WATER DEMAND FORECASTING USING MACHINE LEARNING APPROACH

BY

QENEHELO MAHAMO

201905395

DISSERTATION SUBMITTED TO THE NATIONAL UNIVERSITY OF LESOTHO WATER INSTITUTE IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INTEGRATED CATCHMENT AND WATER RESOURCES MANAGEMENT

NATIONAL UNIVERSITY OF LESOTHO

OCTOBER 2025

Abstract

The increasing challenges related to water security, exacerbated by rapid urbanization, population growth, and climate variability, necessitate accurate and reliable forecasting methodologies to support sustainable water resources planning. This study explores water demand forecasting in Ha-Foso, Lesotho, by evaluating three machine learning models: Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Artificial Neural Networks (ANN). Utilizing time series data sets covering meteorological inputs (2012-2022), population, and water consumption records (2017-2024), the study assesses the influence of climatic and demographic variables, specifically precipitation, maximum and minimum temperatures, and population, on domestic water consumption.

The research first used MLR to assess the influence of population, maximum temperature, minimum temperature, precipitation, and other factors on water demand. Subsequently, the study evaluated the predictive performance of MLR, SVR, and ANN models. Performance was evaluated using performance metrics, including the coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

The regression analysis consistently identified population as the only statistically significant predictor of water demand ($p<0.001$), while climatic variables showed no significant influence during the study period. In the comparative evaluation, the SVR demonstrated the highest accuracy and generalization capacity, outperforming ANN and MLR, with the least error metrics in both the training phase and the testing phase. The 2-year forecast highlighted the distinct behaviours of each model, with the SVR and ANN models providing more moderate growth projections compared to the steep, linear increase predicted by the MLR model.

This study presents the potential of machine learning, particularly SVR and ANN, in addressing the intricate, non-linear relationships inherent in water demand forecasts, delivering precise and actionable water demand forecasts for peri-urban settings in Lesotho. The findings suggest the adoption of advanced architectures and the incorporation of socio-demographic variables to strengthen predictive capacity. The outputs are expected to support utility companies such as Water and Sewerage Company (WASCO) in strategic planning, conservation, and infrastructure investment decisions.

Keywords: Determinants of water demand, water demand prediction, Artificial neural Networks, Machine Learning, Support Vector Regression, Multiple Linear Regression.

Declaration

The work contained in this dissertation was carried out and completed by **Qenehelo Mahamo**, **201905395,** at the National University of Lesotho Water Institute, National University of Lesotho. I hereby declare that this study constitutes my original work and has never been submitted for the award of a degree or diploma to any University. To the best of my knowledge, this dissertation contains no material written by another person except where due reference is made in the dissertation itself.

Signature: **Date: 08/10/2025**

As the candidate's supervisor, I certify the above statement to be correct to my knowledge and have recommended this dissertation for submission.

Dr Liphapang Khaba……………………………… Dae: 08/10/2025

Acknowledgements

The successful completion of this thesis is attributed to the sure help, guidance, and contributions of many individuals and organisations. Therefore, I would like to express my profound gratitude and take this opportunity to acknowledge their significant roles in this endeavour formally.

I am grateful to my supervisor, Dr. Liphapang Khaba, for his contribution to this study.

My sincerest thanks are due to the Water and Sewage Company (WASCO) and the Lesotho Meteorological Services (LMS) for their crucial institutional support. I am particularly grateful to Mr Mpho Mokoatleng of WASCO for his instrumental support role in providing the requisite data. The professionalism and cooperation extended by both organizations were fundamental to achieving the objectives of this study.

I would like to thank the following organizations for the role they played in my MSc in ICWRM at the National University of Lesotho.

- The Government of Lesotho in general and the ReNOKA Movement in particular which initiated and coordinated the implementation of the Integrated Catchment Management Programme in Lesotho.
- The EU and German Government for availing the funds for my full scholarship at the NUL.
- Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) for providing the necessary technical assistance in the implementation of the ICM programme under which the MSc in ICWRM Programme was initiated.
- WaterNet for managing the scholarship fund for MSc in ICWRM.
- The National University of Lesotho in general and the Water Institute in particular for running the MSc in ICWRM Programme.

The pursuit of an academic journey of this nature is sustained by a strong personal support system. I express profound gratitude to my family, particularly my parents, for their steadfast belief in my capabilities and unwavering support throughout this process. Additionally, I extend my appreciation to my friends and colleagues within the program, whose shared experiences of both challenges and triumphs have served as a continuous source of strength and encouragement. Your camaraderie has been invaluable in navigating this academic endeavour.

Finally, I am grateful to God in the Name of Jesus Christ for the grace, strength, wisdom, and perseverance bestowed upon me, for in Him, I live and have my being.

Table of Contents

List of Acronyms

ANN: Artificial Neural Networks

ARIMA: Autoregressive Integrated Moving Average

GDP: Gross Domestic Product

LMS: Lesotho Meteorological Company

LSTM: Long Short-Term Memory

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

ML: Machine Learning

RF: Random Forest

RMSE: Root Mean Square Error

SVM: Support Vector Regression

VC: Vapnik-Chervonenkis

WASCO: Water and Sewage Company

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Population growth is projected to increase globally, with the world population anticipated to reach 8.5 billion by 2030, placing significant pressure on natural resources, particularly water resources (UN, 2017). This escalating demand, coupled with impacts of climate change like increased frequency and intensity of droughts and extreme weather events, threatens water security for approximately half the world's population (IPCC, 2023). The Economic Forum report ranks water insecurity among the top five global risks, highlighting the urgency for sustainable water management (WEF, 2024).

In Sub-Saharan Africa, rapid urbanization, economic growth, and changing consumption patterns are driving a surge in water demand (UN, 2024). To address the water security concerns, many nations have invested in dam construction to support water supply, flood regulation, and hydroelectric power production. Globally, there are 35140 dams, of which 6243 are in Africa, 3170 of which are mainly for water supply. For example, in Egypt, the Aswan Dam was built in 1970 (Zhang & Gu, 2023; Sekamane, et al., 2023).

While dam construction has emerged as a common approach to address water needs, effective strategies for water management necessitate accurate demand forecasting. The understanding of the determinants of water demand is crucial in effective forecasting, policy development, and overall decision-making. These factors encompass population growth, employment and technology, weather and climate, the price of water, and efficiency and conservation programs as major influences on water demand (Billings & Jones, 2008). For instance, studies highlight the role of population size, income, and climate elements, including temperature, humidity, and land use, as major influences on water demand (Anang , et al., 2019; Huang, et al., 2020; Ou, et al., 2023). This demonstrates the importance of understanding the different impacts that different factors have on water demand, thereby enabling accurate forecasting.

Traditionally, water demand forecasting relied on qualitative methods, often based on expert knowledge or judgments (Billings & Jones, 2008). These methods, however, have demonstrated a lack of reliability and empirical analysis. Furthermore, traditional approaches such as Autoregressive Integrated Moving Average (ARIMA) models and multivariate regression analysis assume linear relationships, which may not reflect real-world complexities. To address these limitations, research has increasingly explored non-parametric approaches such as machine learning, which better understand non-linear relations as they excel at identifying intricate learning patterns, leading to improved accuracy when forecasting water demand, moreover can handle diverse and large amount of data, making them highly adaptable to change (Billings & Jones, 2008; Donkor, et al., 2014).

Machine learning is a subfield of artificial intelligence that focuses on developing algorithms capable of learning patterns from data. Consequently, improving accuracy in forecasting various phenomena, including water demand. Research has examined the effectiveness of different methods, such as Support Vector Machines (SVM), Random Forest (RF), Extreme Learning Machines (ELM), and Artificial Neural Networks (ANN), which have presented significant potential owing to their capability to accurately model intricate, non-linear relationships with accuracy and precision (Bennett, et al., 2013; Mouatadid & Adamowski, 2017; Pesantez, et al., 2020).

Lesotho, exhibiting a fast-growing population, is anticipated to increase from approximately 2 million in 2016 to 2.5 million by 2035. The population in urban areas is projected to grow from approximately 700,000 in 2016 to 1 million in 2036, further increasing the burden of demand for water resources (BOS, 2019). While significant investments have been made in water infrastructure, including the construction of the Metolong Dam, whose main purpose is water supply. Effective water resources require accurate demand forecasting to ensure equitable allocation and management of water resources. This presents an opportunity for the utilization of machine learning to forecast water demand.

Despite the global advancements in machine learning for water demand forecasting, Lesotho currently lags in its application. Presently, the Water and Sewage Company (WASCO) relies on historical data to estimate the demand, without any determination of different influences on water demand, neglecting the impact of various factors on water. Therefore, this study aims to forecast water demand for Ha-Foso, Lesotho, using the machine learning approach.

1.2 Problem Statement

Lesotho faces significant challenges in water infrastructure development, exacerbated by the impacts of climate change, which have severely affected water availability (World Bank, 2016). Despite these challenges, the country has made notable progress in improving water access, achieving 86.6% water supply coverage (UN 2023). Initiatives such as the Lowland Water Supply Scheme, the Maseru Peri-Urban Water Supply Project, High North Reservoir Connection Water Project, and the Urban and Peri-Urban Water Supply Project have been implemented to address water shortages that have persisted (2008-2016), however, rapid population growth and water shortages persist (WASCO, 2025).

Ha-Foso, a rapidly growing area in Lesotho, exemplifies these challenges. Despite interventions in the area, the increasing population has placed immense pressure on water resources, resulting in demand outstripping supply (WASCO, 2025). This has led to frequent water supply interruptions, underscoring the pressing need for accurate water demand forecasting to support effective water resource planning and management.

Utility companies require advanced and accurate forecasting approaches capable of addressing the intricate complexities of water demand, such as non-linear relationships between variables like population, temperature, and precipitation. However, in Lesotho, there is a noticeable gap in research on water demand forecasting, particularly in the employment of machine learning techniques. The absence of localised studies and evidence-based forecasting methods hinders the ability of utility companies to plan and allocate water resources efficiently, further exacerbating water shortages.

1.3 Aim and Objectives

The main aim of this study is to forecast the water demand of Ha-Foso peri-urban in Lesotho using a supervised machine-learning approach

1.3.1 Specific Objectives

- To examine the influence of population growth, temperature, and precipitation on water demand
- To compare the performance of Artificial Neural Network, Support Vector Regression, and Multiple Linear Regression on water demand
- To assess the predictive outputs of Artificial Neural Network, Linear Regression, and Support Vector Regression

1.3.1 Research Questions

Outlined below are research questions that were formulated to guide the study toward achieving the above research objectives:

1. What is the influence of population, temperature, and precipitation on water demand?
2. Which variable has the strongest correlation with water demand?
3. What is the best-performing model between ANN, SVR, and MLR?
4. How do the water forecasts generated by the models compare?
5. What are the similarities and differences produced by the model outputs?
6. What are the implications of the outputs produced on water demand and utility companies?

1.4 Significance of the study

This study aims to assess the influence of temperature, precipitation, and population on water demand, moreover, it forecasts water demand using machine learning approaches, informing utility companies of better methods for forecasting, particularly for Lesotho, which currently relies on traditional methods. Machine learning methods present robustness and flexibility through their ability to learn hidden patterns and handle noise, making them indispensable. Understanding how different factors influence water demand will help inform policy formulation to design better conservation measures and allocation means, and ensure better water resource planning and management.

Chapter 2: Literature Review

1.1 Introduction

A thorough understanding of water demand dynamics is crucial for effective sustainable water resource management. This is especially critical in regions experiencing increasing climatic variability, urbanization pressures, and infrastructure constraints. This chapter reviews the existing literature on the determinants of domestic water consumption and the methodologies employed for forecasting it. It aims to contextualize the current study within a broader scholarly framework, identifying key theoretical foundations, methodological approaches, and empirical findings that inform research.

2.2 Theoretical Framework

The study is firmly anchored in Systems Theory, which conceptualizes complex entities as networks of multiple interconnected and interdependent components that interact to form a unified whole. Systems theory, as articulated by Lazlo & Krippner (1998), emphasizes the relationships and dynamics between these components, enabling the identification of behaviours and patterns that arise from their interactions. This perspective is established in the work of Von Bertalanffy (1972), emphasizes the importance of understanding the relationships and synergies within a system to define its boundary. In this view, water demand is posited as a complex system, influenced by multiple interacting factors, both internal and external. By applying the system theory, the study seeks to elucidate the intricate interdependencies that determine water demand (Laszlo & Krippner, 1998; Sloan, 2005).

Complementing this perspective is the computational learning theory, often referred to as machine learning theory. This theoretical framework seeks to understand the fundamental principles governing the computational learning process, drawing from computer science and statistics (Sloan, 2005). Computational Learning Theory provides a mathematical foundation for understanding how machines can learn patterns and relationships from data, enabling the development of automated learning systems. Ultimately, integrating systems theory and computational learning theory allows for a nuanced analysis of water demand, recognizing the need to model the intricate interactions within the water demand system, and leveraging machine learning to capture the non-linear and dynamic relationships between its components.

2.3 Determinants of Water Demand

Water demand is a complex phenomenon driven by complex, non-linear relationships between human society and the natural environment. A comprehensive understanding of these influencing factors is fundamental to water resources planning and management (Billings & Jones, 2008; Donkor, et al., 2014). Moreover, understanding these determinants serves as the foundation for accurate forecasting of future demand (Rinaudo, 2015). Traditionally, water utilities used historical patterns as a basis for predicting water demand, however, these have failed to integrate factors that drive water demand into their values (Niknam, et al., 2014). Recent studies have focused on forecasting water demand in response to the global water crisis, demonstrating the impact of various social, economic, and environmental factors. Critically, these factors exhibit both spatial and temporal variability, as concluded by DeSouzza Groppo et al. (2019), that each region should be studied separately, with an understanding and choice of the appropriate model. For instance, Huang et al. (2020), using demand theory in China, demonstrated the temporal variation of drivers of water demand, revealing the urban population as the dominant force between 2000-2014, shifting to the Gross Domestic Product (GDP) of the tertiary industries between 2015-2016.

Similarly, Lu et al. (2018) identified spatial and temporal variability in water demand determinants across 38 cities in China, utilizing ArcGIS for spatial analysis and Excel for temporal analysis. The findings highlighted the influence of annual temperature and precipitation, available water resources, and the cost of domestic water. Such localized analyses are vital for verifying projections and accounting for specific regional conditions. In some regions, particularly those requiring year-round irrigation, external use of water significantly contributes to water demand, driven primarily by environmental factors like temperature and precipitation (Dias & Ghisi, 2024).

Additionally, Ou et al. (2023) used explainable intelligence to quantify the association between water demand and key determinants, which included social, economic, and environmental factors. Population size was reported to have a key role in shaping demand, especially in serviced cities, while highlighting temperature as a crucial factor in water demand forecasting. Aligning with Anang et al. (2019), who, using the pooling technique, demonstrated the significant impact of population density, gross domestic product, and agriculture on water demand in Malaysia.

Further studies have explored these relationships in specific contexts. In Jeddah city, using statistical analysis and Pearson correlation, a strong correlation between population and water demand was found, and between GDP and water demand, but a weaker correlation with atmospheric factors such as humidity and temperature (Wardak & Abed, 2019). Conversely, Alshaikhli et al. (2021) demonstrated the significant influence of temperature and sunshine hours in Qatar, while observing a strong negative correlation between population density and precipitation. Bashar et al. (2023) found a linear relationship between population and water demand, with water consumption increasing with increasing temperature. Bich-Ngoc et al. (2022), using fixed and mixed effects regression in R, highlighted the statistical significance of income on water demand in Belgium, suggesting that higher-earning households consume higher water volumes due to lifestyle changes. Other studies demonstrate that water alternatives and conservation methods, such as water harvesting and water-efficient fixtures, can decrease water demand (Garcia, et al., 2019; Bich-Ngoc, et al., 2022). In Nigeria, Oyerinde & Jacobs (2021) used multiple linear regression to identify and evaluate the determinants of household water demand, with an explained variance of 0.45, attributed to complexity in water supply systems.

2.3.1 Environmental Factors

Environmental factors have been widely researched in relation to water demand. Zhang et al. (2014), through a literature review, confirmed the significant correlation between environmental factors and water demand. Numerous environmental factors have been explored, including sunshine hours, humidity, evaporation, precipitation, temperature, and available water resources.

2.3.1.1 Temperature

Recent studies have explored the influence of temperature to try to understand water demand. For instance, Ou et al. (2023) found a complex relationship between temperature and water demand, with temperature being influential in industrial cities and agricultural cities. They noted that higher temperatures can lead to higher losses and increased irrigation demand due to evaporation and transpiration, while also potentially increasing industrial water demand for cooling activities. The study highlighted the paradoxical nature of temperature, suggesting that temperatures may lead to decreased demand owing to transpiration and evaporation. Alshaikhli et al. (2021) also identified temperature as the most influential factor in Qatar, a desert region in Asia, with temperatures exceeding 45ºC, where high evaporation and transpiration drive water demand.

2.2.1.2 Precipitation

Fan et al. (2017) analysed determinants of water consumption across 286 municipalities in China, and they observed a significant correlation between precipitation and water usage in high-consumption cities. They proposed that this relationship may stem from increased outdoor activities in rural areas, as well as increased showering and laundry frequency in urban areas due to humid and rainy weather, which necessitates showering and clothing changes. Conversely, Ashoori et al. (2016) demonstrated a decrease in water demand with increasing precipitation in single-family residential households, likely due to reduced outdoor activities, such as garden watering. Alshaikhli et al. (2021) found an insignificant correlation between precipitation and water demand in Qatar, likely due to the region's arid climate and low precipitation levels. Conversely, Muloiwa et al. (2022) and Knox (2020) demonstrate that meteorological determinants may have a negligible impact when there is reliance on indoor water consumption.

It is, however, important to note that environmental factors are not limited to the aforementioned, for example, Alshaikli et al (2021) included humidity and sunshine hours in their study, expressing the interest in such parameters and prioritizing them over precipitation in Qatar. Similarly, Wardak & Abed (2019) included humidity in their study in Saudi Arabia. This further emphasizes that the interest in factors can differ in special and temporal contexts. Ou et al (2019) explored total water resources, which include ground and surface water in their study. While Bich-Ngoc et al (2022) reported the importance of having alternative water sources. This shows that there is a wide array of environmental factors that research can explore, however, it varies with geographic location.

