or cyber system

malfunctions in decision making will suggest suboptimal choices for enhancing system resilience. This hypothesis underscores the significance of data reliability in ensuring the success of the decision-making process. Consequently, this study investigates and addresses the impact of data manipulation on the decision-making process concerning resilience-based options in WDS.

2. *How significantly will the degree of data reliability impact the evaluation of WDS resilience for decision making?*

There are various data imputation techniques using, e.g., statistics, physics-based simulation models, or data-driven models. They have provided reliable performance to fill or interpolate the missing or manipulated data in water system (Mamat et al., 2023; Zanfei et al., 2022) . As the degree of data missing or manipulation is significant, the data imputation performance of the existing techniques will decrease (Jadhav et al., 2019). However, the data imputation models can also provide acceptable performance despite large degree of data missing or manipulation, depending on the models' approach to fill or interpolate the data.

Thus, I hypothesize that the impacts of low data reliability will depend on the performance of data imputation. This study seeks to test existing data imputation models in addressing missing or manipulated data in WDSs and investigate the usefulness of data imputation models in evaluating WDS resilience for decision making under cyber-attacks.

1.4 Thesis Outline

This thesis aims to enhance resilience-based decision-making within WDS. It involves an examination of resilience under diverse conditions, including demand variation and leakage, and assessing how imputed values impact resilience. The focus is on understanding the crucial role of data reliability in influencing resilience sensitivity and decision-making. The two research hypotheses are tested through two tasks:

In Task 1, the resilience of a lab-scale WDS is evaluated across different system options, considering factors like demand variation and leakage. Resilience is calculated with pressure deviations at demand nodes, thereby exploring the influence of various conditions on resilience. Task 2 involves identifying the optimal imputation method for calculating resilience in imputed datasets.

Chapter 2 presents detailed insights into creating the lab-scale WDS, considering various system options, introducing leakage and demand variation scenarios, measuring resilience under normal conditions, and calculating resilience by deviating pressure at demand nodes by +-10%. This chapter addresses the first research question.

Chapter 3 outlines the generation of missing data in the C-Town WDS, ranging from 10% to 50% in tank pressure, utilizing different imputation methods. It evaluates imputation accuracy and calculates resilience for imputed datasets, addressing the second research question.

Finally, Chapter 4 summarizes the key findings of the research and proposes future studies to overcome challenges in enhancing resilience-based decision-making for WDS. The systematic organization of models, methods, and results across these chapters contributes valuable insights to the field, fulfilling the research objectives.

CHAPTER 2

INVESTIGATING THE IMPACT OF RELIABLE DATA ON RESILIENCE-BASED DECISION MAKING IN THE WDS

2.1 Introduction

Ensuring access to safe water, a fundamental human need outlined in Sustainable Development Goal 6, is globally significant. WDSs are pivotal in delivering safe water. Still, their vulnerability to a spectrum of threats, ranging from natural disasters like earthquakes and floods to cyber-attacks, poses a significant challenge to their robust functionality (Joshi et al., 2020b; Kalra et al., 2022; Sagarika et al., 2015). Recent incidents, such as water pipe damage resulting from the 2023 earthquake in Turkey and Syria, underscore the fragility of WDSs. The escalating frequency of extreme weather events, driven by the climate crisis, intensifies these challenges (Aryal et al., 2022; Thakali et al., 2016). Despite the critical role of resilience in designing and operating WDSs, achieving a consensus on its definition and measurement remains a significant challenge.

Resilience-based strategies, focusing on minimizing system losses and rapidly recovering to normal conditions, are becoming increasingly recognized as a means of improving the sustainability of many systems, including water systems, under uncertain circumstances (Pickett et al., 2014). Engineering resilience is often defined in two independent but interconnected ways: attribute-based and performance-based. The first relates to the system and includes design concepts like duplication and connectivity that allow for a successful reaction to any anomalies. On the other hand, performance-based resilience addresses the system's agreed-upon performance in dealing with certain risks; it frequently follows operational standards and is prescriptive (Butler et al., 2014). A thorough grasp of the resilience of the system is necessary for

the continuous study of how a system's characteristics contribute to satisfying performance criteria before, during, and after disruptions.

Incorporating smart systems, such as intelligent water and electricity distribution networks, has emerged as a pivotal strategy to bolster resilience in critical infrastructure. These advanced systems leverage cutting-edge technologies and real-time data analysis to enhance efficiency, reliability, and responsiveness. Smart WDSs, for instance, utilize sensors and monitoring devices to continually assess water quality, detect leaks, and promptly address potential issues, thereby reducing water loss and ensuring a dependable water supply. Similarly, smart electricity grids optimize energy distribution, reroute power during outages, and augment grid stability, thus fortifying infrastructure against disruptions. The integration of such smart systems not only reinforces the robustness of critical infrastructure but also amplifies its adaptability and recovery capabilities, making a substantial contribution to overall resilience in the face of diverse challenges.

In the context of system resilience analysis, the reliability and sensitivity of system data emerge as pivotal factors. Khetwal et al. (2022) emphasized the significance of assessing tunnel resilience, underscoring the sensitivity of parameters such as traffic volume, fire suppression systems, maintenance, and operational variables through simulation modeling. Simic et al. (2023) extended this focus to water resource systems, advocating for dynamic analysis via multi-scenario simulations to enhance understanding and forecast resilience under varying scenarios. Furthermore, Jonnalagadda et al. (2023) highlighted the critical role of data reliability through sensitivity analysis, employing data sets of varying reliability levels to inform system resilience updates. To validate their model, a benchmark problem involving a South Carolina, USA highway network is employed, demonstrating a systematic approach to quantify and reduce

uncertainties. The benchmark results underscore that incorporating monitoring and inspection data for key variables can significantly enhance the accuracy of seismic resilience predictions for the network. These studies collectively reinforce the importance of data quality and sensitivity in the assessment and enhancement of system resilience.

The reliability and accuracy of data are foundational elements in the effective analysis of systems and the evaluation of resilience. Without dependable data, it is challenging to comprehensively understand system behaviors and vulnerabilities, hindering the development of strategies for resilience enhancement. Accurate data serves as the foundation for modeling and simulation, allowing for the identification of potential stressors and the evaluation of system responses. Furthermore, reliable data is instrumental in measuring the effectiveness of resilience strategies and informing real-time decision-making during crises. In essence, the precision of data is paramount in ensuring that resilience-enhancing measures are targeted, evidence-based, and capable of effectively mitigating potential threats.

One of the promising approaches to achieving higher resilience is a decentralized system. By decentralizing critical components in the WDS, the system becomes more flexible and responsive, thereby lessening the impact of disruptions and ensuring a resilient water supply network (Kalbar and Lokhande, 2023; Shin et al., 2018). Drawing on existing research, implementing decentralized detention systems to manage extreme flooding events regionally highlights the crucial role of comprehensive resilience strategies (Ngo et al., 2018). The use of decentralized microgrid energy management in the power sector to enhance reliability and performance (Alstone et al., 2015), and the adoption of decentralized systems in healthcare for improved equity, efficacy, and resilience (Abimbola et al., 2019) further emphasizes

decentralization as a notable example of fortifying WDS against the diverse uncertainties they may face.

In WDSs, incorporating diverse water resources like rainwater harvesting and desalination exemplifies a decentralized water supply approach. This decentralized system, integrated with existing centralized WDSs to form hybrid systems, is vital in minimizing disruptions during failures, including those resulting from cyber-physical attacks or water scarcity due to drought (Chhetri and Tamang, 2019; Ghimire et al., 2023b; Liang et al., 2023; Shrestha et al., 2020b). By strategically balancing contributions from centralized and decentralized components based on prevailing failure conditions, these hybrid systems enhance water availability, reduce leakage, cut costs associated with long-distance water delivery, and alleviate the strain on centralized water sources and treatment systems. Numerous studies highlight the advantages of alternative water systems within centralized frameworks, with integrated hybrid systems significantly reducing potable water consumption and wastewater flow. Despite these advancements in existing literature, quantitative assessments of the contribution of hybrid water supply systems to resilience still need to be improved.

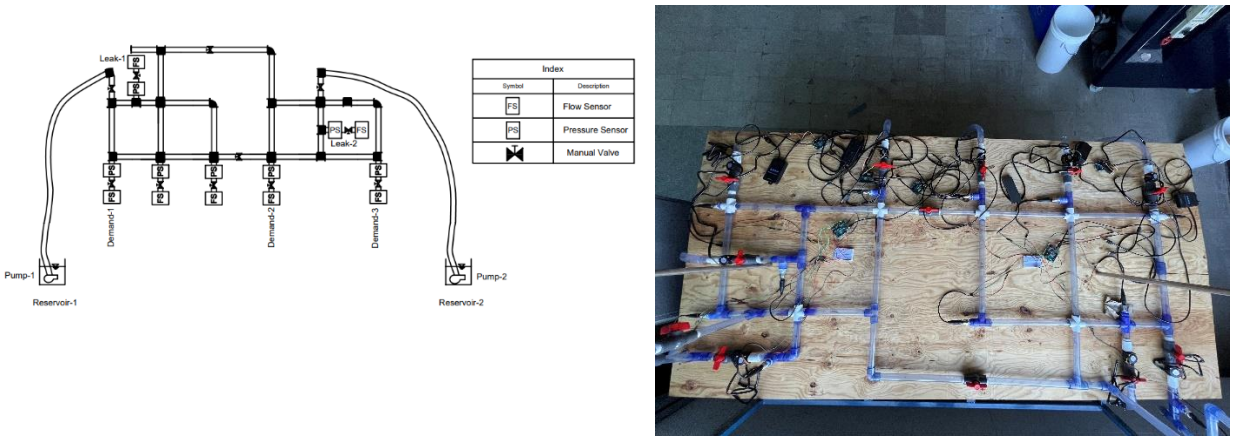
This study investigated the fundamental role of accurate and reliable data in enabling a comprehensive analysis of systems and evaluating the resilience of different water supply options to various disruptive events. A lab-scale WDS was used to showcase centralized and decentralized water supply strategies under various disruptive events. The contributions of this study include defining the resilience effect of a WDS at different decentralization levels and providing insights into how variations influence resilience-based decision-making in demand node pressure.

2.2 Methods

For a resilience-based strategy, a WDS combining centralized and decentralized water systems were designed, which can alter from a centralized WDS to decentralized WDS or vice versa – a hybrid WDS. Then, a physical lab-scale model of the WDS was built and tested by measuring water flow and pressure through sensors.

2.2.1 Prototype of Lab-Scale physical Hybrid WDS

Figure 1 depicts a schematic of a hybrid WDS that combines centralized and decentralized elements. The system includes dual water sources supplied from PONDFORSE 24V DC Ultra Quiet Submersible Water Pumps, PVC pipes, flow control valves, and three demand nodes. The configuration details are shown in Figure 1, and operational conditions determine flow rates and pressures. The model integrates nine flow sensors and ten pressure sensors strategically placed for measurement. Arduino UNO connects sensors to obtain data, and each scenario undergoes a stable phase before recording pressure and flow values every second for two minutes. The water flow rates and pressures were measured using flow and pressure sensors, and an Arduino UNO microcontroller board automatically saved the data in a txt file. The recorded information from Arduino Uno was later transformed into a CSV format using Python. The average values are then used to analyze system performance, specifically assessing resilience for various hybrid WDS configurations and failure scenarios.



(a) (b)

Figure 1: A hybrid WDS (a) Schematic diagram (b) Lab Scale prototype (Adapted (Babu Ghimire et al., 2023))

2.2.2 Different operational choices of the Hybrid WDS Model

This study identified six operational setups for the hybrid WDS model, varying decentralization levels and pump operations (see Table 1). In the first configuration (Option 1), characterized as a centralized system, the water supply relied solely on Pump 1. The degree of decentralization increased with additional water supply from the second source facilitated by Pump 2. Detailed operational options are outlined in Table 1, with a consistent water flow rate of approximately 585 liters per hour from one or both sources.

Table 1: Operational options depending on decentralization levels of the hybrid WDS.

