ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

School of Civil and Environmental Engineering Graduate Studies MSc in

Water Supply and Environmental Engineering

Application of Genetic algorithm and Artificial Neural Network in Water Consumption Forecasting and Driving Factors determination

(The case of Addis Ababa Ethiopia)

A Thesis Submitted to the School of Graduate studies in Partial

Fulfillment of the Requirements for the degree of Master of Science in Civil and Environmental Engineering

(Water Supply and Environmental Engineering)

By. Terefe Tilahun Kebede

Thesis Advisor: Mebruk Mohammed (Dr. –Ing)

Addis Ababa University

Addis Ababa, Ethiopia

2024 G.C

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATES STUDIES

SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

This is to certify that the thesis prepared by Terefe Tilahun, entitled: Application of Genetic algorithm and Artificial Neural Network in Water Consumption Forecasting and Driving Factors determination

(The case of Addis Ababa Ethiopia)

Submitted in partial fulfilment of the requirements for the degree of science in civil Engineering (Water Supply and Environmental Engineering) that complies with regulations of the university to meets the accepted standards concerning originality and quality.

APPROVAL BY EXAMINER COMMITTEE

Name	Signature	Date
1. Terefe Tilahun		02/12/2024
Advisor		
2. Mebruk Mohammed (Dr.-Ing)		02/12/2024
Internal Examiner		
3. Agizew Nigussie (PhD)		02/12/2024
External Examiner		
4. Zerihun Getaneh (PhD)		/12/2024
Chair of Department or		
Graduate Programmer Coordinator		
5. __________	__________	

Abrham ... (Dr.)
Dean, School of Civil and
Environmental Engineering

School of Civil & Environmental Engineering
Addis Ababa University Institute of Technology

i

DECLARATION

I hereby declare that this thesis is my original work and has not previously been submitted for a Degree or master's program at this or any other University

Name: Terefe Tilahun

Signature

Date: 02/12/2024

This thesis has been submitted for examination with my approval as a university advisor.

Name: Mebruk Mohammed (Dr. –Ing)

Signature:

Date: 02/12/2024

ii

ABSTRACT

Accurate understanding of water consumption is paramount for effective resource management, infrastructure planning, and ensuring a reliable water supply. Understanding and determining the driving factors of water consumption has become a key challenge. In order to address this issue, the study employs Genetic Algorithm (adopted for both linear and nonlinear regression) along with an ANN model consisting of one hidden layer with either one or five nodes. The GA and ANN models were used to predict water consumption in Addis Ababa city and analyze the driving factors behind water consumption. The model was developed using input data such as water consumption time series, average temperature, population, construction activity, relative humidity, economic development, number of livestock, industrial development, and holiday/festival. Monthly data on water consumption and meteorology (from 2015 (June) to 2023(August)) were gathered from Addis Ababa city water and sewerage authority, as well as the National Meteorological Agency. Sensitivity analysis is used in the forecasting process to choose the most important explanatory factors.

Four different models were developed and their performances were assessed using two metrics: root mean squared error (RMSE) and Normalized root mean squared error (NRMSE). The linear regression GA model achieved an RMSE value of 0.355 Mm3/month and an NRMSE value of 0.0451. On the other hand, the nonlinear regression GA model yielded an RMSE value of 0.339 Mm3/month and an NRMSE value of 0.0430. Moving on to the ANN model with one hidden node, it achieved an RMSE value of 0.325 Mm^3/month and an NRMSE value of 0.0413. Lastly, the ANN model with five hidden nodes achieved the lowest RMSE value of 0.3195 Mm3/month and the lowest NRMSE value of 0.0405. Among the four models, the ANN model with five hidden nodes performed the best, according to these results.

The study conducted using ANN model with five hidden nodes revealed that the primary factors influencing water consumption in Addis Ababa city are population, relative humidity, and industrial activity. Population was identified as the most significant factors, with an effect size of 18.73%, followed by relative humidity at 16.65% and industrial activity at 12.77%. The additional factors play a substantial role (ranging from 1.1% for average temperature to 11.3% for livestock), to the point that neglecting them in water consumption calculations could result in inaccuracies when forecasting future demand trends.

It is recommended that future predictions of water consumption in Addis Ababa city take into account nine driving factors: population, average temperature, construction activity, relative humidity, economic development, agricultural activity, industrial development, holiday/festival, and precipitation.

Key Words: Addis Ababa City, Artificial Neural network, Genetic Algorithm, Water Consumption Driving factors,

ACKNOWLEDGEMENTS

Glory be to God, the Most Gracious and Most Merciful, who endowed humanity with power, knowledge, and intelligence at creation. I wish to gratefully thank my advisor, Mebruk Mohammed (Dr.-Ing.), for all of his aid, support, patience, and constructive criticism. The methodology of this study thesis from start to finish. I would also like to thank all of my friends and colleagues, with whom I have enjoyed sharing, learning, and laughing. It's been fun. Finally, but just as importantly, I would like to thank everyone who, in one way or another, contributed to the success of this thesis.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF FIGURES

ACRONYMS

AAPDC	Addis Ababa City Administration Plan and Development Commission
AAWAS	Addis Ababa Water and sewerage authority
APE	Absolute Percentage Error
ANN	Artificial Neural Network
CSA	Central Statistic Authority
GA	Genetic Algorithm
MATLAB	Matrix Laboratory
NRMSE	Standardize Root Mean Squared Error
RMSE	Root Mean Squared Error
R 2	Coefficient of determination

1. Introduction and Background

1.1 Introduction

Water, the lifeblood of our communities, is essential for survival and economic development. Precise predictions of water consumption are crucial for managing this precious resource effectively, planning infrastructure, and guaranteeing a dependable supply. In booming urban areas like Addis Ababa, Ethiopia, where population surges and urbanization are putting a strain on water resources, forecasting water consumption has become a critical challenge. Modern technologies, such artificial neural networks (ANN) and Genetic algorithms (GA), are being used to overcome this difficulty by increasing the precision of water consumption forecasts and identifying the elements that influence them.

Meeting customer demand is the most crucial aspect of designing and running a water distribution system. Accurate short-term water consumption forecasts are crucial for efficient system operation and management. Planning a reliable water supply system requires anticipating and projecting future municipal water usage. Forecasting water usage has become a vital tool for many tasks, such as monitoring, constructing, and running water systems. Issues with urban water management are covered, along with the right dimensions and operation of water treatment plants, dams, and pipeline ability. There are plans for new innovations and system expansions.

Ethiopia's capital city, Addis Abeba, is rapidly changing due to an increase in population and urban development. This growth translates to a constantly rising consumption for water resources in the city. To effectively manage this precious resource, Addis Ababa needs sophisticated forecasting tools. Genetic Algorithms (GA) and Artificial neural network (ANNs), are powerful technique within the realm of machine learning, are proving to be a game-changer. GAs excels at modeling intricate and nonlinear connections within time series data, making them ideally suited for tackling water consumption forecasting challenges in dynamic urban environments like Addis Ababa.

Genetic algorithms belong to the class of optimization algorithms, which find the optimum solution or solutions that minimize or maximize a particular function for a given computer problem. They are made up of regression equations, both nonlinear and linear, that process data and provide predictions. When used to forecast water consumption in Addis Ababa, various aspects are taken into account, including the city's size, population characteristics, the type and scale of businesses and industries, the environment, and the price of supplies (Altunkaynak, Özger et al. 2005). Data on every day values of climate factors, such as temperature in Celsius and precipitation in millimeters, are available to us in addition to water consumption figures. The way that people use water is influenced by all of these variables. Due to its direct impact on many water sources, like showers and water for gardens, temperature is particularly significant (Herrera, Torgo et al. 2010).

Artificial neural networks (ANNs) are a powerful machine learning approach that learns from data, unlike traditional methods that rely on pre-defined algorithms. This data-driven approach allows ANNs to automatically discover the relationships between input parameters and the desired output, such as water quality predictions. An important feature of ANNs is their capacity to continuously adjust to new data. As water quality data evolves, the ANN model can be updated to reflect these changes, ensuring ongoing accuracy. Generally, an ANN is composed of three layers: Layer of input: Data is fed into the system through this layer. Hidden layer(s): Here, networked nodes are used to process and analyze the data. Layer of output: Results are delivered here, scaled in accordance with the input data.

In this research four models were developed. The two-model developed by genetic algorithm (linear regression model and nonlinear regression model) and the last two model develop by ANN (with one hidden node and five hidden nodes). The input factors for the four models in this model development include the following: population, average temperature, relative humidity, precipitation, economic growth, industrial activity, construction activity, holiday and festival, and livestock count. The quantity of water actually used is also an output. As a result, the optimal water consumption is ascertained by a process of trial and error. This facilitates the new operators' understanding of how the process control parameters relate to one another (Baxter, Stanley et al. 2002).

This study proposes to do a detailed investigation of the relationship between water consumption and weather indicators, socioeconomic factors, and industrial factors. It does this by using relatively new techniques for water consumption prediction models, particularly Genetic algorithms and artificial neural networks (ANNs). Regarding the historical time series of weather and water use data were collected from the Ethiopian metrology agency and AAWSA's respectively, industrial activity, economic growth and construction activity were collected from Addis Ababa plane commission office and number livestock, population and number of holiday and festival data collected from Addis Ababa agriculture office, Addis Ababa statically agency office and Ethiopian calendar respectively.

1.2 Statement problem

Water consumption forecasting is a crucial aspect of urban planning and resource management, especially in rapidly growing cities like Addis Ababa, Ethiopia. The availability of an adequate and reliable water supply is essential for the well-being and development of the city's residents.

Water consumption forecasts is a key component of decisions made in sustainable development and urban planning. Research on urban water consumption has primarily concentrated on capitalization or the development of new water sources. However, there is a lack of fundamental knowledge regarding water consumption based on type/purpose in Ethiopian cities, including Addis Ababa. The method used for disaggregating water consumption data seems to be unspecified. Understanding the detailed consumption patterns within a city is crucial for accurate future consumption predictions. Water scarcity poses a significant challenge that impacts economic progress. Therefore, precise forecasting of water consumption is essential to ensure proper water supply management and address these issues. It is imperative to overcome all the aforementioned obstacles, particularly by establishing a robust model for disaggregating water consumption.

The challenge of meeting the necessary requirements for a water supply system, including pressure, quantity, and quality, is exacerbated by the intricate interplay between water supply and demand. Conventional forecasting techniques, primarily reliant on population projections, often struggle to account for the intricate and nonlinear nature of these variables. This limitation can result in ineffective distribution of water resources and infrastructure investments. Water consumption patterns are influenced by a range of factors such as climate, socioeconomic conditions, policies, urbanization, economic activities, agricultural practices, industrial operations, and strategic considerations. Given the variability of these factors across different locations and time periods, it is imperative to develop tailored models to forecast water consumption and consumption issues specific to each city.

Currently, numerous demand forecasting models exist in the literature, yet most are tailored to specific locations and circumstances, limiting their applicability to those regions. While these studies may offer relevance and transferability, the issue of water usage, both domestically and non-domestically, has grown more intricate due to the influence of climate and socio-economic changes, which are increasingly unpredictable, among other factors mentioned earlier. These relationships are not straightforward. It is crucial to select and comprehend the most suitable factors for a particular consumption disaggregation scenario. The study utilizes genetic algorithm (GA) models and artificial neural network (ANN) models to pinpoint crucial factors and utilize them in understanding water consumption by type in Addis Ababa city. The aim is to determine an appropriate water consumption disaggregation for Addis Ababa city.

1.3 Objectives of the Study

1.3.1 General objective

The general objective of this research is to develop and implement a Genetic algorithm and ANN model for accurate and efficient water consumption disaggregation to understand the main driving factors of water consumption in Addis Ababa, Ethiopia, with the aim of improving water supply management and distribution.

1.3.2 Specific objective

- To design and train GA and ANN model for water consumption disaggregation, taking into account various influencing factors.
- To give recommendations for the effective use of both GA and ANN model in improving water resource management in Addis Ababa city.
- To determine and evaluate the driving factors that influence water demand/consumption in Addis Ababa.

1.4 Research Questions

This serves as the basis for the following research questions, which this study aims to address after it is finished.

- Can genetic algorithm and ANN model be effectively used to disaggregate water consumption factors in Addis Ababa?

- How can the application of GA and ANN in water consumption disaggregation contribute to the sustainability and reliability of the water supply in Addis Ababa, Ethiopia?
- What are the significant factors that influence water demand/consumption in Addis Ababa city?

1.5 Variables

Independent Variables:

- Average temperature,
- precipitation,
- Relative humidity,
- Population,
- Economic growth,
- Construction activity,
- Seasonal factors (holidays, festivals),
- Industrial development rate and
- Number of livestock

Dependent Variable:

- Water consumption

1.6 Scope of the study

To meet the goals of the research, a number of projects were completed.

I. Geographical Focus: Ethiopia's capital city, Addis Ababa, is the subject of the study. It examines the elements that drive water consumption in this urban area and forecasts requirements related to water usage.

II. Time Frame: The study covers a specific time frame, such as the past decade or a period of historical water consumption data, and determine the contribution of the driving factors for the future prediction of water consumption,

III. Data Collection and Analysis: This research involve collected historical water consumption data, including factors that influence consumption (e.g., population growth, climate, economic factors). The study also examines the existing infrastructure, water sources, and management practices.

IV. Genetic algorithm (GA) and ANN: The primary focus of this study the application of genetic algorithm and ANN as a forecasting tool. This includes the selection of GA and ANN architectures, training algorithms, and data preprocessing techniques that are relevant to the specific context of Addis Ababa.

V. Predictive Variables: The study was considering various factors that influence water consumption, such as population growth, weather patterns, industrial activities, Agriculture activity, and urban development. The selection of relevant variables and their incorporation into the GA and ANN model were been explored.

VI. Model Validation: The research was evaluating the accuracy and reliability of the GA and ANN based water consumption -forecasting model. This was involve assessing the model's performance using appropriate validation techniques, such as cross-validation or split-sample testing.

VII. Comparison with Traditional Methods: The study may also compare the GA and ANN-based approach with traditional statistical and mathematical models that are currently used for water consumption forecasting and determination of driving factors in Addis Ababa city.

VIII. Policy and Management Implications: The findings been used to provide insights into the potential improvements in water resource management and infrastructure planning in Addis Ababa. This may include recommendations for optimizing water distribution, resource allocation, and sustainability measures.

IX. Ethiopian Context: The research was taken into consideration the specific challenges and opportunities related to water resources management in Ethiopia, including any relevant government policies or regulations.

X. Feasibility and Limitations: The study acknowledge the feasibility and limitations of implementing GA and ANN model-based water consumption forecasting in a real-world context, including data availability, computational resources, and technical expertise.

XI. Interdisciplinary Approach: The research may involve collaboration with experts in the fields of hydrology, urban planning, and data science to ensure a comprehensive analysis of the water consumption -forecasting problem.

XII. Recommendations: The study conclude with recommendations for policymakers, water utilities, and other stakeholders in Addis Ababa, suggesting how the findings can be practically applied to enhance water supply and demand management.