2.3.2 Socio-economic factors

Socio-economic factors are among the most significant determinants of water demand, though their influence can vary dramatically depending on geographic, technological, and cultural contexts. The most commonly cited factors include population, income, and GDP.

2.3.2.1 Population

Traditionally, utilities have relied on population-based forecasting, based on the rationale that increasing population leads to increased water demand (Billings & Jones, 2008). This is supported by studies like Ou et al. (2023), who found a statistically significant relationship between population and water demand in China. The authors attribute this influence primarily to factors such as urbanization and socio-economic development, which collectively shape consumption patterns and resource allocation. Similarly, Huang et al. (2020) identified population as the primary driver of water demand between 2000-2014, shifting to the tertiary industry between 2015-2016 due to economic development. Anang et al. (2019) noted a positive relationship between population density and agriculture on water demand, highlighting that with a growing population, food production is expected to increase, subsequently increasing water demand.

However, this relationship is not universally direct. A contrasting view is offered by Alshaikhli et al. (2021), who found a negative correlation between population and water demand in Qatar, identifying GDP and temperature as the significant drivers of water demand. Thus, while population is a fundamental baseline for demand, its impact can be mediated by the surrounding economic and environmental context.

2.3.2.2 Income

Water use responds to personal income, and water customers respond to how water is billed, therefore, water savings are realized when residential customers are billed based on consumption. Typically, higher-income households are expected to use more water (Billings & Jones, 2008). For instance, Bich-Ngoc et al. (2022) demonstrated the statistical significance of income on water demand, suggesting lifestyle changes, including the adoption of water-utilizing appliances such as dishwashers, the creation of gardens and pools, and the increased capacity to purchase more water, which contribute to higher water consumption in higher-income households. This is corroborated by the findings of Sant'Ana & Mazzega (2018), who observed higher water consumption in higher-income households in Brazil, attributing this to the prevalence of dwellings with amenities that enhance living standards, such as improved bathroom fixtures, kitchen faucets, and external faucets, as well as increased frequency and duration of water use. Interestingly, Anang et al. (2019) revealed a nonlinear relationship between income and water demand, suggesting that while consumption rises with income, the rate of increase may change at different income levels. This highlights that income influences not just the quantity of water used, but also the patterns and efficiency of its consumption. They note that while a country's wealth may provide amenities that facilitate an increase in water consumption, an optimum will be reached, where comfort declines due to limited water resources.

2.3.2.3 GDP

The association linking GDP and water demand is complex and multifaceted, with empirical studies revealing diverse trends. For example, Xiangmei et al. (2021) reported a declining trend in water consumption with increasing GDP in over half of the regions they studied. They attribute this phenomenon to two potential factors: firstly, economic development can facilitate the adoption of more efficient water-saving technologies. Secondly, rising prices of water associated with economic growth can discourage water consumption, promoting water saving. Conversely, other regions have observed an increase in consumption alongside GDP growth, primarily due to the expansion of water-intensive industries with a rising economy. Huang et al. (2020) also identified the significant role of GDP as a driver of water demand in China between 2015-2016, emphasizing its dynamic nature with different influences. They specifically noted the association with the expansion of the tertiary industry, consequently increasing water demand. Anang et al. (2019) further demonstrated the statistical significance of GDP in relation to water resources, suggesting that while initial economic development may be supported by sufficient water resources, a point can be reached where further economic activity negatively impacts water resources. However, they acknowledge that increased GDP can simultaneously create opportunities for investment in innovative water-saving technologies. Therefore, the association between GDP and water demand is not unidirectional, however, it is mediated by technological advancements, pricing mechanisms, and the structure of economic development.

2.3.2.4 Household size

Household size is widely recognized as an important factor influencing residential water demand. Billings & Jones (2008) explicitly recommend its inclusion in water demand forecasting models to better understand its influence. This is supported by literature, for instance, Motho et al. (2022) reported that a larger household size is a significant factor influencing water demand, explaining that larger households typically require larger amounts of water for daily domestic activities such as cooking, bathing, laundry, and others. The result is consistent with Ogunbode & Ifabiyi (2014) research, which demonstrated the strong statistical significance of household size as a driver of water demand, indicating a positive correlation between household size and water consumption. A review by Dias & Ghisi (2024) further corroborates this relationship, citing numerous studies that have established household size as a key factor in explaining variations in residential water demand.

2.3.2.5 Water pricing and availability

Economic principles suggest that the price of water is a key determinant of its demand. This is acknowledged by Lu et al. (2019), who report that water price is affected by a wide array of factors, including accessibility, abundance, and residents' income. Their findings show that regions with abundant water resources often exhibit higher water consumption, which they attribute to a lack of water-saving awareness driven by a low perceived scarcity of water. Ashoori et al (2016) highlighted that residents are responsive to price and the importance of knowing the impact of elasticity in Los Angeles. Further emphasizing that during drought periods, higher prices have effected lower demand. This, therefore, reflects an interlinkage between awareness and the availability of water.

Beyond primary economic indicators, other demographic characteristics also shape water demand. For example, education levels and gender roles can have a notable impact. As shown by Abdullah et al. (2019), who observed in their study area that households with higher literacy levels tend to have higher water demand. In the same study, they highlighted the role of women, who often bear the primary responsibility for water collection and management, influencing household consumption patterns. This is also observed by Motho et al (2022), who reported a positive association between gender and water consumption. These factors, while sometimes harder to quantify, add another layer of complexity to understanding and forecasting water demand.

The reviewed literature reflects the complexity of the interconnectedness of socio-economic factors influencing demand. For example, population growth is closely linked to population density and household size, while income and GDP together shape both individual consumption habits and broader industrial activity. Additionally, the impact of economic instruments such as water pricing is influenced by social dimensions such as education and public awareness of resource scarcity. While this review addresses key drivers, it is by no means exhaustive. Other critical factors, such as technological advancements, governance and policy, and land-use patterns, also have an influence. It is therefore important to note that no single factor operates in isolation but interact within a broader socio-environmental system.

2.4 Water Demand Forecasting Approaches

Water demand forecasting is crucial in effective planning and management, serving as the foundation for informed decision-making. These can inform capacity expansion, revenue planning, system operation management, and optimization (Billings & Jones, 2008). Various methods exist, but their suitability depends on the forecasting horizon. However, there is no universally accepted timeframe for forecasting. Donkor et al. (2014) show that it can be long-term, medium-term term or short-term. With long-term spanning over two years, medium being less than two years to three months, and less than three months being short-term. Contrasting with Billings & Jones (2008), who show that long-term forecasting horizons are for horizons of a decade or more, medium for years to a decade, and short-term for a year to two, while very short-term is for hours, days, weeks, and months (Billings & Jones, 2008; Donkor, et al., 2014; Rinaudo, 2015).

Forecasting methods can be divided into two primary categories: linear (Exponential smoothing, Autoregressive integrated moving averages), and non-linear approaches (ANNs, Support Vector Machines, Genetic Algorithms, and Expert systems) (De Souza Groppo, et al., 2019). Qualitative methods, relying on expert knowledge or judgment, also exist, but their subjective nature makes them less reliable for long-term biases and limitations in capturing trends (Billings & Jones, 2008; Donkor, et al., 2014).

Time series or extrapolation methods, which include averaging, trend analysis, exponential smoothing, and Autoregressive Integrated Moving Average (ARIMA), utilize historical data and associated error terms. They are practical and straightforward, and require no knowledge of the internal process of the system (Billings & Jones, 2008). However, with these models, there is no account of exogenous variables for external variables like climate and price, demographic, economic, and technology, which significantly influence water demand (Donkor, et al., 2014). Furthermore, ARIMA models, while useful for short-term forecasting, struggle with multi-step predictions and assume linear relationships, which is often not the case in water demand variables (Hao, et al., 2022; Shu, et al., 2024). Similarly, there are multivariate regression models, incorporating multiple dependent variables, however, similar to time-series methods, they often assume linear relationships and may not be fitting for complex, long-term prediction (Billings & Jones, 2008; Donkor, et al., 2014).

Non-parametric methods, particularly machine learning models such as fuzzy ANNs, SVR, and RF, have gained prominence in recent years. They determine complex relationships between variables and generalise the unseen data, requiring fewer assumptions. Their ability to model non-linear relationships with high precision makes them attractive to water planners (Solomatine, et al., 2009; De Souza Groppo, et al., 2019) . It is imperative to acknowledge that forecasting and prediction inherently involve a degree of uncertainty. Consequently, utility companies must engage in robust decision-making processes that prioritize the incorporation of uncertainty factors into their strategic planning and operational frameworks (Srdjevic, et al., 2015; Marchau, et al., 2019).

2.5 Overview of Machine Learning Models

Machine Learning (ML) is increasingly emerging as a pivotal instrument within environmental sciences, particularly in water demand forecasting and predictive modelling. The application of ML techniques enables researchers and practitioners to analyse complex datasets, uncover patterns, and improve the accuracy of water consumption predictions (Janiesch, et al., 2021). The field encompasses diverse algorithms, including regression models, instance-based algorithms, decision trees, Bayesian methods, and ANNs. Among the most frequently used ML methods are decision trees, ANN, Support Vector Machines (SVM), Support Vector Regressors (SVRs), and ensembles. While there can be anomalies, most research demonstrates that ML models yield better results compared to classical techniques (Villarin & Rodriguez-Galiano, 2019; García-Soto, et al., 2024) . ML can be divided into three distinct learning paradigms, Reinforced Learning, Unsupervised learning, and Supervised Learning (Nishant, et al., 2020; Ghobadi & Kang, 2023).

Reinforced Learning Technique

Reinforcement Learning (RL) is a distinctive paradigm in AI where an agent learns optimal behaviours in the environment by trial-and-error. RL does not depend on human-generated data for training (Singh, et al., 2021). Instead, the agent learns from feedback in the form of reward signals based on its actions, enabling it to refine its decision-making over time. The learning approach mimics human learning processes and has proven highly effective in sequential decision-making tasks. In water resources management, reinforcement learning can be used to simplify the process of modelling, thus assisting in locating the optimal solution (Kåge, et al., 2025). Although its application in engineering is still being explored, reinforcement learning is predominantly used in fields such as robotics, economics, and intelligent vehicles. In the context of water resources, it has been explored in wastewater prediction, water quality prediction, and hydrology (Shakya, et al., 2023).

Unsupervised Learning Techniques

Unsupervised learning techniques are utilized to reveal hidden patterns in unlabelled datasets (Usama, et al., 2019). These techniques depend exclusively on the inherent structure and feature characteristics of data (Naeem, et al., 2023). Unsupervised learning proves particularly beneficial for tasks involving clustering and feature extraction. By eliminating the need for labelled data and manual feature engineering, they offer greater flexibility and automation in data analysis (Zamri, et al., 2022). While unsupervised learning excels in handling complex data, it often lacks the precision of supervised methods due to the absence of explicit guidance. Nonetheless, unsupervised learning plays a critical role in exploratory data analysis, especially when handling vast volumes of unlabelled or unprocessed data. Common algorithms in this category include K-means clustering and Principal Component Analysis (PCA) (Zamri, et al., 2022; Papavasileiou, et al., 2025).

These methods are used in different aspects in the context of water resources management. For instance, Ghobadi & Kang (2023) demonstrate that supervised methods are employed in water distribution to identify various customer patterns, determine the spatiotemporal distribution of precipitation, detect leaks in distribution mains, and monitor non-revenue water.

Supervised learning Techniques

Supervised learning algorithms are specifically designed to establish a predictive relationship between input features and the target variable. The type of output determines the model's category, where if the target is categorical, the model performs classification, and if numerical, the model is used for regression. The learning process relies on labelled data, where the algorithm trains a statistical model to predict outcomes for unlabelled instances. Commonly used supervised learning techniques include decision trees, which include RF, SVMs, SVRs, ANNs, CNNs, and LSTMs, among others (Jovel & Greiner, 2021; Ghobadi & Kang, 2023).

Supervised learning algorithms are applied in different contexts in water resources management. Ghobadi & Kang (2023) demonstrate in a review the different research fields for different prediction supervised algorithms, these include and are not limited to streamflow, soil moisture, water quality, leakage detection, and water demand

With the rising crisis in water security, it has become of grave importance to ensure the sustainable use of water to ensure that demand meets supply. The global population is rapidly increasing, given that business continues as usual, water stress will only increase (Zubaidi, et al., 2020). Recent research has delved into regression models, particularly for water consumption and demand prediction, as they are better suited as they learn from historical labelled data to map input features (population, income, temperature, etc) to an exact and expected output (Donkor, et al., 2014; Ghalehkhondabi, et al., 2017). Furthermore, Ahmed et al. (2024) and Ghobadi & Kang (2023) demonstrate in a review the different applications of ML in water resources management. These include groundwater level forecasting, flood management, water quality, water drainage systems, wastewater treatment, water demand and consumption, agricultural water requirements, stream-flow forecasting, and hydropower operation, among others.

2.5.1 Regression Trees

Regression trees (RT) offer an alternative to the limitations of multivariate regression. They operate by subdividing data into smaller data sets to allow for more manageable interactions, simplifying their analysis. They are characterized by their straightforward nature and interpretability, and minimal computational requirements with allowance for graphical representation. Compared to other ML methods, they are often considered white-box, enabling their ease of interpretation of the relations between variables. Several RT techniques exist, for example, the Iterative Dichotomiser 3 (ID3), Classification and Regression Tree (CART), and Random Forest (RF). However, regression trees can be very susceptible to noisy data. RF overcomes this by averaging multiple trees, however, this undermines the quality of the forecast (Liu, 2023) . RF enables the quantification of feature importance, revealing variables with the greatest influence on predicting water demand. However, the increased complexity of RF models can make interpretation more challenging. Most notably, RF has been widely applied in recent water demand forecasting studies, as it combines the results of multiple RT algorithms to predict the target variable more accurately. For instance, Villarin & Rodriguez-Galiano (2019) compared CART and RF models to identify key drivers in Spain, finding that RF performed better, although both models identified similar importance. Kulaczkowski & Lee (2024) successfully use RF for urban water demand in Italy, proving its effectiveness and robustness.

2.5.2 Support Vector Regression Models (SVR)

Support Vector Regression (SVR) applies the principles of Support Vector Machines (SVM) to regression problems. It is distinguished by a data-driven approach that constructs the regression function using a limited subset of the training data, known as support vectors

(Ghalehkhondabi, et al., 2017). Contrary to conventional methods that minimize the overall deviation between predicted and observed values across all data points, SVR is based on the principle of structural risk minimization, aiming to enhance generalization by balancing model complexity and prediction accuracy. (Basak, et al., 2007; Awad & Khanna, 2015).

The fundamental feature of SVR is in the application of kernel functions, such as the radial basis function, which project input data into a higher-dimensional feature space. Allowing it to model complex, non-linear relationships effectively, by fitting an optimal linear hyperplane in the transformed space (Cortes & Vapnik, 1995). The model's complexity is managed through the Vapnik-Chervonenkis (VC) dimension control, which helps mitigate overfitting. Consequently, SVR has demonstrated robust generalization capabilities, especially where input data may be limited.

Several studies highlight the efficacy of SVR in water demand forecasting. Ibrahim et al (2020) observed that SVR performed better than the traditional ARIMA model in short-term water demand in Kuwait. Similarly, Herrera et al. (2010) identified SVR to be the most accurate model when compared against ANN, RF, Projection Pursuit Regression, and Multivariate adaptive regression splines. In a study focused on agricultural water demand, Soebroto et al. (2022) concluded that SVR was superior to both ANN and Multiple linear regression for water demand forecasting for agriculture. However, SVR is not universally superior. Research by Mu et al. (2020) and Hao et al (2022) showed that deep learning models, specifically Long-Short Term Memory (LSTM), attained better accuracy than SVR, particularly when integrated with wavelet decomposition.

2.5.3 Artificial Neural Networks

ANNs are machine learning models modelled after the human brain's neural architecture. They are composed of layers of interconnected processing nodes or artificial neurons. Data enters the network through the input layer and is then successively transferred to one or more hidden layers, where the learning is executed. Ultimately, data is transferred to the output layer, which produces the results. ANNs excel at learning and representing the complex, non-linear relationships inherent in many real-world processes (Ghalehkhondabi, et al., 2017). Their primary advantages include the ability to handle large and discontinuous datasets and their robustness in data-scarce environments (Solomatine, et al., 2009; Niknam, et al., 2014)

The utilization of ANNs in water demand forecasting has been widely validated in the literature. Adamowski & Karapataki (2010) reported that ANNs significantly outperformed multiple linear regression for short-term water demand forecasting. Similarly, Vijai and Sivakumar (2018) compared ANNs with MLR, Exponential Gaussian Process Regression, Least Squares-SVM, Extreme Learning Machines, and Deep Neural Networks. They found that ANNs had the best performance metrics. Behboudian et al (2014) successfully used a feed-forward back-propagated multi-perceptron ANN to predict long-term water demand in Iran, demonstrating its superior performance compared to conventional regression methods.

Despite their strengths, ANNs have limitations. Their performance can be compromised by small datasets (Ghalehkhondabi, et al., 2017; Oyebode & Ighravwe, 2019; Lee & Derrible, 2020). Additionally, Antunes et al. (2018) note that a greater number of features can impact model performance, either increasing or decreasing their performance. Zubaidi et al. (2020) observed better performance of backtracking-search ANN over stand-alone in Gauteng, South Africa. Moreover, Menapace et al. (2021) showed that simpler architectures may struggle with larger forecasting horizons. Additionally, Tiwari & Adamowski (2014), as well as Ferrah et al. (2019), noted limitations related to the smoothing of data.

Hence, advanced deep learning architectures may offer better performance. For instance, Hao et al (2022) and Alikhani & Moeini (2025) used wavelet-coupled models in their studies. Guo et al. (2018) also found that Gated Recurrent Unit Network (GRUN), which has a deeper architecture than traditional ANN, provided more reliable predictions, suggesting that increased network depth may enhance predictive accuracy.

2.6 Model Performance Evaluation Metrics

The performance of predictive models relies heavily on performance metrics, which are crucial in quantifying the variance between the prediction and observed data (Donkor, et al., 2014). These metrics serve as an aid for researchers and practitioners in evaluating, contrasting, enabling rigorous comparison, assessment, and refinement of algorithms and models (Zhang & Garcia, 2015). The selection of evaluation metrics is linked to the nature of the task. Regression tasks often employ Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination(R^2). In contrast, classification tasks typically rely on metrics such as accuracy, precision, recall, and F1, among others, (Miller, et al., 2024; Varoquaux & Colliot, 2023). However, these will not be covered in the scope of this research, only the aforementioned regression metrics will be explored.