Operational condition	Options (Portions of each water source in water supply, %)		Pump 1 (l/h)	Pump 2 (l/h)	Total Supply (l/h)
	Water source 1	Water source 2			
Complete centralization	100	0	585.00	0.00	585.00
Partial decentralization	90	10	526.50	58.50	585.00
	80	20	468.00	117.00	585.00
	70	30	409.50	175.50	585.00
	60	40	351.00	234.00	585.00
Complete decentralization	50	50	292.50	292.50	585.00

2.2.3 Scenarios mentioning disruptive events

This study classified three disruptive events in a WDS: a base scenario representing normal operations, scenarios with demand variation, and those simulating physical failures like leakage. These scenarios aimed to depict various system failures or disturbances, as detailed in Table 2. Across all scenarios, water flow rates from sources were constant at around 585 l/h, and the flow supplied to each of the three demand nodes remained equal at approximately 195 l/h. Demand variation scenarios, like Scenario D1, examined the impact of significant changes in water consumption at demand nodes on WDS performance. For example, in Scenario P1, a 5%

leakage represented water leakage from a pipe at 29.25 l/h, assessing how leakage conditions influence system resilience. Operational failure scenarios explored unintended or intentional deactivation of WDS actuators, focusing on pump 1. At the same time, in decentralized configurations, only pump-2 was operational, delivering water from the second source to all demand nodes.

Table 2: Disruptive event scenarios considered in this study.

Scenario		Description
Normal operation	Base scenario	Normal operation conditions with no disruptions
Demand variation	Scenario D1	A decrease in demand at Node 2 (Demand-2) by 20%
	Scenario D2	A decrease in demand at Node 2 by 20% and an increase in demand at Node 3 (Demand-3) by 15%
Physical failure	Scenario P1	5% water leak at Leak-1 pipe (shown in Figure 1)
	Scenario P2	5% water leak at Leak-2 pipe (shown in Figure 1)
	Scenario P3	5% water leak at Leak-1 and Leak-2 pipes

2.2.4 Pressure Variation Scenario

Following a comprehensive analysis of various disruptive events and collecting pressure and flow data at demand nodes, the study incorporates additional scenarios focusing on pressure variation of ±10% across different demand nodes. This consideration aims to assess the sensitivity of resilience in selecting decision options. The scenarios, denoted as D1, D2, D3, and D1, D2, D3, involve fluctuating pressure levels by ±10% under different decentralization configurations. These scenarios are designed to explore how pressure variations impact the system's resilience at various decentralization levels, providing valuable insights for decision-making processes.

2.2.5 Evaluation of Resilience

This study utilized Todini's resilience measure (Todini, 2000) to calculate resilience in different disruptive event scenarios for hybrid WDS operational options, as represented in equation (1). Commonly employed for quantifying WDS resilience (Shin et al., 2018), the measure is defined as the fraction of available energy surplus at nodes, internally dissipated to meet demand and head requirements, over the maximum energy surplus in the network.

$$R = \frac{\sum_{i=1}^{n} q_i^*(h_i - h_i^*)}{\sum_{j=1}^{r} Q_j H_j + \sum_{k=1}^{p}\left({}^{P_k}/_{\gamma}\right) - \sum_{i=1}^{n} q_i^* h_i^*} \quad (1)$$

where q_i^* and h_i^* are the design demand and head required at node i; h_i is the available head at node i; Q_j is the flow from j^{th} reservoir; H_j is the total head in j^{th} reservoir; P_k is the energy supplied to the network from k^{th} pump; γ is the specific weight of water, and n, r, and p are the number of nodes, reservoirs, and pumps, respectively, in the WDS. Considering the small size and dimensions of our physical model and its preliminary demonstration, the design flow (q_i^*) and head (h_i^*) values for each demand node were assumed to be 150 L/H and 0.42 m, respectively. The method of creating various scenarios during operational disruption conditions are explained in the following sections.

2.3 Result and Discussion

2.3.1 For Disruptive Event Scenario

The study conducted a thorough assessment of the hybrid WDS resilience, comparing it across operational choices and disruptive scenarios. Resilience levels, illustrated in Figure 3, indicate a significant increase with higher decentralization in the hybrid WDS. In normal operations (Figure 2a), the fully decentralized option showed 74% higher resilience than the centralized, demonstrating superior response to disruptions.

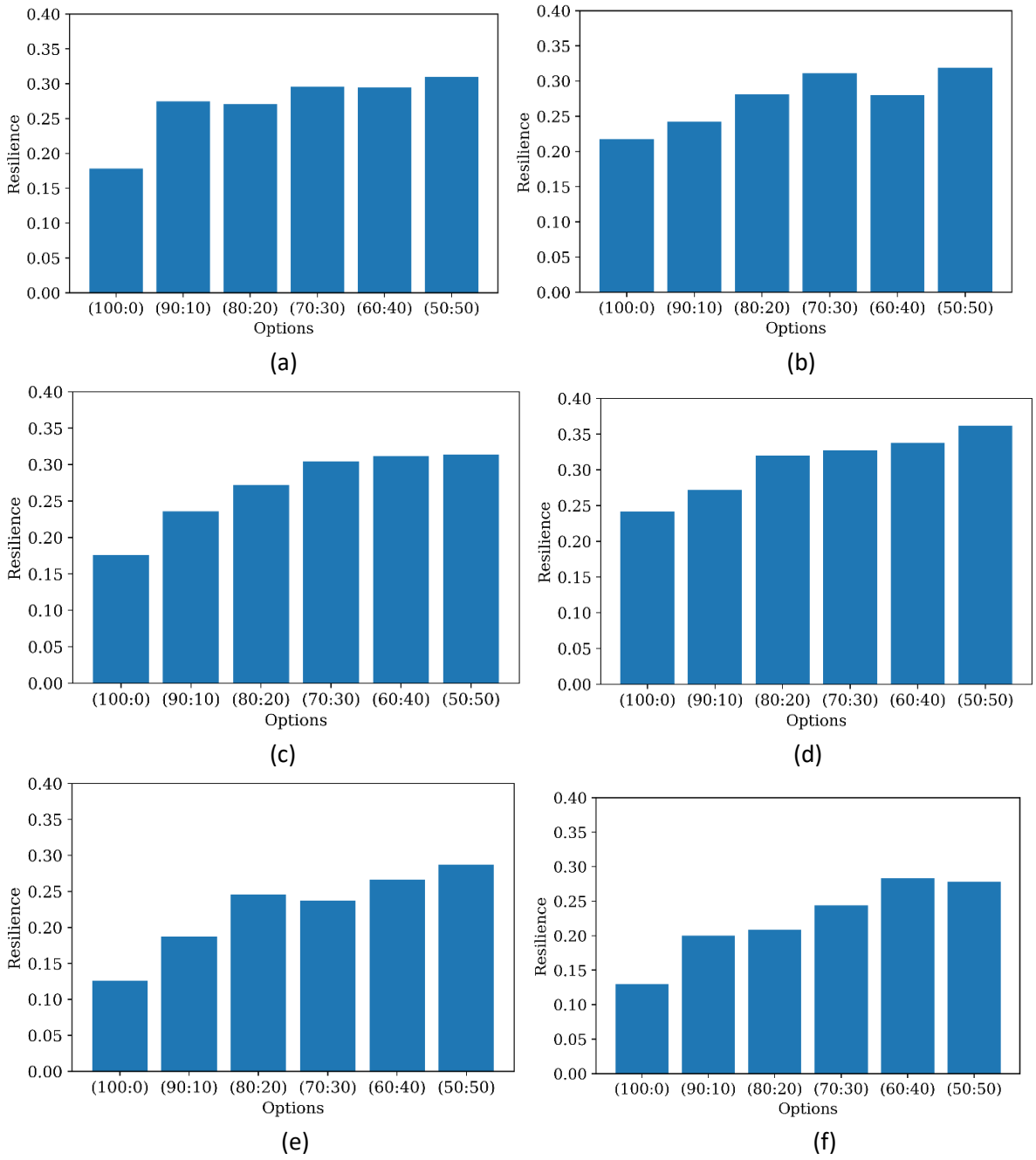


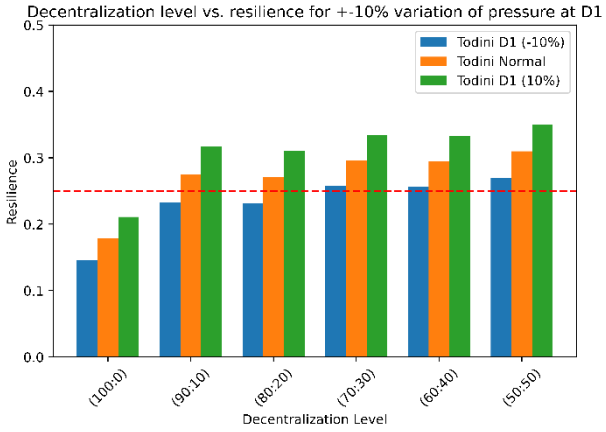
Figure 2: The variation of resilience depending on system options and disruption scenarios: (a) Base scenario; (b) Scenario D1; (c) Scenario D2; (d) Scenario P1; (e) Scenario P2; and (f) Scenario P3.

Figures 3b and 3c revealed decentralized WDS outperforming centralized by 47% and 49% in scenarios D1 and D2. Similarly, Figure 2d, 2e, and 2f highlighted that leakage reduced system

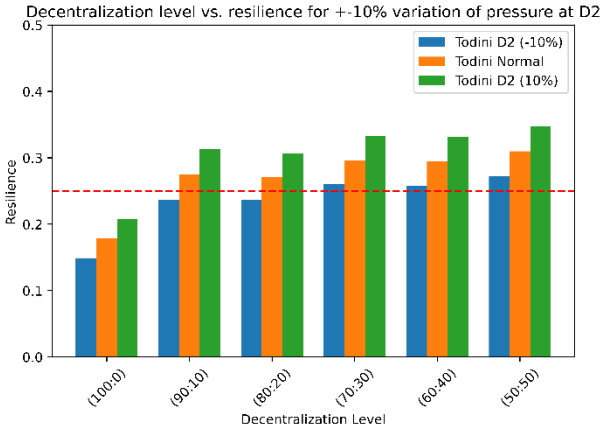
resilience, but this impact lessened with increased decentralization. Leakage effects varied based on proximity to water sources, emphasizing decentralization's role.

2.3.2 For Pressure deviation at Demand Nodes

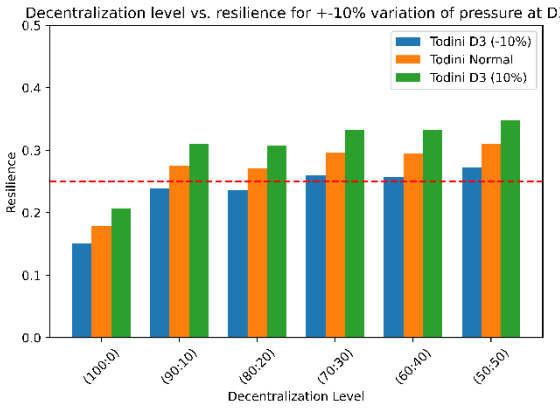
In the second part of the result the variation of resilience value due to change in pressure at different scenarios is considered. Three scenarios were explored in the investigation of water distribution system resilience under varying conditions, each shedding light on the interplay between decentralization, system resilience, and the impact of manipulated pressure values. The common thread across all scenarios was the identification of a resilience threshold set at 0.25. This threshold served as a critical indicator, signifying an optimal level of decentralization for effective system performance. The study examined pressure manipulations in four conditions: at demand node d1, d2, d3 individually, and collectively at demand nodes d1, d2, and d3.