XIII. Future Research: The study may also identify potential areas for future research, such as the integration of advanced technologies like IOT (Internet of Things) and machine learning for real-time water consummation forecasting and driving factors determination.

1.7 Significance of the study

I. Water Scarcity Mitigation: Addis Ababa, like many urban centers in Ethiopia and around the world, faces water scarcity challenges due to population growth, urbanization, and changing climate patterns. Accurate water consumption forecasting is essential for effective resource management, helping to mitigate water shortages and ensure a sustainable water supply.

II. Urban Planning: The study's findings can aid city planners and policymakers in making informed decisions about infrastructure development, such as water supply and distribution systems. It can also assist in optimizing resource allocation and investment in water-related projects.

III. Resource Efficiency: Accurate water consumption forecasting and determine the driving factors can lead to better resource allocation and reduced waste, resulting in more efficient water usage. This has economic and environmental benefits, as it reduces the need for resource-intensive water infrastructure and the energy required for water treatment and distribution.

IV. Improved Resilience: Increasing the water supply system's resilience to extreme weather and climate change requires accurate forecasting of water usage. It can help prepare for and manage crises and interruptions in the water supply.

V. Sustainable Development: Sustainable water resource management is essential for achieving the Sustainable Development Goals (SDGs) as set forth by the UN. Goal 6—clean water and sanitation—is advanced by the data-driven plan presented in this report, which aims to increase access to safe, clean water.

VI. Technological Advancements: The employment of ANN and genetic algorithms in water consumption forecasts and driving factor analysis is a technological advance that can serve as a template for related research both locally and internationally. It provides examples of how cutting-edge technology is being used to solve urgent environmental problems.

VII. Local Relevance: Focusing on Addis Ababa, the study addresses a specific regional need, taking into account the unique socio-economic, environmental, and demographic factors of the city. It provides actionable insights for a specific geographic area while demonstrating the broader potential of the methodology.

VIII. Research Contribution: The research paper shows how to use ANN and genetic algorithms in practice for water consumption forecasts, adding to the body of knowledge in the fields of artificial intelligence and water resources management. Researchers, scholars, and professionals in relevant fields can use this as a reference.

1.8 Research Design Framework

Research Design Frameworks:

1. Introduction:

- Give a summary of the research issue,
- Describe the goals of the research and
- Explain why the study is important.

2. Literature Review: -

- Review existing literature on water consumption forecasting methods.
- Discuss the strengths and weaknesses of traditional models.
- Explore previous applications of genetic algorithms and ANN in similar contexts.

3. Research Methodology and data Collection and Preprocessing

Research Methodology

- Explain the data collection process: how historical water consumption data and relevant environmental variables gathered,
- Describe the genetic algorithm and ANN parameters,
- Explain how the genetic algorithm and ANN customized for water consumption forecasting
- Discuss how model performance evaluated, considering metrics like Root Mean Square Error (RMSE)

Data Collection and Preprocessing:

- Present the data sources and collection methods,
- Show the characteristics of the collected data, including a summary of historical water consumption patterns in Addis Ababa.
- Describe the procedures used to purify and preprocess the information, taking into account how to deal with outliers and values that are absent.
- ANN Model and Genetic Algorithm:
- Describe the parameters of the GA and ANN design that was selected.
- Describe the model training procedure, taking into account the selection of optimization techniques and activation functions.
- Validation and Model Performance split the dataset into training testing sets and validate the model performance against historical data.
- Utilize suitable metrics to assess the efficacy of the GA and ANN models.

4. Result and Discussion:

- Interpret the results and discuss their implication for water resource management in Addis Ababa,
- Address the research objectives and
- Compare the findings with the existing literature and provide insights into the unique challenges of water consumption forecasting in Addis Ababa.

5. Conclusion:

- Summarize the key findings of the study and
- Discuss the practical applications and policy implications of using GA and ANN in water consumption forecasting for Addis Ababa.

6. Recommendations:

- Make suggestions for Addis Ababa's water resource management based on the study's findings and

- Make recommendations for possible directions for further study and development in the forecasting of water consumption.

7. References

- Cite all the sources used in the research.

8. Appendices:

- Add any more resources you have, such as data sources, code snippets, and further research information.

1.9 Summary of the Research Design Framework

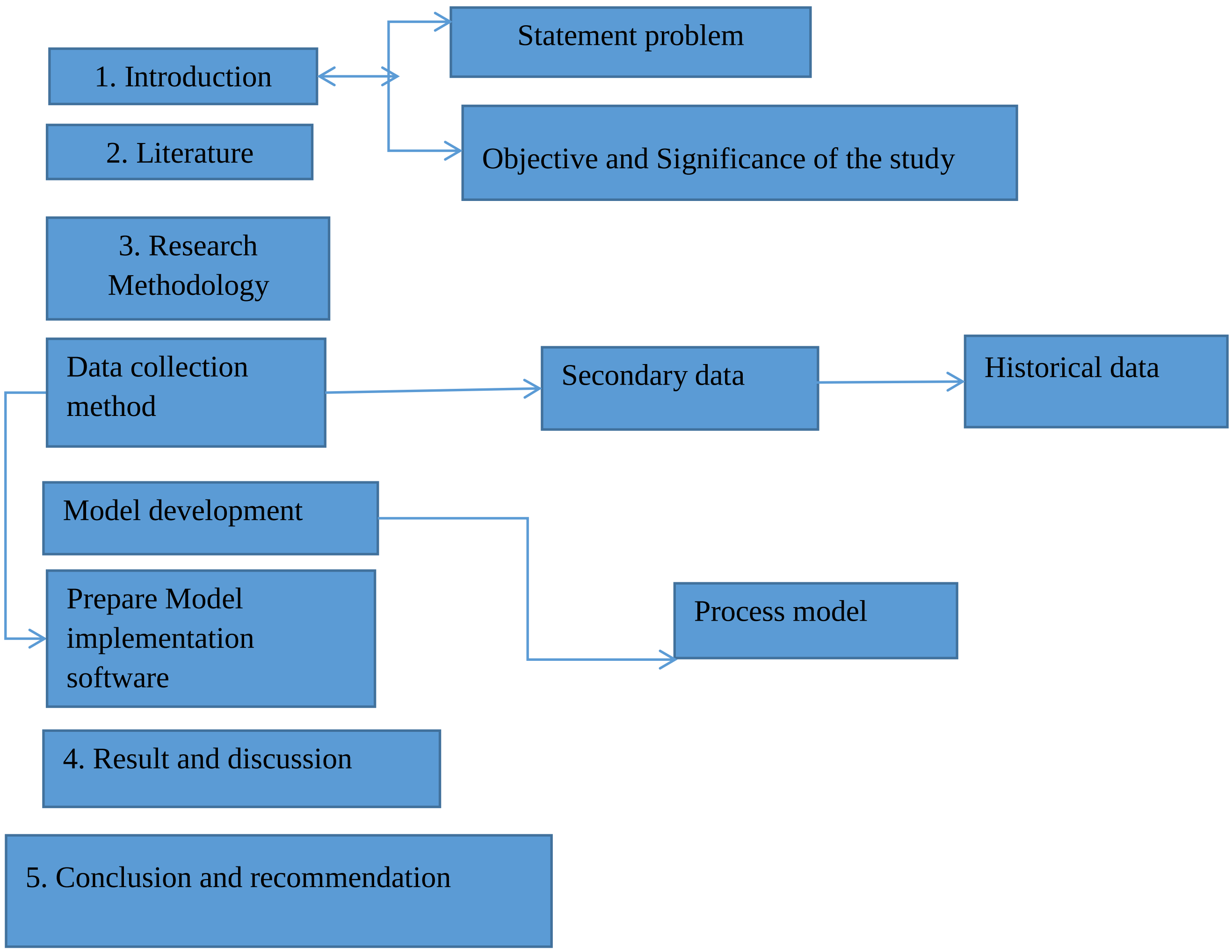

Figure 1-1 Structure of Research

2. Literature Review

2.1 Introduction

Following is a review of pertinent literature on topics like evaluating the current status of water consumption, forecasting water consumption and consumption criteria, using genetic algorithms and artificial neural networks to forecast water consumption, and making use of detailed design documents.

The examination of the literature reveals that modeling water consumption and other aspects of the water resources system are frequently predicted using GA and ANN methodologies.

Many methods of forecasting water consumption make the assumption that variables have a linear relationship with one another; nevertheless, the relationships between the variables that characterize the fluctuation in water consumption frequently show nonlinear behavior. It has been demonstrated that using GA with ANN to analyze a range of nonlinear time series events, such as electrical load forecasting and water consumption forecasting, is a successful method.

The choice of forecasting technique for water consumption prediction is influenced by various aspects, including the extent and additional properties of the utility and its supply territory, as well as the anticipated applications of the forecast data. The amount of work needed to forecast usually increases with the utility and the importance of the decisions that will be impacted by the forecast outcomes (Billings and Jones 2011).

The most common method for estimating consumption for the next day or month used to be to use the profile from the day before or month, or the identical day from the preceding week or month. This method's primary flaw is that it fails to take into account the underlying factors that influence consumption patterns on a certain day or month. It is plausible that the confluence of societal and meteorological elements contributing to the recorded consumption on a certain day may not be valid on the day the forecast is required.

2.2 Water consumption /demand forecasting methods

Accurate data on the present population, trends, and water sales and revenues are essential for any forecasting method to employ.

While it aims to explain the underlying processes, actual-based modeling, also known as knowledge-driven modeling, is the conventional modeling of physical processes. However, the so-called data-driven models, which significantly draw from Artificial Intelligence (AI) approaches, are dependent on data that describes input and output features and have a limited understanding of the modeling process. Because this approach may abstract and generalize the process, it frequently complements physically-based models. According to one conclusion, data-driven modeling leverages findings from overlapping domains like artificial neural networks (ANN), data mining, rule-based techniques like expert systems, genetic algorithm concepts, rule-induction, and machine learning systems.

The most popular method, time series forecasting, is based on directly identifying trends seen in previous data on water consumption. To model water use, researchers have employed a variety of techniques, including genetic algorithms (GA), autoregressive integrated moving averages (ARIMA), exponential smoothing, and regression. Some have incorporated weather, demographic, and economic variables into their estimates of water demand. Identifying the relationship between the amount of consumption and important meteorological variables, GA was utilized to forecast water consumption (Hartley 1995).

methods of water demand /consumption forecasting: -

- **Per capita use**: - it determines by Multiply future population by per capita use. The data use for this method Population projections and Historical billing and population data. This method advantage is simple data and simple calculation and the disadvantage of this method also assumes there is a strong correlation between population and demand-driven factors.
- **Extrapolation: -** Predicate future water use by extrapolating past data. 10 to 20 average annual demand data require. This approach uses straightforward statistical techniques and requires little data.
- **Unit use per customer type: -** It determine by Project growth in demand by customer category. The method also requires Water billing by customer type, Demographic data by customer type and Projected growth for each customer. The advantage of this method is Disaggregation addresses more trends in water use and the disadvantage is Call for more thorough data and analysis.

Most recent research on the topic of water consumption forecasting has focused on using mathematical algorithms to registries of past consumption numbers and meteorological variables that have a causal relationship. In situations where extraneous non-cyclic factors have negligible or no effect, the accuracy of the fundamental mathematical models functions as intended. Two instances of water consumption customs are artificial intelligence and time series models.

GA and ANN are the two main ideas that make up artificial intelligence. When modeling and simulating a system with one or more variables, GA and ANN have been shown to be beneficial (Bozokalfa 2005).

2.3 Genetic algorithm (GA)

The functioning of immune systems and biological evolution seem to be influenced by mechanisms that are somewhat approximated by evolutionary algorithms. Fundamental to these evolutionary systems is the concept of a genotype population that constitutes components of a high dimensional (Baeck, Fogel et al. 2018).

It is a type of scanning approach that simulates natural evolution and is applied to optimization and search issues to produce practical answers. Evolutionary algorithms, which use methods inspired by natural selection, mutation, inheritance, and crossover to solve optimization problems, are a subclass of genetic algorithms.

A GA is an optimization algorithm that uses parallel random search to imitate both the theory of biochemical evolution and the natural genetic mechanism. Numerous fields have utilized this algorithm (Huang, Zhang et al. 2022). The unique feature of GA is that it incorporates the "survival of the fittest" idea of biological evolution into the parameter optimization process. Based on the selected fitness function, individuals are screened utilizing three basic techniques (i.e., crossover, mutation, and selection). Individuals with high fitness values are retained, while those with low fitness values are eliminated. In this way, not only do the new people gain the knowledge of the older generation, they also have an advantage over them. The three aforementioned actions are repeated till the requirements are satisfied.

2.3.1 Components, Structure, & Terminology of GA

Since most of the terminology used in genetic algorithms is taken from biology, they are intended to mimic biological processes. The terms used in genetic algorithms, however, pertain to far simpler entities than their biological counterparts (Mitchell 1995). The fundamental elements shared by nearly all genetic algorithms include a population of chromosomes, a fitness function for optimization, the choice of which chromosomes will replicate, which chromosomes will overlap to form the generation that follows, and if chromosomes in the new generation will randomly mutate.

2.3.2 Selection of input parameters

Research is necessary to determine the set of theoretically available input variables to use as pattern inputs, and this is a crucial step in the model-building process. In data-driven models such as GA, the input variable is very important because it determines the majority of the model's performance (Fernando, Maier et al. 2005). The model is susceptible to not converge to the ideal value soon or at all when the input parameters are chosen poorly, as is frequently the case. As a result, choosing the appropriate input parameter values is essential.

The corresponding parameters are those that reduce redundancy or identify which parameters have no proper relationship to the output and should be excluded (Bowden, Dandy et al. 2005). In order to accomplish this, the model is run with all parameters set and without performing a sensitivity analysis. The relationship between the output and the data's accessibility is the main factor considered while choosing the input parameters.

2.4 ANN Model

Artificial neural networks, often known as ANNs, are a type of machine learning technique that mimic the way the human brain uses an interconnected structural network to retain information. (Haykin 2009). In 1943, McCulloch and Pitts made the initial neural network proposal (Baxter, Stanley et al. 2002). The mechanism involves modifying the synaptic weight value between the network's nodes or neurons. An intricate picture of the complex link between input and output can be created by a neural network because of its capacity to learn from past data.

A wide range of problems, including forecasting, prediction, and procedure management involving multiple and independent fluctuations of several inputs—variables that may be unknown or challenging to characterize—have been addressed by ANN. This is the reasoning behind the naming of the statistical (black box) model (Loucks and Van Beek 2017). Furthermore, ANNs can be used in broader contexts when they are exposed to distinct input designs that were not included in the preparation information set (Hunt, Sbarbaro et al. 1992). A system's capacity for generalization is its capacity to react to input that has never been observed before, in contrast to the human brain's capacity for inductive reasoning to produce order.

2.4.1 Basic structure and component of ANN

The main parts of an artificial neural network (ANN) are the propagation rule, learning rule, transfer function, connection weight between neurons, and perceptron, or processing units (Baxter, Stanley et al. 2002).

2.4.1.1 Layer and neurons

A neuron is a neural network's main building block. The activation function of a non-linear neuron receives the input signal Xi multiplied by the specified weight value Wi and its sum. Wi, summation, and the activation function are the three crucial components (Balakrishnan and Weil 1996).