Systematically quantifying a model's performance metrics is instrumental for effective optimization (Van Thieu, et al., 2023). This framework is crucial for diagnosing and addressing challenges that may arise during model development, such as overfitting, where a model excels on trained data but falters on unseen data, and underfitting, where a model fails to capture the true structure and trends within the data (Ying, 2019). Moreover, beyond predictive accuracy, measuring uncertainty is crucial for assessing the informational richness of data. The quantification of uncertainty influences the reliability of the information and its applicability in decision-making processes. By effectively quantifying the uncertainty, one gains valuable insight into the error characteristics of the measurement model (Van Thieu, 2024).

Despite their usefulness, performance metrics are not without limitations. A notable challenge is their interdependence with one another. A significant challenge is their interdependence with one another. For instance, Tian et al. (2016) illustrate that MSE can be mathematically decomposed in various ways, revealing its relationship with other statistical measures such as the correlation coefficient and bias. This interconnectedness can lead to redundancy in evaluation and ambiguity in interpretation. Moreover, the lack of consensus or standardized guidelines regarding the optimal number of metrics to employ often results in researchers reporting numerous metrics, potentially obscuring clear performance assessment.

The coefficient of determination (R^2), or explained variance, is favoured for its natural scale, ranging from 0 to 1. An R^2 of 0 indicates no predictive power, while a value close to 1 suggests a robust fit to the data. However, its application in predictive tasks, particularly out-of-sample settings, can be misleading, as a high correlation between actual and predicted values does not equate to high predictive power (Varoquaux & Colliot, 2023).

Moreover, an over-reliance on R^2 can lead to overfitting, as it tends to increase with more parameters regardless of their predictive value. Therefore, it is important to use a combination of metrics to gain a holistic understanding of the performance and error distribution of the model (Miller, et al., 2024; Varoquaux & Colliot, 2023) .

Accordingly, absolute errors of measure, such as the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), are frequently used to evaluate model performance. MAE is advantageous for its simplicity in interpretation, as it represents the size of the prediction errors. Additionally, it is less sensitive to outliers compared to the RMSE because it does not square the error terms, placing less weight on deviations. Consequently, the MAE is consistently less than the RMSE. In contrast, the RMSE demonstrates higher sensitivity to outliers because of the squaring of errors before averaging. While this can be a drawback if outliers are indicative of poor performance, RMSE is often favoured in mathematical applications because it eliminates the need for absolute values by squaring the errors instead. A lower RMSE generally indicates a better model fit, noting that RMSE is scale-dependent, making it more suitable for comparing models on the same dataset or datasets with similar scales. (Hodson, 2022; Chai & Draxler, 2014).

The debate regarding the appropriate error metric often centres on the underlying distribution error. MAE is often considered more suitable for describing uniformly distributed errors, while RMSE is theoretically more appropriate for normally distributed errors. However, Chai & Draxler (2014) argue that RMSE is a suitable metric when errors are Gaussian, but also suggest that MAE can be a robust alternative. Some have argued that RMSE can be ambiguous in its interpretation, leading to preferences for MAE in certain contexts (Willmott & Matsuura, 2005). However, it has been emphasised that each metric may be optimal under specific assumptions and correct application, however, neither RMSE nor MAE alone is sufficient in practice (Hodson, 2022).

The Mean Absolute Percentage Error (MAPE) is another extensively utilized metric, especially in forecasting, appreciated for its scale independence and interpretability as a percentage error. This feature enables comparison across datasets of varying scales. However, it can yield infinite and undefined results when the actual values are zeros or near zero (Kim & Kim, 2016). Additionally, unlike RMSE and MAE, which rely on squared errors, MAPE, being percentage-based, can be scale sensitive, resulting in a more significant penalty for positive errors compared to negative errors (Jierula, et al., 2021).

Several studies have highlighted the importance of specific metrics, identifying RMSE, MAE, MAPE, and R^2 among others as some of the most broadly used (Donkor, et al., 2014; Niknam, et al., 2014). Ultimately, these performance-based metrics are designed to assess how effectively a model achieves the objectives defined by the specific learning tasks. Table 2-1 presents the performance metrics as applied in different literature

Table 2-1: Overview of various studies, models, and performance metrics employed

Authors	Model	Performance Metrics
Pachinn et al. 2019	ANN-WDF	MAE, RMSE
Bheboudian et al. 2014	ANN- MLP- FFP	RMSE, MAPE, MAE
Hao et al. 2022	ANN, SVR, Long-Short Term Memory (LSTM)	RMSE, MAE, MAPE, and R^2
Zanfei et al. 2023	ANN-MLP	MAPE, R^2 and MAE
Adamowski et al, 2012	MLR, ARIMA, ANN, MNLR	R^2, RMSE, RRMSE
Shu et al, 2024	GA-BP, ELM, GPR, SVR	R^2, MAPE, MAE, RMSE
Gracia-Soto et al (2023)	Extreme Gradient Boost, Deep Neural Network, SARIMA, K-Nearest Neighbour, RF	R^2, RMSE, MAE, MAPE
Villarin & Rodrigas 2019	CART, RF	R^2, RMSE
Adamowski & Karapataki 2010	MLR, ANN	R^2, RMSE, AARE, Max ARE
Mouatadid & Adamowski 2017	ANN, SVR, ELM, MLR	R^2, RMSE

Ideally, the evaluation should reflect how the model would perform when applied in a real-world context. Therefore, it is essential to link model performance to the practicality of its outputs and its ability to produce plausible future scenarios (Makridakis, et al., 2018). Rathnayaka et al. (2017) show that water systems should be designed to meet maximum demand, highlighting the importance of capturing key features such as seasonality and peak demands in models (House-Peters & Chang, 2011).

Chapter 3: Methodology

2.1 Introduction

This chapter presents the research methodology used to examine the factors influencing water demand and forecast future demand using a machine learning approach, specifically Artificial Neural Network (ANN), Support Vector Regression (SVR), and Multiple Linear Regression (MLR). The methodology is grounded in statistical analysis complemented by a machine learning framework utilizing time series data. The dependent variable (water consumption) is regressed against a series of independent variables, including Precipitation, Population, and Maximum and Minimum temperatures.

3.2 Study Area

The study focuses on Ha-Foso, located in the Berea district of Lesotho at geographic coordinates 29.2069º S, 27.8780º E. Figure 3-1 depicts the WASCO service area in Ha-Foso Berea, Lesotho.

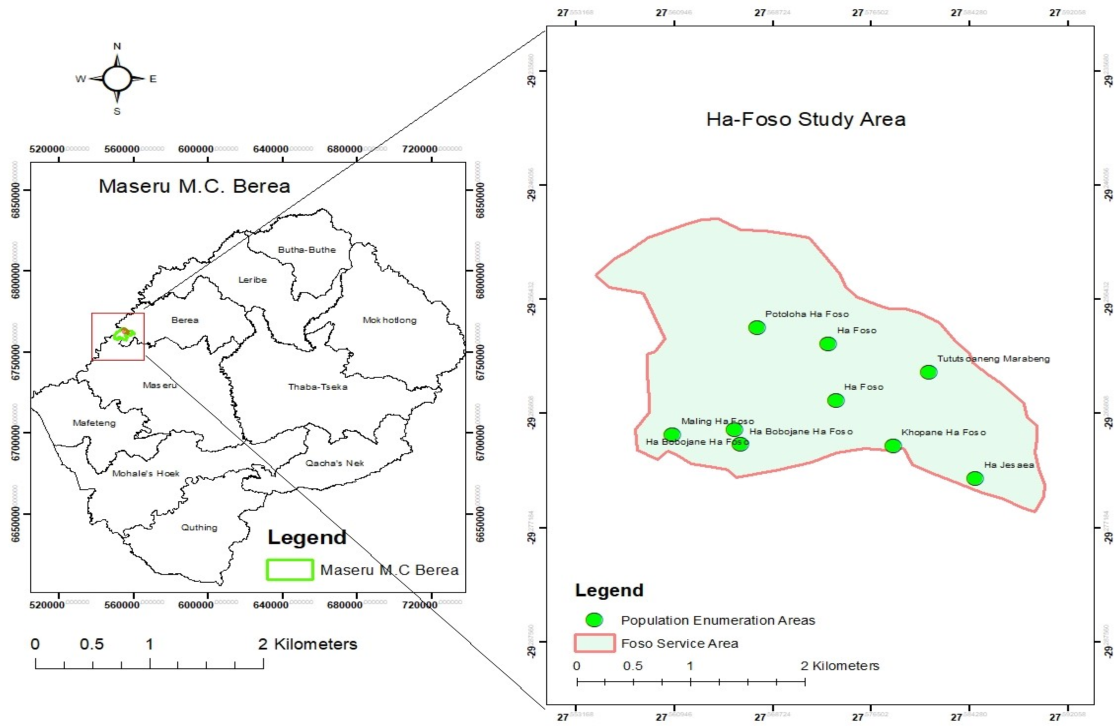

Figure 3-1: The location of Ha- Foso WASCO service area within the Berea district, Lesotho

Ha-Foso is an emerging peri-urban area situated approximately 15 KM from the Maseru district. This region falls within the lowlands agroecological zone. It is associated with the middle Mohokare sub-catchment, characterized by a temperate climate, that experience an average annual precipitation of approximately 700 mm between October and April.

Historically, Ha-Foso was predominantly an agricultural region, characterized by subsistence farming practices. However, in recent decades, it has undergone a significant transformation into a residential area due to factors such as land scarcity within the main urban centres. This rapid and largely unplanned urbanization has resulted in a substantial increase in population density. According to the 2016 census conducted by the Bureau of Statistics, the recorded population of Ha-Foso was recorded at 2530 (BOS, 2019). This accounts for communities of Potoloha- Ha-Foso, Khopane, Maling, Ha-Jesea, Ha Bobojane, and Tututsoaneng Marabeng, which are WASCO-serviced areas.

Ha-Foso falls under the administrative jurisdiction of the Maseru M.C. Berea Municipal Council, which covers an area of 4014 ha with a topography ranging from 1473 to 1740 meters above sea level. The council area exhibits a diverse land use/land cover (LULC) pattern, comprising built-up areas (1475.10 ha), cropland (1665.77 ha), trees (81.928 ha), water bodies (45.706 ha), wetlands (27.965 ha), shrubland (138.869 ha), grassland (465.218 ha), bare surfaces (17.209 ha), and gullies (96.556 ha) (FAO, 2023).

The ongoing residential shift, coupled with increasing population growth, GDP, higher income, gender, and education, necessitates a thorough investigation into water demand dynamics and the implementation of accurate predictive models to inform resource management and planning strategies in the region. Inadequate planning in this context may lead to challenges such as intermittent water supply due to demand exceeding supply, among other related problems. Therefore, Ha-Foso emerges as a pertinent study site to examine the dynamics of water demand in a rapidly urbanizing area within the Lesotho context.

3.3 Data Sources and Attributes

The datasets utilized in this study were compiled in tabulated Excel spreadsheets and the shapefile format. Monthly water consumption data spanning 2017-2024 were obtained from the Water and Sewerage Company (WASCO) for 2061 metered houses in the Ha-Foso Area. Climate data, including monthly precipitation and maximum and minimum temperatures from 2012-2022, was sourced from Mejametalana station, Lesotho Meteorological Services

(LMS). Population data for Ha-Foso was extracted from the 2016 census enumeration areas provided by the Bureau of Statistics. It was spatially analyzed using ArcGIS and a shapefile to determine the population of the study area. The Bureau of Statistics population projections report of 2016-2036 reported an annual population growth rate of 0.65%.

Due to constraints in data availability, only precipitation and temperature were included as climatic variables, while other potentially relevant climatic factors, such as wind speed, sunshine hours, and humidity, were excluded from the analysis. Moreover, population was used as the principal demographic factor, therefore, this study serves as a baseline, from which its findings and gaps can inform future studies. Data from Mejametalana station, the closest station to the study area, were used for temperature and precipitation data. A detailed summary of data types, sources, temporal scales, and specific datasets utilized in this study is outlined in Table 3-1.

Table 3-1: Summary of data attributes

Data Type	Source	Period	Scale	Variable used
Historical weather records	Lesotho Meteorological Services	2012-2022	Monthly	Max Temperature
				Min Temperature
				Precipitation
Historical water consumption	Water and Sewage Company	2017-2024	Monthly	Water consumption (KL/Months)
Historical Population	Bureau of Statistics	2016	Annual	Population
Population Projections	Bureau of Statistics	2016-2036	Annual	Population Growth Rate

3.4 Data Cleaning and Pre-processing

Missing data was addressed using interpolation and removal, depending on the extent of missing data. Data with implausible values were identified and treated by removal to ensure data quality (Billings & Jones, 2008; Ridzuan & Wan Zainon, 2019). To guarantee that the input variables had a uniform influence on the models, the data was normalized using the Min-Max method. This technique compresses values into a standard range. The consolidated data was divided into training and testing sets at a ratio of 80:20, to evaluate their ability to generalize unseen data, where training (2018-01-18 to 2023-02-2020) and testing data (2023-03-20 to 2024-12-20).

The water consumption dataset had missing data. While imputation techniques are commonly employed, they introduce a degree of uncertainty based on the chosen estimation method. In this study, a decision was made to retain the observed data without imputation to prioritise learning from the actual consumption patterns, even with the gaps. This method assumes that the data segments represent the inherent water trends. However, it is acknowledged that this approach may limit the ability of the model to reproduce long-term trends and may introduce bias if the missing data are randomly distributed. Alternative strategies such as imputing with time-series methods or using models robust to missing data were considered but deemed less suitable given the percentage of missing and the potential for introducing further distortions.

R programming language version 4.3.3 was utilised to perform data cleaning and pre-processing. R is a widely used open-source programming language, known for statistical computing and graphical presentation. It has a vast collection of packages that extend the function of the language, making it ideal for statistical analysis and tool development (Dizon, et al., 2023)

3.5 Model Approach

The influence of population, maximum and minimum temperatures, and precipitation on water demand was investigated using Multiple linear regression. Following the findings, the results guided the input for Multiple Linear Regression, Support Vector Regression, and Artificial Neural Networks. A comparative analysis was done, and the best-performing model for water demand prediction was selected as subsequently described

3.5.1 Analysis of Determinants of Water Demand

To identify the most influential variables for forecasting, an MLR analysis was conducted (Aho, et al., 2016). MLR analysis was used to examine the association between a set of independent variables (meteorological elements and population growth) and the dependent variable (water consumption) during the model development phase. MLR is a parametric statistical approach that simultaneously utilizes two or more independent variables to predict a single continuous dependent variable. This approach enables the quantification of associations and causal links between the predictors and the outcome of the variable (Bangdiwala, 2018). The focus on such models is the interpretation of the slope (regression coefficients or estimates), where if it equals zero, it indicates no linear relationship between the dependent and independent variables. The intercept shows the expected value when all dependent variables are zero. Therefore, in this study, the estimates were used to assess the influence of independent variables on water consumption, and p-values were utilized to assess statistical significance, where a threshold of less than 0.001 indicated a robust model, as well as significant variables (Brandner, 2016).

The equation is generally expressed as follows:

$$y = \beta_{\circ} + \beta_1 X_1 + \beta_2 X_3 + \cdots \beta_n X_n + \varepsilon \qquad \textit{Equation 1}$$

Where:

y is the dependent variable,X_i the independent variable, β_i the parameter $\beta_{\circ}$ the intercept and ε the error

The reliability of the MLR model was validated by testing statistical assumptions, which include normality, homoscedasticity, multicollinearity, and autocorrelation (Starbuck, 2023). To address assumptions, violations, and confirm findings, the analysis was supplemented by semi-logarithmic and robust regression models. The variables that consistently demonstrated statistical significance across these models were selected as inputs for the predictive forecasting models.

Variance inflation factors (VIFs) were examined to address multicollinearity. A VIF value of 1 indicates no multicollinearity, while values ranging from 1 to 5 indicate modest correlation, and values ranging from 5 to 10 suggest high correlation (Shrestha, 2020). VIF technique is represented mathematically by the equation below:

$$VIF = \frac{1}{1-R^2} = \frac{1}{Tolerance} \qquad \textit{Equation 2}$$

Homoscedasticity was assessed using the Breusch-pagan test in the R environment. Where a p-value < 0.05 indicates that the assumption of homoscedasticity is not violated. Normality was assessed using the Shapiro-Wilk test, where a p-value < 0.05 suggests that residuals do not significantly deviate from the normal distribution (Starbuck, 2023; Flatt & Jacobs, 2019).

3.5.2 Predictive Model Development

This study adopted a medium-term forecasting horizon, projecting water demand over a 2-year (2025-2026) period. Medium-term horizons can be used for tactical planning and seasonal allocations.

3.5.2.1 Multiple Linear Regression

The standard LR model was constructed using the *lm* () function in R. It serves as the baseline parametric model, which assumes the relationship between the predictors and water consumption is linear. Feature Engineering involved deriving temporal variables such as months and years, and incorporating seasonal patterns to enhance the model's ability to capture periodic fluctuations in water demand.

3.5.2.2 Support Vector Regression

SVR is an advanced machine learning algorithm that performs regression by operating within a high-dimensional feature space. This is a transformed version of data where complex patterns and relationships become easier to detect. Therefore, by working in this feature space, SVR can model both simple and complex relationships more effectively, leading to higher predictive accuracy. It aims to reduce the generalized error by using structural risk minimization (Basak, et al., 2007; Awad & Khanna, 2015). Through kernel functions, SVR captures non-linear relationships by transforming the original data and identifying linear patterns in the new space (Cortes & Vapnik, 1995) . The model's sparse resolution relies on selecting training data points, known as support vectors, to define the regression function. Its complexity is managed through Vapnik-Chervonenkis control, which helps with overfitting.

The model was developed with the *e1071* package in the R environment. The model's key configurations included training using the kernel Radial Basis Function to capture non-linear relationships. To optimize performance and prevent overfitting, a 5-fold cross-validation with grid search was employed to identify the best values for cost (50), gamma (0.001), and epsilon (0.1) parameters. Temporal features, including year, month, and lagged consumption values, were engineered to capture trends and seasonal dependencies.