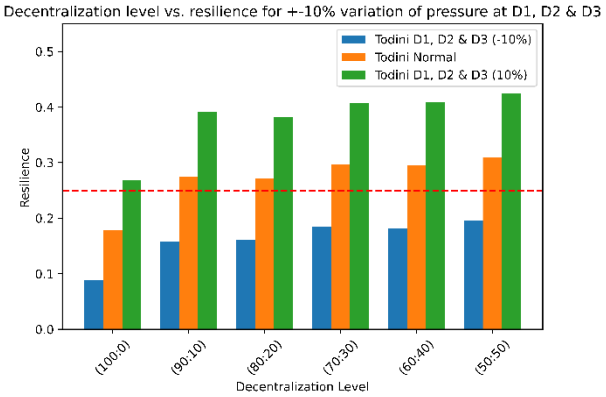
(a)



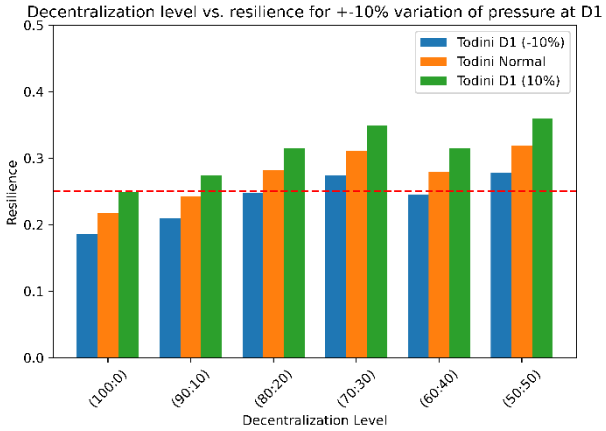
(b)


Decentralization level vs. resilience for +-10% variation of pressure at D3
Todini D3 (-10%)
Todini Normal
Todini D3 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

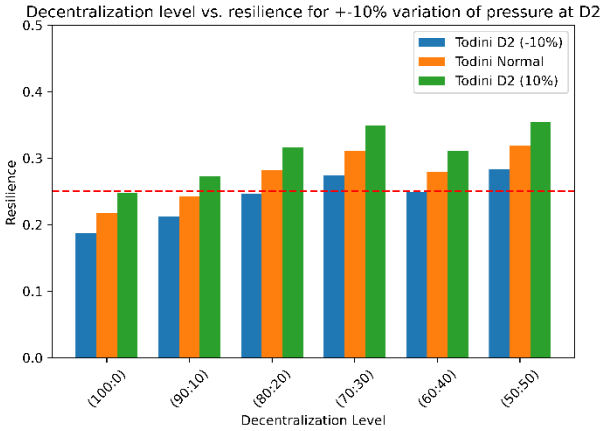
(c)


Decentralization level vs. resilience for +-10% variation of pressure at D1, D2 & D3
Todini D1, D2 & D3 (-10%)
Todini Normal
Todini D1, D2 & D3 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

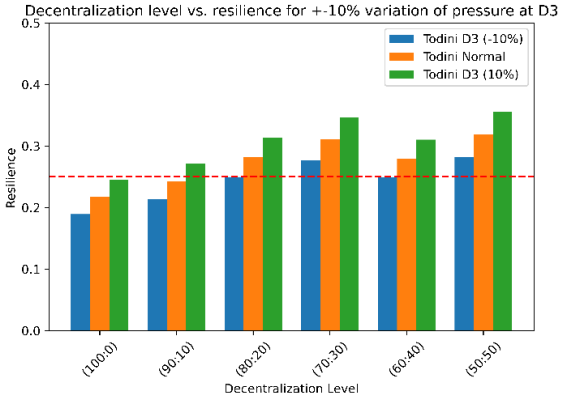
(d)


Decentralization level vs. resilience for +-10% variation of pressure at D1
Todini D1 (-10%)
Todini Normal
Todini D1 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

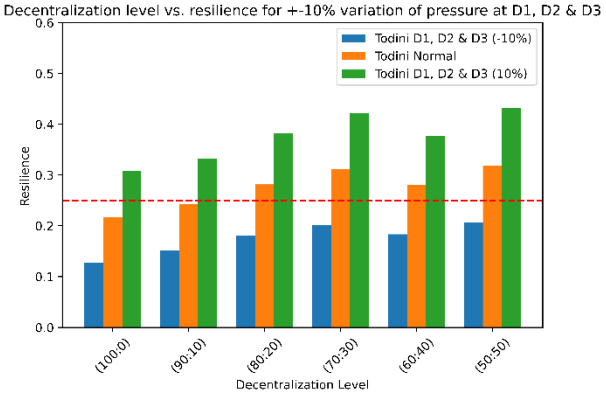
(e)


Decentralization level vs. resilience for +-10% variation of pressure at D2
Todini D2 (-10%)
Todini Normal
Todini D2 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

(f)


Decentralization level vs. resilience for +-10% variation of pressure at D3
Todini D3 (-10%)
Todini Normal
Todini D3 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

(g)


Decentralization level vs. resilience for +-10% variation of pressure at D1, D2 & D3
Todini D1, D2 & D3 (-10%)
Todini Normal
Todini D1, D2 & D3 (10%)
Resilience
0.0
0.1
0.2
0.3
0.4
0.5
0.6
(100:0)
(90:10)
(80:20)
(70:30)
(60:40)
(50:50)
Decentralization Level

(h)

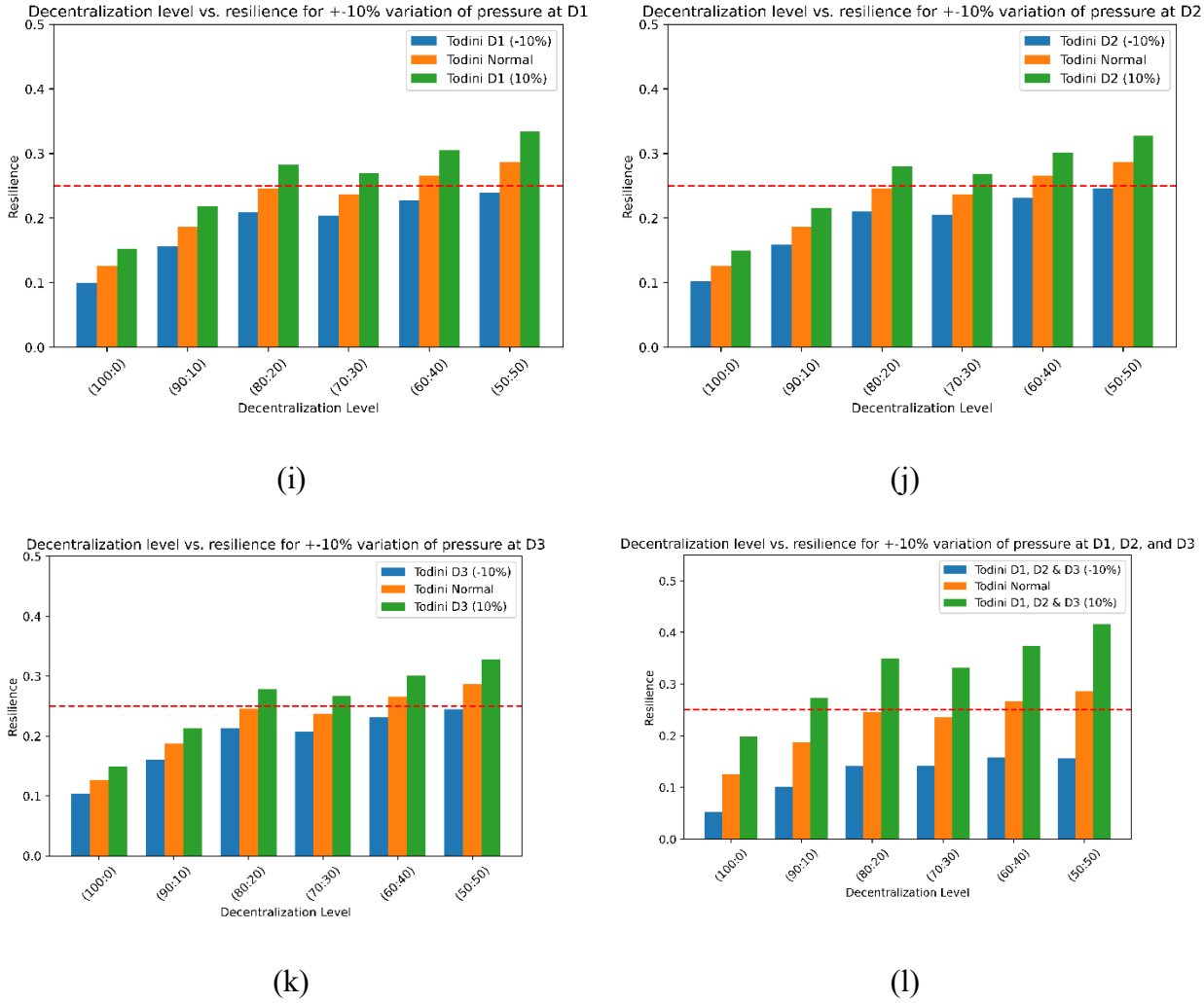


Figure 3: Decentralization Level vs Resilience altering pressure by ±10% for normal, demand variation and leakage condition at (a), (e), (i) d1, (b), (f), (j) d2, (c), (g), (k) d3, and (d), (h), (l) d1, d2, and d3

However, the paper recognizes the inherent trade-off between decentralization and construction costs. As decentralization levels increase, so does the cost of construction. Therefore, the findings advocate for a balanced approach, emphasizing the significance of achieving the threshold resilience value at the lowest possible decentralization level.

1. **Normal Condition Scenario:**
    - Optimal system option for resilience (0.25) and cost-effectiveness is 90:10.

- Manipulated pressures at demand nodes influenced deviations from the optimal 90:10 option.
- Increased pressures by 10% maintained 90:10 for the first three conditions and shifted to 100:0 for the fourth condition. Decreased pressures by 10% resulted in the optimal 70:30 for the first three conditions.

2. **Demand Node D2 Reduced by 20% Scenario:**
    - Optimal system option for resilience (0.25) is 80:20.
    - Alternatives (70:30, 60:40, 50:50) met the resilience threshold but proved cost-intensive compared to 80:20.
    - Increased pressures by 10% maintained 90:10 for the first three conditions and shifted to 100:0 for the fourth condition. Decreased pressures by 10% resulted in the optimal 70:30 for the first three conditions.
3. **5% Leak at Leak1 Scenario:**
    - Optimal system option for resilience (0.25) is 60:40.
    - Another option, 50:50, met the resilience threshold but was more expensive than 60:40.
    - Increased pressures by 10% resulted in 80:20 for the first three conditions and 90:10 for the fourth condition. Under pressure decrease by 10% at demand nodes, no system option fulfilled the resilience threshold criteria.

In all scenarios, deviations from optimal options were influenced by manipulated pressure values, underscoring the importance of accurate data interpretation for resilient based decision-making for effective performance and cost-effective selection of system options.

2.4 Conclusion

In conclusion, this research focused on enhancing the resilience of a WDS to disruptive events. The study advocated for increased decentralization as a strategy to improve WDS resilience. A hybrid centralized and decentralized WDS was explored, demonstrating increased resilience with higher decentralization levels under various disruptive event scenarios. The lab-scale simulation provided valuable insights into the hydraulic performance of the hybrid WDS, addressing normal operating conditions, demand variations, physical failures, and operational failures.

Furthermore, in all three scenarios with manipulated pressure, the study reveals that the initially identified optimal system option, chosen among different decentralization levels, can be changed due to manipulated pressure values at various demand nodes. These alterations in resilience values may lead the system manager to make incorrect choices in selecting the system option. Hence, the importance of reliable data is emphasized to ensure the accurate and informed selection of the best system option.

The study emphasized the need for further exploration to advance the practical application of a hybrid centralized and decentralized WDS in terms of resilience-based decision-making, including long-term performance assessments, strategic interactions between water sources, and cost-effective diversification and decentralization strategies. These insights provide valuable guidance for designing and operating resilient WDSs, urging stakeholders to consider diversification and decentralization in current water distribution systems.