$Z = \sum WiXi$ Equation 2.1

$Y = f(z)$ Equation 2.2

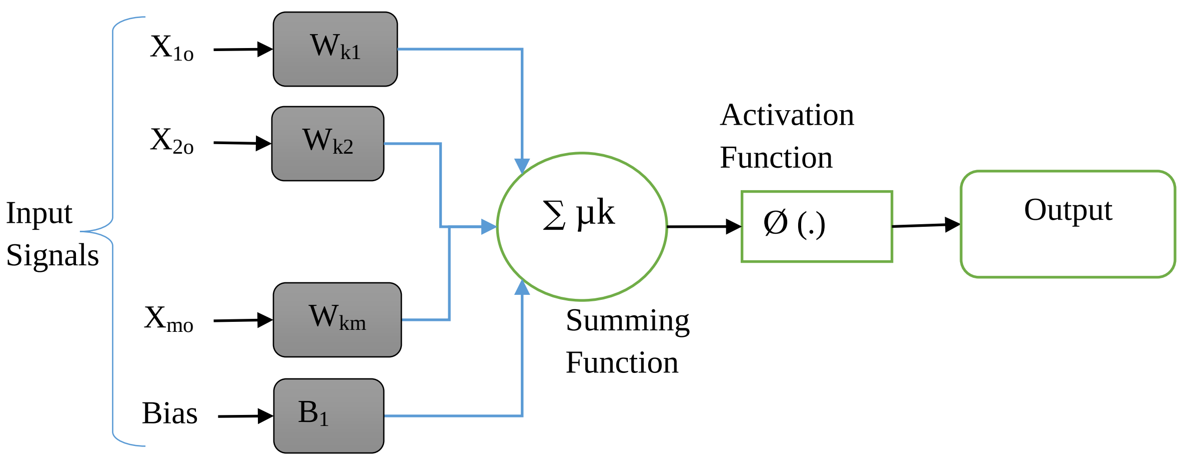

Figure 2-1 nonlinear neuron model: Source (Fuangkhon 2014)

Brain cells make up the three levels of a multilayer perceptron. The neurons of the input layer, which is the first layer, are not real working components; rather, they are nodes that receive signals from external sources. The input value is simply changed to enable the network to handle ranges of [0,1] or [-1,1] (Baxter, Stanley et al. 2002). The middle layer is referred to as the hidden layer as the user cannot see it. Its function is to transfer data from the input layer or preceding concealed layer to the output layer of the subsequent neighboring layer. In order to reduce the network's stiffness and boost convergent ability while using fewer link weights, it is possible to employ many hidden layers (Maier and Dandy 2000). Receiving information from the hidden layer and providing the user with a predicted value for the output are the responsibilities of the node that makes up the output layer (Loucks and Van Beek 2017). One output variable is all that the ANN models can predict with more accuracy, although several output variables can be modelled (Baxter, Stanley et al. 2002).

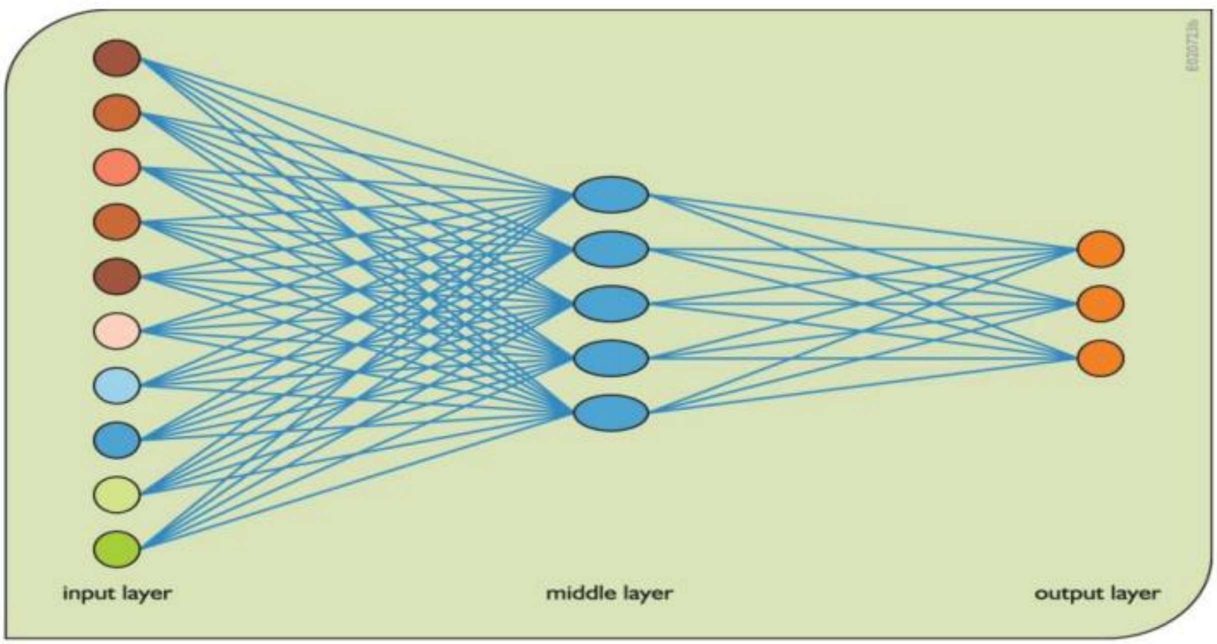

Figure 2-2 A representation of a standard multilayer artificial neural network with the hidden layer (s) and the output layer (with three output sources) each having ten distinct inputs each:(Loucks and Van Beek 2017)

2.4.2 Weights

In a neural network, every neuron is connected to every other neuron by connections that are stronger or have more weight. Weighted or strong connections connect each neuron in the neural network to every other neuron. These connections are first established at random and can be changed. It is always initially set at random and then modified.

2.4.3 Activation function

The neural network uses activation functions to calculate the weighted total of input and bias. Using one of the many different kinds of activation functions that are available for use in neural network architecture is the best way to handle regression problems. These functions include the simpler bias axon function as well as the more complex logistic, hyperbolic tangent, binary, and

Gaussian transfer functions (Ravn, Hansen et al. 2000). The primary purpose of a transfer function is to restrict the output value to the intervals [0-1] and [-1,1]. The data inputs were immediately scaled from -0.9 to 0.9 (Haykin 2009). It is regarded as a single-layer artificial neural network (ANN) model since the linear activation functions are merely one-degree polynomials. Its inability to accurately replicate complex data is the reason it is not widely used. After multiplying the input by each neuron's weight, it produces an output signal that is proportionate to the input (Mohri, Rostamizadeh et al. 2018). The most widely used transfer functions in neural networks are non-linear functions like the sigmoid function, which combines the logistic and hyperbolic tangent transfer functions (Maier and Dandy 2000). The sigmoid transfer function helps lessen the computational load that the network encountered during back propagation, and it is primarily utilized in feed-forward neural networks. The growth is smooth from 0 to 1. The hyperbolic function is bounded by - 1 to 1, yet it has the same form as the sigmoidal function. It's referred to as the zero-centered function and the tan function. The asymmetric functionality makes it possible for the boundary to be negative in relation to positive, which speeds up the network's training.

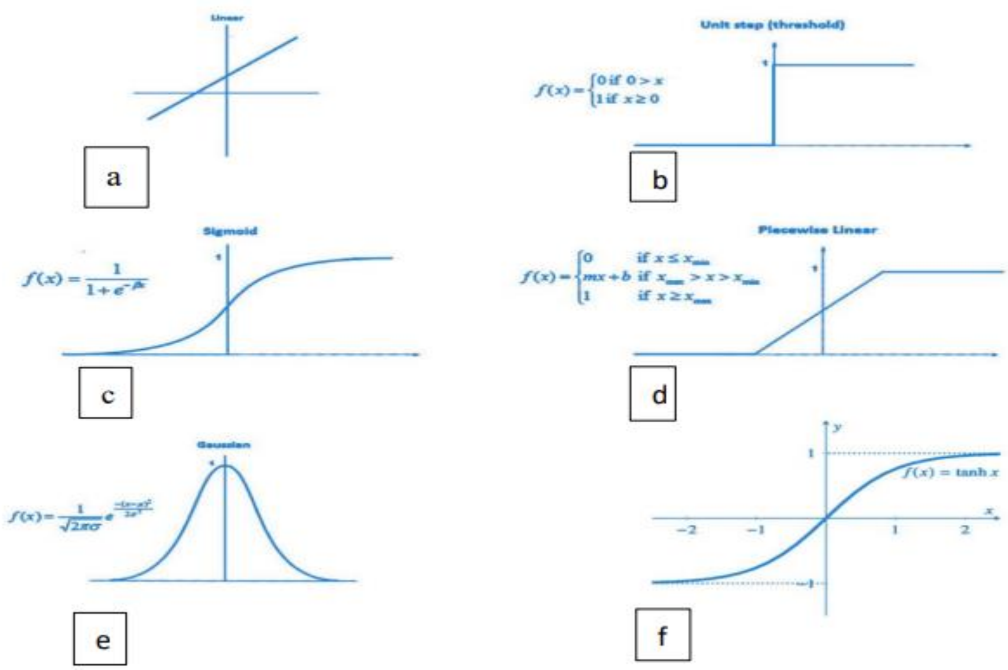

Figure 2-3 Activation functions: a) tangential; b) sigmoid; c) piecewise linear; e) Gaussian; and f) linear (created with Adobe Photoshop)

2.4.4 The learning guideline

Neural networks learn by adjusting connections between artificial neurons, similar to how biological neurons strengthen or weaken connections. This adjustment process is guided by learning rules, which determine how the strengths of these connections, called weights, are modified during training. Neural networks employ four basic categories of learning rules:

- **Error correction learning:** The most common type involves changing the weights in accordance with the discrepancy between the network's real and predicted outputs.
- **Boltzmann learning:** By using this rule, the network can learn unsupervised and identify patterns in the incoming data.
- **Hebbian learning:** Inspired by how biological neurons learn, this rule strengthens connections between neurons that are active at the same time.
- **Competitive learning:** This rule is used in unsupervised learning for tasks like feature detection. It encourages competition among neurons, with only the "winning" neuron strengthening its connections for a particular input (Basheer and Hajmeer 2000).

2.4.5 Training function

Training is achieved through a training algorithm. This algorithm acts as the overall method that guides the network's learning process. Its primary function is to modify the network's weights and biases. By iteratively fine-tuning these values, the training algorithm allows the network to learn how to map specific inputs to desired outputs.

2.5 ANN Model development

Building an artificial neural network model with the MATLAB toolbox, MATLAB code, or other neuro solution application requires following a few fundamental steps. The user should generally go step by step. The fundamental ones include gathering and analyzing data, statistical analysis, choosing input and output parameters, choosing an architecture, training, optimizing network parameters, and assessing network performance and stability (Baxter, Stanley et al. 2002). On the other hand, choosing parameters may come after training by figuring out which model parameters are crucial or sensitive. It can be difficult to determine which input parameters are crucial, redundant, or unnecessary until the very end, depending on the methodology employed. Before considering input, output, and crucial factors, the procedures for gathering and evaluating data must be taken into account. The trial-and-error method is the most widely used way for figuring out the optimal neuronal value and layout, transfer function, learning rule, weight initialization, learning rate, and momentum (Maier, Morgan et al. 2004).

2.6 Data analysis

This step focuses on preparing the raw data for effective use in training the neural network. It follows these key practices outlined by (Baxter, Stanley et al. 2002):

1. **Data Exploration and Outlier Detection:** Before feeding data into the network, it's crucial to meticulously examine it. This involves checking for outliers (extreme values) that might skew the training process. Furthermore, measurements of central tendency (such as mean or median) are computed to comprehend the overall distribution of the data. To understand each parameter's qualities, each one in the data is likewise thoroughly characterized.

2. **Normalization:** Raw data inputs often benefit from normalization. The data is transformed by this process into a particular range, usually ranging from 0 to 1 or between -1 and 1. Normalization makes sure that every feature has a same scale, which improves network speed.

3. **Data Splitting:** Three separate sets are carefully selected from the gathered data:

- **Training Set (largest portion):** This set forms the backbone of the training process. The network is exposed to this data and learns from it to establish the mapping between inputs and outputs. (e.g., 50% or 60% of the data)

- **Validation set (medium portion):** utilized during training to adjust the hyper parameters (learning rate, momentum, etc.) of the network. The validation set helps prevent overfitting by allowing adjustments to avoid the network memorizing the training data instead of learning the underlying patterns. (e.g., 30% or 20% of the data)

- **Test Set (smallest portion):** This unseen data is employed for the final evaluation of the network's generalization capability. The network’s performance on this set reflects its ability to handle new, unseen data effectively. (e.g., 20% or 15% of the data).Common data splitting ratios mentioned in literature include 5:3:2 or 3:1:1, as suggested by (Baxter, Stanley et al. 2002)and (Rodriguez and Sérodes 2004). The ideal ratio may change based on the particular issue at hand and the size of data.

- **Input Selection (Dimensionality Reduction):** If two input data points appear functionally similar, it might be redundant to include both as separate inputs. In these situations, the model is streamlined and may become more efficient because only one of the two is selected as an input to the network.

2.7 Assessment of prediction accuracy

Two criteria that are frequently used in water consumption forecasting are utilized in order to examine the forecast accuracy of prediction models. These are the mean absolute percentage error (MAPE) and the root mean square error (RMSE). The following are the mathematical formulas for these metrics (Donkor, Mazzuchi et al. 2014):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y'_i)^2}{n}}) \qquad \text{Equation 2.3}$$

$$MAPE = \frac{1}{n} \times \sum_{i=1}^{n} \frac{|Yi - Y'i|}{Yi} \qquad \text{Equation 2.4}$$

Where,

At time i, the observed values are represented by Yi and Y'i, respectively, while the number of observed values or forecast values is denoted by n.

2.8 Factors that influence water consumption / demand

Four categories comprise the elements that affect water consumption: economic, sociodemographic, climatic, and regulatory/ordinance. Water consumption models ought to consider the impact of climate change by incorporating variables such as temperature and precipitation (Cabral, Loureiro et al. 2014). According to Ferreira, consumption outside is highly susceptible to temperature and other aspects of the atmosphere. The next variable that will impact water usage.

Economic: -

- Pricing arrangement: water basic component, billing term, and residual block cost

Socio demographic: -

- Population: Total individuals living in the home and the number of dependents in each family
- Education/Knowledge- percentage of the total population that is composed of graduates, 12-year-olds, and university students
- Income: Real, median, per capita, income per household member, average household income, and monthly income

Climate: -

- Temperature- Mean monthly temperature, mean maximum daily temperature, average annual temperature (°C), and average summertime temperature
- Rainfall- Monthly totals, average monthly precipitation, cumulative monthly rainfall, and annual rainfall (mm) throughout the summer
- Relative humidity

Rules and Regulations: -

Regulation and limits on water: - Total number of hours per day of restrictions

3. MATERIAL AND METHODOLOGY

3.1 Overview of the Research Area

In 1889, Addis Ababa became a city. Located at the foot of the Entoto Mountains, the city is part of the watershed of the Awash Basin. The population of Addis Abeba is 2,739,551 people, both urban and rural, according to the 2007 national statistics authorities' population census. In the recent years Addis Ababa's city population has been increased by more than 4% over year with the current estimated city's population standing at about 5.2 million.