The SVR model's nonlinear form $f(w,b)$ is as follows:

$$f(w,b) = w \cdot \emptyset(x) + b \qquad \textit{Equation 3}$$

Where w and b are parameter vectors of the functions, x denotes the input vector and $\emptyset(x)$ expresses the nonlinear function.

3.5.2.3 Artificial Neural Networks

ANNs are algorithms designed to simulate the learning process of the human brain. Known for their ability to capture non-linear relationships and to handle large data sets. A feedforward multilayer perceptron ANN was developed using the *neuralnet* package in R (Razak, et al., 2023). The multilayer perceptron (MLP) layered feedforward network is the most widely used and has been used successfully in water demand studies, as it also boasts simplicity. As such, it was utilised in this study (Ghalehkhondabi, et al., 2017; Pacchin, et al., 2019; Lee & Derrible, 2020).

The network was configured with an input layer, two hidden layers (10 and 5 neurons), and an output layer. The default hyperbolic tangent activation function was used, and the network weights were trained and iteratively adjusted using the backpropagation algorithm, with a learning threshold of 0.01. The study employed 5-fold cross-validation to identify the best-performing model structure (Liu, 2020; Hao, et al., 2022). Figure 3-2 below is the structure of ANN

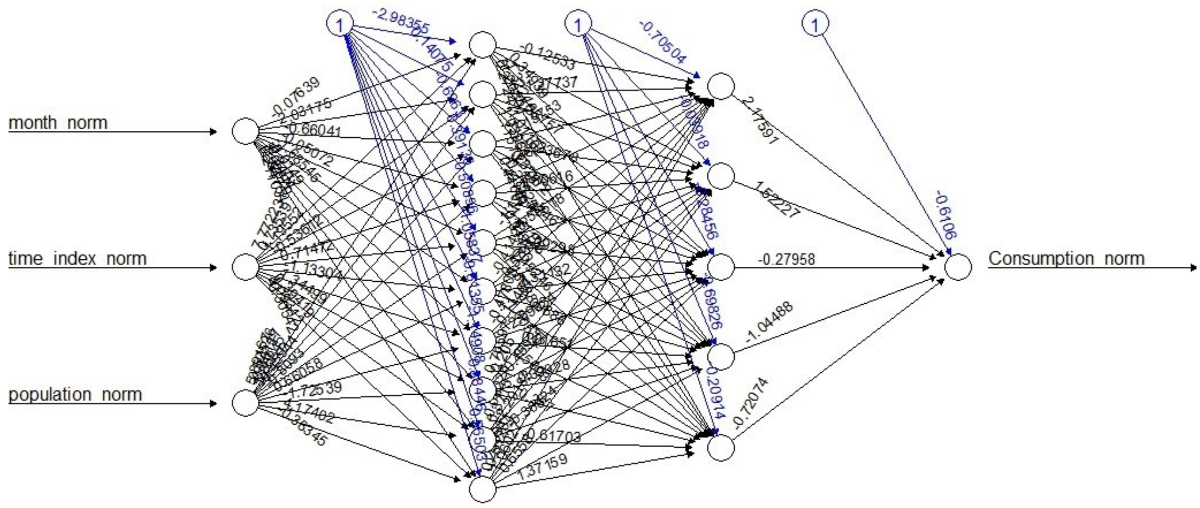

Figure 3-2: Graphical depiction of the ANN structure in R

3.5.3 Model Performance Evaluation

Model performance assessment was executed through a multi-metric validation framework employing the coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean

Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) to ensure a comprehensive assessment of model efficacy to select the best performing model (Billings & Jones, 2008; Behboudian, et al., 2014; Iqelan, 2017; Liu, 2020). The metrics were chosen on the basis of their common use and simple interpretability. For Error metrics, lower values indicate superior performance, and an R^2 closer to one is the best in the training and testing phase. Outlined below are the Metrics employed in model comparison:

The coefficient of determination (R^2) was used to measure the proportion of explained variance. In contrast. Adjusted R^2 incorporates the number of explanatory variables in the model, hence, it is typically lower than R^2, offering a more balanced view of model performance. Model selection often prioritizes statistical significance (Brandner, 2016).

Range between 0 to 1, with values closer to 1 representing a greater proportion of variance explained by the model. Values near 0 indicate that the model does not effectively capture the variability of observed data, resulting in unreliable predictions. It is given by:

$$R^2 = \frac{n(\sum y_{obs}y_{sim}) - (\sum yobs)(\sum ysim)}{\sqrt{[n\sum yobs^2 - (\sum yobs)^2][n\sum ysim^2 - (\sum ysim)^2]}} \qquad \textit{Equation 4}$$

Where $ysim$ denotes the model's predicted value, and $yobs$ the actual value, and n the sample size. Table 3-2 presents a threshold of R2 values

Table 3-2:R^2 threshold values

Performance Rating	**R^2**
Very strong relationship	>0.8
Strong relationship	0.6-0.79
Moderate relationship	0.4-.59
Weak relationship	0.2-0.39
Very weak relationship	<0.19

Source: (Shilehwa, et al., 2019)

MAE was also used for model validation, offering a measure of the average magnitude of errors in a set of forecasts. MAE puts less weight on the rare large deviations, and it is the absolute value of the variance between the forecasted and the observed values. It quantifies the average error magnitude of the prediction, and is calculated as:

$$MAE = \frac{1}{n}\sum\left|y_{sim-y_{obs}}\right| \qquad \textit{Equation 5}$$

RMSE was used to assess simulation errors across different models. It describes the concentration of data around the line of best fit, lower values indicate lower residual. Unlike MAE, RMSE places higher emphasis on larger errors, with lower RMSE values indicating better concentration of data points around the fitted line and, consequently, lower residual errors. It is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_{sim-y_{obs}})^2} \qquad \textit{Equation 6}$$

MAPE is a metric often presented in percentage terms, it expresses the percentage error concerning the actual values by dividing each measured error by the actual value of its corresponding observation. It is given by:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{obs-\hat{Q}}\right|}{y_{obs}} \times 100\% \qquad \textit{Equation 7}$$

Where n represents the total number of months, i the sequence number of each month, Q_i is the observed water consumption. By applying these metrics, the study ensures a comprehensive assessment of the predictive capabilities of the developed water demand forecasting models, allowing a well-informed selection of the most accurate and reliable model for the specific context of this research. Figure 3-3 denotes the workflow of the study

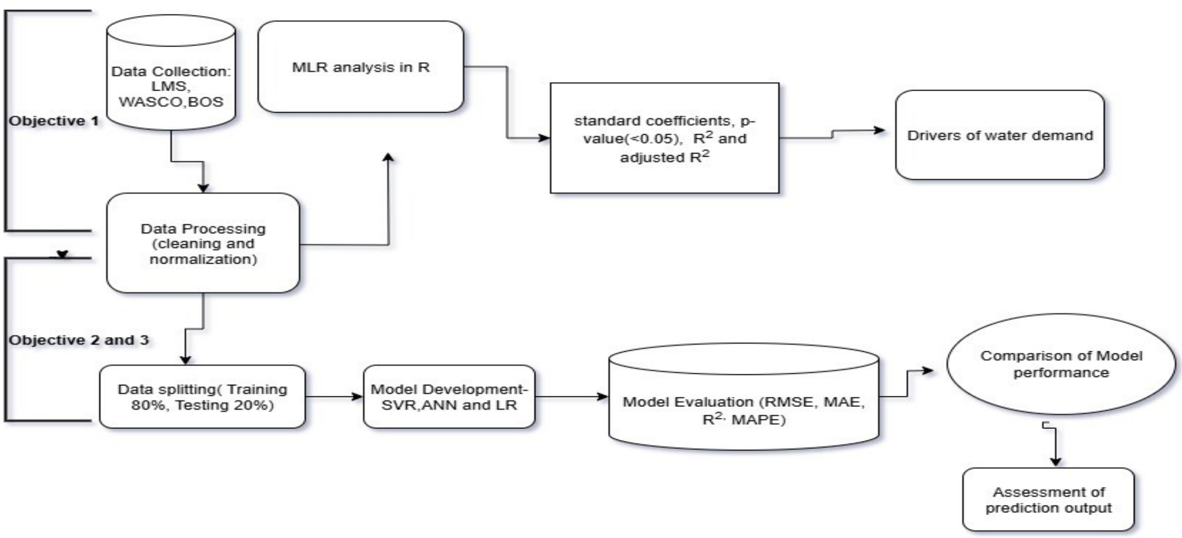

Figure 3-3:Schematic diagram denoting the study workflow

Chapter 4: Results

3.1 Introduction

The study assessed the influence of precipitation, maximum and minimum temperature, and population on water demand. Then, a model performance comparative assessment of MLR, SVR, and ANN was conducted. Ultimately, the outputs of the models were assessed for practical implications on water utility companies. Therefore, the findings of the statistical conducted in the investigation of the factors influencing consumption and the performance of the models chosen for the study.

4.2 Descriptive Statistics and Data Exploration

An initial distribution of each variable, including measures of tendency, Standard deviation, and skewness, is described in Table 4-1. The statistics show precipitation to be highly skewed, this is because some days or months experience no precipitation, while those that do experience very high precipitation. For this reason, normalization was done through log transformation. Concerning Maximum and minimum temperatures, they were not highly skewed and hence no transformation was done. Subsequently, an MLR analysis was conducted to examine the impact of precipitation, population, and maximum and minimum temperatures on water demand.

Table 4-1: Summary of descriptive statistics

Variable	Dataset mean value	Median	Standard Deviation	Skew
Precipitation (mm)	6.22	0.00	15.9	3.24
Max-Temperature(^{0}C)	23.66	23.97	5.13	-0.31
Min-Temperature(^{0}C)	9.04	10.00	5.8	-0.4

Figure 4-1 depicts a visualization of data. Further showing how the data was distributed throughout the years in the different months. For rainfall, it shows

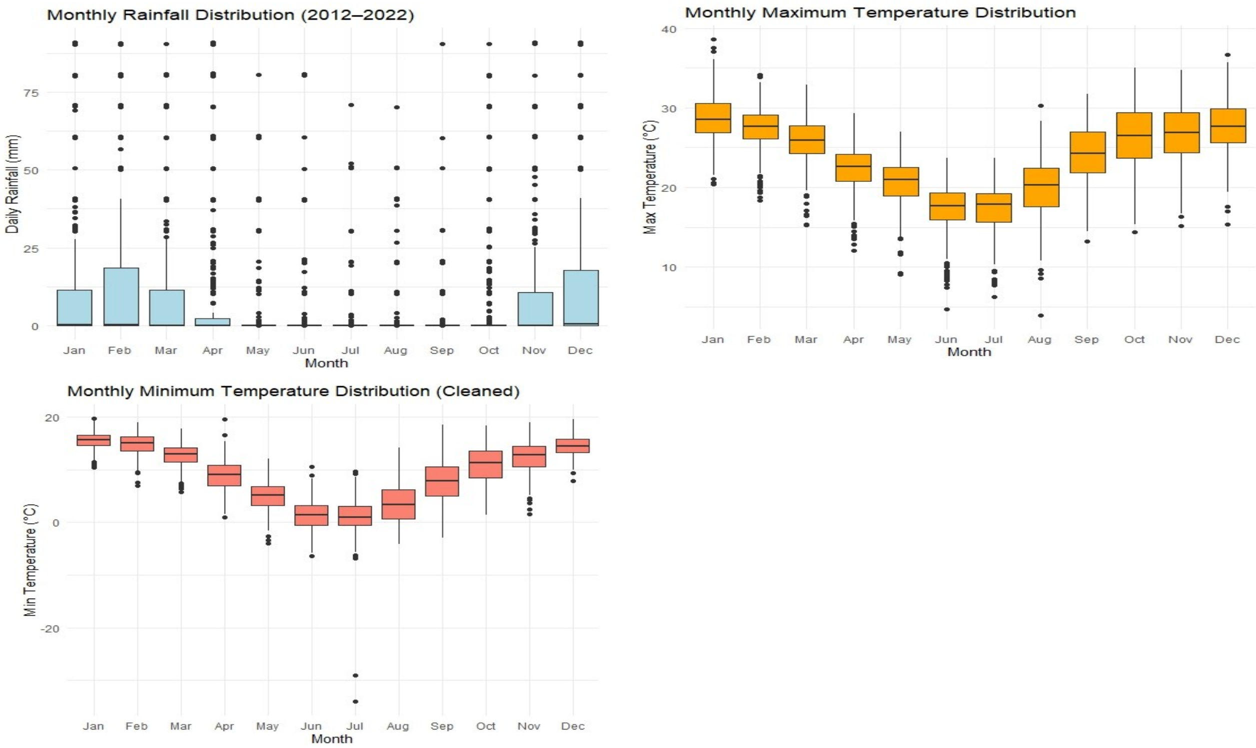

Figure 4-1: Visualization of Climate data

4.3 Analysis of Factors Influencing Water Demand

The preliminary results of the multiple linear regression analysis are presented in Table 4-2. The initial MLR model explained 70.4% of the variance in water consumption (R^2 of 0.704, adjusted R^2 of 0.666), and the model was generally statistically significant ($p< 0.001$). Among the independent variables, population exhibited a statistically significant positive influence on water consumption ($p< 0.001$, standard error of 380.0), indicating that an approximate increase of 3175.9 KL of water consumption accompanies every 1-unit increase in population.

Table 4-2: Multiple linear regression results of influence on water consumption

Variable	Estimate (KL)	Std error	P-Value
Intercept	6340.3	367.6	<0.001
Precipitation(mm)	198.3	427.6	0.646
Max-Temperature(^{0}C)	701.3	705.7	0.328
Min Temperature(^{0}C)	-872.0	752.6	0.255
Population	3175.9	380.0	<0.001

R^2	0.704
Adjusted R^2	0.6661
P-value	<0.001

Conversely, precipitation (p= 0.646), maximum temperature (p=0.328), and minimum temperature (p= 0.255) did not demonstrate statistically significant relationships with water demand in this model.

The relationships are further illustrated in the correlation analysis presented in Figure 4-2. The scatter plot of population versus water consumption showed a statistically significant positive association, while precipitation displayed a near-horizontal trend, suggesting negligible correlation. Similarly, maximum and minimum temperatures exhibited weak or no noticeable linear relationships with demand.

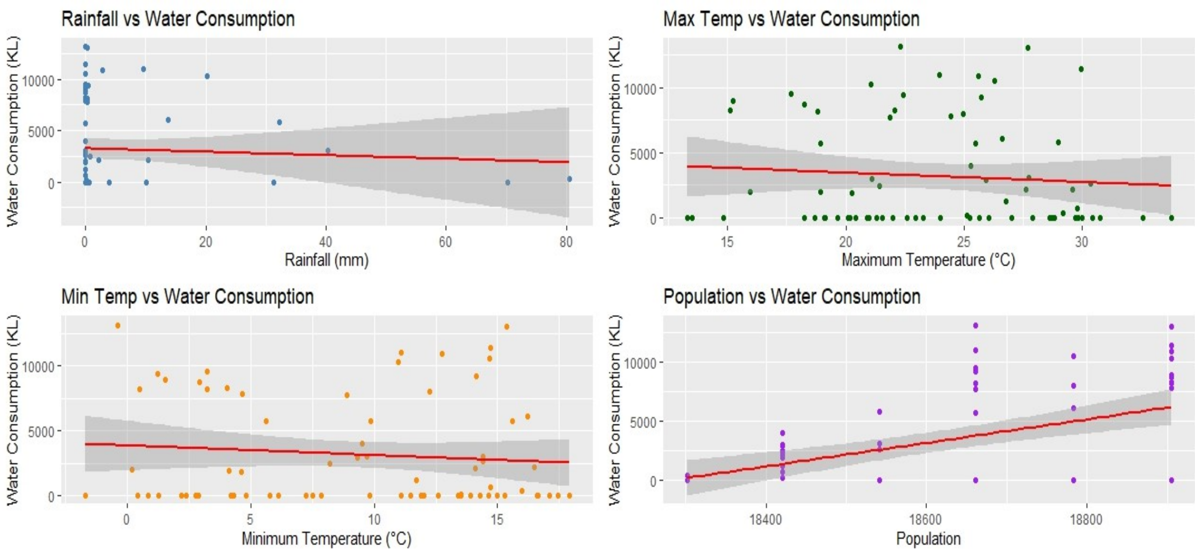

Figure 4-2: Correlation analysis

Furthermore, diagnostic tests were run to ensure that assumptions were not violated. The results showed that the assumptions of normality and absence of multicollinearity were not violated, however, homoscedasticity and absence of autocorrelation were found to be violated. Table 4-3 below outlines a summary of the diagnostic results of the simple multiple linear regression model.

Table 4-3: Multiple Linear Regression Diagnostic Results

Assumption	Test	p-value	
Normality	Shapiro-Wilk test	0.26	W=0.963
Homoscedasticity	Breusch-Pagan test	0.080	BP= 8.34
Auto-Correlation	Durbin-Watson Test	0.004	DW= 1.22
Multicollinearity	VIF technique	All<5 Population= 1.04 Max=3.58 Min=4.08 Precipitation=1.31	

To address violations of assumptions, a semi-logarithmic model was adopted, where water consumption was log-transformed. This transformation aimed to stabilize the variance and linearize the relationships between the independent and dependent variables. The summary statistics are summarised in Table 4-4.

Table 4-4: Summary Statistics for the Semi-log Model

Variable	Estimate	Std Error	P-Value
Intercept	8.4570	0.146	<1.82e-06
Precipitation(mm)	0.1592	0.1217	0.200
Max temperature(^{0}C)	0.1592	0.2006	0.434
Min Temperature(^{0}C)	-0.3135	0.2142	0.153
Population	0.7075	0.1081	<0.001
R^2	0.6154		
Adjusted R^2	0.5658		
P-value	<0.001		

The Semi-log result demonstrated a change in the explained variance, with an R^2 of 0.6154. The model remained statistically significant ($p<0.001$). Consistent with the MLR model, population emerged as the most statistically significant predictor of water demand ($p< 0.001$).

Similar to the MLR model, precipitation p =0.200), maximum temperature (p=0.434), and minimum temperature p=0.153) remained statistically insignificant in the semi-log model.

To ensure the validity of the regression analysis, diagnostic tests were conducted to assess whether fundamental assumptions of multiple linear regression were met. Table 4-5 summarises the tests employed and their corresponding p-values for the Semi-log model regression. Tests were done on each assumption, indicating that the assumption of homoscedasticity (p=0.424), the assumption of significant autocorrelation (p=0.10), and the absence of high multicollinearity (VIF < 5) were satisfied. However, the Normality assumption was not satisfied, with a p= 0.0007.