CHAPTER 3

EFFECTS OF DIFFERENT PERCENTAGES OF IMPUTATED DATA ON WDS RESILIENCE

3.1 Introduction

Missing data or manipulated, which refers to the absence of correct data values for a given variable in an observation, is a pervasive issue in various research domains and is recognized as a common challenge encountered in the analysis of real-world datasets (Hernández-Pereira et al., 2015). In scientific research, missing data can present serious problems, mainly when there is no obvious pattern or cause for the missing data (Garciarena and Santana, 2017). This can lead to smaller sample sizes, potential biases (Beaulieu-Jones and Moore, 2016), and weakened results validity (Sterne et al., 2009). The common issue of missing data can hinder data analysis, study, and visualization, negatively impacting real-world studies (Ssali and Marwala, 2008) . Missing data can occur due to a range of reasons, such as data collection problems, equipment faults, incomplete manual data entry, and non-participation (Razavi-Far et al., 2020) or attendance in data gathering. Despite researchers' careful control over data measurement and recording in experimental and survey data, missing data can still occur due to uncontrollable factors (Kyureghian et al., 2011). Acknowledging this issue can help researchers better address missing data during data analysis and interpretation.

Data imputation has emerged as a successful strategy across diverse research domains, including social sciences, medical research, engineering, environmental research, business, and finance, among others. These techniques range from traditional methods (such as deletion and single imputation) to more modern and advanced methods such as multiple imputation and machine learning techniques. The application of data imputation has facilitated the interpretation,

analysis, and advancement of research outcomes in these fields. For instance, Durrant (2005) has demonstrated the practical selection of various data imputation methods in social science research, while Jerez et al. (2010) has exhibited the superior performance of machine learning-based imputation techniques for predicting patient outcomes compared to statistical imputation methods. In the field of marine systems, Cheliotis et al. (2019) has introduced a novel data condition and performance imputation technique for enhancing energy efficiency. Additionally, Quinteros et al. (2019) has illustrated the efficacy of imputation techniques in reconstructing actual air quality datasets. Moreover, Cheng et al. (2019) have proposed an imputation algorithm that displays high precision and stability in predicting financial distress across varying degrees of missing data and noise.

Based on the review of the available scientific literature, it is apparent that the application of data imputation techniques in the field of water research has primarily focused on hydrological variables such as stream flow data (Oriani et al., 2016), water quality data (Nieh et al., 2014; Rodríguez et al., 2021), ground water data (Evans et al., 2020; Sarma & Singh, 2022), as well as demand forecasting (Bata et al., 2020; Zanfei et al., 2022). It is important to note that, as far as the author is aware, there is a need for more studies addressing the imputation of missing data in WDS and analyzing the system's resilience based on the imputed value. The emphasis is on determining the best-imputed method among the selected ones and calculating the resilience using an imputed dataset that almost resembles normal condition resilience.

WDS’s security and integrity face significant challenges due to missing data in various regions, hindering accurate analysis of system performance. The absence of crucial data makes it difficult to assess the actual operational status of the WDS, impacting decision-making processes. The inability to obtain accurate information about the system's performance may lead

to misallocations of resources and potential disruptions in managing the water supply. Addressing and mitigating the impact of missing data in different regions are essential to ensure WDSs' reliability and security.

This study selected the C-town network as the research network to examine the effects of cyber-physical attacks on WDS. Missing values are created on pressure readings of tank values located at different network regions. A simulation period of one day (24 hours) was employed for the analysis. Various multiple imputation methods, including classification and regression techniques, predictive mean matching, linear regression ignoring model error, and linear regression with predicted values, were implemented using RStudio for data imputation.

The main objective of this paper is to evaluate the impact of missing data imputation on the resilience of the WDS, considering different imputation percentages. After removing the missing data and applying the imputation methods, the study aims to assess the extent to which the imputed datasets resemble the original water distribution scenario. The assumption posits that once the percentage of imputation is increased from 10% to 50%, a higher percentage of imputed missing values (ranging from 10% to 50%) will yield a higher deviation of resilience from the normal condition. By investigating the efficacy of different imputation techniques, this research aims to contribute to the understanding of data recovery and restoration in the context of missing data on WDS and explain the concept of the effect of resilience more clearly.

3.2 Methodology

The methodology portion can be divided into four sections. 1) Creating normal condition datasets, 2) Creating varying percentages of missing data in tank pressure data, 3) Imputation of missing values using various imputation approaches, 4) Checking dataset accuracy of methods of imputation, and 5) Checking resilience values for different percentage of imputation.

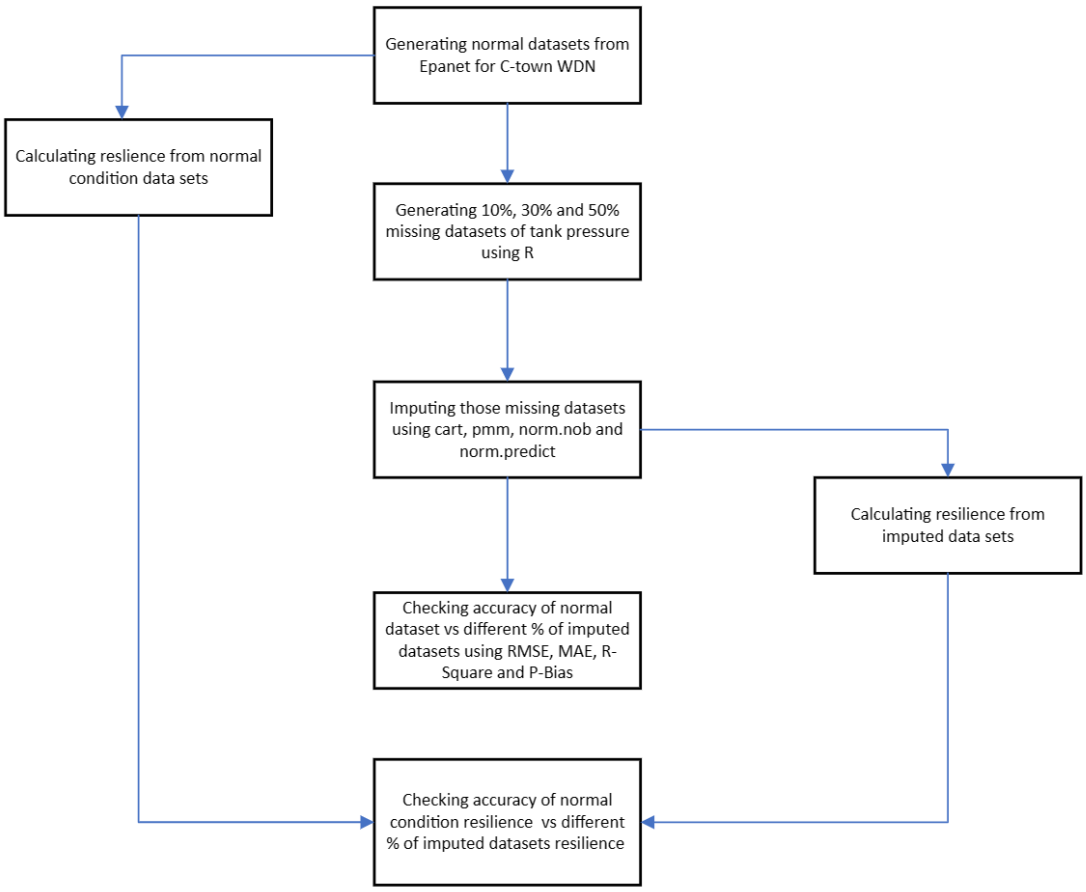


Figure 4: Detail Methodology of Data Imputation Process

3.2.1 Creating Normal condition datasets.

This study used the C-town network in EPANET software to produce a thorough 24-hour dataset for WDS. The analysis included introducing varied percentages of missing data into the pressure data of seven tanks simulating different scenarios. The datasets representing normal operation were derived from EPANET, serving as a baseline for systematic evaluation. This approach facilitated a detailed examination of how missing data influences the performance and dynamics of WDSs in diverse conditions.

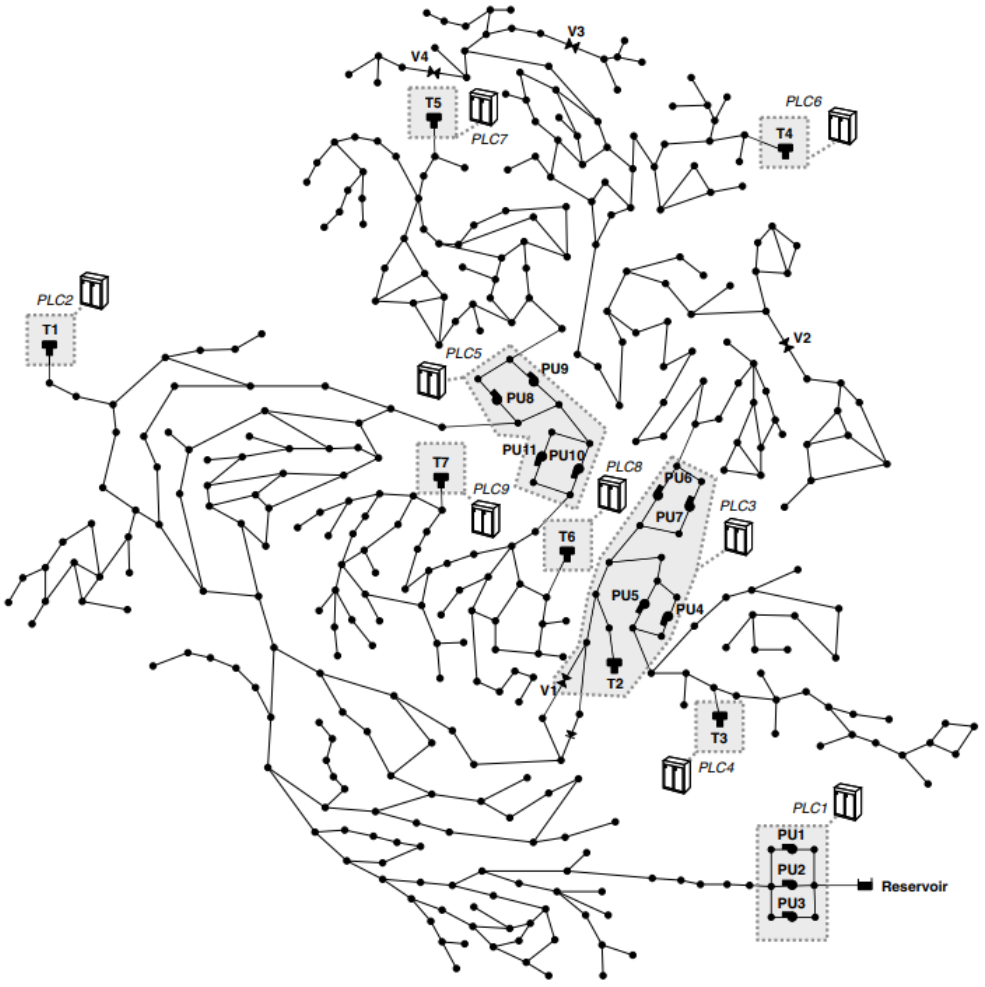


Figure 5: Graphical representation of C -Town WDS (Adapted (Taormina et al., 2017))

3.2.2 Creating varying percentages of missing data in cyber-attack data.

In this part, RStudio was employed to introduce missing values into the 24-hour dataset. Different percentages of missing values, randomly ranging from 10% to 50%, were inserted into the pressure values of seven tanks.

3.2.3 Imputation of missing values using various imputation approaches

Data analysis requires careful consideration of missing information, as there is a distinction between empty and missing values. Empty values cannot be assigned, while missing

values exist but may not be available in the dataset. It is crucial for data miners to distinguish between the two types of values to avoid misinterpretation. To address missing data, it's crucial to understand why the data is missing. (Little & Rubin, 2019; RUBIN, 1976) identified three missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (NMAR).

MCAR is the highest level of randomness, where missing values occur entirely at random and independent of any variables considered. MAR involves missing data probabilities that depend on observed information in the dataset. On the other hand, MNAR occurs when the probability of missing data is dependent on unobserved values of the variable due to the sensitivity of the response variable.