3.1.1 The Research Area's Location

The latitudes of Addis Ababa are 23°21′ N and 23.35°N, and the longitudes are 85°20′ E and 85.33°.

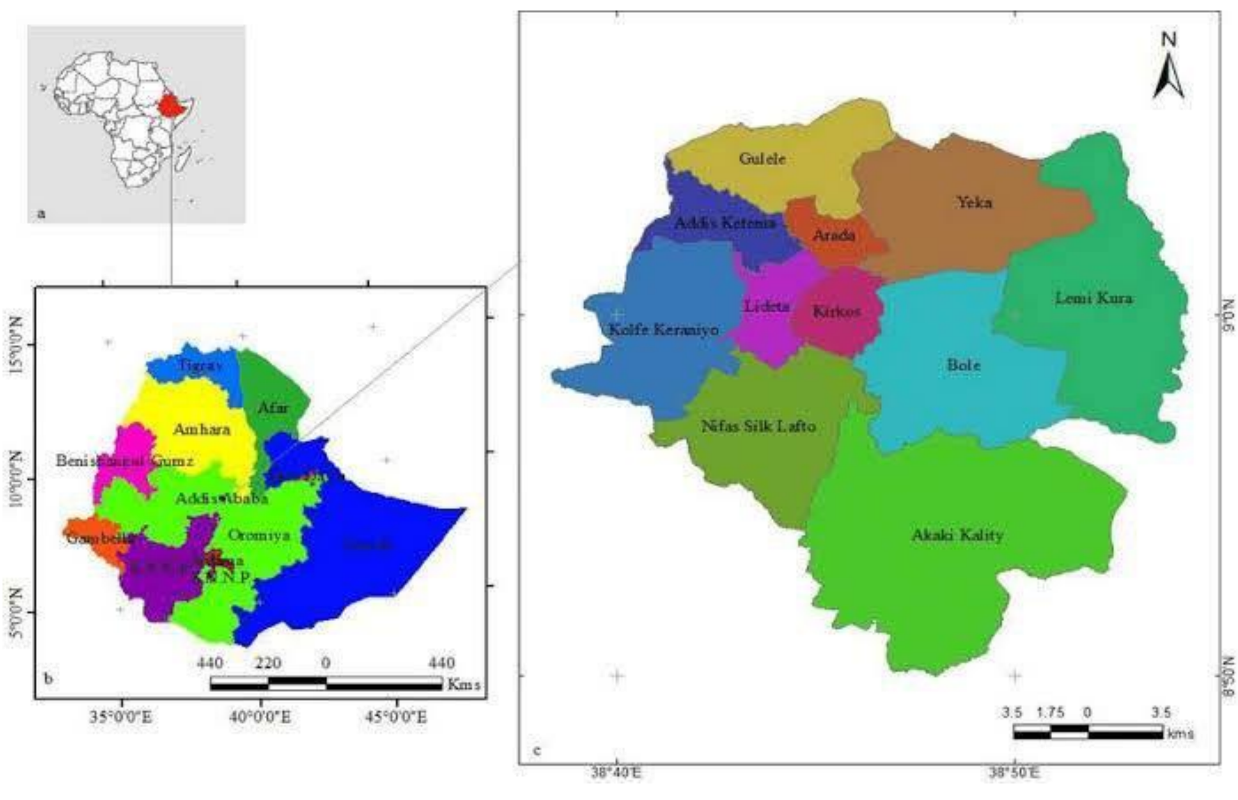

Figure 3-1 location of study area

3.1.2 Climate characteristics of the study area

In the city, the mean annual maximum temperature is about 24 °C, while the mean annual minimum temperature is about 12 °C. The largest monthly mean rainfall (around 260 mm) occurs in July and August. The mean annual rainfall in Addis is approximately 1255 mm (Ababa 2007).

3.1.3 Study Area Land Use and Land Cover

There are 527 km2 that make up the city of Addis Ababa. The city's land use data shows that 45% of the area is used for residential purposes, with the remaining 20%, 25%, 5%, and 5% going toward green space, roads, open spaces, and other land uses, respectively (Fenta 2014).

3.2 Data collection

Data is gathered across various categories including water consumption, social factors, economic factors, and industrial and climate factors. Table 3.1 provides a detailed description of all secondary data sources and the organizations associated with them.

Table 3-1 input and output data source

No	Data	Unit	Source
1	Water consumption data	M3/month	Addis Ababa Water and Sewerage Authority
2	Temperature	° c	Ethiopian Metrological Agency
3	Relative humidity	%	Ethiopian Metrological Agency
4	Precipitation	mm	Ethiopian Metrological Agency
5	Industrial activity	%	Addis Ababa City Administration Plan and Development Commission (AAPDC)
6	Economic activity	%	AAPDC
7	Construction activity	%	AAPDC
8	Number of livestock	Number	Addis Ababa Agricultural office
9	Population	Number	Addis Ababa statically agency office.
10	Holiday & festival	Number	Ethiopian Calendar

This data collected from 2015 (June) to 2023 (August).

According to the chosen study region, the following variables are thought to be driving or influential factors in water consumption. Although there may be additional influences, only those components are included in this analysis due to resource constraints and data availability.

Dependent variable

- WC: water consumption (m3/month)

Independent variables:

- IA: industrial activity (%),

- EG: economic growth (%),
- POP: number of population (number),
- TEMP: average temperature (o} C),
- PRE: precipitation for month (mm),
- CA: construction activity (%),
- RH: relative humidity for month (%),
- LS: Number of livestock for month and
- H&F: holiday and festival

3.2.1 Industrial development

The kind of enterprises that are present in the city mostly determine how much water they need. Depending on how quickly their output is rising, different industries, such as sugar refineries, breweries, cotton mills, and paper mills, require varied amounts of water. Therefore, it is assumed that their annual percentage growth represents their water use.

3.2.2 Holidays/festivals

The amount of water consumed is influenced by the occurrence of holidays and festivals during the month. In this study, holiday and festival information is utilized as a factor to account for fluctuations in consumption. The data was gathered by pinpointing these special occasions in the Ethiopian calendar within this research period (June 2015 to August 2023) for analysis.

3.2.3 Economic developments

The city's economic development has an impact on the pattern of water usage. The economic growth of the city, as expressed as a percentage derived from AAPDC, corresponds with the historical data required for the model.

3.2.4 Agricultural activities

Livestock presence in urban areas has a direct impact on water consumption levels in cities. The data regarding livestock numbers was obtained from the Addis Ababa Agriculture Office for the analysis period.

3.3 Filling of Missing Meteorological Data

Meteorological information gathered every day, such as temperature, relative humidity, and precipitation. The monthly data was obtained by analyzing this daily data. Sometimes, an equipment malfunction or the absence of an observer might result in the rainfall total at a particular station for a specific day or days being missing. In these situations, it might be necessary to approximate the value using the data of the neighboring stations in order to estimate the missing meteorology amount. The following techniques were typically used to calculate the temperature, relative humidity, and missing rainfall data.

The first three stations that is, station X are chosen based on how close they are to and how far away they are from the station with the missing record. At this point, data is collected for these three stations (i.e., 1, 2, and 3) on the day that station X's data is absent. The average annual rainfall data for each of the four stations must also be known. If the average yearly rainfall at each of these three index stations differs by less than 10% from the average yearly rainfall of the station X (i.e., the station with missing data), then we can estimate the quantity by taking an easy arithmetic average of the precipitations (corresponding to the missing precipitation) at the three index stations. The average annual rainfall at stations 1, 2, 3, and X is represented by N1, N2, N3, and Nx, while the corresponding precipitation data for the day for which the data is missing at station X is represented by P1, P2, P3, and Px , then we obtain

$px = \frac{p1+p2+p3}{3}$ ……. [Given that N1, N2, and N3 differ by 10% of Nx] Equation 3.1

Nonetheless, the normal ratio approach is applied if the average yearly precipitation at any of the three sites deviates more than 10% from the station under consideration. The three index stations' average yearly precipitation ratios 000are used to weight the amounts at each station in this manner. Therefore, in this instance, the missing precipitation data Px will be provided by

$px = \frac{1}{3}(P1\frac{Nx}{N1} + P2\frac{Nx}{N2} + P3\frac{Nx}{N3})$ ….. [Provided any of N1, N2, and N3 differ from Nx by more than 10%] source (Garg 2005) Equation 3.2

Simple arithmetic average was used to fill in the gaps in the research area's meteorological variables based on the previously discussed methods.

3.4 Materials and Software

A number of historical data such as like Temperature, Rainfall, Relative humidity, economic development, agriculture activity, industrial activity, holiday and festival and construction activity were used for this research. These data and were analyzed using:

- Geographic Information System (GIS) software for spatial analysis and mapping

- Excel spreadsheet for data analysis and manipulation Software such as easy fit to GA and ANN modeling
- XLSTAT for outlier analysis
- MATLAB for coding and adoption of the GA and ANN analysis for consumption disaggregation analysis.

3.4.1 XLSTAT for outlier analysis

XLSTAT is a powerful statistical software add-on designed to work seamlessly within Microsoft Excel. It offers a comprehensive suite of statistical tools, including a robust set of functionalities specifically designed for outlier analysis. It is a valuable tool for researchers, analysts, and anyone working with data in Excel who needs to identify and address potential outliers in their datasets. Use this software to evaluate the model's outliers in the input and output data for this study.

3.5 Data Quality Analysis

Data on water use are frequently incorrect containing missing values, seasonal variations, and trends. For this reason, proper water consumption disaggregation analysis depends on the efficient cleaning and preparation of raw data. Given this, compiling a useful set of input data to be employed in model construction is the primary goal of this preprocessing. Data on water consumption is available from June 2015 to August 2023. It isn't expected that the average, maximum, and lowest flow values, nor the number of domestic and nondomestic uses, will exhibit abnormal variations in the trend gradient of consumption values. The presence of outliers affects these unusual consumption values.

Outlier detection analysis is conducted on all input data, with a focus on detecting outliers in water consumption. The data collected reveals a range of water consumption varying from 4,895,988 m3/month to 9,611,861 m3/month, and an average consumption of 7,830,544 m3/month. In Addis Ababa city, the number of customers for domestic and non-domestic uses ranges from 41,545 to 618,316. Analysis indicates a noticeable difference in consumption patterns across different months (refer to Figure 3.2). A hyperbolic equation is identified as the best-fit model for consumption, with an R2 value of 0.4792, indicating a weak correlation in the raw data. Hence, thorough data cleaning is essential.

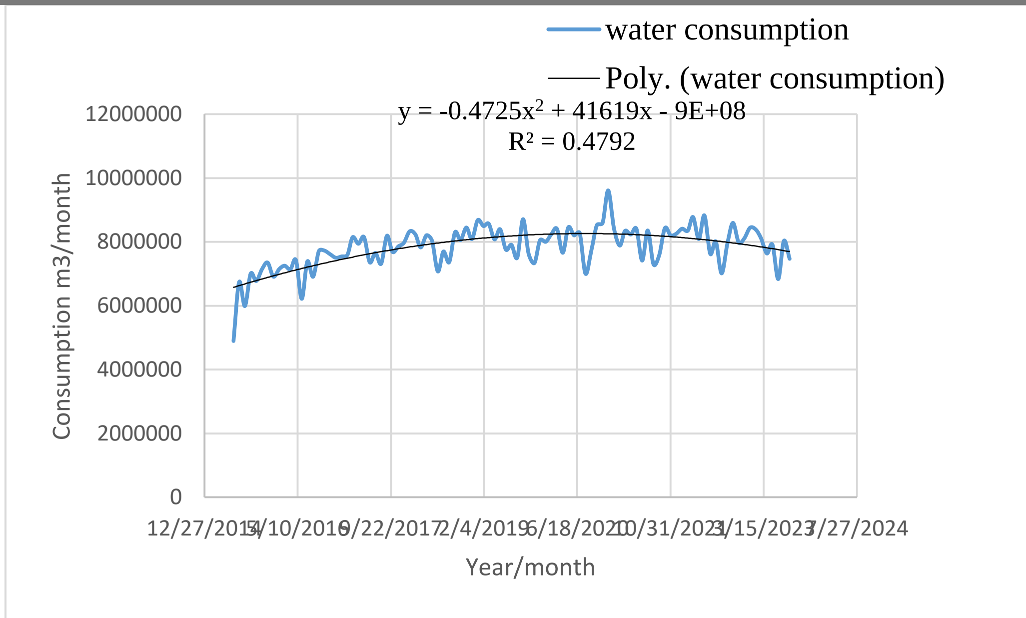

Figure 3-2 monthly total billed water consumption

Any observation that differs from the majority of water consumption for any reason including recoding errors, unavoidable circumstances, and faults in data collection is considered an outlier. It is this problem that makes fitting the distribution to the data difficult.

Assessing the data's quality is the aim of statistical research. The technique of reviewing the data to look for anomalies and inconsistencies and improving its quality is known as data quality checking.

Identification of Dixon test outliers was carried out both before to and following data entry and analysis. Determining if the highest or lowest value meets the criteria for being an outlier is the aim of this test. It is assumed in this test that the data correspond to a sample taken from a population that is regularly distributed. To assist in identifying potential outliers, XLSTAT displays Z-scores. The following formula shows how Z-scores and the standardized sample are related.

$$Zscore = \frac{Xi - Xmean}{s} \qquad \text{Equation 3.3}$$

Where,

s= is standard deviation

Xi= observed data

Xmean = mean of the observed data

The value of z between negative two and two (-2, 2) are deemed to be acceptable. All result of the outlier by Dixon test to see in the appendix 1.

The detection of outliers for the billed water usage and climatic variable of Addis Ababa city prior to data entry is displayed in Figures 3.3 to 3.6. The months of August 2015, September 2015, and December 2020 were the ones with the highest consumption figures, which were considered outliers. To get the clean flow time series data, the discovered outliers were averaged along with the outlier data from the flow time series.