Table 4-5: Semi-log Model Diagnostic Results

Assumption	Tests	p-value	
Normality	Shapiro-Wilk test	0.0007	0.874
Homoscedasticity	Breusch-Pagan test	0.424	3.87
Auto-Correlation	Durbin-Watson Test	0.10	1.55
Multicollinearity	VIF technique	<5 Population= 1.04 Max=3.58 Min=4.08 Precipitation=1.31	

Despite the application of semi-log transformation, diagnostic tests revealed that key assumptions were still violated, particularly the assumption of normality. Moreover, influential outliers were identified. To address these challenges and to attain more reliable estimates, a robust regression model using M-estimators was employed. Table 4-6 below presents the results of the robust regression model.

Table 4-6: Summary of Robust regression Model results

Term	Estimate	Standard Error	t-value
Intercept	8.5220	0.0787	108.30
Precipitation(mm)	0.0878	0.0709	0.2381
Max temperature(^{0}C)	0.0047	0.1511	0.5815

Min Temperature(⁰C)	-0.1736	0.1611	-1.0772
Population	0.6370	0.0814	7.8295

The robust regression results, similar to the simple Linear regression model and the Semi-Log MLR model, also demonstrate population as the most significant predictor with a t-value of 7.8295, which is above 2. While all other predictors have t-values <2, making them insignificant.

4.4 Predictive Model Performance

4.4.1 Linear regression Model

The performance of the Linear Regression model in predicting water demand is presented in Table 4-7. During the training phase, LR recorded an RMSE of 1700.48 KL, MAE 1366.94 KL, MAPE 62.02%, and R^2 of 0.78. However, during the testing phase, the error metrics increased, notably, RMSE increased to 1727.07 KL, MAE to 1454.87 KL, and MAPE increased to 13.74 %, while R^2 declined to 0.65.

Table 4-7: Linear Regression model performance results

	RMSE(KL)	MAE(KL)	MAPE %	R^2
Training	1700.48	1366.94	62.02	0.78
Testing	1727.07	1454.87	13.74	0.65

Following the presentation of performance metrics, Figure 4-3 depicts the LR model's comparison between actual and predicted water consumption values over time. The plot generally depicts a temporal alignment between the actual (blue line) and predicted (red line) values, particularly in capturing the trend, demonstrating a general upward trajectory. In the initial years, the graphs show overestimations and then continue to underestimate from 2019 to mid-2022, then begin to overestimate till 2025.

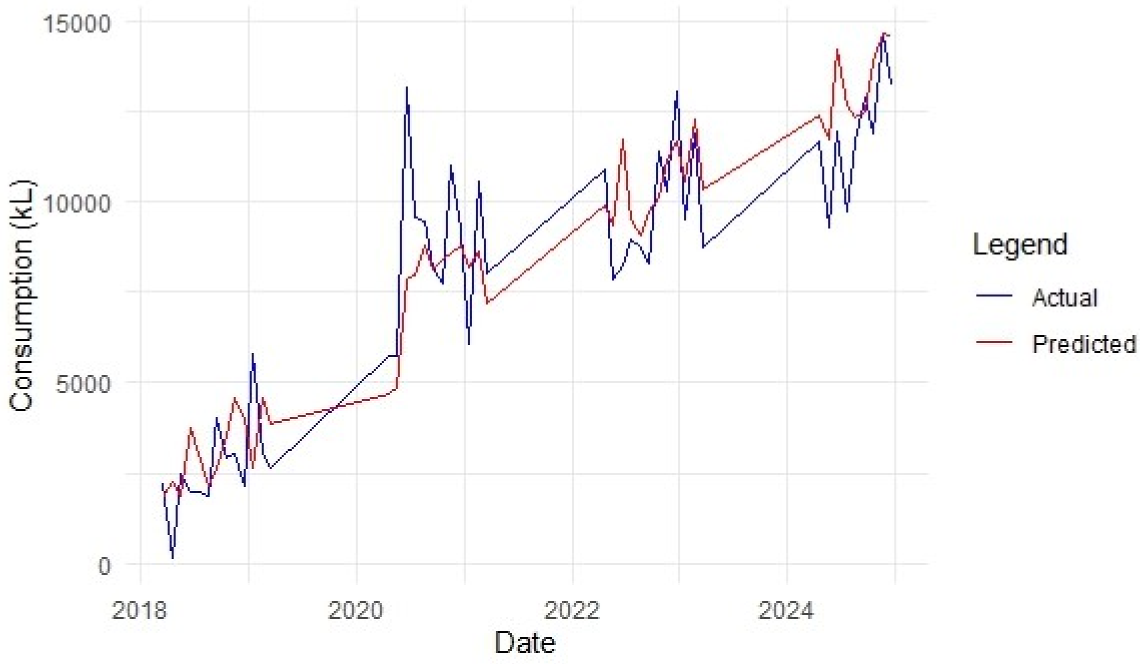

Figure 4-3: Actual vs Predicted Plot for Linear Regression Model

4.4.2 Support Vector Regression

SVR model performance metrics are shown in Table 4-8. In the training phase, SVR recorded RMSE of 1009.98 KL, MAE of 717.39 KL, MAPE of 11.08% and R^2 of 0.88. During the testing phase, the performance increased significantly, RMSE dropped to 786.98 KL, MAE to 690.75 KL, and MAPE to 6.03 %. However, the R^2 value slightly dropped to 0.86.

Table 4-8: SVR performance evaluation results

	RMSE(KL)	MAE(KL)	MAPE%	R^2
Training	1009.98	717.39	11.08	0.88
Testing	786.98	690.75	6.03	0.86

Figure 4-4 below illustrates the comparison between actual and predicted water consumption using the SVR model. The model captures the increasing trend of consumption, demonstrating an increasing trend. The model overestimates in the initial years from 2018 to mid-2020, it is picking up on the trend and the sharp peaks, the predicted line closely tracking the observed lines, however, with some overestimations and underestimations. For instance, between mid-2021 and mid-2022, and mid-2023.

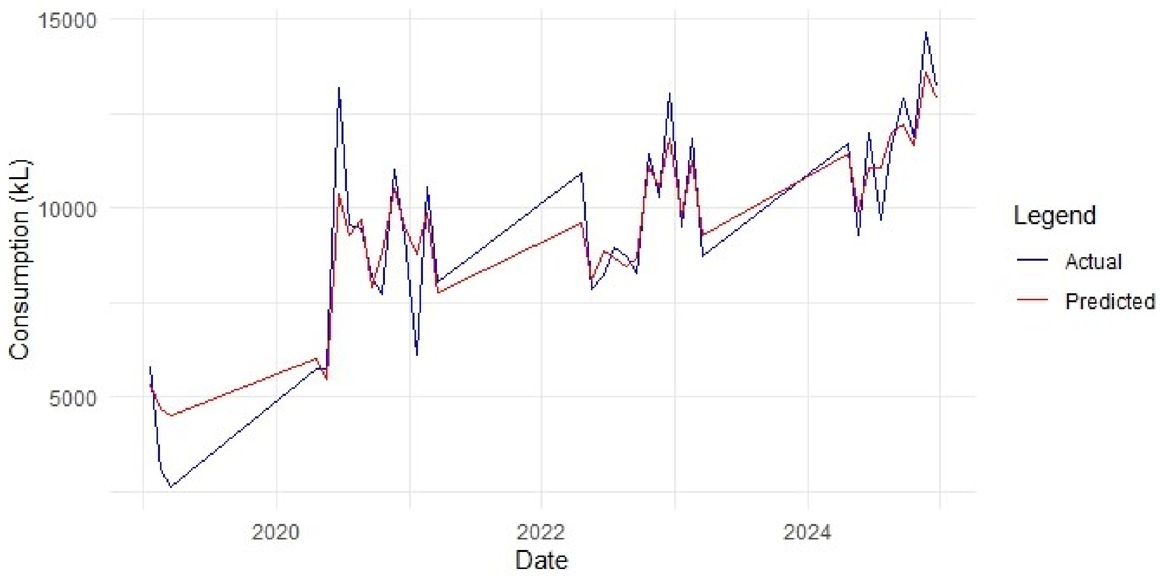

Figure 4-4: Actual vs Predicted Water Consumption plot for the SVR model

4.4.3 Artificial Neural Networks

Table 4-9 presents the findings of the ANN model. In the training phase, the model achieved an RMSE of 1576.01 KL, MAE of 1233.87 KL, MAPE of 41.92%, and R^2 of 0.83. In the testing phase, performance deteriorated with increasing RMSE of 1467.73 KL, MAE of 1313.81 KL, however, it decreased to 11.19%. Similarly, the R^2 decreased to 0.31.

Table 4-9: ANN performance Evaluation results

	RMSE(KL)	MAE(KL)	MAPE%	R^2
Training	1576.01	1233.87	41.92	0.83
Testing	1467.73	1313.81	11.19	0.31

Figure 4-5 presents the comparison between the actual and predicted water consumption over time as predicted by the ANN model. While the model captures the general upward trajectory of the water consumption, the predicted values appear smoother than the actual values. The model follows the broader consumption pattern, however, it underrepresents the fluctuation and peaks demonstrated by the actual values.

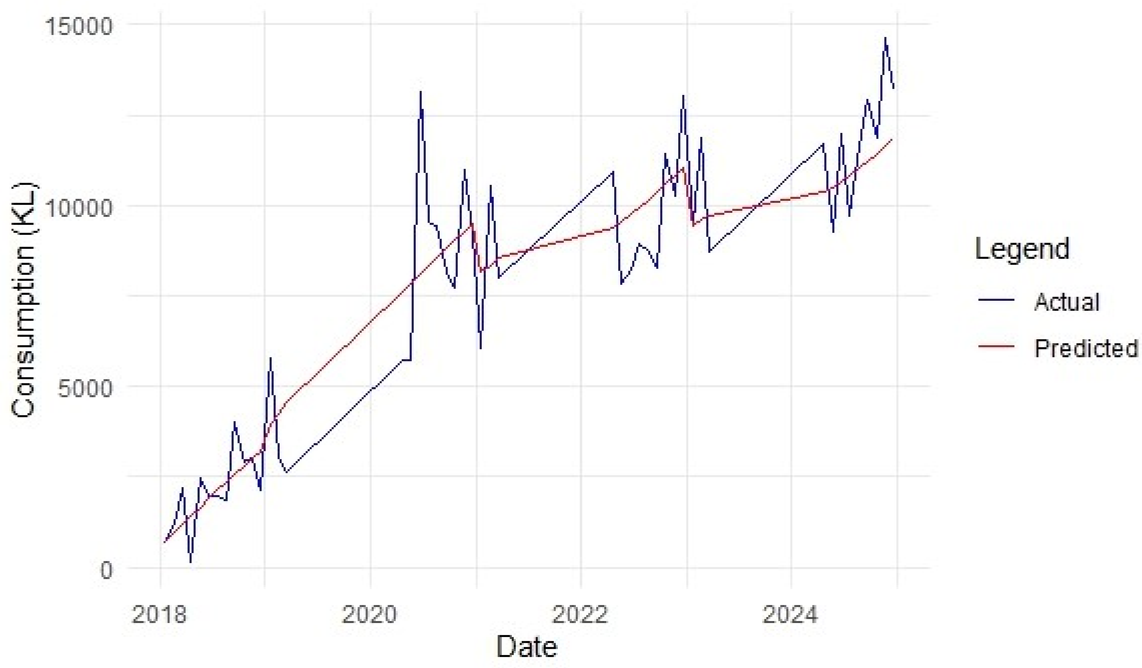

Figure 4-5: Actual vs predicted demand for the ANN model

Table 4-10 provides a consolidated comparison of the performance metrics of LR, ANN, and SVR across both training and testing phases. Results show that Multiple linear regression lags in performance with the highest RMSE, MAE, and MAPE, as well as the lowest R^2 in the training phase. The bolded metrics demonstrate the best performance across the models for all phases, respectively.

Table 4-10: Model performance Metrics summary of results

	RMSE (KL)	MAE(KL)	MAPE%	R^2
		TRAINING		
LR	1700.48	1366.94	62.02	0.78
SVR	**1009.98**	**717.39**	**11.08**	**0.88**
ANN	1576.01	1233.87	42.92	0.83
		TESTING		
LR	1727.07	1454.87	13.74	0.65
SVR	**786.98**	**690.75**	**6.03**	**0.86**
ANN	1467.73	1313.81	11.19	0.31

The results indicate that during the training phase and testing phase, SVR achieved the lowest error metrics (RMSE, MAPE, MAE) and the highest R^2. ANN came second with the least error metrics in the training and testing phase, however, R^2 dropped significantly from 0.83 to 0.31. The LR model exhibited the lowest performance with the highest error metrics in the training and testing phases.

Figure *4-6* illustrates the comparative performance of LR, ANN, and SVR predicted water demand against the observed water demand. The observed data (dashed blue) shows a fluctuating pattern as seen in reality, oscillations caused by behaviour and seasons, and climatic variations. The ANN model (Black) demonstrated the least estimates, moreover showing the least seasonality. The SVR model(green) demonstrated more intermediary estimates, with values within the observed water consumption range. The LR model exhibits mild variation in the observed data, however overpredicts, with values high above the observed data range.

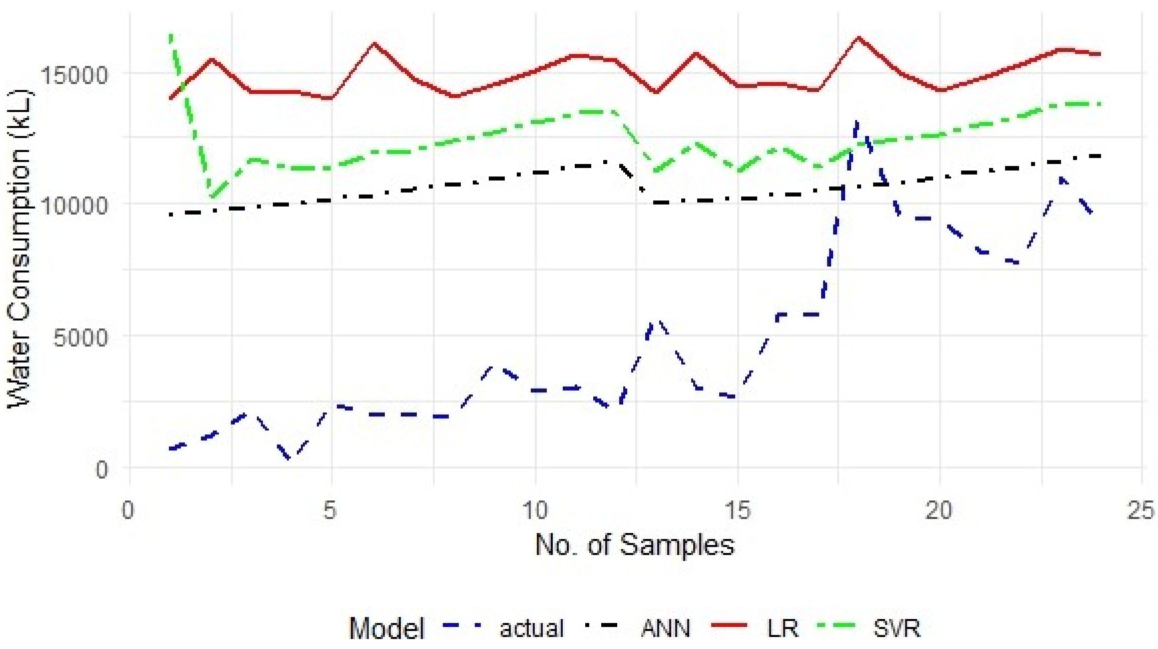

Figure 4-6: Actual vs Predicted for all models

Figure 4-7 visually compares RMSE and MAE. The SVR model demonstrated the least RMSE and MAPE. The trend showed a decrease from the training phase to the testing phase for the SVR and the ANN. The MLR, however, demonstrates an increasing trend from training to testing, with the highest RMSE and MAE values.

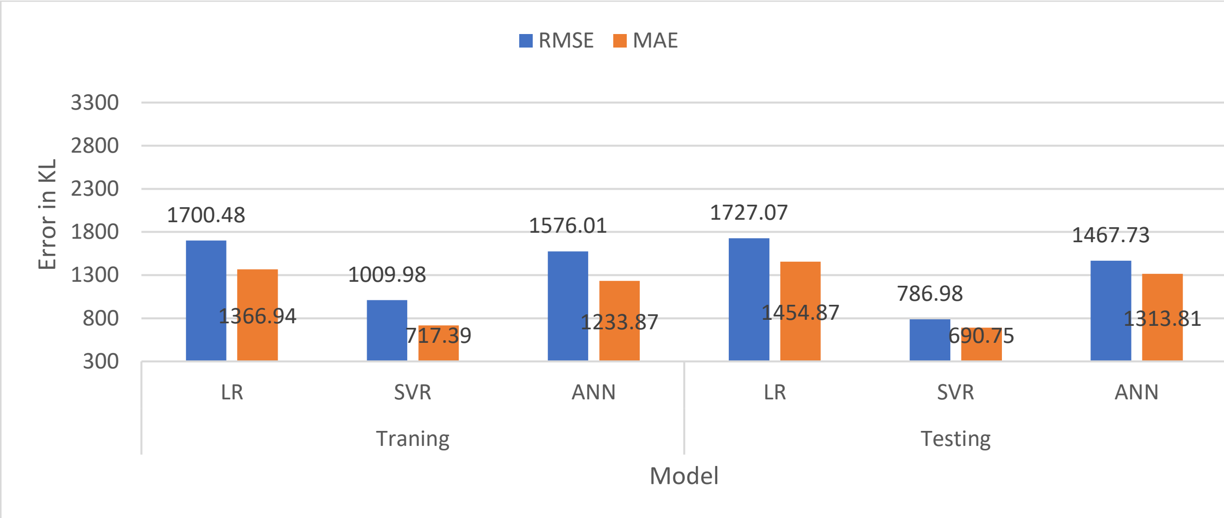

Figure 4-7: RMSE and MAE distribution

Figure 4-8 illustrates a comparison of R^2 values. All models exhibited R^2 values above 0.70 in the training phase. In the testing phase, a wider range of R^2 was observed, with values ranging from 0.31 to 0.65. All models demonstrated a decrease in R^2 from the training phase to the testing phase.