To achieve the goals of the research, the author experimented with different multiple imputation, to address the challenge of missing data, which lacks a perfect solution (Wolpert and Macready, 1997). To address this, the author employed classification and regression trees (cart), predictive mean matching (pmm), linear regression ignoring model error (norm.nob), and linear regression with predicted values (norm.predict) assuming MCAR and evaluated them using four metrics: Normalized RMSE, Normalized MAE, Normalized R-Square and Normalized PBIAS.

Multiple Imputation (MI) is accomplished by generating several complete datasets, each with a different set of imputed values, which are then analyzed separately using standard statistical methods. Finally, the results of each analyzed dataset are combined. According to Azur et al. (2011), using a MI technique allows for better measurement of statistical uncertainty than single imputation methods. The different methods of multiple imputation used in this study is explained below:

Classification and regression methods: CART multiple imputation use decision trees to impute missing values by dividing the data based on the values of other variables. It builds a decision tree for each variable with missing data and predicts the missing values using the known values of other variables. CART multiple imputation provides a robust and flexible approach for coping with missing data by repeatedly constructing several imputed datasets and integrating uncertainty, allowing researchers to achieve trustworthy estimates and valid conclusions in their investigations.

Predictive mean matching: Predictive mean matching imputation replaces missing data in a dataset by locating similar donor cases and using their observed values. The imputed values are chosen from the observed values of the donor cases with the closest predicted mean.

Linear regression ignoring model error: It involves using a predefined model to generate multiple estimates for the missing data. Multiple datasets are produced as a result of this process, each including imputed values for the missing data. Following that, each of these datasets is independently examined using standard statistical methods. A single set of estimates that takes into account the uncertainty brought on by the missing data is then created by combining the results from the various analyses.

Linear regression with predicted values: It entails generating multiple imputed values for the missing items depending on the observed data. The norm.predict method makes the assumption that the data follow a normal distribution. It incorporates the imputation process's inherent uncertainty by imputing the missing values by selecting random samples from a predicted normal distribution. With this method, the variability and uncertainty present in the missing data are preserved, allowing for more precise and trustworthy statistical analysis.

3.2.4 Checking dataset accuracy

The accuracy of both observed and imputed values for different imputation techniques was assessed using the mean normalized root mean square error (Mean-NRMSE), mean normalized mean absolute error (Mean-NMAE), mean normalized root mean square error (Mean-NR-Square), and mean normalized percent bias (Mean-NPBIAS) equations. By normalizing the disparities between the observed and imputed values, these metrics allowed a thorough evaluation of the overall accuracy. In order to evaluate imputation techniques, first the imputed value is normalized using mean max normalization. Mean max normalization is a method for scaling numeric data that involves subtracting the minimum value from each data point and then dividing the result by the difference between the data set's maximum and minimum values. The resulting values are then scaled to range from 0 to 1. Following are the equations used to calculate various metrices for each variable:

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{2}$$

$$NMAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i) \tag{3}$$

$$NR - Square = 1 - \frac{\sum_{i=1}^{n}((P_i - A_i)^2}{\sum_{i=1}^{n}(A_i - M)} \tag{4}$$

$$N - PBIAS = 100 * \frac{\sum_{i=1}^{n}(A_i - P_i)}{\sum_{i=1}^{n} A_i} \tag{5}$$

where n is the total number of observations, P_i is the normalized imputation value for ith missing value and A_i is the normalized true value for ith missing value, M is the normalized mean of true value.

After obtaining the NRMSE value for each variable in the dataset, the Mean NRMSE is calculated by adding all NRMSE values and dividing by the total number of variables, yielding an average NRMSE value for the dataset as shown below:

$$Mean\ NRMSE = \frac{\sum_{i=1}^{n} NRMSE}{m} \tag{6}$$

where m is the total number of variables in the dataset.

The following formula is used to calculate Mean NMAE, Mean NR-Square, and Mean N-PBIAS as of Mean NRMSE:

$$Mean\ NMAE = \frac{\sum_{i=1}^{n} NMAE}{m} \tag{7}$$

$$Mean\ NR-Square = \frac{\sum_{i=1}^{n} NR-Square}{m} \tag{8}$$

$$Mean\ N-PBIAS = \frac{\sum_{i=1}^{n} N-PBIAS}{m} \tag{9}$$

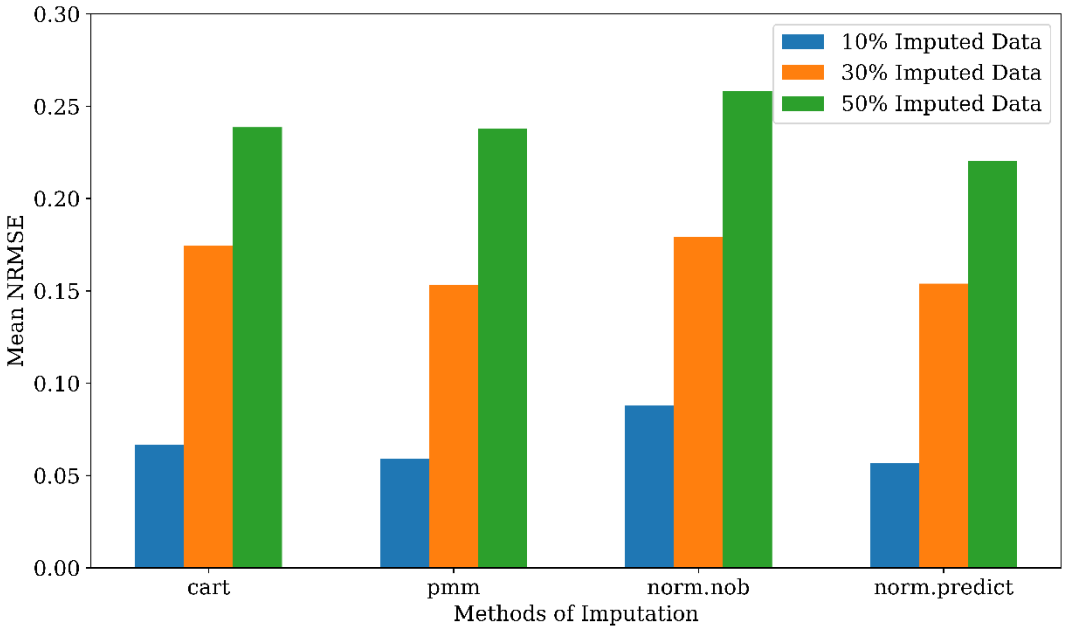
3.2.5 Resilience Calculation under various conditions

Following the computation of imputed values across a range of missing data percentages, resilience is subsequently assessed under diverse conditions using equation 1. The evaluation encompasses varying percentages of missing values and distinct imputed datasets generated through various imputation methods. This comprehensive analysis aims to illustrate the efficacy

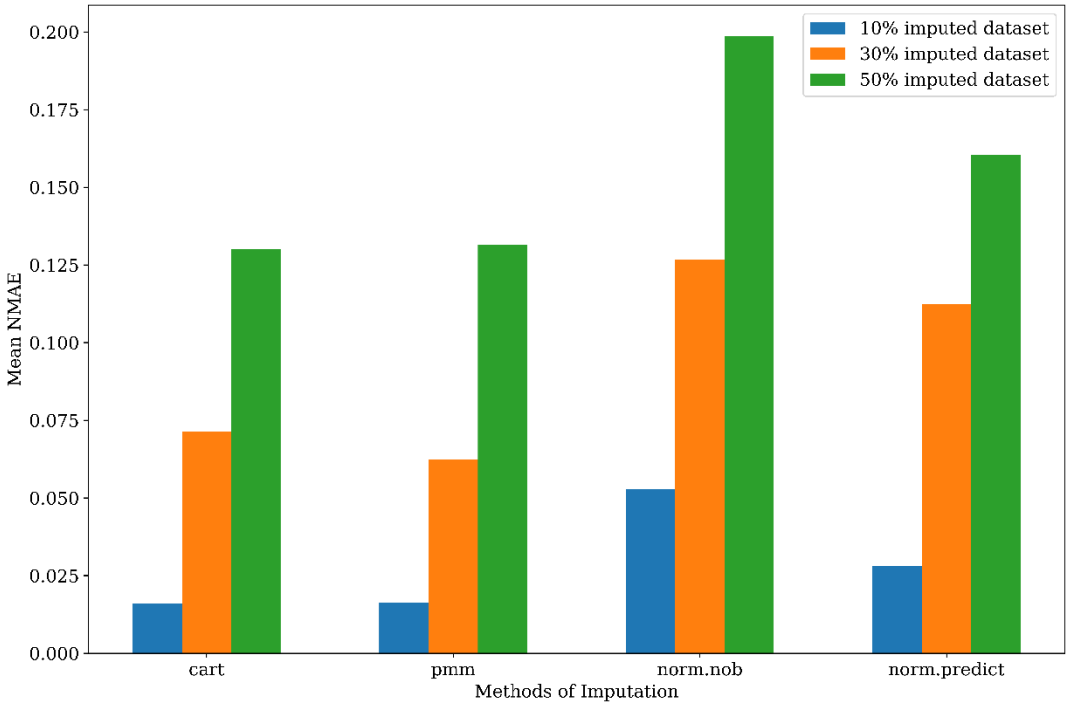
of imputation techniques in enhancing resilience calculations compared to scenarios where datasets remain unimputed.

3.3 Results and Discussion

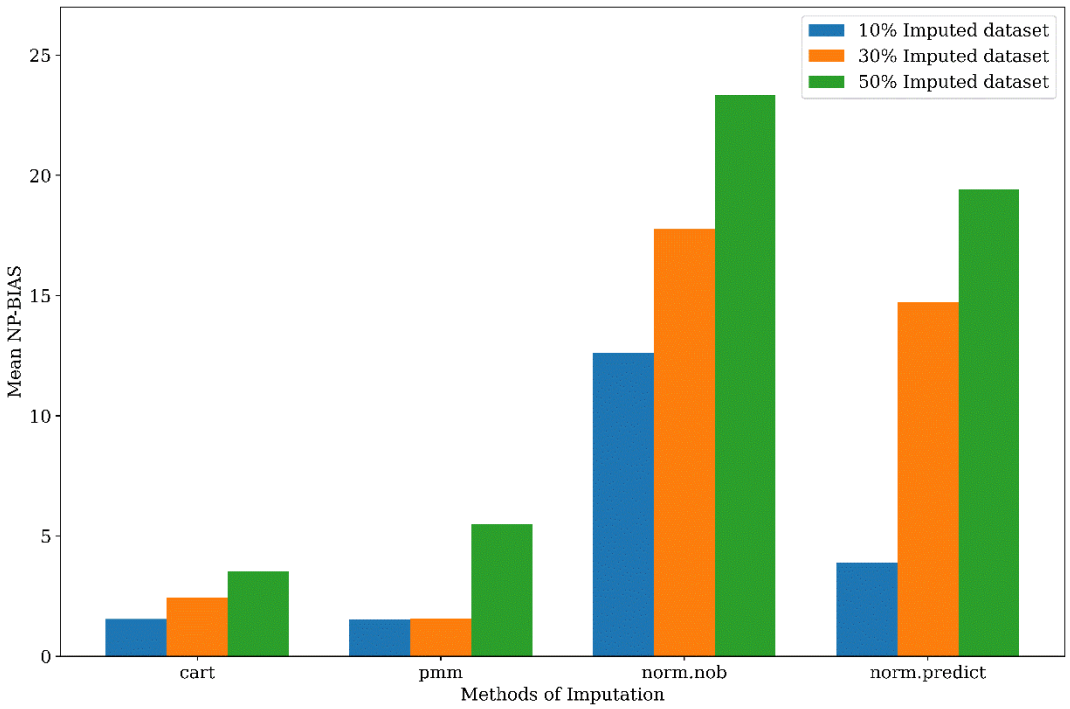
The multiple imputation technique used a unique strategy to resolve imputation uncertainty by repeating a single imputation numerous times. With this method, the imputation uncertainty was considered in an effort to provide a more precise estimation of missing data. The multiple imputation method required assessing each "m" imputed datasets after imputing the incomplete dataset "m" times. A final result was created by combining the outcomes of various analyses. The "mice" package in R was used to implement the Multivariate Imputation by Chained Equations (MICE) approach in this work. Several imputation methods, including cart, PMM, norm.nob, and norm.predict, were used in this methodology for imputation of varying percentages of missing data in two cyber physical attack scenarios with regard to their normal condition data.