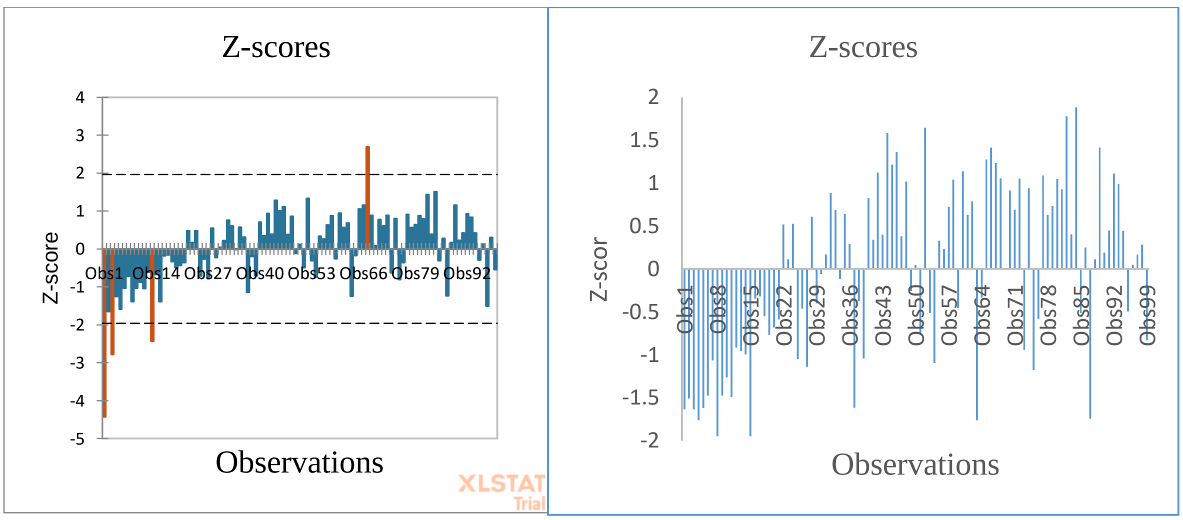

Figure 3-3 water consumption with and without outlier

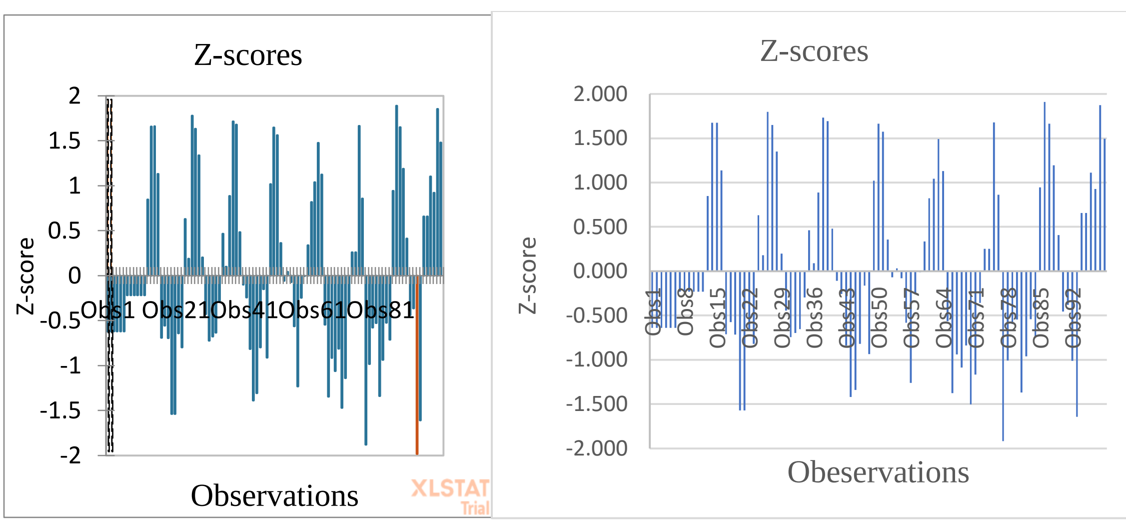

Figure 3-4 Relative humidity with and without outlier

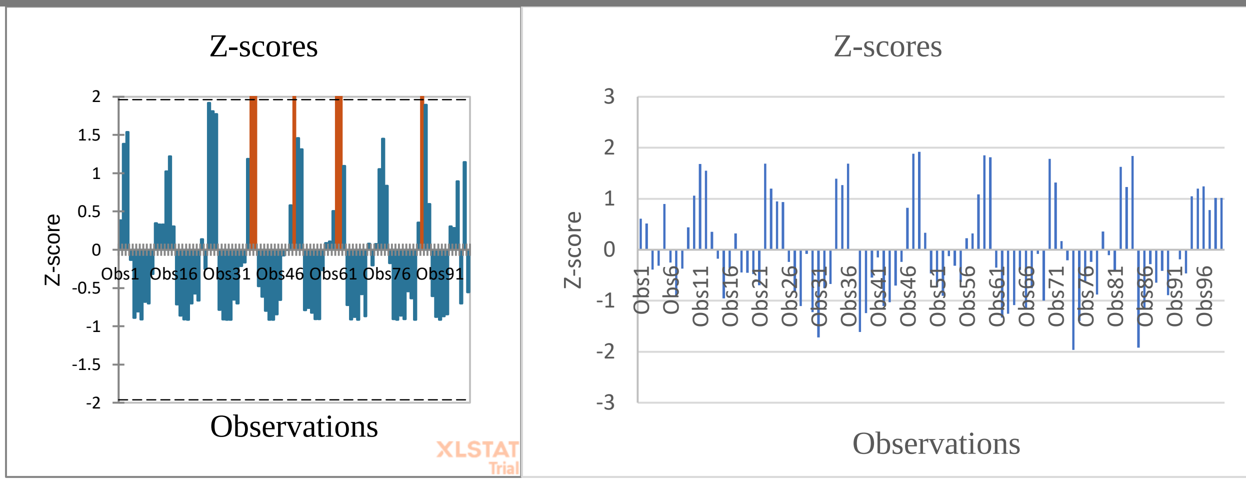

Figure 3-5 precipitation with and without outlier

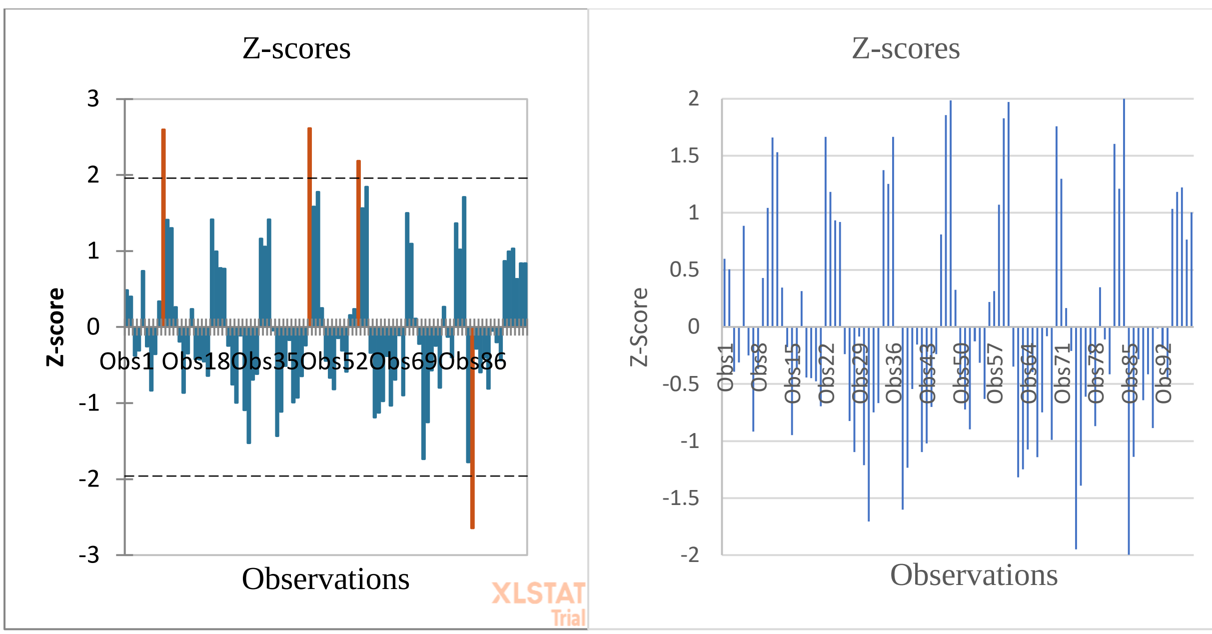

Figure 3-6 average temperature with and without outlier

Figures 3.3, 3.4, 3.5, and 3.6 demonstrate the absence of any outliers in the data, indicating its suitability for direct utilization in subsequent analysis.

3.6 Software for modeling

The software utilized for the creation procedures of the GA and ANN models was MATLAB version 2013a. This software provides the capability to write code based on both linear and nonlinear regression, allowing for modifications within a unified environment. Additionally, the

MATLAB Toolboxes can be employed to create GA (linear and nonlinear regression) and ANN (one hidden layer with one and five nodes) models within this software.

3.7 Input variables preparation for GA and ANN model

When determining the model's input parameters, special consideration was given to the selection of the parameters that have a direct impact on the amount of water consumed. Accessible data sources include a time series covering 99 months of water use (August 2015 to September 2023), average temperature, population, relative humidity, economic growth, industrial activity, construction activity, number of animals, and holidays and festivals. Since the input data was measured in different units, normalization was necessary. To convert the dataset to dimensionless units, the data matrix including variables measured in different units must be normalized. Normal distribution and min-max procedures are two different types of normalization techniques. For this research use min.-max. Method by using the following equation.

$$Y = \frac{(Xi - Xmin)*(Ymax - Ymin)}{Xmax - Xmin} + \text{Ymin} \qquad \text{Equation 3.4}$$

Here, xi stands for the raw dataset, ymin and ymax are the new values that are assigned in accordance with the activation function, and y is the normalization value. Xmin and Xmax are the original minimum and maximum values (achievable from the statistical information in the raw data). The prediction's input and output parameters were standardized to fall between -1 and 1. The model output will be converted back to its original dimensions by de-normalizing the model result.

Table 3-2 input and output data determination of the model

Parameter	Symbol	Parameter use in LR GA model	Parameter use in NLR GA model	Parameter use in ANN model
Average temperature	TEMP	X1	X1,X10	X1
Precipitation	PRE	X2	X2,X11	X2
Population	POP	X3	X3,X12	X3
Relative Humidity.	RH	X4	X4,X13	X4
Economic development	EG	X5	X5,X14	X5
Number of livestock	NL	X6	X6,X15	X6

Construction activity	C.A	X7	X7,X16	X7
Industrial development	I.A	X8	X8,X17	X8
Holidays/festivals	H&F	X9	X9,X18	X9
Other factor/Bias	OF	X0	X0	B
Water consumption (output)	WC	WC	WC	WC

To generalized the input and output of GA model by the following framework

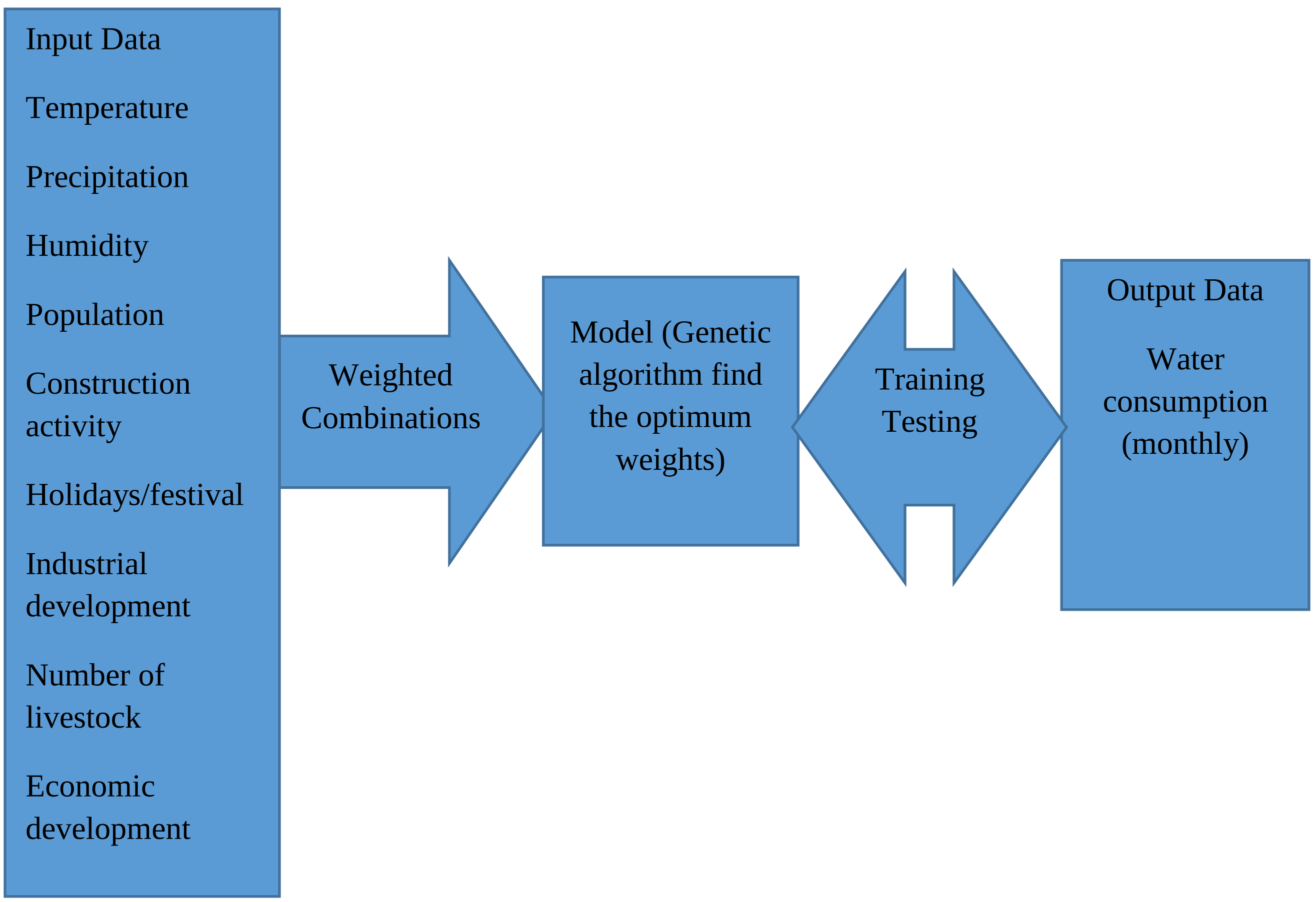

Figure 3-7 Methodological framework of GA model

In order to prepare data for function fitting problems, genetic algorithms divide the data into two matrices: the goal matrix (water consumption), or 'co,' and the input matrix (x). Nine inputs, or components, make up each column of the input matrix and each one represents the value of the input data Similarly, each pertinent column of the target matrix contains a single element that represents the water usage. The data was split into three pieces (training, validation, and test) using a random sample technique. The comment box below provides an explanation of the code used for this method.

```
%Input data
ainp1=xlsread('Input','NND');
% dividing a data
data=ainp1;
n=size(data,1);
idx=randperm(n);
trian_idx=idx (1: round (0.7*n));
val_idx=idx(round (0.7*n)+1:round(0.85*n));
test_idex=idx(round (0.85*n)+1:end);
train_data=data (train_idx, :);
val_data=data (val_idx, :);
test_data=data (test_idx, :);
modinp=trian_data(:,3:11);
[ndata,nc]=size(modinp);
Out1(water consumption) =trian_data(:12);
```

3.8 Genetic algorithm for water consumption modeling

The idea behind a genetic algorithm is derived from evolutionary computation search algorithms, which suggests that the most suited members of a population have a higher chance of surviving and procreating. In a genetic algorithm (GA), a chromosome is a set of variables that, by sorting over the space of possible chromosomal values, indicate a recommended solution to the problem the GA is trying to solve. This method treats a network's hyperparameters like an individual's chromosome. The model is composed of both linear and nonlinear regression equations, with root mean square error (RMSE) serving as the objective function.