Figure 4-8: R^2 trend among models in the training and testing phases

Figure 4-9 illustrates the MAPE trends across the models. The MAPE for all models decreased from the training to the testing phase. The LR model demonstrated the highest MAPE in all phases, followed by the ANN model, while the SVR model had the least MAPE in both the training and testing phases.

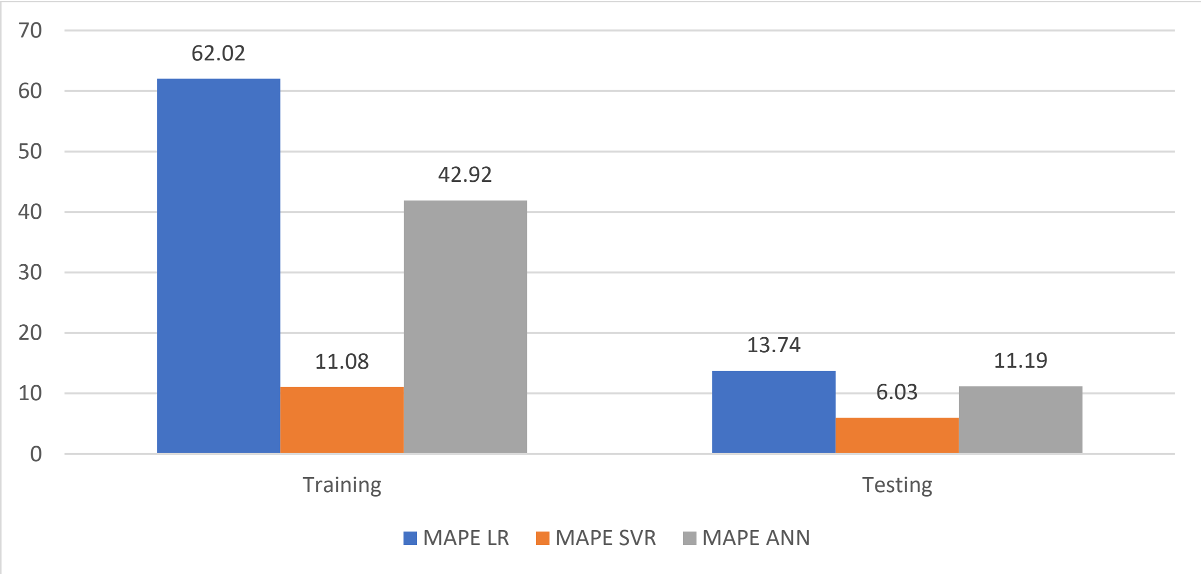

Figure 4-9: Comparison of the MAPE trend among models

4.5 Water Demand Forecast (2025-2026)

The detailed predictive outputs for the 2025-2026 forecasting horizon are presented in Table 4-11. Across the forecasting horizon, all three models, the LR models, the SVR model, and the ANN model, exhibited an overall increasing trend in water consumption, though they differ in their predicted magnitudes and seasonal behaviours.

The LR model maintains the highest values of predicted consumption, projecting a steady upward trajectory, albeit with fluctuations showing seasonality. the LR forecast begins at 14510.94 KL in January 2025 and rises to a peak of 21337. 53 KL by December 2026. This monthly increase shows consistency despite a minor variation in April-June 2025. Conversely, the SVR model produced the lowest prediction. Its predictions ranged from 16441.89 KL in January 2025 to a high of 13724.34 KL in December 2026. The SVR forecast displays a moderate and controlled growth trend compared to the LR model, moreover demonstrating more intermediate estimates between the LR and ANN models.

The ANN model also forecasts a general upward trend, with predicted values rising from 9,595.55 KL in January 2025 to 11,863.81 KL in December 2026. Notably, a slight dip in projected demand occurred at the start of 2026 (January), after which the trend resumed to a pattern similar to the previous year.

Table 4-11: Predicted Monthly water demand (KL) values for the period of 2025-2026

Date	LR	SVR	ANN
01/01/2025	13972.85	16441.89	9595.55
01/02/2025	15523.25	10274.80	9720.18
01/03/2025	14203.63	11699.97	9856.66
01/04/2025	14315.86	11341.02	10004.04
01/05/2025	14011.77	11345.49	10164.84
01/06/2025	16095.82	11955.60	10338.79
01/07/2025	14751.96	11983.08	10526.31
01/08/2025	14046.52	12414.63	10725.7
01/09/2025	14512.94	12654.17	10935.53
01/10/2025	15056.62	13066.65	11153.56
01/11/2025	15660.67	13457.85	11376.69
01/12/2025	15399.91	13441.65	11602.25
01/01/2026	14205.15	11225.27	9999.26
01/02/2026	15755.55	12291.42	10100.58
01/03/2026	14435.92	11274.67	10215.42
01/04/2026	14548.16	12119.67	10343.53
01/05/2026	14244.06	11420.29	10487.26
01/06/2026	16328.12	12261.14	10646.5
01/07/2026	14984.26	12423.58	10821.46

01/08/2026	14278.81	12630.82	11010.48
01/09/2026	14745.23	13017.48	11211.9
01/10/2026	15288.92	13326.88	11423.31
01/11/2026	15892.97	13796.81	11641.58
01/12/2026	15632.20	13724.34	11863.81

Figure 4-10 presents the predicted monthly water consumption (KL) for the period January 2025- December 2026 as generated by LR, SVR, and ANN Models. Across all models, a general upward trend in water demand is observed over the two-year forecasting period. The LR model generally consistently yields the highest consumption estimates. The SVR model starts on the highest yield, drops, and then gradually increases, peaking in November and December and dropping in January. Fluctuations can be seen between January 2025 and May 2026. The ANN offers more underestimated and smoothed out predictions compared to SVR and LR. However, it captures a similar trend as SVR in the increased consumption in December 2025, which then drops in January 2026, capturing a dip around January, after which consumption gradually increases

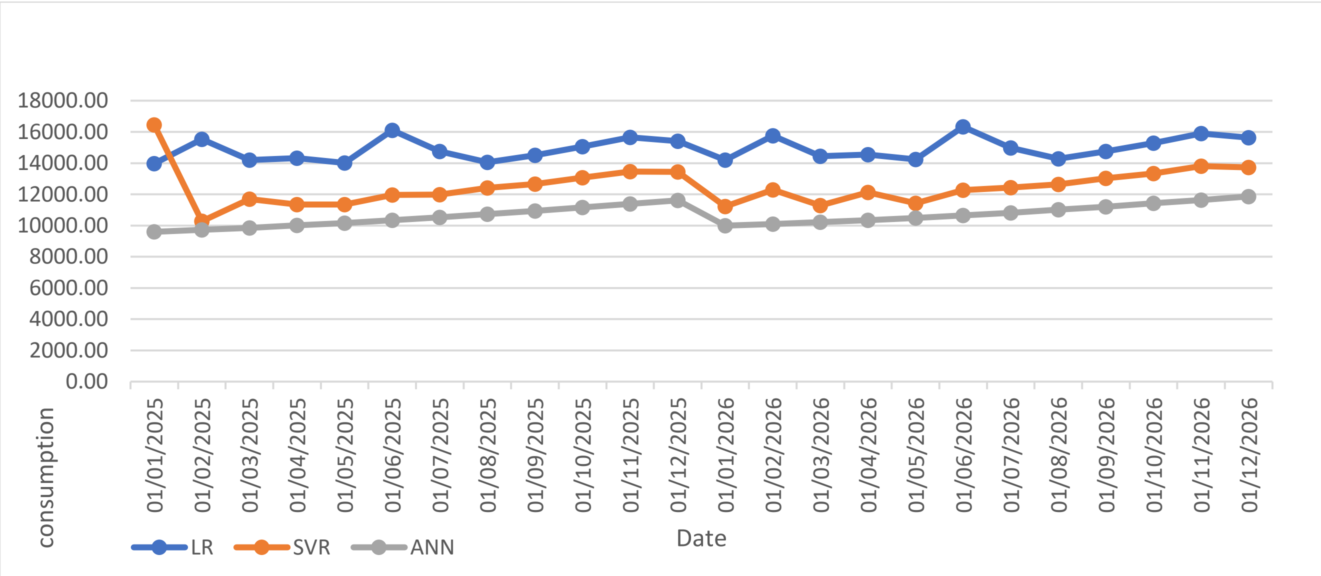

Figure 4-10: Graphical representation of the trend in predicted consumption for all models

Chapter 5: Discussion

5.1 Introduction

This study examines the influence of precipitation, maximum and minimum temperatures, and population on water demand, directly addressing the first objective. Secondly, it provides a comparative performance assessment of the LR model, the SVR model, and the ANN model in forecasting water demand. Finally, concluding with an analysis of the forecasts generated by each model, assessing their practical predictive behaviours and the practical implications of the outputs for utility planning, operations, and water management. This study details a critical discussion of the findings obtained from the study. The results are analysed and situated within the context of existing literature to interpret their implications for water demand management.

5.2 Influence of Determinants on Water Demand

Understanding factors driving water demand is crucial in water resources planning and management, creating the basis for accurate and meaningful forecasting (Billings & Jones, 2008; Rinaudo, 2015). The simple MLR and semi-log model yielded an R^2 of 0.704 (adjusted R^2 of 0.666) and 0.615, respectively. These results are higher than the range of 0.404-0.413 reported by Bich-Ngoc et al. (2022) for the study conducted in Belgium. Other studies by Fiorillo et al. (2021) and Aho et al. (2016) have reported varying model fit. For instance, Fiorillo et al. (2021) reported an R^2 value of 0.51-0.66 in different regression models investigating the influence of weather on water demand, while Aho et al. (2016) reported an R^2 of 0.434 (adjusted R^2 of 0.403) for their regression model. An R^2 of 0.704 signifies a strong relationship and model fit (Shilehwa, et al., 2019) This suggests that the variables selected for the analysis explain the variance effectively, although partially. Suggesting that some variables which were not included in this study could explain the variance, therefore, future studies should include more variables for a more comprehensive analysis, with a look into more predictors. For example, demographic and socioeconomic predictors such as income, household size, Gross Domestic Product (GDP), gender among others, these were seen to explain influence in other studies, for instance, Sant ‘Ana & Mazzega (2017) suggested that housing characteristics, number of residents, income, and family and consumption patterns should be considered as they are pertinent in prediction models. Oyerinde & Jacobs (2022) also indicated the inclusion of vast variables to better understand the influence of water consumption.

This study further determines population as a driver of water demand across all three models, the simple MLR, semi-log MLR, and the Robust MLR, demonstrating population as the single statistically significant determinant of water demand, with $p<0.001$ in Ha-Foso as per the time frame analyzed. Bich-Ngoc et al. (2022) also observed that the number of adults and children was found to be statistically significant with $p<0.001$. The linkage with population and number of adults and children is that, with a greater number of adults and children, the population increases. Similarly, Ou et al. (2023) emphasized population size as a critical determinant of water demand, particularly in serviced cities, where it accounted for approximately 50% of the water demand. Anang et al (2019) also identified population size as a significant predictor of water consumption. These observations are further supported by Billings & Jones (2008), who assert that traditional forecasting methodologies such as linear regression models are based on the rationale that increasing population leads to increased demand. Consequently, urban expansion, coupled with increased migration and residential development in the study area, likely explains the high significance of the population as a predictor. This is corroborated by Lechesa (2023), who reported that Ha-Foso is experiencing rapid growth due to the high demand for housing settlements. Showing that arable land is increasingly being converted to housing settlements. This is consistent with the findings of this study, which have identified population as the main driver.

The significance of population as a driver of water demand in this study is further affirmed by reports from WASCO (2023), which have documented persistent water shortages in the study area. The shortages have been attributed to the increasing population, which outpaces supply, despite efforts to augment supply through projects such as the Lowland Water Supply Scheme, the Maseru Peri Urban Water Supply Project, and the High North Reservoir Connection Water Project. For instance, GDP has been observed to influence water demand, in studies like Xiangmei et al. (2021), although with a negative impact, showing that with increasing GDP, water efficient amenities can be afforded. Huang et al. (2020) and Anang et al (2019) report in the positive impact of GDP, suggesting an increase in water demand with increasing GDP due to lifestyle changes.

The influence of temperature, encompassing both maximum and minimum daily values, on water consumption was examined in this study. Contrary to the general hypothesis positing a positive correlation between increased maximum temperatures and increased water consumption, the results across all three MLR models indicated a statistically insignificant relationship. Specifically, the coefficient for the maximum temperature coefficient was 198.3KL, accompanied by a p-value of 0.328. While the positive coefficient suggests a potential increase of 198.3 KL in water consumption per unit increase in maximum temperature, the high p-value suggests that this relationship is not statistically significant. This finding implies that, based on the analysed data, fluctuations in maximum temperature do not reliably predict changes in water consumption in the Ha-Foso area during the study period.

This contrasts with results from Ou et al. (2023), who observed temperature as the second most critical driver of water demand in industrial cities and agricultural cities. However, their analysis revealed that the effect of temperature was not consistently significant in their study, highlighting the context-dependent nature of this relationship. The authors emphasise the need for a thorough analysis of the complex dynamic linking temperature and water demand, acknowledging its varying impact depending on the city type and context (Ou, et al., 2023).

While a theoretical expectation for an increase in temperature to increase water demand is plausible according to Billings & Clive (2008), this study reports similar results to Ashoori et al. (2016), who found that the influence of maximum temperature on water demand is not statistically significant. They suggest that further research, potentially with a seasonal focus and over longer periods to capture climatic variations, is necessary to illustrate the specific role of temperature in driving water consumption patterns in this region.

In contrast, Alshaikhli et al. (2021) reported temperature as the most significant factor influencing water demand in their study in Qatar. As previously noted, the determinants of water consumption are both temporally and spatially specific. Therefore, the strong effect of temperature on water demand in Qatar is plausible, whereas in the Ha-Foso study area, characterized by different climatic conditions and water usage behaviours, it may indeed vary. Moreover, it is possible that the connection between temperature and water consumption in the Ha-Foso area is non-linear, or that the available data used showed limited variation, hence the partial explanation of the models as suggested by Ashoori et al. (2016).

Similar to maximum temperature, minimum temperature emerged as a statistically insignificant predictor of water consumption in this study. With a coefficient of -872.0KL and a p-value greater than 0.001, indicating a negative correlation between minimum temperature and water consumption, suggesting a potential decrease in water consumption with an increase in minimum temperature. Suggesting that as minimum temperature increases, there is a tendency for water consumption to decrease, although not statistically reliable. Colder minimum temperatures may lead to increased water use for warming, especially in the morning when temperatures are low. Conversely, when minimum temperatures are higher, such activities may decrease, thus decreasing water consumption. These results are similar to those of Surendra et al. (2014), who found that the minimum temperature had a weak correlation with water demand. While there appears to be a negative correlation. Further research could explore this relationship using longer periods of data that demonstrate the trend of minimum temperature's influence on water consumption.

Several context-specific factors may explain the lack of statistical significance for maximum and temperature as a predictor in Ha-Foso. The temperature variation in the study period might not be extreme enough to trigger significant behavioural shifts in water consumption. Moreover, households in Ha-Foso may not be engaged with water-intensive activities such as swimming pools or irrigation of the lawn, which increase water consumption. Although the results of this study do not demonstrate maximum temperature and minimum temperature as significant predictors, more studies can be done to understand consumption patterns on a long-term basis.

Across all three MLR functional models, precipitation was found to be statistically insignificant (p-value >0.05) with a coefficient of approximately 198.3 KL. This implies that every 1mm rise in precipitation is linked with an estimated increase of 198.3 KL in water consumption, however, this relationship is not statistically supported. Ashoori et al (2016) similarly reported a statistically insignificant relationship between water consumption and precipitation for multi-family households, attributing it to reduced outdoor water use during rainy periods, such as lawn watering and car washing, thus lowering overall consumption. Similarly, Alshaikli et al. (2021) observed a negative correlation between precipitation and water demand. Moreover, Ebad & Yilmaz (2019) also found precipitation to be ineffective as a determinant for water consumption. However, these findings differ from those of Fan et al. (2017), who reported a positive and statistically significant relationship between precipitation and water consumption for high consumption cities, with a p-value<0.01. Showing that urban households demand more water due to kitchen use, laundry, and showering.

In the case of Ha-Foso, the lack of significant impact of precipitation and maximum and minimum temperatures may be explained by independence on indoor use as opposed to outdoor use. This is corroborated by the results of Muloiwa et al. (2022), who reported that for the peri-urban area of Thohoyandou in South Africa, indoor water consumption outpaces outdoor. Moreover, Knox (2020) highlights that if water consumption is based more on indoor use, it is generally related to demographic, socio-economic, and behavioural habits of the residents. Showing that water consumption remains fairly constant indoors, with little to no evidence of seasonal fluctuation. Emphasizing that the change in water use patterns for indoor activities between summer and winter is generally considered insignificant.

Moreover, this can be explained by limited alternative water sources, suggesting a continued reliance on tap water. Therefore, promoting water conservation initiatives such as rainwater harvesting may be beneficial. This is explained by Bich-Ngoc et al (2022), who indicated that the presence and access to supplementary water sources through systems such as rainwater harvesting can noticeably decrease the demand for piped water. Moreover, Garcia et al. (2019) reported that installing fixtures that reduce consumption, such as rainwater harvesting, significantly reduces water consumption.

It can thus be deduced that across all three models, population has consistently emerged as a statistically significant predictor of water demand, with p-values well below the conventional significance threshold for the study. Demonstrating it as the biggest influence on water demand. In contrast, climatic variables such as precipitation and maximum and minimum temperatures failed to demonstrate statistical significance, with p-values well above 0.05. These findings are consistent with other prior studies that emphasise the dominant influence of demographic factors over climatic variables on water demand (Ou et al., 2023; Anang et al., 2019). Accordingly, the subsequent predictive modelling phase of this study focuses exclusively on population as the sole predictor variable.

5.3 Comparative Analysis of Model Performance

The evaluation of predictive model performance is crucial for determining the effectiveness of various methodologies for predicting water demand (Liu, 2020; Behboudian, et al., 2014; Adamowski & Karapataki, 2010). The selection of an optimal model necessitates a rigorous assessment of established performance metrics. Specifically, the model demonstrating superior performance is characterized by low values of RMSE, MAE, and MAPE, coupled with a high R^2, signifying a strong explanatory power.