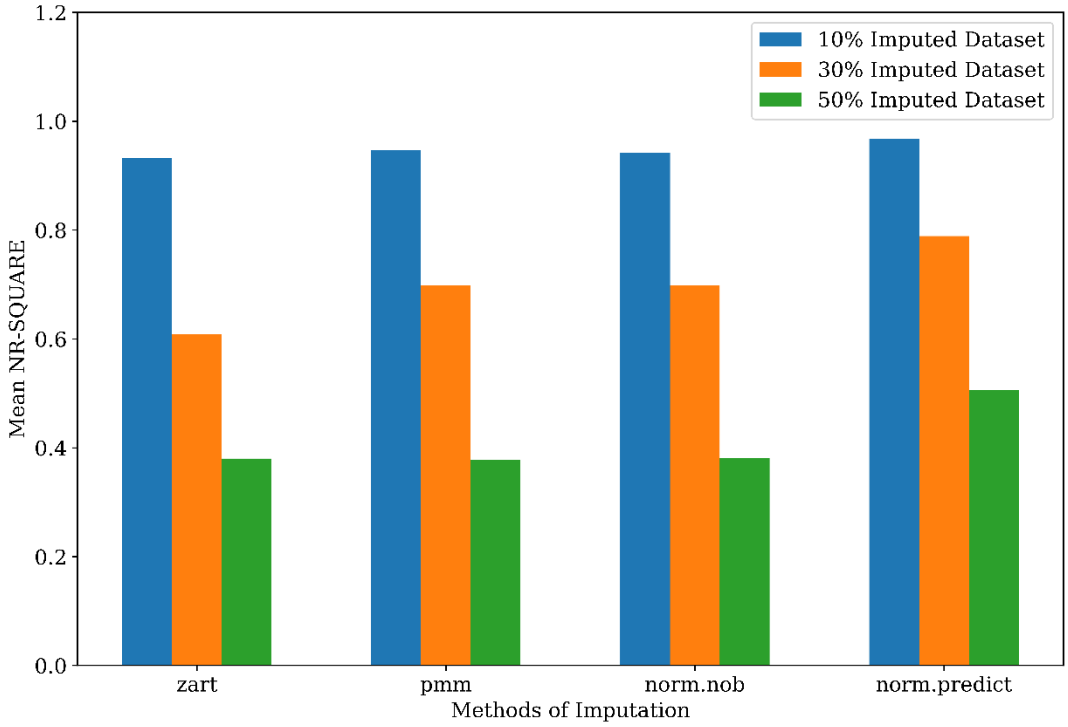
(a)


10% imputed dataset
30% imputed dataset
50% imputed dataset
Mean NMAE
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
cart
pmm
norm.nob
norm.predict
Methods of Imputation

(b)


10% Imputed dataset
30% Imputed dataset
50% Imputed dataset
Mean NP-BIAS
25
20
15
10
5
0
cart
pmm
norm.nob
norm.predict
Methods of Imputation

(c)



(d)

Figure 6: Performance evaluation of different imputation methods for different percentage of missing data using (a) Mean NRMSE (b) Mean NMAE (c) Mean NPBIAS (d) Mean NR-SQUARE

The table presents the performance indicators, specifically NRMSE, NMAE, NR-SQUARE, and NP-BIAS, for different imputation methods at varying percentages of imputation (10%, 30%, and 50%). The methods of imputation include "cart," "pmm," "norm.nob," and "norm.predict." Looking at the NRMSE values, the "cart" method generally exhibits higher error metrics across all imputation percentages than other methods. The lowest NRMSE values are observed with the "norm. predict" method, indicating better accuracy in predicting missing values. In terms of NMAE, the "cart" method shows relatively low error values, especially at 10% imputation. However, the "norm.nob" method demonstrates a noticeable increase in error as the percentage of imputation rises, suggesting a sensitivity to the imputation process. Examining

NR-SQUARE, the "norm.predict" method consistently outperforms others, exhibiting higher values, indicating a better fit to the observed data. The "cart" method, on the other hand, shows a significant decrease in NR-SQUARE as the percentage of imputation increases, implying a diminishing model fit. Lastly, NP-BIAS values reveal the bias introduced by each imputation method. The "norm.nob" method stands out with considerably higher NP-BIAS values, especially at 30% and 50% imputation. This suggests a tendency for this method to introduce bias in the imputed values.

In summary, the choice of imputation method and the percentage of imputation significantly affect performance indicators, emphasizing the importance of careful consideration in handling missing data in WDS datasets. The analysis revealed that a single imputation method is not always suitable, and its effectiveness depends on the specific conditions and the nature of available data. Additionally, the examination of performance indicators concludes that as the percentage of imputation increases, accuracy and resemblance to normal condition datasets decrease. This distortion underscores the limitations of imputing missing data in WDS, highlighting the need for ongoing improvements in imputation methods and the critical importance of selecting the most appropriate approach to mitigate the impact of missing data effectively.

Table 3: Rank of imputation method for different percentage of missing dataset using NMRMSE.

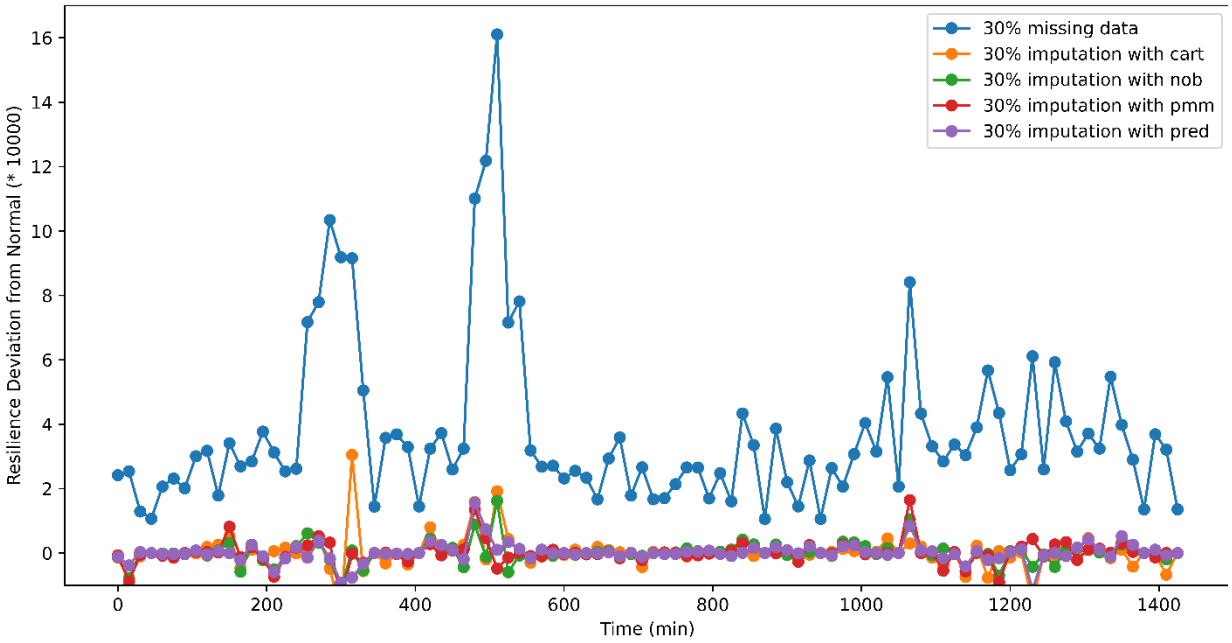
Imputation Method	10% Imputation	30% Imputation	50% Imputation	Rank by Mean	Rank by Mode
cart	3	3	3	3	3
pmm	2	1	2	1.7	2
norm.nob	4	4	4	4	4
norm.predict	1	2	1	1.3	1
Alternative hypothesis:			Kendall		
Wt=0.9111		Chi-Squared=8.2			p-value=0.04205

Table 3 displays a ranking of imputation methods for different percentages of missing data using NRMSE. The imputation methods are ranked higher if they have lower values for Mean NRMSE, Mean NMAE, and Mean NPBIAS, and vice versa. However, for Mean NR-Square, the ranking is higher for imputation methods with higher values, and vice versa. At the bottom of Table 3, Kendall's statistics are used to test the agreement among the ranking of imputation methods when different imputation percentages are done for the same dataset.

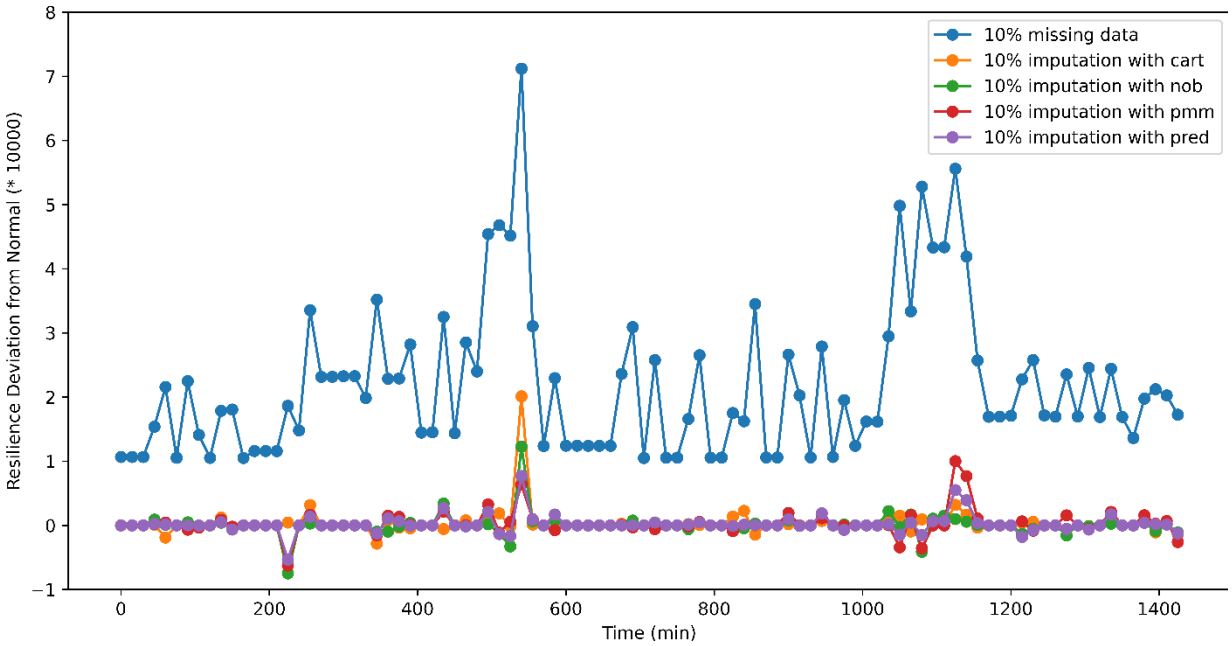
The researcher used Kendall's W test statistics to test two hypotheses about the consistency of each imputation method's performance, with the null hypothesis stating no agreement and the alternative hypothesis stating agreement among rankings of different imputation methods. The W statistic for Table 2 was close to one, and the p-value was significant at a 5% level of significance, implying that the null hypothesis was rejected in all cases. A chi-square test with n-1 degrees of freedom was used to determine the statistical significance of Kendall's W.