Genetic algorithms are a subset of optimization algorithms, which means that its purpose is to identify the ideal solution or solutions, maximizing or minimizing a specific function, to a given computing issue. GA is a subclass of evolutionary algorithms that leverages crossover, inheritance, mutation, selection, and other evolutionary biology-inspired strategies. To construct GA models that provide accurate water consumption predictions and identify the influencing elements, a number of aspects need to be taken into account.

To determine who is most suitable to be the first and second parents to go through the crossover process, three selection operators are used in conjunction with the Normalized Root Mean Square Error (NRMSE) as the fitness function. The GA optimizes the coefficient of the input variable while attempting to minimize the error between the model's output and the actual output.

In the first iteration, the initial coefficient is randomly chosen, and each coefficient's output as well as the error are calculated. Next, parent selection, reproduction, and mutation are applied to GAs updated the coefficient. After that, the modified coefficient is the subject of the subsequent iteration (Gündüz and Akkoyunlu 2020). For the analysis, two GA optimized models were created.

3.8.1 Multiple Linear regression GA model

Modeling the link between a dependent variable and one or more independent variables using the multiple linear regression method (what you're basing the prediction on). The model assumes a straight-line relationship between the variables.

$$WC = X1 * (IA) + X2 * (EG) + X3 * (POP) + X4 * (TEMP) + X5 * (PREC) + X6 * (RH) + X7 * (LS) + X8 * (H\&F) + X9 * (CA) + X0 \quad \text{Equation 3.5}$$

The goal of this model is for GA to optimize the Xi (i = 0, 9) values in order to accurately predict the measured consumption. All nine inputs are normalized in this GA as outlined above.

3.8.2 Nonlinear regression GA model

Nonlinear regression method deals with situations where the relationship between variables is not linear. It uses more complex functions like polynomials, exponentials, or trigonometric functions to capture the relationship. For this research use exponential, function.

$$WC = X1 * (IA^{X10}) + X2 * (EG)^{X11} + X3 * (POP)^{X12} + X4 * (TEMP)^{X13} + X5 * (PREC)^{X14} + X6 * (RH)^{X15} + X7 * (LS)^{X16} + X8 * (H\&F)^{X17} + X9 * (CA)^{X18} + X0 \quad \text{Equation 3.6}$$

Both models the value of ‘Xi (i=0,18)’ is determine by the Principe of genetic algorithm. The goal of this model is for GA is also to optimize the Xi values in order to accurately predict the measured consumption.

3.9 The GA Codes

Comment box: MATLAB code for GA analysis by multiple linear regression equation coefficients optimization:

```
%% Multiple linear regression
obf=@(x)sqrt(mean(((x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,
3)+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm
(:,7)+ x(1,8)*indnorm(:,8)+(x(1,9)*indnorm(:,9)+x(1,10))-outdnorm(:,1)).^2));
```

```
Ai1=diag(-1*ones(1,10)); bi1=zeros(10,1);% constraint that makes all the nine variables to be > 1
Ai1(10,10)=0;
[x,fval1,exitflag1,output1,population1]= ga(obf,10,Ai1,bi1,[],[]);
xmodr=x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)+x(1,8)*indnorm(:,8)+(x(1,9)*indnorm(:,9)+x(1,10)
```

Comment box: MATLAB code for GA analysis by nonlinear regression equation coefficients optimization

```
%% Nonlinear regression
obf=@(x)sqrt(mean(((x(1,1)*(indnorm(:,1).^x(1,10))+x(1,2)*(indnorm(:,2).^x(1,11))+x(1,3)*(indnorm(:,3).^x(1,12))+x(1,4)*(indnorm(:,4).^x(1,13))+x(1,5)*(indnorm(:,5).^x(1,14))+x(1,6)*(indnorm(:,6).^x(1,15))+x(1,7)*(indnorm(:,7).^x(1,16))+x(1,8)*(indnorm(:,8).^x(1,17))+x(1,9)*(indnorm(:,9).^x(1,18))+x(1,19))-outdnorm(:,1)).^2));
Ai1=diag(-1*ones (1,19)); bi1=zeros(19,1);
Ai1(19,19) =0
Ai1(2,2)=1
Ai1(4,4) =1
[x, fvall,exitflag1,output1,population1]= ga(obf,19,Ai1,bi1,[],[]);
xmodr=x(1,1)*(indnorm(:,1).^x(1,10))+x(1,2)*(indnorm(:,2).^x(1,11))...
+x(1,3)*(indnorm(:,3).^x(1,12))+x(1,4)*(indnorm(:,4).^x(1,13))+x(1,5)*(indnorm(:,5).^x(1,14))+x(1,6)*(indnorm(:,6).^x(1,15))+x(1,7)*(indnorm(:,7).^x(1,16))+x(1,8)*(indnorm(:,8).^x(1,17))+x(1,9)*(indnorm(:,9).^x(1,18))+x(1,19);
```

3.10 Training and testing the GA model

The model is prepared for training when the crossover GA number and testing parameters have been set. Random sampling is used to separate the samples into training, validation, and test sets. The model was trained using the training set. The test set offers an entirely impartial way to measure the accuracy of the model. In order to generalize the solutions generated by its outputs, a GA model must be trained by performing the necessary ordinated steps for adjusting the synaptic weights (Xi i = 1...) and thresholds of its weight (X0).

It is now possible to calculate the trained GA model's root mean square error (RMSE) in relation to the testing samples.by using the subsequent linear and nonlinear regression, with the GA model testing serving as the goal function.

% Multiple Linear regression GA model

```
obf=@(x)sqrt(mean(((x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)+ x(1,8)*indnorm(:,8)+ x(1,9)*indnorm(:,9)+x(1,10))-outdnorm(:,1)).^2));
```

```
%% Nonlinear regression GA model
obf=@(x)sqrt(mean(((x(1,1)*(indnorm(:,1).^x(1,10))+x(1,2)*(indnorm(:,2).^x(1,11))+x(1,3)*(indnorm(:,3).^x(1,12))+x(1,4)*(indnorm(:,4).^x(1,13))+x(1,5)*(indnorm(:,5).^x(1,14))+x(1,6)*(indnorm(:,6).^x(1,15))+x(1,7)*(indnorm(:,7).^x(1,16))+x(1,8)*(indnorm(:,8).^x(1,17)) + x(1,9)*(indnorm(:,9).^x(1,18))...
 +x(1,19))-outdnorm(:,1)).^2));
```

Performance is evaluated by calculating the mean squared error and presenting it on a logarithmic scale. Typically, the mean squared error decreases as the model undergoes iterations The training, validation, and test sets' respective performances can be shown. To be used in the examination of the test set are the model weight values that produce the lowest RMSE value for the training sets.

3.11 Artificial Neural Network (ANN) model

An artificial neural network's architecture specifies the connections between its many processing units, or neurons. The way information moves throughout the network is determined by these connections, which resemble brain synapses. The network is able to discover intricate patterns and links in the training data by methodically organizing these connections.

3.11.1 Artificial Neural Network Structure

The design plan of the model, similar to an architectural blueprint, outlines how it's constructed. Recently, researchers have unveiled various approaches to designing these internal structures, known as network architectures. One successful technique for water consumption forecasting is the well-established Multi-Layer Perceptron (MLP) with a feed-forward configuration, which can be like to a specific floor plan within the overall design.

The models in this study are built using a particular neural network design called a multi-layer perceptron (MLP), which has one hidden layer and between one and five nodes. The decision to select these nodes was made based on the possible inaccuracy in each network, with the goal of reducing the root mean square error between the measured and model water consumption data. Finding the crucial elements influencing water usage was the first step.

After then, they used a process of trial and error to improve the neural network's remaining crucial parts so that they would be specifically tailored to the situation at hand. These included the number of neurons in the hidden layer, the number of hidden layers (limited to one hidden layer with one to five nodes in this instance), the activation function that the neurons used, the learning rate, a momentum term (introduced to potentially accelerate the model's convergence), and the number of neurons. The MATLAB neural network tool's default variables for momentum and learning rate are used.

The general goal of the hidden layer, which is composed of nodes (neurons), is to convert input layer data into a weighted value in order to transfer it from the input layer to the output layer or the subsequent hidden layer. the one and five hidden nodes, respectively, seen in figures 3.8 and 3.9 below.

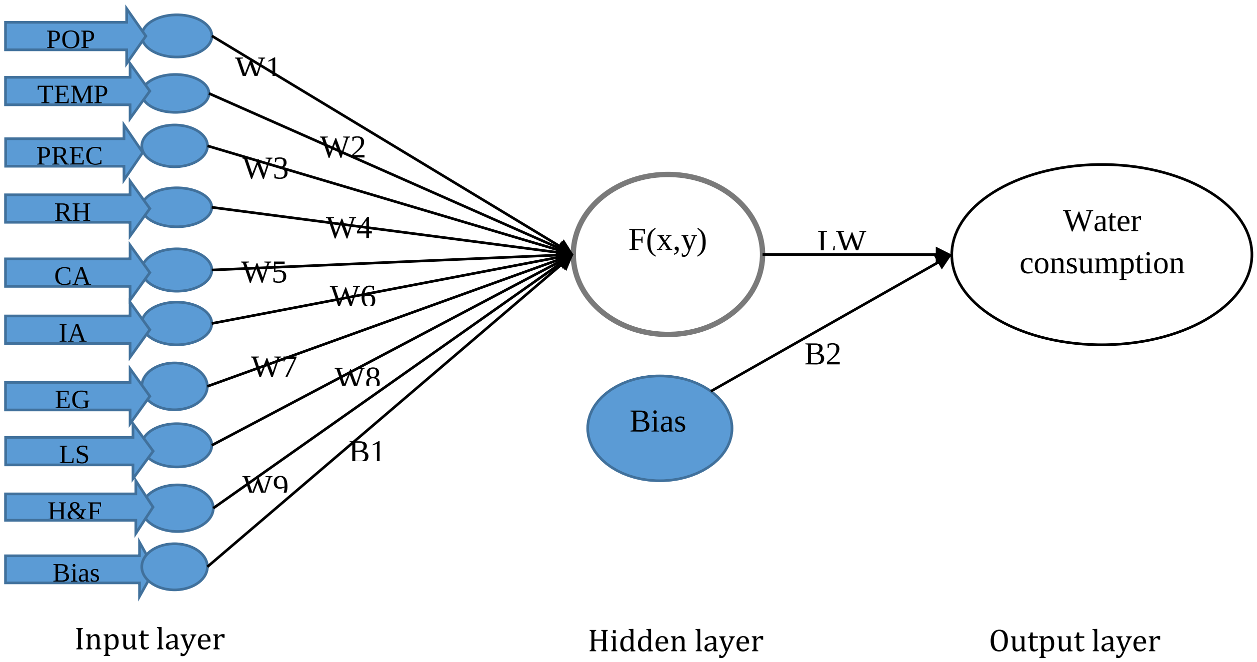

Figure 3-8 Artificial Neural Network Structure with one hidden node

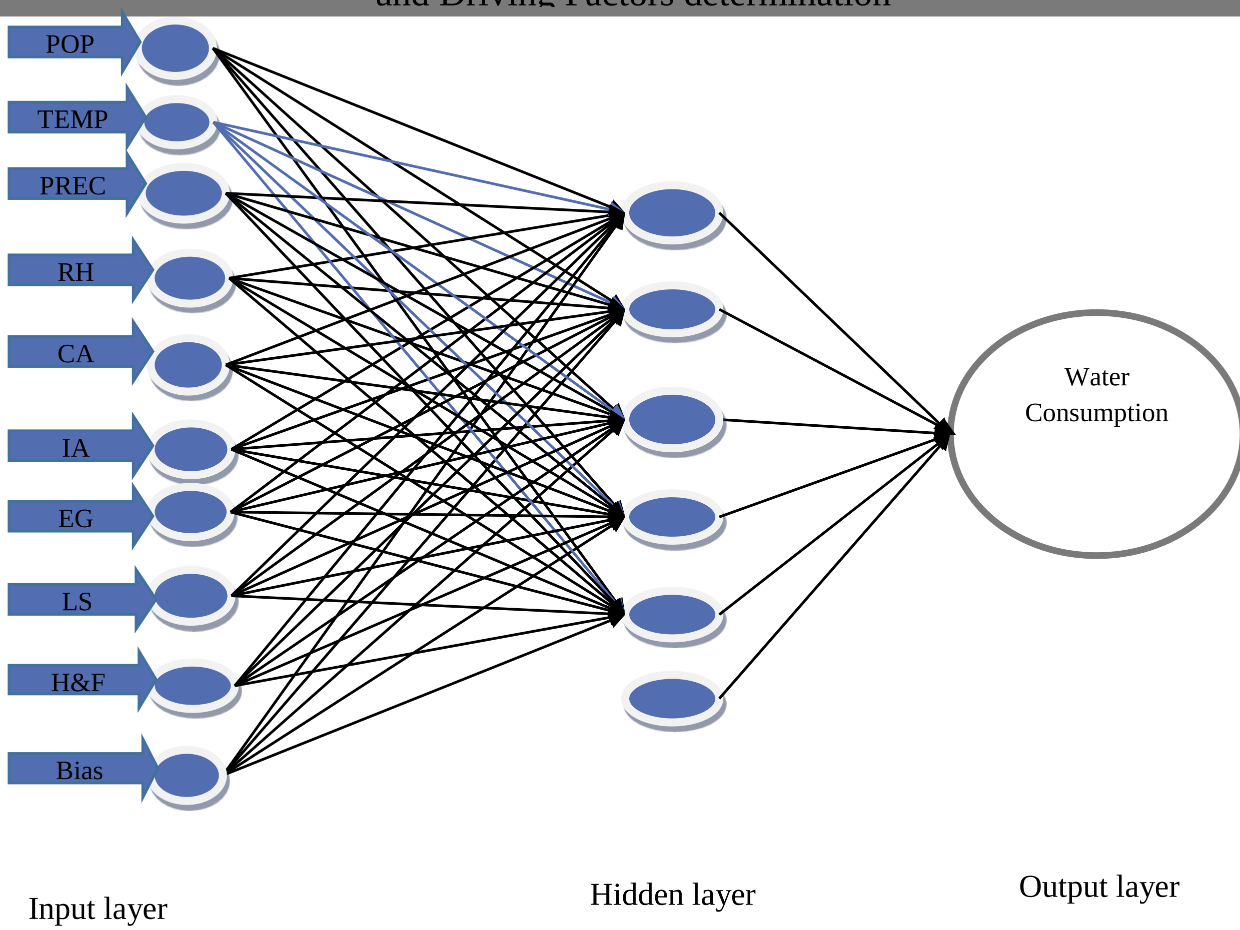

Figure 3-9 Artificial Neural Network Structure with five hidden nodes

In both models, as can be seen in the image above, there are nine input neurons and one output neuron. But after some trial and error, it was discovered that these architectures with their one and five hidden neurons, respectively were the best ones.