The SVR model demonstrated the best generalizability, achieving the best performance metrics in the training phase. Moreover, demonstrating consistency with the least metrics in the testing phase, highlighting its robustness on unseen data. This is consistent with the results of Soebroto et al. (2022), who also found in their study that SVR performed better than ANN. Their study showed that SVR can easily be effective in forecasting data with a limited number of features. Similarly, SVR was found to be suitable in comparison to Genetic Programming and ANN in a study by Zanjani et al. (2024). Hererra et al. (2010) also found SVR to have better performance over ANN, Random Forest, multivariate adaptive regression, and Projection pursuit when predicting hourly urban water demand. Similarly, Candelieri (2017) reported SVR to be better suited in predicting water demand in Italy, citing that SVR is better suited due to variability in water usage patterns. This performance on unseen data is confirmed by Ghalehkhondabi et al. (2017), who showed that the effectiveness of SVR is in its risk minimization and its robustness under limited data. Tamang & Shukla (2019) show that SVR also works well with small datasets in their study, with ANN surpassing SVR in accuracy, confirming that SVR works effectively with limited datasets. However, Vijai & Sivakumar (2018) reported differing results, with ANN surpassing SVR in their study. Similarly, Antunes et al. (2018) reported on the unexpected poor performance of SVR compared to ANN, K- K-nearest neighbor (KNN), RF, and ARIMA. Hao et al. (2022) reported in their study that wavelet decomposition improved LSTM, such that in their study, wavelet-coupled LSTM outperformed all other models, showing its ability to improve models, such that they can better capture peaks and predictive efficiency than stand-alone models.

The ANN model also demonstrated strong performance, as the second-best performing model, consistently exhibiting the lowest performance metrics after SVR in both the training and testing phases, demonstrating an R^2 of 0.83 in the training phase. Similar results were reported in existing studies, where ANN demonstrated a strong fit on training data, for example, Hao et al. (2022) report R^2 values of 0.886 and 0.763 for their ANN models. Alikhani & Moeini (2025) report a value of 0.767 for ANN. However, a deterioration was observed during testing, where values dropped from 0.83 to 0.31. Suggesting the model is unable to fit the unseen data. Such performance deterioration was also noted by Oyebode & Ighravwe (2019), who report a drop in R^2 from 0.7236 to 0.6614, and Alikhani & Moeini (2025) also observed a deterioration in R^2 from 0.76 to 0.68, indicating a decline in performance in terms of fit. Although ANN did not achieve the best performance in this study, they have demonstrated superior performance and suitability in studies like Behboudian et al. (2014), where ANN was found to be suitable for long-term forecasting. Similarly, Adamowski & Karapataki (2010) reported that ANN demonstrated superior results in their studies.

A common challenge in machine learning is when models memorize training data patterns without capturing generalizable trends, thus overfitting (Zubaidi, et al., 2020; Ghalehkhondabi, et al., 2017). The ANN model demonstrated signs of overfitting, with a significant drop in the R^2 from the training to the testing phase. Oyebode & Ighravwe (2019) reported overfitting of models, which was indicated by stellar performance in the training phase however, the performance declined in the testing phase. They attributed the issue to the inclusion of irrelevant and redundant input variables. This, however, does not fit with this study, as only population was used as an input variable. Therefore, the challenge may be the limited number of variables. This can be corroborated by Antunes et al. (2018), who stated that model performance can increase with the increasing number of features.

The models demonstrated smoothing and underestimation of peaks, especially the ANN model. Tiwari & Adamowski (2014) noted that ANNs may struggle with simulating high-demand values or capturing non-stationary data patterns without a sufficiently large and diverse training set. Furthermore, factors such as suboptimal network architecture, insufficient hyperparameter tuning, and sensitivity to data can contribute to poor generalization, leading to either overestimation or underestimation (Zubaidi, et al., 2020; Lee & Derrible, 2020). Manapace et al. (2021) suggest that more articulate architectures are needed to handle longer forecasting horizons to prevent overfitting. While Tiwari & Adamowski (2014) note that ANN models may struggle with non-stationary data or simulating peak demand values without sufficiently large, diverse training datasets. Therefore, future research should focus on increasing features while also exploring more complex architectures. Similarly, Soebroto et al. (2022) reported that an ANN trained with back propagation is not effective at training in comparison to SVR, MLR, and ANN. Suggesting that they require random weights to be iteratively adjusted. This, therefore, implies that for this study, the performance of the ANN model can be greatly increased with deeper architectures and an exploration of advanced training algorithms, to find one that best suits the data.

While ANNs are celebrated for their adaptability and ability to learn complex relationships, their effectiveness is dependent on large and diverse data sets (Tiwari & Adamowski, 2014). Walker et al. (2015) also reported that ANN's robustness is dependent on extensive training data. In data-sparse or highly variable environments, as is often the case in developing countries, these models may overfit, capturing noise rather than meaningful patterns. This further explains the performance observed in the ANN.

By comparison, the LR model exhibited the weakest performance across both training and testing phases. Despite an improvement in MAPE from 59.96% to 22.6% in the test phase, it simultaneously demonstrated an increase in RMSE (1806.36 KL to 2659.55 KL) and MAE (1484.07 KL to 2438.29 KL), alongside a drop in R^2 in the test phase (0.77 to 0.64). Donkor et al. (2014) reported on the inadequacy of LR models in capturing non-linear dynamics. Similarly, Adamowski & Karapataki (2010) reported underperformance in the LR model, showing that LR's reliance on linear assumptions restricts its ability to capture complex relationships between variables. Vijai & Sivakumar (2018) also identified MLR as the least effective model when compared to ANN-based approaches in a water demand forecasting study. While linear regression does have the capacity to forecast, for complex systems such as water demand, its unable to accurately and sufficiently capture the underlying trends.

It is essential to recognize that the superiority of any model is context-dependent. The performance is shaped by the attributes of the dataset, forecasting objective, and both spatial and temporal scales. For example, Mouatadid & Adamowski (2017) found Extreme Learning Machines (ELM) to outperform ANN, SVR, and MLR in their application, while Guo et al. (2018) demonstrated the advantages of deep learning models such as Gated Recurrent Unit Networks for their enhanced memory functions. Moreover, to mitigate the limitation of standalone models, researchers advocate for hybrid models. for instance, Adamowski et al. (2012) found that hybrid or coupled models, such as Wavelet-ANN, outperformed traditional techniques, including MLR and ARIMA, emphasizing the potential of coupled models for improved prediction. Alikhan & Moeini (2025) demonstrated that equipping ANN with a Wavelet transformation function increases performance and prediction accuracy.

While the findings presented in this study strongly suggest superior performance of machine learning models over MLR in the context of this specific water demand forecasting problem, literature shows that the choice of the most appropriate model necessitates a thorough evaluation, considering the unique data attributes and the objectives specific to the forecasting task. The capability of machine learning models to forecast water demand is demonstrated, albeit with limitations.

5.4 Forecast Implications and Practical Applications

The value of a forecasting model is determined not only by statistical accuracy on historical data but also by its capacity to produce plausible and actionable future scenarios. While quantitative performance metrics are essential for model validation, qualitative analysis of forecast outputs is equally critical for evaluating a model's practical utility and its inherent structural biases (Makridakis, et al., 2018). As House-Peters & Chang (2011) emphasize, the ability of a model to capture key features such as seasonality and peak demands directly impacts its applicability in water management decisions.

The forecast outputs of all the models in this study, while indicating an upward trend, deviate in characteristics, highlighting the critical tradeoffs between models, accuracies, and the ability to capture essential system dynamics, such as seasonality and non-linearity. The LR model predicted the highest consumption values, projecting a steep and consistently increasing trajectory. This aligns with Adamowski et al. (2012), who highlighted that MLR models tend to over-forecast, particularly during peak demand periods, due to their inability to capture complex interactions and seasonal variability. This overestimation reflects the model's fundamental limitation in traditional regression models, which assume linear relationships between variables. This is corroborated by Donkor et al. (2014) and Rinaudo (2015) who report on the failure of simplistic methods to account for moderating influences like improvements in water use efficiency or shifts in consumer behavior, both of which are non-linear. For a utility company such as WASCO, basing infrastructure investment on inflated projections could lead to premature or oversized development of capital-intensive projects, resulting in inefficiencies and potential stranded assets.

In contrast, the SVR and ANN models' estimates generally predicted moderate growth trends. However, the outputs exhibited a smoothing effect, flattening the sharp peaks and troughs associated with seasonal demand. While SVR and ANN models provide more accurate estimates of the average demand in terms of the metrics, their failure to capture peak demand periods is a critical limitation. This limitation was observed by Zubaidi et al. (2020), showing that while machine learning models are superior at capturing non-linear trends, they may struggle to replicate the high-frequency variability and extreme values. This limitation may affect water supply as water supply systems must be designed and operated to meet maximum, not average demand (Rathnayaka, et al., 2016). This smoothing effect is a well-documented challenge in machine learning applications.

The SVR model's estimates were intermediate compared to both LR and ANN. The average estimates it provides offer a reliable basis for planning than relying on ANN, with the least estimates, or the LR model, with the most inflated estimates. The limitations observed across all three models highlight the inadequacy of relying on a single deterministic forecast.

Contemporary water resources management increasingly advocates for planning under deep uncertainty, where the goal is to identify robust strategies that perform well across a varied range of plausible features (Marchau, et al., 2019). In Lesotho's context, this would entail using different model outputs to construct low, medium, and high growth scenarios. These can be used to stress-test policy and infrastructure options, ensuring the selected strategies are adaptive and resilient, regardless of which future unfolds (Srdjevic, et al., 2015). These observed limitations point toward the promise of hybrid forecasting approaches that combine the advantages of different modelling approaches.

Seasonal peaks, such as increased consumption during the summer months, were not efficiently represented in either SVR or ANN forecasts. This underperformance may partially be attributed to the use of the constant annual population growth rate, which oversimplifies the complex and often non-linear nature of demographic changes (Caswell, 2019). Additionally, the models' learning may have been affected by noise and missingness in data, as highlighted by (Walker, et al., 2015). Notably, the SVR and ANN forecast outputs showed some drops in December and a slight decrease in January, possibly reflecting real-world scenarios in Lesotho, such as the festive season, which might reduce certain water-consuming activities, due to people going into the rural areas for holidays and coming back in January. Furthermore, Farah et al. (2019) and Pesantez et al. (2020) similarly report that ANN-based forecasts often preserve general trends at the expense of peak accuracy, due to this smoothing effect.

Mu et al. (2020) emphasize that performance variations between predictions also stem from underlying architectural discrepancies across models. Their work suggests that refining model structure can significantly improve performance. Consequently, future research should explore more adaptive architectures and incorporate more exogenous predictors to capture non-linear dynamics effectively. Antunes et al. (2018) report that the inclusion of such variables has been shown to improve prediction accuracy in some studies, although others, like Pesantez et al. (2020), note only marginal improvements, indicating the importance of contextual relevance and variable sensitivity analysis.

Chapter 6: Conclusions and Recommendations

6.1 Conclusion

The thesis aimed to use Multiple Linear Regression (MLR) to assess the influence of population, maximum and minimum temperature and precipitation on water consumption. The analysis, which tested simple, semi-log, and robust functional forms of MLR, consistently identified population as the only statistically significant predictor of water demand. Climatic variables, including maximum temperature, minimum temperature, and precipitation, did not demonstrate a statistically significant influence on consumption within the study period. The final model explained approximately 70% of the variance in water demand, suggesting that other unexamined socio-economic variables contribute to consumption patterns.

The study further explored a comparative performance evaluation between Linear Regression (LR), Artificial Neural Network (ANN), and Support Vector Regression (SVR) models on the basis of a multi-metric system, which included the RMSE E, MAE, MAPE, and R^2. In this regard, SVR demonstrated superior performance, with the least metrics. While the ANN performed well during training, it showed signs of overfitting, whereas the LR model was the least accurate. This study affirms the potential of machine learning models in forecasting water demand, demonstrating that machine learning approaches are better suited for handling the complex non-linear relationships. However, it is equally acknowledged that machine learning models have limitations such as under estimation and smoothing.

Finally, an assessment of each model's 2-year forecast highlighted distinct behaviors of each model. The LR model projected the steepest linear increase, reflecting its structural limitations as a linear model. In contrast, the SVR and ANN models forecasted a more moderate and subtle growth trend, which reflected the non-linear interactions present in water demand systems. However, both machine learning models exhibited a smoothing effect, struggling to capture the sharp peaks and troughs of seasonal demand. For Water utilities like the Water and Sewerage Company (WASCO), this shows that optimization of models is imperative, to ensure prediction accuracy and robust models. Moreover, it emphasizes the need for quality data and the selection of the most appropriate model to achieve prediction accuracy and reliability.

6.2 Recommendations

Based on the study's findings and limitations, future research should address the shortcomings of this research. Firstly, the unavailability of data should be addressed, for example, the constant population growth rate may have under or overestimated future projections. Furthermore, to develop a more comprehensive model, studies should incorporate a wider range of predictors. These could include socio-economic factors such as water pricing, Gross Domestic Product, household size, and income, among others. While this study employed standard ANN and SVR, future efforts could explore more complex network architectures, as well as other models such as Random Forests. Additionally, the use of other normalization techniques, more advanced architectures, and activation functions, as well as rigorous hyperparameter tuning, is also recommended to mitigate overfitting. More than standalone models, hybrid models should be explored to augment the limitations of stand-alone models

Beyond supervised forecasting, unsupervised learning techniques like clustering should be employed. This could help segment consumers into distinct groups based on their consumption patterns, enabling utility companies like WASCO to design targeted tariff structures and more effective demand-side management strategies

It is recommended that utility companies like WASCO adopt advanced forecasting methods like those explored in this thesis. Moving beyond estimates based purely on historical averages will support more effective resource planning. Coupled with a broader scope, the study should be replicated in other peri-urban and urban areas within Lesotho to inform policy development and to have a more localized understanding of water consumption.

The company can also invest in capacity building, ensuring a robust department that can use these methods for planning and water resources management. A data sharing platform may also be established between BOS and WASCO to be able to share data for reliable forecasting.

References

Abdullah, A., Ibn Abubakar, B. S. U. & Jones, A. N., 2019. Pattern of residential water demand analysis for Maiduguri Metropolis, North-Eastern Nigeria. *Arid Zone Journal of Engineering, Technological & Environmental,* 15(2), pp. 292-303.

Adamowski, J., Fung Chan, H., Prasher, S O., Ozga-Zielinski, B. & Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research,* 48(1).

Adamowski, J. & Karapataki, C., 2010. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. *JOURNAL OF HYDROLOGIC ENGINEERING,* Volume 15, pp. 729-743.

Aho, M. I., Akpen, G. D. & Ivue, P., 2016. Determinants of residential per capita water demand of Makurdi Metropolis. *Nigerian Journal of Technology,* 35(2), pp. 424-431.

Alikhani, M. R. & Moeini, R., 2025. Predicting the urban water demand by equipping intelligent-based with discrete wavelet transform function. *Applied Water Science,* 15(2).

Alshaikhli, M., Aqeel, S., Valdeolmillos, N., Fathima , F. & Choe, P., 2021. A Multi-Linear Regression Model to Predict the Factors Affecting Water Consumption in Qatar. *IOP Conference Series: Earth and Environmental Science,* 6(1), p. 012004.

Anang, Z., Padli, J., Rashid, N. K. A., Aliphia, R. M. & Musa, H., 2019. Factors Affecting Water Demand: Macro Evidence in Malaysia. *Jurnal Ekinomi Malaysia,* 53(1).

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A. & Oliveira, M. S., 2018. Short-term water demand forecasting using machine learning techniques. *Journal of Hydroinformatics,* 20(6), pp. 1343-1366.

Ashoori, N., Dzombak, D. A. & Small, M. J., 2016. Modeling the Effects of Conservation, Demographics, Price, and Climate on Urban Water Demand in Los Angeles, California. *Water Resources Management,* 30(14), pp. 5247-5262.

Awad, M. & Khanna, R., 2015. *Efficient Learning Machines.* 1 ed. California: Apress Berkeley.

Bangdiwala, S. I., 2018. Regression: multiple linear. *International Journal of Injury Control and Safety Promotion,* 25(2), pp. 232-236.

Basak, D., Pal, S. & Patranabis, D. C., 2007. Support Vector Regression. *Neural Information Processing-Letters and Reviews,* 11(10), pp. 203-224.

Bashar, A. M. et al., 2023. Investigation of factors affecting rural drinking water consumption using intelligent hybrid models. *Water Science and Engineering,* 16(2), pp. 175-183.

Behboudian, S., Tabesh, M., Falahnezhad, M. & Ghavanini, F. A., 2014. A long-term prediction of domestic water demand using preprocessing in artificial neural network. *Journal of Water Supply: Research and Technology - AQUA,* 63(1), pp. 31-42.

Bennett, C., Stewart, R. A. & Beal, C. D., 2013. ANN-based residential water end-use demand forecasting model. *Expert Systems with Applications,* 40(4), pp. 1014-1023.

Bich-Ngoc, N., Prevedello, C., Cools, M. & Teller, J., 2022. Factors influencing residential water consumption in Wallonia, Belgium. *Utilities Policy,* 74(101281).

Billings, R. B. & Jones, C. V., 2008. *Forecasting Urban Water Demand.* 2nd ed. California: American Water Works Association.

BOS, 2019. *Kingdom of Lesotho 2016 Lesotho Population and Housing CENSUS Volume VI: Population Projections,* Maseru: Government of Lesotho.

Brandner, H., 2016. Identifying the Influential Factors of the Temporal Variation of Water Consumption: A Case Study using Multiple Linear Regression Analysis.

Candelieri, A., 2017. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. *Water,* 9(3), p. 224.

Caswell, H., 2019. Sensitivity analysis of nonlinear demographic models. In: *Sensitivity Analysis: Matrix Methods in Demography and Ecology.* Cham, Switzerland: Springer Cham, pp. 199-252.

Chai, T. & Draxler, R. R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature. *Geoscientific Model Development,* 7(3), pp. 1247-1250.

Cortes, C. & Vapnik, V., 1995. Support-vector Networks. *Machine Learning,* Volume 20, pp. 273-297.

De Souza Groppo, G., Costa, M. A. & Libânio, M., 2019. Predicting water demand: A review of the methods employed and future possibilities. *Water Science and Technology: Water Supply,* 19(8), pp. 2179-2198.