Consequently, the imputation method's ranking remains consistent, regardless of the performance indicator or the proportion of missing data in the dataset. This alignment in rankings across various imputation methods aligns with the findings proposed by (Jadhav et al., 2019). This conclusion applies to the study's numeric datasets, which had varying percentages of missing data. Furthermore, the norm.nob imputation method exhibited the highest Mean NRMSE, while norm.predict demonstrated the lowest Mean NRMSE across various imputation percentages, suggesting superior performance for norm.predict. But, for other performance indicator, considering Figure 6, it becomes evident that the norm.predict stands out as the most favorable choice when assessing performance indicators Mean NR-Square. However, when considering Mean NP-BIAS and Mean NMAE, cart and pmm emerge as the superior choices among the selected imputation options. This result demonstrates the importance of employing a combination of diverse imputation methods chosen based on various performance indicators. Such an approach is essential for achieving precise imputation of missing data and identifying the most practical combination of imputation methods. It's essential to recognize that this study focuses on numerical datasets. The chosen imputation methods work well for this study because they only deal with numeric data. However, different imputation methods would be necessary if the study involved categorical datasets(Ishaq et al., 2023; Nishanth & Vadlamani, 2016). The best imputation method can vary based on the specific situation and the type of data being handled.

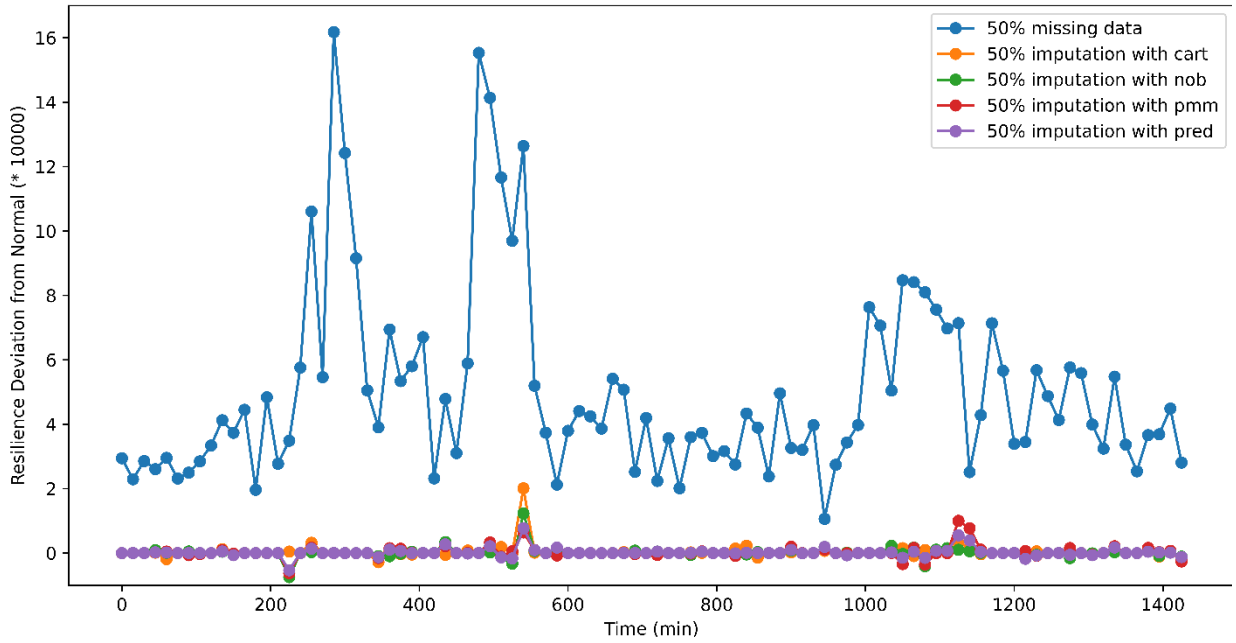
The second portion of the result demonstrated the change in resilience value for different percentages of missing values from 10% to 50%.


30% missing data
30% imputation with cart
30% imputation with nob
30% imputation with pmm
30% imputation with pred
Resilience Deviation from Normal (* 10000)
Time (min)

(a)


10% missing data
10% imputation with cart
10% imputation with nob
10% imputation with pmm
10% imputation with pred
Resilience Deviation from Normal (* 10000)
Time (min)

(b)



(c)

Figure 7: Resilience deviating from Normal condition resilience for unimputed and imputed datasets for (a) 10%, (b) 30% and (c) 50% datasets.

Figure 7 illustrates the resilience deviation in both unimputed and imputed datasets across various imputation percentages. In Figure 7(a), the resilience deviation in time series data calculated from different imputation methods closely aligns with the resilience under normal conditions. However, a significant difference is apparent when assessing the deviation between unimputed dataset resilience and normal condition resilience. This shift in deviation intensifies as the percentage of missing data increases, as evident in Figure 7(b) and (c). Notably, implementing appropriate imputation methods results in negligible resilience deviation for different percentages of imputed datasets compared to normal condition datasets.

3.4 Conclusion

In summary, compared to normal condition datasets, this investigation in imputed datasets across various percentages underscores a growing deviation in performance indicators—

NRMSE, NMAE, N-R-Square, and PBIAS—as the imputation percentage increases. Analyzing the resilience deviation across different percentages of imputed and unimputed datasets highlights the necessity for robust imputation methods that closely approximate the original datasets. Utilizing these values becomes imperative for achieving resilience in imputed datasets that closely align with resilience under normal conditions.

To find the optimal imputation method from the selected options, experiments were conducted by introducing varying percentages of missing pressure data from seven tanks at different positions. The performance of imputation methods, evaluated through diverse indicators, demonstrated a high level of consensus in ranking imputation methods across datasets and missing value percentages, as indicated by Kendall's coefficient of concordance, W, approaching 1. Norm.predict emerged as the most favorable imputation method based on mean NRMSE performance indicator, however when considering different performance indicator for selection best imputation, different imputation methods were found best for different percentages of imputed datasets compared to normal datasets.

Expanding this study to include more substantial amounts of missing values in various pressure and flow parameters across different demand nodes would enable a more comprehensive analysis of resilience deviation under other imputation methods. Determining the most effective imputation method and ensuring proximity to original values is critical for advancing our comprehension and management of missing data in practical scenarios. This research bears substantial implications for enhancing system performance and refining resilience-oriented decision-making within water distribution systems.

CHAPTER 4

CONCLUSIONS AND RECOMMENDATIONS

This thesis systematically addresses research questions, each aligned with logical hypotheses in line with the overall goals and objectives of the thesis. Two specific hypotheses were investigated: i) the impact of manipulated data on resilience-based decision-making and ii) the significance of manipulated or missing data in evaluating the resilience of WDS.

In the first phase, the study focused on the influence of manipulated data on resilience-based decision-making. The research supported hypotheses confirming that reliable pressure and flow data, combined with an energy-surplus-based resilience measure, enhance resilience in decentralized water supply systems under both normal and failure conditions. Notably, a 10% pressure deviation substantially affected system resilience, revealing vulnerabilities and potential inefficiencies. The study highlighted the critical role of accurate pressure values, emphasizing the necessity for robust cybersecurity measures to protect critical infrastructure.

Transitioning to the second part, the investigation explored the impact of the percentage of missing data on the reliability of resilience. The norm.predict imputation method proved effective, validating an alternate hypothesis. The study revealed that resilience calculated from different percentages of imputed datasets closely resembled resilience from normal condition datasets. In contrast, unimputed datasets exhibited increased resilience deviation as the percentage of missing data increased. This underscored the importance of robust imputation methods for a comprehensive system understanding and effective resilience analysis.

In conclusion, the research aimed to enhance water distribution system resilience to disruptive events, advocating for increased decentralization as one of the option. The study explored a hybrid centralized and decentralized system, demonstrating heightened resilience with

higher decentralization levels under various disruptive scenarios. The investigation delved into the impact of pressure variation on resilience, highlighting trade-offs in manipulating pressure values with implications for resilience and associated costs.

An analysis of imputed datasets at different percentages revealed increasing deviations in performance indicators as the percentage of missing data increased. Time series plots of resilience demonstrated the deviation of various percentages of imputed datasets, unimputed datasets, and normal datasets. These findings underscored the importance of imputation methods in effectively recovering missing data and calculating the resilience of imputed datasets, closely resembling normal condition resilience.

As a critical reflection on limitations and avenues for future research, it is essential to acknowledge that this study focused on specific failure scenarios and simulated datasets. Future research could broaden dimensions by incorporating contamination, pipe bursts, intermittent water supply, and unauthorized consumption. Validation of the study's findings with real-world failure datasets is crucial for enhancing the model's performance and applicability. Additionally, exploring a broader range of cyber-attack scenarios adjusting failure durations and magnitudes contributes to a more comprehensive understanding of resilience and sensor placement strategies in WDS.

REFERENCES

Abimbola, S., Baatiema, L., & Bigdeli, M. (2019). The impacts of decentralization on health system equity, efficiency and resilience: A realist synthesis of the evidence. *Health Policy and Planning*, *34*. https://doi.org/10.1093/heapol/czz055

Alstone, P., Gershenson, D., & Kammen, D. M. (2015). Decentralized energy systems for clean electricity access. *Nature Climate Change*, *5*(4), Article 4. https://doi.org/10.1038/nclimate2512

Aryal, A., Acharya, A., & Kalra, A. (2022). Assessing the Implication of Climate Change to Forecast Future Flood Using CMIP6 Climate Projections and HEC-RAS Modeling. *Forecasting*, *4*(3), Article 3. https://doi.org/10.3390/forecast4030032

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49.

Babu Ghimire, A., Parajuli, U., Bhusal, A., Parajuli, A., Banjara, M., & Shin, S. (2023). Investigating a Diversified and Decentralized Water Distribution System to Enhance Water Supply Resilience to Disruptive Events. *World Environmental and Water Resources Congress 2023*, 941–951. https://doi.org/10.1061/9780784484852.087

Bata, M. H., Carriveau, R., & Ting, D. S.-K. (2020). Short-Term Water Demand Forecasting Using Nonlinear Autoregressive Artificial Neural Networks. *Journal of Water Resources Planning and Management*, *146*(3), 04020008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001165

Beaulieu-Jones, B. K., & Moore, J. H. (2016). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Biocomputing 2017* (pp. 207–218). WORLD SCIENTIFIC. https://doi.org/10.1142/9789813207813_0021

Bhandari, S., Jobe, A., Thakur, B., Kalra, A., & Ahmad, S. (2018). Flood Damage Reduction in Urban Areas with Use of Low Impact Development Designs. *World Environmental and Water Resources Congress 2018*, 52–61. https://doi.org/10.1061/9780784481431.006

Bhusal, A., Kalra, A., & Shin, S. (2023). Resilience effect of decentralized detention system to extreme flooding events. *Journal of Hydroinformatics*, 1. https://doi.org/10.2166/hydro.2023.176

Butler, D., Farmani, R., Fu, G., Ward, S., Diao, K., & Astaraie-Imani, M. (2014). A New Approach to Urban Water Management: Safe and Sure. *Procedia Engineering*, *89*, 347–354. https://doi.org/10.1016/j.proeng.2014.11.198

Cao, G., Gu, W., Lou, G., Sheng, W., & Liu, K. (2022). Distributed synchronous detection for false data injection attack in cyber-physical microgrids. *International Journal of Electrical Power & Energy Systems*, *137*, 107788. https://doi.org/10.1016/j.ijepes.2021.107788

Cheliotis, M., Gkerekos, C., Lazakis, I., & Theotokatos, G. (2019). A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. *Ocean Engineering*, *188*, 106220. https://doi.org/10.1016/j.oceaneng.2019.106220

Cheng, C.-H., Chan, C.-P., & Sheu, Y.-J. (2019). A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Engineering Applications of Artificial Intelligence*, *81*, 283–299. https://doi.org/10.1016/j.engappai.2019.03.003

Chhetri, A., & Tamang, L. (2019). *DECENTRALIZATION OF WATER RESOURCE MANAGEMENT: ISSUES AND PERSPECTIVES INVOLVING PRIVATE AND*

COMMUNITY INITIATIVES IN DARJEELING TOWN, WEST BENGAL. *39*, 240–255. https://doi.org/10.32381/ATNAGI.2019.39.02.6

Durrant, G. B. (2005). *Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review*.