The transfer function is used to compute the weighted total of the inputs, the bias value, and the output value restriction within the transfer function's range. The log sigmoid, pure linear, and hyperbolic tangent sigmoid transfer functions are among the many accessible transfer functions. However, the transfer function is typically selected through a trial-and-error procedure. Two pure linear transfer functions were employed in this work to generate the two models. The ability of these functions to establish the relative weights of each input in the comprehensive assessment of consumption of water resulted to this decision. The selection of five neurons for the hidden layer, as illustrated in figure 3.9's second option, was made with the primary goal of optimizing the neural network model's performance. The interpretation of the weights linking the hidden and input layers to the output layer was not taken into account. As seen in figure 3.8, the logical hidden number of neurons is set to one, meaning that the weights joining the inputs to the hidden neurons reflect the inputs' respective contributions to the total output (consumption). The weights connecting the input layer to the output layer are multiplied by the hidden to output weight to generate this contribution.

These are the transfer function algorithms:

'Purlin' $f(x) = x$

$WC = b2 + LWI * (b1 + IW1 * XN)$ Equation 3.7

Where

B1= bias value between the hidden layer and the input layer

B2= bias value between the output layer and the hidden layer

IWI= Weight value between the hidden layer and the input layer.

LWI= Weight value between the output layer and the hidden layer.

XN=input variable

An overview of the creation of the ANN model

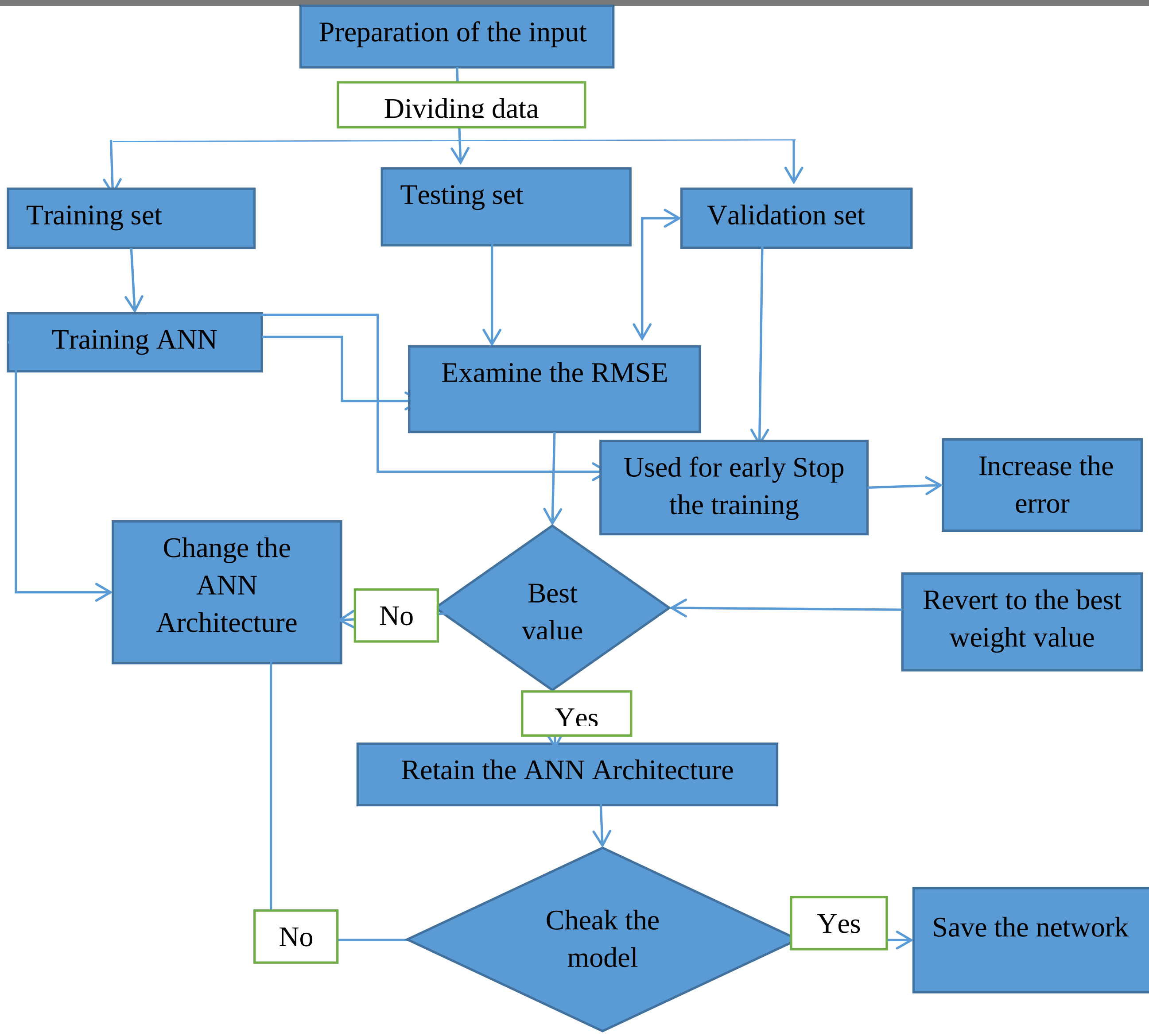

Figure 3-10 general framework of ANN

3.11.2 Network Training

To bring the actual (predicted) outputs of the network closer to the desired (measured) outputs, the weights are changed during network training. Training, testing, and validation subsets of the 99-month data set are randomly selected. The training set is used to estimate the unknown connection weights, the testing set is used to determine the best network topology and/or when to terminate training to prevent overfitting, and the validation set is used to evaluate how well the trained model generalizes. The best MLP design with the highest coefficient of determination and the lowest mean square error can be found using a few training procedures, including the Bayesian regularization back

3.11.3 Validation the neural network

The same data set that was used for training is used as the input for the model's validation; targets are not supplied during this phase, and the trained network now forecasts the outputs based on the weight matrix that was created during training. In this case, back propagation does not happen and mistakes are not computed. Only during the training phase does back propagation take place..

3.11.4 Testing the neural network

With testing data set the model is tested and the overall model performance is measured by Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Normalized Root Mean Squared Error (NRMSE). In this paper, the MSE Root Mean Squared Error (RMSE), and Normalized Root Mean Squared Error (NRMSE) are used to evaluate the prediction performance of the model.

$$Mean\ Absolute\ Deviation\ (MAD) = \frac{1}{N}\sum |Y - Y't|$$

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \frac{100}{N}\sum \frac{|Yt - Y't|}{Yt}$$

$$Mean\ Square\ Error\ (MSE) = \frac{1}{N}\sum |Y - Y't|^2$$

$$Root\ Mean\ Square\ Error\ (RMSE) = \sqrt{MSE}$$

$$Normalized\ Root\ Mean\ Square\ Error\ (NRMSE) = \frac{RSME}{\bar{Y}}$$

$$Accuracy = 100 - MAPE$$

N = number of sample data

Y't=true information from the nth sample

Y't = model result for sample and $\bar{Y}$ = average consumption

3.12 Model Selection

The optimal training algorithm is chosen based on the global process's average error and regression parameter as part of the four algorithms' validation process. Three subsets of the available dataset are used to train the ANN, nonlinear regression, and linear regression: model training (70%) is used to calculate the root mean square error (RMSE).and changing the model's parameters; the ideal number of inputs and outputs is found using test data (15%) and validation data (15%).

Two methods to divide the data into representative subsets, a self-organizing map (SOM) and a genetic algorithm (GA) were developed. These methods were compared to the conventional methodology that is commonly used in the literature, which divides the data in an arbitrary way (Bowden, Maier et al. 2002).

Consequently, the standard approach technique of simple random sampling serves as the foundation for the data splitting methods used in this study. This approach is semi-deterministic; the training, test, and validation datasets are formed by randomly selecting every nth sample from the initial point. To perform systematic sampling, the data in this study are first sorted in increasing values along the output variable dimension. The user then enters the percentages of test and training data, which are used to determine the sampling interval. After that, random starting points are selected, followed by the drawing of training samples and test samples.

Finally, sampled data is added to the training, testing, and validation sets. The following metrics are widely used to measure calibration and validation: root mean square error (RMSE), mean square error (MSE), and coefficient of determination (R-squared). Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Normalized Root Mean Squared Error (NRMSE) are the three metrics utilized in this study to evaluate the model's performance.

3.12.1 Normalization and de-normalization

The input value needs to be standardized, and then the resulting output from the artificial neural network (ANN) must be converted back to its original value through a process called demoralization.

This is achieved by applying the following formula:

$$WCD = \frac{\text{Min (out1)} + (\text{Max(out1)} - \text{Min(out1)}) * (\text{WC(:,1)}}{(\text{lfmax} - \text{lfmin})} - \text{lfmin} \qquad \text{Equation 3.8}$$

Where

WC= predicated water consumption value for normalization of the data

WCD=predicated water consumption value for de-normalization of the data

Min (out1) = minimum value of out1 (water consumption)

Max (out1) = maximum value of out1 (water consumption)

1fmax=1

1fmin=-1

Comment Box: MATLAB code for normalization and de-normalization.

```
% Normalization of input and output data
inmax = max(modinp); % the highest values across all data
inmin=min(modinp); % the lowest numbers over the whole
1fmax=1; %the highest number that should be given to the logistic activation
function
1fmin=-1;% he lowest number that should be entered for the logistic
activation function
indnorrm=zeros(ndata,nc); % normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i= 1:ndata
    for  j=1:nc
        indnorrm (i,j)=1fmin+(1fmax-1fmin)*(modinp(I,j)-inmin(j)/...
            (inmax(j) inmin(j);
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*(out1(i,1)-min(out1))/...
            (max(out1)-min(out1));
End
% dnorm, is the reverse normalization.
dnorm=min(out1)+ (max(out1)-min(out1))*(xmodr(:,1)-lfmin)/(lfmax-lfmin)
```

4. RESULTS AND DISCUSSION

Water consumption pertains to the volume of water that is charged to customers via the household water distribution system. This water is utilized for a number of reasons, like as industrial, agricultural, residential, and other non-residential use. As consumption is monitored over a period of time (m3/months), the dependent variable is a time series, and one method for water consumption aggregated by consumption is through the utilization of GA and ANN. GA can be employed to optimize the contribution once the weights and corresponding inputs are defined as either linear or nonlinear regression models that describe the observed consumption. An additional option is to use ANN to create an input-output relationship and produce weights for each input's contribution to the total observed consumption data.

Four models were developed to forecast water consumption and determine the driving factors of water consumption using one dependent and nine independent variables. These models include a linear regression GA model, a nonlinear regression GA model, and an ANN with one hidden layer consisting of one and five nodes.

4.1 GA Multiple Linear regression model

A linear regression model is proposed, assuming that the dependent and independent variables have a linear relationship. This model is developed using the theory of genetic algorithms. The goal (fitness) function is used to construct the code in MATLAB software, together with the linear model equation and root mean square error (RMSE). by using many trails to determine the value of X, the coefficient of the input variable. Based on equation 3.5 determine the coefficient of the input factors.

$$WC = X1(IA) + X2(EG) + X3(POP) + X4(TEMP) + X5(PREC) + X6(RH) + X7(LS)\, X8(H\&F) + X9(CA) + X0$$

For which the coefficient of the input data are parameters that need to be approximated but are not known.

To forecast future water consumption, the values of the unknown parameters must be optimized. In this research, sample data from September 2015 to August 2023 is utilized to identify and estimate the model parameters.

A GA code shown in chapter three is used to obtain the required weights (contributions) of each input parameters. Appendix 3 displays the constructed GA Multiple Linear Regression Model.

The generated Multiple linear regression model was

$$\mathrm{WC} = 0.0000085 * (\mathrm{IA}) + 0.1266 * (\mathrm{EG}) + 0.3252 * (\mathrm{POP}) + 0.0074(\mathrm{TEMP}) - 0.173 * (\mathrm{PREC}) - 0.1555 * (\mathrm{RH}) + 0.2115 * (\mathrm{LS}) + 0.2500 * (\mathrm{H\&F}) + 0.0127 * (\mathrm{CA}) + 0.2842$$

Each coefficient represents the impact of its corresponding independent variable on the dependent consumption. The coefficients (weights) seen in the above equation are values that represent best the observed consumption for those input variables (see Appendix 2). Figure 4.1 shows the model result compared to the observed consumption of water and also show the error of the actual and model value for water consumption.

Understanding the model output requires attention to two key factors: the magnitude and the sign (+/-) of the coefficients. The magnitude reflects the level of impact each parameter has on overall consumption, while the sign indicates the direction of that impact in relation to the input parameter. The correlation between population and water consumption is strong, indicated by a higher positive coefficient. Analysis of the coefficients reveals that population has the greatest influence on total water consumption in Addis Ababa. Factors such as average temperature, industrial activity, construction activity, number of livestock, economic growth, and holidays all positively contribute to consumption. On the other hand, precipitation and relative humidity show a negative correlation, indicating that consumption decreases with increasing values of these variables.

All else being equal, the base level of water consumption (when all variables are 0) was predicted to be positive 0.2842 units. A positive coefficient signifies that a rise in the variable results in an increase in consumption. Conversely, a negative coefficient indicates that an increase in the variable results in a decrease in consumption. The size of the coefficient indicates the intensity of the relationship: Higher coefficients indicate stronger effects, while lower coefficients indicate weaker effects.