Dias, T. F. & Ghisi, E., 2024. Urban Water Consumption: A Systematic Literature Review. *Water,* 16(6).

Dizon, S. L., Mamalias, A. Q. & Lincopinis, D. R., 2023. *A Review on R Programming Language,* Pagadian City: ResearchGate.

Donkor, E., Asce, S. M., Mazzuchi, T. A., Refik, S. & Roberson, J. A., 2014. Urban Water Demand Forecasting: Review of Methods and Models. *Journal of Water Resources Planning and Management,* 140(2).

Ebad , O. M. & Yilmaz, V., 2019. *Investigation of variables that affect the urban water consumption over districts of Koya City.* Samsun, Turkey, SETSCI conference proceedings.

Fan, L., Gai, L., Tong, Y. & Li, R., 2017. Urban water consumption and its influencing factors in China: Evidence from 286 cities. *Journal of Cleaner Production,* Volume 166, pp. 124-133.

FAO, 2023. *Lesotho: Land cover atlas 2017–2022,* Maseru: FAO.

Farah, E., Abdallah, A. & Shahrour, I., 2019. Prediction of water consumption using Artificial Neural Networks modelling. *In MATEC Web of Conferences,* 295(01004).

Fiorillo, D., Kapelan, Z., Xenochristou, M., De Paola, F. & Giugni, M., 2021. Assessing the Impact of Climate Change on Future Water Demand using Weather Data. *Water Resources Management,* 35(5), pp. 1449-1462.

Flatt, C. & Jacobs, R. L., 2019. Principle Assumptions of Regression Analysis Testing,Techniques and statistical reporting of Imperfect data Sets. *Advances in Developing Human Resources,* pp. 484-502.

Garcia, J., Salfer, L. R., Kalbusch, A. & Henning, E., 2019. Identifying the drivers of water consumption in single-family households in Joinville, Southern Brazil. *Water (Switzerland),* Volume 11.

García-Soto, C. G., Torres, J. F., Zamora-Izquierdo, M. A., Palma, J. & Troncoso, A, 2024. Water consumption time series forecasting in urban centers using deep neural networks. *Applied Water Science,* 14(2).

Ghalehkhondabi, I., Ardjmand, E., Young, W. A. & Weckman, G. R., 2017. Water demand forecasting: review of soft computing methods. *Environmental Monitoring and Assessment,* 189(313).

Ghobadi, F. & Kang, D., 2023. Application of Machine Learning in Water Resources Management: A Systematic Literature Review. *Water,* 15(4).

Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R. & Zhu, X., 2018. Short-Term Water Demand Forecast Based on Deep Learning Method. *Journal of Water Resources Planning and Management,* 144(12).

Hao, W., Cominola, A. & Castelletti, A., 2022. *Comparing Predictive Machine Learning Models for Short- and Long-Term Urban Water Demand Forecasting in Milan, Italy.* Milan, Elsevier B.V..

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-Garcı´a, R., 2010. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology,* 387(1-2), p. 141–150.

Hodson, T. O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development,* 15(14), p. 5481–5487.

House-Peters, L. A. & Chang, H., 2011. Urban water demand modeling: Review of concepts, methods and organizing principles. *Water Resources Research,* 47(5).

Huang, Z., Liu, J., Mei, C., Wang, H., Shao, W. & Li , Z., 2020. Analysis of driving forces of water demand Jing-Jin-Ji district in recent years based on water demand field theory. *Science China Technological Sciences,* Volume 63, pp. 2593-2605.

Ibrahim, T., Omar, Y. & Maghraby, F. A., 2020. Water Demand Forecasting Using Machine. Puna, India, AISSMS Institute of Information Technology.

IPCC, 2023. *Climate Change 2023: Synthesis Report,* Geneva: IPCC.

Iqelan, B. M., 2017. Comparison of Parametric and Nonparametric Techniques for Water Consumption Forecasting. *International Journal of Scientific & Engineering Research,* 8(1).

Janiesch, C., Zschech, P. & Heinrich, K., 2021. Machine learning and deep learning. *Electron Markets,* Volume 31, pp. 685-695.

Jierula, A., Wang, S., Oh, T. & Wang, P., 2021. Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences,* Volume 11, pp. 1-21.

Jovel, J. & Greiner, R., 2021. An Introduction to Machine Learning Approaches for Biomedical Research. *Frontiers in Medicine,* Volume 8, p. 771607.

Kåge, L., Milić, V., Andersson, M. & Wallén, M., 2025. Reinforcement learning applications in water resource management: a systematic literature review. *Frontiers in Water,* Volume 7, p. 1537868.

Kim, S. & Kim, H., 2016. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting,* Volume 32, pp. 669-679.

Knox, A. J., 2020. Modeling water demand for residential households, Stellenbosch University: Unpublished Master's Thesis.

Kulaczkowski, A. & Lee, J., 2024. Harnessing the Power of Random Forest for Precise Short-Term Water Demand Forecasting in Italian Water Districts. *Engineering Proceedings,* 69(1), p. 81.

Laszlo, A. & Krippner, S., 1998. Chapter 3 Systems theories: Their origins, foundations, and development. *Advances in Psychology,* Volume 126, pp. 47-74.

Lechesa, M. E., 2023. Land formalization and influence of customary chiefs on recent land policy implementation: case of Ha- Foso, Berea Lesotho. *Journal of the Asia-Japan Research Institute of Ritsumeikan University,* Volume 5, p. 154.

Lee, D. & Derrible, S., 2020. Predicting Residential Water Demand with Machine-Based Statistical Learning. *Journal of Water Resources Planning and Management,* 146(1).

Liu, G., 2023. *Machine Learning for Short-Term Water Predictions,* Exter: University of Exter.

Liu, H., 2020. *Municipal Water Demand Forecasting in the Short and Long Term with ANN and SD Models,* Australia: University of Alberta.

Lu, S., Gao, X., Li, W., Jiang, S. & Huang, L., 2018. A study on the spatial and temporal variability of the urban residential water consumption and its influencing factors in the major cities of China. *Habitat International,* Volume 78, pp. 29-40.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one,* 13(3).

Marchau, V. A., Walker, W. E., Bloemen, P. J. & Popper, S. W., 2019. *Decision making under deep uncertainty from theory to practice.* Switzerland: Springer Nature.

Menapace, A., Zanfei, A. & Righetti, M., 2021. Tuning ANN Hyperparameters for Forecasting Drinking Water Demand. *Applied Sciences,* 11(9), p. 4290.

Miller, C., Portlock, T., Nyaga, D. M. & O'Sullivan, J. M., 2024. A review of model evaluation for machine learning in genetics and genomics. *Frontiers in Bioinformatics,* Volume 4.

Motho, M., Kolawole, O. D., Motsholapheko, M. R. & Mogomotsi, P. K., 2022. Influence of household demographic and socio-economic factors on water demand in Ngamiland District, Botswana. *Water Science,* 36(1), pp. 48-59.

Mouatadid, S. & Adamowski, J., 2017. Using extreme learning machines for short-term urban water demand forecasting. *Urban Water Journal,* 14(6), pp. 630-638.

Mount, G., 2021. *Advancing into Analytics From Excel to Python and R.* 1st ed. USA: O'Reilly Media, Inc..

Muloiwa, M., Dinka , M. O. & Nyende-Byakika, S., 2024. Analysis of domestic water consumption in peri-urban South Africa: The case study of Thohoyandou in Limpopo province, South Africa. *Africa Journal of Science, Technology, Innovation and Development,* 14(6), pp. 1546-1559.

Muloiwa, M., Dinka, M. O. & Nyende-Byakika, S., 2022. Analysis of domestic water consumption in peri-urban South Africa: The case study of Thohoyandou in Limpopo province, South Africa. *African Journal of Science,* 14(6), pp. 1546-1459.

Mu, L., Zheng, F., Tao, R., Zhang, Q. & Kapelan, Z., 2020. Hourly and Daily Urban Water Demand Predictions Using a Long Short-Term Memory Based Model. *Journal of Water Resources Planning and Management,* 146(9).

Naeem, S., Ali, A. & Ahmed, M. M., 2023. An Unsupervised Machine Learning Algorithm: Comprehensive Review. *International Journal of Computing and Digital Systems,* 13(1), pp. 911-921.

Niknam, A., Zare, H. K., Hosseininasab, H., Mostafaeipour, A. & Herrera, M., 2014. A Critical Review of Short-Term Water Demand Forecasting Tools—What Method Should I Use?. *Sustainability,* 14(9).

Nishant, R., Kennedy, M. & Corbett, J., 2020. Artificial Intelligence for Sustainability: Challenges, opportunities and a research agenda. *International Journal of Information Management,* 53(8).

Ogunbode, T. & Ifabiyi, T., 2014. Determinants of domestic water consumption in a growing urban centre in Osun State, Nigeria. *African Journal of Environmental Science and Technology,* 8(4), pp. 247-255.

Ou, Z., He, F., Zhu, Y. & Wang, L., 2023. Analysis of driving factors of water demand based on explainable artificial intelligence. *Journal of Hydrology: Regional Studies,* 47(101).

Oyebode, O. & Ighravwe, D. E., 2019. Urban water demand forecasting: A comparative evaluation of conventional and soft computing techniques. *Resources,* 8(3).

Oyerinde, A. O. & Jacobs, H. E., 2022. Determinants of household water demand: a cross-sectional study in South West Nigeria. *Journal of Water Sanitation and Hygiene for Development,* 12(2), pp. 200-207.

Pacchin, E., Gagliardi, F., Alvisi, S. & Franchini, M., 2019. A Comparison of Short-Term Water Demand Forecasting Models. *Water Resources Management,* Volume 33, pp. 1481-1497.

Papavasileiou, P., Giovanis, D G., Pozzetti, G., Kathrein, M., Czettl, C., Kevrekidis, I G., Boudouvis, A G. & Bordas, S P.A., 2025. Integrating supervised and unsupervised learning approaches to unveil critical process inputs. *Computers and Chemical Engineering,* Volume 192, p. 108857.

Pesantez, J. E., Berglund, E. M. & Kaza, N., 2020. Smart meters data for modeling and forecasting water demand at the user-level. *Environmental Modelling and Software,* 125(104633).

Rathnayaka, K., Malano, H. & Arora, M., 2016. Assessment of sustainability of urban water supply and demand managemnet options: A comprehensive approach. *Water,* 8(12), p. 595.

Razak, T. R., Jarimi, H. & Ahmad, E. Z., 2023. *Simplified Artificial Neural Network Configuration in R Programming for Predictive Modelling.* Kuala Lumpur, Malaysia, Institute of Electrical and Electronics Engineers Inc..

Ridzuan, F. & Wan Zainon, W., 2019. *A review on data cleansing methods for big data.* Surabaya (Indonesia), Elsevier B.V..

Rinaudo, J. D., 2015. Long-Term Water Demand Forecasting. In: Grafton, Q., Daniell, K., Nauges, C., Rinaudo, J D. & Chan, N. eds. *Springer.* Dordrecht: Springer, pp. 239-268.

Sant'Ana, D. & Mazzega, P., 2018. Socioeconomic analysis of domestic water end-use consumption in the Federal District, Brazil. *Sustainable Water Resources Management,* 4(4), pp. 921-936.

Sekamane, T., Nel, W. A. J., McKay, T. J. & Tantoh, H. B., 2023. Community Perceptions of the social impacts of the Metolong Dam and Reservoir in Lesotho. *Land Use Policy,* 125(106495).

Shakya, A. K., Pillai, G. & Chakrabarty, S., 2023. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications,* Volume 231, p. 120495.

Shilehwa, C. M., Makhanu, S. K. & Khaemba, A. W., 2019. Application of Artificial Neural Network Model In Forecasting Water Demand: Case of Kimilili Water Supply Scheme, Kenya.. *International Journal of Civil Engineering,* 6(9), pp. 6-11.

Shrestha, N., 2020. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics,* pp. 39-42.

Shu, J., Xia, X., Han, S., He, Z., Pan, K. & Liu, B., 2024. Long-term water demand forecasting using artificial intelligence models in the Tuojiang. *PLOS One,* 19(5).

Singh, B., Kumar, R. & Singh, V. P., 2021. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review,* 55(2), p. 945–990.

Sloan, R., 2005. Special issue on computational learning theory. *Journal of Computer and System Sciences,* pp. 433-434.

Soebroto, A. A., Cholissodin, I., Pratiwi, D. E. & Putra, G. P. W., 2022. Comparative Study of SVR, Regression and ANN Water Surface Forecasting. *HABITAT,* 33(01), pp. 86-92.

Solomatine, D., See, L. M. & Abrahart, R. J., 2009. Data-Driven Modelling: Concepts, Approaches and Experiences. In: R. J. Abrahart, L. M. See & D. P. Solomatine, eds. *Practical Hydroinformatics.* Berlin: Springer, pp. 17-30.

Srdjevic, B., Medeiros, Y. D. & Srdjevic, Z., 2015. Comparative evaluation of water resources management scenarios using multicriteria decision-making techniques. *Water Resources,* Volume 29, pp. 1473-1490.

Starbuck, C., 2023. *The Fundamentals of People Analytics With Applications in R.* Switzerland: Springer.

Surendra, H. J., Ganesh, S. P., Pruthvi Raj, V. & Deka, P. C., 2014. Analysing the rainfall and temperature influence municipal water consumption using regression technique. *International Journal of Researcg in ENgineering and Technology,* 3(3).

Tamang, A. & Shukla, S., 2019. *Water demand prediction using support vector machine regression.* Banglore, India, IEEE, pp. 1-5.

Tian, Y., Nearing, G. S., Eters-Lidard, C. D., Harrison, K. W. & Tang, L., 2016. Performance Metrics, Error Modeling, and Uncertainty Quantification. *Monthly Weather Review,* 144(2), pp. 607-613.

Tiwari, M. K. & Adamowski , J. F., 2015. Medium-Term Urban Water Demand Forecasting with Limited Data Using an Ensemble Wavelet–Bootstrap Machine-Learning Approach. *Journal of Water Resources Planning and Management,* 141(2).

UN, 2017. *World Population Prospects-Data Booklet,* New York: United Nations.

UN, 2024. *The United Nations Water Development Report: Water for Prosperity and Peace,* Paris: UNESCO.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L., Elkhatib, Y., Hussain, A. & Al-Fuqaha, A., 2019. Unsupervised Machine Learning for Networking: Techniques, applications and Research Challanges. *IEEE Access,* Volume 7, pp. 65579-65615.

Van Thieu, N., 2024. PerMetrics:A framework of perfromance metrics for machine learning models. *Journal of Open Source software,* 9(95), p. 6143.

Van Thieu, N., Barma, S. D., Van Lam, T., Kisi, O. & Mahesha, A., 2023. Groundwater level modeling using Augmented Artificial Ecosystem Optimization. *Journal of Hydrology,* Volume 617, p. 129034.

Varoquaux, G. & Colliot, O., 2023. Evaluating machine learning models and their diagnostic value. *Machine Learning for Brain Diosorders,* Volume 5, pp. 601-630.

Vijai, P. & Sivakumar, P. B., 2018. Performance comparison of techniques for water demand forecasting. *Procedia Computer Science,* Volume 143, pp. 258-266.

Villarin, M. & Rodriguez-Galiano, V. F., 2019. Machine Learning for Modeling Water Demand. *Journal of Water Resources Planning and Management,* 145(5).

Von Bertalanffy, L., 1972. The History and Status of General Systems Theory. *Academy of Management Journal,* Volume 15, pp. 407-426.

Walker, D., Creaco, E., Vamvakeridou-Lyroudia, L., Farmani, R., Kapelan, Z. & Savi´c, D., 2015. Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks. *Procedia Engineering,* Volume 119, pp. 1419-1428.

Wardak, H. A. & Abed, S. Y., 2019. *Analysis of Factors Affecting Water Consumption in Jeddah City.* Jeddah, Saudi Arabia, IEEE.

WASCO, 2025. *WASCO.* [Online]
Available at: https://www.wasco.co.ls/wasco-launches-new-water-transmission-pipeline/
[Accessed 16 February 2025].

WEF, 2024. *Global Risks Report,* Switzerland: World Economic Forum.

Willmott, C. J. & Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research,* 30(1), pp. 79-82 .

World Bank, 2016. *World Bank Group.* [Online]
Available at: https://www.worldbank.org/en/country/lesotho/publication/climate-change-key-for-lesothos-domestic-and-industrial-water-security-agricultural-production-and-regional-water-transfers
[Accessed 2025 February 2025].

Xiangmei, M., Leping, T., Chen, Y. & Lifeng, W., 2021. Forecast of annual water consumption in 31 regions of China considering GDP and population. *Sustainable Production and Consumption,* Volume 27, pp. 713-736.

Ying, X., 2019. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series,* 1168(2).

Zamri, N., Pairan, M. A., Azman, W., Abas, S. S., Abdullah, L., Naim, S., Tarmudi, Z. & Gao, M., 2022. A comparison of unsupervised and supervised machine learning algorithms to predict water pollution. *Procedia Computer Science,* Volume 204, pp. 172-179.

Zanfei, A., Menapace, A. & Righetti, M., 2023. An artificial intelligence approach for managing water demand in water supply systems. *IOP Conference Series: Earth and Environmental Science,* Volume 1136, p. 012004.

Zanjani, M., Bozorg-Haddad, O., Zanjani, M., Arefina, A., Pourgholam-Amiji, M. & Loáiciga, H. A., 2024. Estimating the riverine environmental water demand under climate change with data mining models. *Natural hazards,* 120(12), p. 11451–11464.

Zhang, A. T. & Gu, V. X., 2023. Global Dam Tracker: A database of more than 35000 dams with location catchment and attribute information. *Scientific Data,* 10(111).

Zhang, D., Ni, G., Cong, Z., Chen, T. & Zhang, T., 2014. Statistical interpretation of the daily variation of urban water consumption in Beijing, China. *Hydrological Sciences Journal,* 59(1), pp. 181-192.

Zubaidi, S., Ortega-Martorell, S., Al-Bugharbee, H., Olier, I., Hashim, K. S., Gharghan, S. D., Kot, P. & Al-Khaddar, R., 2020. Urban Water Demand Prediction for a City That Suffers from Climate Change and Population Growth: Gauteng Province Case Study. *Water,* 12(7), p. 1885.