Evans, S., Williams, G. P., Jones, N. L., Ames, D. P., & Nelson, E. J. (2020). Exploiting Earth Observation Data to Impute Groundwater Level Measurements with an Extreme Learning Machine. *Remote Sensing*, *12*(12), Article 12. https://doi.org/10.3390/rs12122044

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, 52–65. https://doi.org/10.1016/j.eswa.2017.07.026

Ghimire, A., Banjara, M., Bhusal, A., & Kalra, A. (2023a). Evaluating the Effectiveness of Low Impact Development Practices against Climate Induced Extreme Floods. *International Journal of Environment and Climate Change*, *13*, 288–303. https://doi.org/10.9734/ijecc/2023/v13i81953

Ghimire, A., Faruk, O., Shadia, N., Parajuli, U., & Shin, S. (2023b). Evaluating the correlation of SPI, SPEI, and SSI with climatic and socioeconomic factors for drought monitoring. *Journal of Environmental Engineering and Science*, 1–9. https://doi.org/10.1680/jenes.23.00070

Hernández-Pereira, E. M., Álvarez-Estévez, D., & Moret-Bonillo, V. (2015). Automatic classification of respiratory patterns involving missing data imputation techniques. *Biosystems Engineering*, *138*, 65–76. https://doi.org/10.1016/j.biosystemseng.2015.06.011

Ishaq, M., iftikhar, L., Khan, M., & Khan, A. (2023). *Machine Learning Based Missing Values Imputation in Categorical Datasets*.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, *33*, 1–21. https://doi.org/10.1080/08839514.2019.1637138

Javaid, M., Haleem, A., Singh, R. P., Rab, S., & Suman, R. (2021). Significance of sensors for industry 4.0: Roles, capabilities, and applications. *Sensors International*, *2*, 100110. https://doi.org/10.1016/j.sintl.2021.100110

Javed, M., Tariq, N., Ashraf, M., Khan, F., & Imran, M. (2023). Securing Smart Healthcare Cyber-Physical Systems against Blackhole and Greyhole Attacks Using a Blockchain-Enabled Gini Index Framework. *Sensors*, *23*, 9372. https://doi.org/10.3390/s23239372

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Jonnalagadda, V., Lee, J. Y., Zhao, J., & Ghasemi, S. H. (2023). Quantification and Reduction of Uncertainty in Seismic Resilience Assessment for a Roadway Network. *Infrastructures*, *8*, 128. https://doi.org/10.3390/infrastructures8090128

Joshi, N., Rahaman, Md. M., Thakur, B., Shrestha, A., Kalra, A., & Gupta, R. (2020a). Assessing the Effects of Climate Variability on Groundwater in Northern India. *World Environmental and Water Resources Congress 2020*, 41–52. https://doi.org/10.1061/9780784482964.005

Joshi, N., Tamaddun, K., Parajuli, R., Kalra, A., Maheshwari, P., Mastino, L., & Velotta, M. (2020b). Future changes in water supply and demand for Las Vegas valley: A system dynamic approach based on CMIP3 and CMIP5 climate projections. *Hydrology*, *7*(1), 16.

Kalbar, P., & Lokhande, S. (2023). Need to adopt scaled decentralized systems in the water infrastructure to achieve sustainability and build resilience. *Water Policy*, *25*. https://doi.org/10.2166/wp.2023.267

Kalra, A., Thakur, B., & Gupta, R. (2022). Analyzing the Relationship between the Pacific Ocean SST and Streamflow of Two Drought Sensitive Watersheds within Northern California. *World Environmental and Water Resources Congress 2022*, 472–480. https://doi.org/10.1061/9780784484258.043

Khetwal, S., Gutierrez, M., & Pei, S. (2022). Sensitivity analysis of road tunnel resilience through data-driven stochastic simulation. *Intelligent Transportation Infrastructure*, *1*. https://doi.org/10.1093/iti/liac003

Kyureghian, G., Capps, O., & Nayga, R. M. (2011). A Missing Variable Imputation Methodology with an Empirical Application. In D. M. Drukker (Ed.), *Advances in Econometrics* (Vol. 27, pp. 313–337). Emerald Group Publishing Limited. https://doi.org/10.1108/S0731-9053(2011)000027A015

Liang, X., Konstantinou, C., Shetty, S., Bandara, E., & Sun, R. (2023). Decentralizing Cyber Physical Systems for Resilience: An Innovative Case Study from A Cybersecurity Perspective. *Computers & Security*, *124*, 102953. https://doi.org/10.1016/j.cose.2022.102953

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Mamat, N., Fatin, S., & Mohd Razali, S. F. (2023). Comparisons of Various Imputation Methods for Incomplete Water Quality Data: A Case Study of The Langat River, Malaysia. *Jurnal Kejuruteraan*, *35*, 191–201. https://doi.org/10.17576/jkukm-2023-35(1)-18

Ngo, T., Yoo, D., & Kim, J. (2018). Decentralization-based optimization of detention reservoir systems for flood reduction in urban drainage areas. *Urban Water Journal*, *15*, 1–8. https://doi.org/10.1080/1573062X.2018.1508600

Nieh, C., Dorevitch, S., Liu, L. C., & Jones, R. M. (2014). Evaluation of imputation methods for microbial surface water quality studies. *Environmental Science: Processes & Impacts*, *16*(5), 1145–1153. https://doi.org/10.1039/C3EM00721A

Nishanth, K., & Vadlamani, R. (2016). Probabilistic Neural Network based Categorical Data Imputation. *Neurocomputing*, *218*. https://doi.org/10.1016/j.neucom.2016.08.044

Oriani, F., Borghi, A., Straubhaar, J., Mariethoz, G., & Renard, P. (2016). Missing data simulation inside flow rate time-series using multiple-point statistics. *Environmental Modelling & Software*, *86*, 264–276. https://doi.org/10.1016/j.envsoft.2016.10.002

Parajuli, R., Nyaupane, N., & Kalra, A. (2017). Analyzing Future Flooding under Climate Change Scenario using CMIP5 Streamflow Data. *AGU Fall Meeting Abstracts*, H11E-1216. https://ui.adsabs.harvard.edu/abs/2017AGUFM.H11E1216P/abstract

Pickett, S. T. A., McGrath, B., Cadenasso, M. L., & Felson, A. J. (2014). Ecological resilience and resilient cities. *Building Research & Information*, *42*(2), 143–157. https://doi.org/10.1080/09613218.2014.850600

Pokhrel, I., Kalra, A., Rahaman, M. M., & Thakali, R. (2020). Forecasting of future flooding and risk assessment under CMIP6 climate projection in neuse river, North Carolina. *Forecasting*, *2*(3), 323–345.

Quinteros, M. E., Lu, S., Blazquez, C., Cárdenas-R, J. P., Ossa, X., Delgado-Saborit, J.-M., Harrison, R. M., & Ruiz-Rudolph, P. (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmospheric Environment*, *200*, 40–49. https://doi.org/10.1016/j.atmosenv.2018.11.053

Razavi-Far, R., Cheng, B., Saif, M., & Ahmadi, M. (2020). Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, *187*, 104805. https://doi.org/10.1016/j.knosys.2019.06.013

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., & Gorgoglione, A. (2021). Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. *Sustainability*, *13*(11), Article 11. https://doi.org/10.3390/su13116318

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sagarika, S., Kalra, A., & Ahmad, S. (2015). Evaluating the Relationship between Western U.S. Streamflow and Pacific Ocean Climate Variability. *World Environmental and Water Resources Congress 2015*, 999–1008. https://doi.org/10.1061/9780784479162.097

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *Sn Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Sarma, R., & Singh, S. K. (2022). A Comparative Study of Data-driven Models for Groundwater Level Forecasting. *Water Resources Management*, *36*(8), 2741–2756. https://doi.org/10.1007/s11269-022-03173-6

Shin, S., Lee, S., Burian, S. J., Judi, D. R., & McPherson, T. (2020). Evaluating Resilience of Water Distribution Networks to Operational Failures from Cyber-Physical Attacks. *Journal of Environmental Engineering*, *146*(3), 04020003. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001665

Shin, S., Lee, S., Judi, D. R., Parvania, M., Goharian, E., McPherson, T., & Burian, S. J. (2018). A Systematic Review of Quantitative Resilience Measures for Water Infrastructure Systems. *Water*, *10*(2), Article 2. https://doi.org/10.3390/w10020164

Shrestha, A., Rahaman, M. M., Kalra, A., Jogineedi, R., & Maheshwari, P. (2020a). Climatological drought forecasting using bias corrected CMIP6 climate data: A case study for India. *Forecasting*, *2*(2), 59–84.

Shrestha, A., Rahaman, M. M., Kalra, A., Thakur, B., Lamb, K. W., & Maheshwari, P. (2020b). Regional climatological drought: An assessment using high-resolution data. *Hydrology*, *7*(2), 33.

Simic, V., Stojkovic, M., Milivojevic, N., & Bacanin, N. (2023). Assessing water resources systems' dynamic resilience under hazardous events via a genetic fuzzy rule-based system. *Journal of Hydroinformatics*, *25*(2), 318–331. https://doi.org/10.2166/hydro.2023.101

Ssali, G., & Marwala, T. (2008). Computational intelligence and decision trees for missing data estimation. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 201–207. https://doi.org/10.1109/IJCNN.2008.4633790

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and

clinical research: Potential and pitfalls. *BMJ*, *338*, b2393. https://doi.org/10.1136/bmj.b2393

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., & Ostfeld, A. (2017). Characterizing cyber-physical attacks on water distribution systems. *Journal of Water Resources Planning and Management*, *143*(5), 04017009.

Thakali, R., Kalra, A., Mastino, L., Velotta, M., & Ahmad, S. (2016). *Assessment of Vulnerability to Climate Change Effects on City of Las Vegas Urban Stormwater Infrastructure*. http://thrivingearthexchange.org/wp-content/uploads/2017/01/Poster_AGU_Final.pdf

Todini, E. (2000). Looped water distribution networks design using a resilience index based heuristic approach. *Urban Water*, *2*, 115–122. https://doi.org/10.1016/S1462-0758(00)00049-2

Varshini, G., & Latha, S. (2023). Detection of Data Integrity Attack Using Model and Data-Driven-Based Approach in CPPS. *International Transactions on Electrical Energy Systems*, *2023*, 1–24. https://doi.org/10.1155/2023/6098519

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Zanfei, A., Menapace, A., Brentan, B. M., & Righetti, M. (2022). How Does Missing Data Imputation Affect the Forecasting of Urban Water Demand? *Journal of Water Resources Planning and Management*, *148*(11), 04022060. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001624

VITA

Graduate School
Southern Illinois University Carbondale

Amrit Ghimire

ghimire451@gmail.com

Tribhuvan University, Nepal
Bachelor of Engineering, Civil Engineering, December 2018

Thesis Title: Impact of Data Reliability on Resilience-Based Decision Making in a Water Distribution System

Major Professor: Dr. Sangmin Shin

Publications:

1) Babu Ghimire, A., Parajuli, U., Bhusal, A., Parajuli, A., Banjara, M. and Shin, S., Investigating a Diversified and Decentralized Water Distribution System to Enhance Water Supply Resilience to Disruptive Events. In *World Environmental and Water Resources Congress 2023* (pp. 941-951).
2) Parajuli, U., Bhusal, A., Babu Ghimire, A. and Shin, S., Comparing HEC-HMS, PCSWMM, and Random Forest Models for Rainfall-Runoff Evaluation to Extreme Flooding Events. In *World Environmental and Water Resources Congress 2023* (pp. 1250-1262).
3) Ghimire, A.B., Faruk, O., Shadia, N., Parajuli, U. and Shin, S., 2023. Correlation of drought indices with climatic and socio-economic factors in San Diego, USA. *Journal of Environmental Engineering and Science*, *40*, pp.1-12.
4) Bhusal, A., Ghimire, A.B., Thakur, B. and Kalra, A., 2023. Evaluating the hydrological performance of integrating PCSWMM and NEXRAD precipitation product at different spatial scales of watersheds. *Modeling Earth Systems and Environment*, pp.1-14.

5) Ghimire, A.B., Banjara, M., Bhusal, A. and Kalra, A., 2023. Evaluating the Effectiveness of Low Impact Development Practices against Climate Induced Extreme Floods. *International Journal of Environment and Climate Change*, *13*(8), pp.288-303.