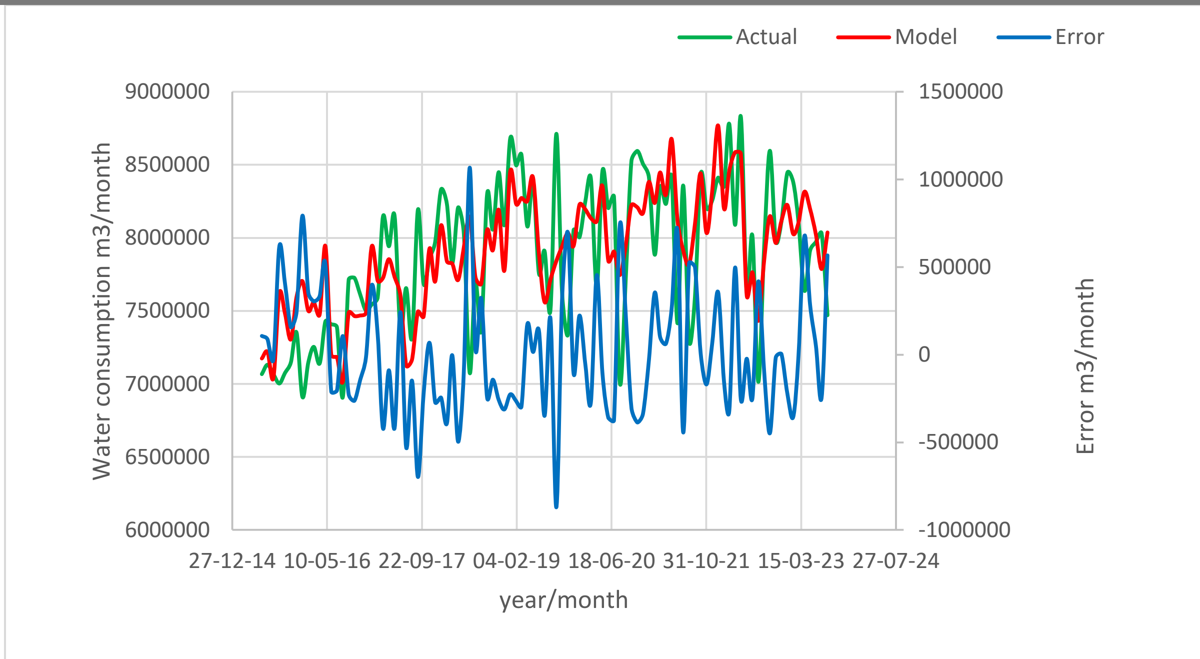

Figure 4-1 Comparing water consumption between the real and modeled amounts using the linear GA Model (see Appendix 4 for tabular data)

4.2 GA Nonlinear regression model

The code developed for water consumption using GA for non-linear regression Model can be seen in appendix 5. The dependent variable and the independent variables do not follow a linear relationship in the model. More intricate relationships that are not amenable to being depicted by a straightforward straight line are instead supported by the model. Based on equation 3.6 determine the coefficient and the power of the input factors

$$\mathrm{WC} = \mathrm{X1}(\mathrm{IA}^{\mathrm{X10}}) + \mathrm{X2}(\mathrm{EG})^{\mathrm{X11}} + \mathrm{X3}(\mathrm{POP})^{\mathrm{X12}} + \mathrm{X4}(\mathrm{TEMP})^{\mathrm{X13}} + \mathrm{X5}(\mathrm{PREC})^{\mathrm{X14}} + \mathrm{X6}(\mathrm{RH})^{\mathrm{X15}} + \mathrm{X7}(\mathrm{LS})^{\mathrm{X16}} + \mathrm{X8}(\mathrm{H\&F})^{\mathrm{X17}} + \mathrm{X9}(\mathrm{CA})^{\mathrm{X18}} + \mathrm{X0}$$

Assuming that the coefficient of the input data are unknown parameters that must be estimated by genetic algorithm.

$$\mathrm{WC} = 0.00001 * (\mathrm{IA})^{1.4642} + 0.1658 * (\mathrm{EG})^{0.2578} + 0.2001 * (\mathrm{POP})^{1.22} + 0.0303 * (\mathrm{TEMP})^{1.047} - 0.269 * (\mathrm{PREC})^{0.365} - 0.00001 * (\mathrm{RH})^{1.0457} + 0.007823 * (\mathrm{LS})^{1} + 0.078145 * (\mathrm{H\&F})^{0.00001} + 0.3399 * (\mathrm{CA})^{0.3647} + 0.2016$$

The coefficients that show how the independent and dependent variables are related to one another.

They are employed to scale the impact of the independent variables, increased to specific powers, in the equation above. The terms $((IA)^{1.4642})$,$(POP)^{1.22}$, and so on, represent the independent variables raised to different powers. As a result, nonlinear interactions between the independent and dependent variables can be captured by the model. Figure 4.2 displays the outcome of the nonlinear GA model in comparison to the observed consumption data. The residual between the output of the model and the measured consumption data is depicted in the same picture.

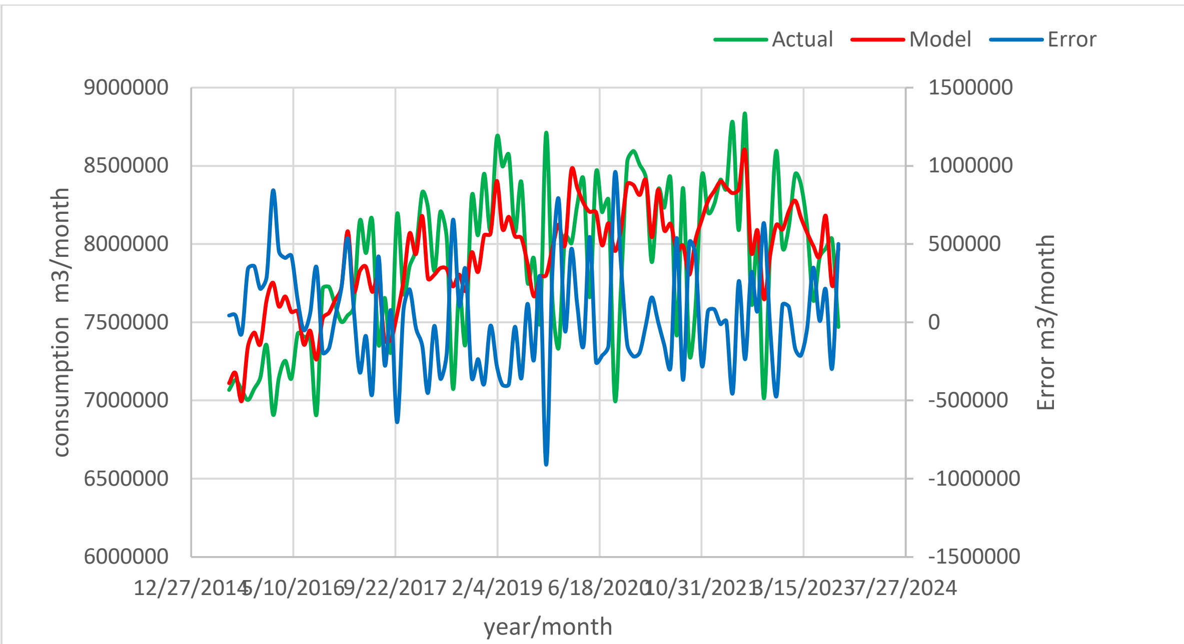

Figure 4-2 Comparison of the actual and modelled water consumption using non-linear GA model (see appendix 6 for tabular model output)

The contribution of each input to the overall water consumption cannot be easily determined based on the weight assigned to each coefficient. In order to accurately assess the contribution, it is necessary to reevaluate the multiplier and powers associated with each input. The following steps is utilized for this analysis:

1. Assign Factors Based on Exponents: We differentiate the influence of terms based on their exponents:

- Exponent ≥ 2: Factor = 2 (stronger influence due to exponential growth)
- Exponent between 1 and 2: Factor = 1.5 (intermediate influence)

- Exponent = 1: Factor = 1 (linear term, maintains coefficient value)
- Exponent < 1 (here, some are negative): Factor = 0.5 (reduced weight, with an additional consideration for negative signs)

2. Calculate Weight Contribution per factors**:**

Multiply the absolute value (to handle negative coefficients) of each term's coefficient by the corresponding factor from step 1:

Table 4-1 percentage contribution of nonlinear regression GA model

factors	Coefficient	Abs. Value	Exponent	factor	Weight	percentage Contribution
IA	0.00001	0.00001	1.4642	1.5	0.000015	0.001528
EG	0.1658	0.1658	0.2578	0.5	0.0829	8.446467
POP	0.2001	0.2001	1.22	1.5	0.30015	30.58151
AVE.TEMP	0.0303	0.0303	1.047	1.5	0.04545	4.630783
PREC	-0.269	0.269	0.365	0.5	0.1345	13.70386
RH	-0.00001	0.00001	1.0457	1.5	0.000015	0.001528
LS	0.007823	0.007823	1	1	0.007823	0.797065
H&F	0.078145	0.078145	0.00001	0.5	0.039073	3.980996
CA	0.3399	0.3399	0.3647	0.5	0.16995	17.31577
Other factors	0.2016	0.2016	1	1	0.2016	20.5405
sum					0.981476	

The result of the model show that Population has the highest weight percentage contribution for water consumption predication due to its larger coefficient of the impact but industrial activity and relative humidity are less contribution on water consumption predication.

4.3 ANN

For many tasks, an ANN with one hidden layer is an effective and basic architecture that strikes a reasonable compromise between complexity and performance. Weights are adjusted as data flows over the network to minimize the discrepancy between the predicted and actual output. The importance of weights and biases in Artificial Neural Networks (ANNs) using MATLAB cannot be emphasized because they dictate the network's behavior and learning process. The design of the model consists of an input layer with nine inputs and a bias node, an output layer with one node representing the output (in this case, water usage), and a hidden layer with one to five neurons and a bias node.

The strength of connections between neurons in various levels is represented by weights. Greater weight denotes a stronger input influence on the linked neuron's output. Usually, random values are used to initialize weights, which are then modified during training.

The constant value that each neuron adds to the weighted sum of its inputs is known as bias. serves as a threshold, changing the neuron's activation function. Appendix 7 displays the general input/output settings for the neural network toolbox in the MATLAB environment.

One crucial element in ANN is the activation function utilized in each neuron. In order to interpret the network's results in a tangible way, the activation function was designed to be purely linear, represented as 'purelin' in the MATLAB neural network platform.

4.3.1 ANN model with one hidden node

For one hidden layer node architecture with pure linear transfer function the following equation will be developed.

$$Y_{\text{linear}} = \text{B2} + \text{LW1} * (\text{ B1} + \text{IWI} * \text{Xn})$$

Y-linear=predicated water consumption value

XN=input variable

B1= bias value between the hidden layer and the input layer

B2= bias value between the output layer and the hidden layer

IWI= Weight value between the hidden layer and the input layer.

LWI= Weight value between the output layer and the hidden layer.

ANN model with one hidden node code to see at the appendix 7

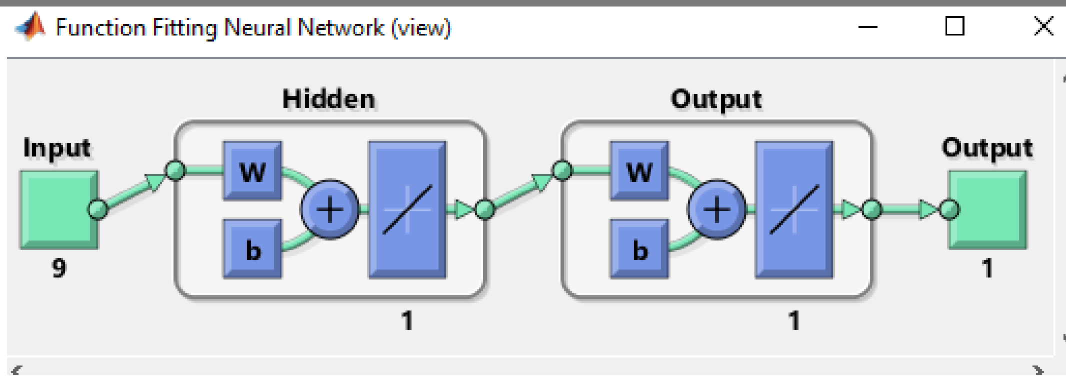

Figure 4-3 Input, hidden and output layer of ANN with one hidden node

Table 4.2 displays the outcome, which includes the bias weight at these layers as well as the input to hidden and hidden to output weights. The outcomes for this neural network architecture are displayed in Figure 4.4.

Table 4-2 weight and bias values

Inputs	IW1	LW1	B1	B2	LW1*B1	LW1*IW1	B2+LW1*B1
POP	-11.4628	-0.03495	-0.07821	0.242652	0.002733	0.400589	0.245386
PREC	0.221474					-0.00774	
AVE.TEMP	-0.43166					0.015085	
RH	8.97677					-0.31371	
H&F	-7.81571					0.273134	
CA	-4.05068					0.141559	
LS	-4.62073					0.16148	
IA	6.228377					-0.21766	
GR	-7.80325					0.272699	
other factors	0.245386					0.245386	

Based on these weights the weighted contribution of each input is as depicted in the following equation.

$$WC = -0.2177 * (IA) + 0.2727 * (EG) + 0.4006 * (POP) + 0.0150 * (TEMP) - 0.008 * (PREC) - 0.3137 * (RH) + 0.1615 * (LS) + 0.2731 * (H\&F) + 0.1415 * (CA) + 0.2454$$

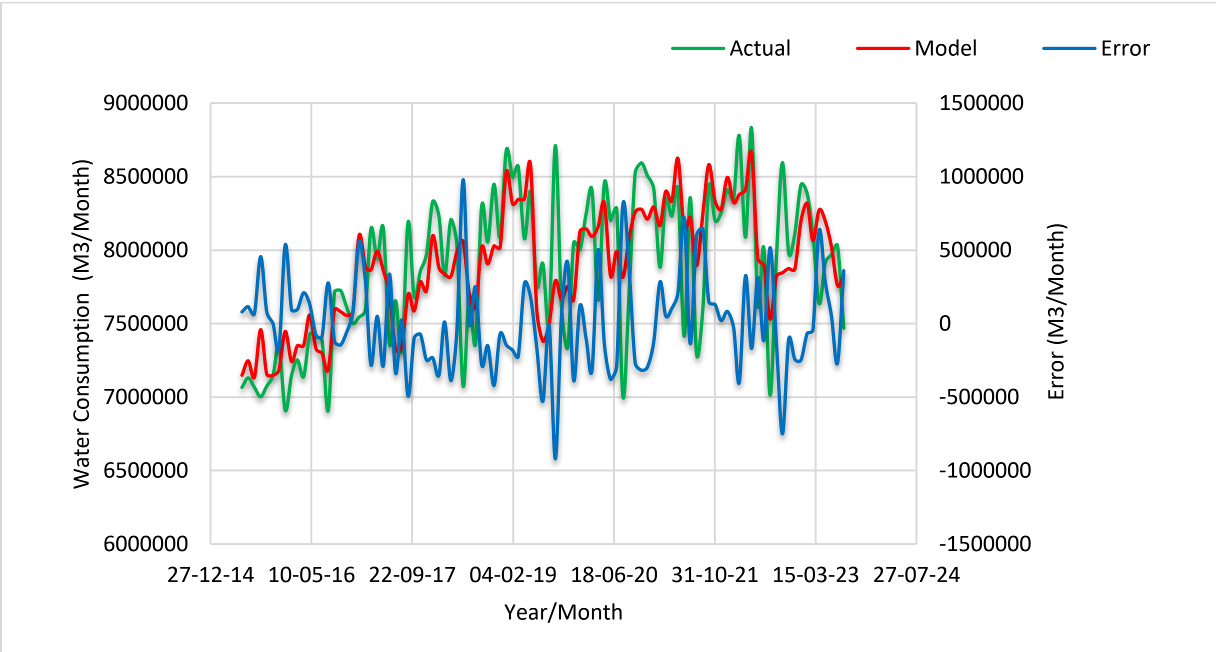

Figure 4-4 Analyzing and contrasting the ANN model's predicted and real water consumption with one hidden node (see appendix 8 for tabular data)

4.3.2 ANN model with five hidden nodes

The network has nine input layer, one hidden layer with five nodes in it, and an output layer for predictions. Training a model with five hidden nodes in its hidden layer requires significantly more computational resources (processing power and time) compared to a single hidden node model. This is because, as the number of layers increases exponentially, every neuron in a hidden layer calculates something based on its inputs. Additionally, the model needs to store and update the weights associated with each connection between neurons. With five hidden nodes, there are many more connections and weights compared to a model with just one hidden node. As a result, training a five-layer model can take significantly longer and require more powerful hardware compared to a simpler model. The code of ANN model with five hidden nodes is the same to ANN model with one hidden nodes code but the difference only changes the number of node (one change five).

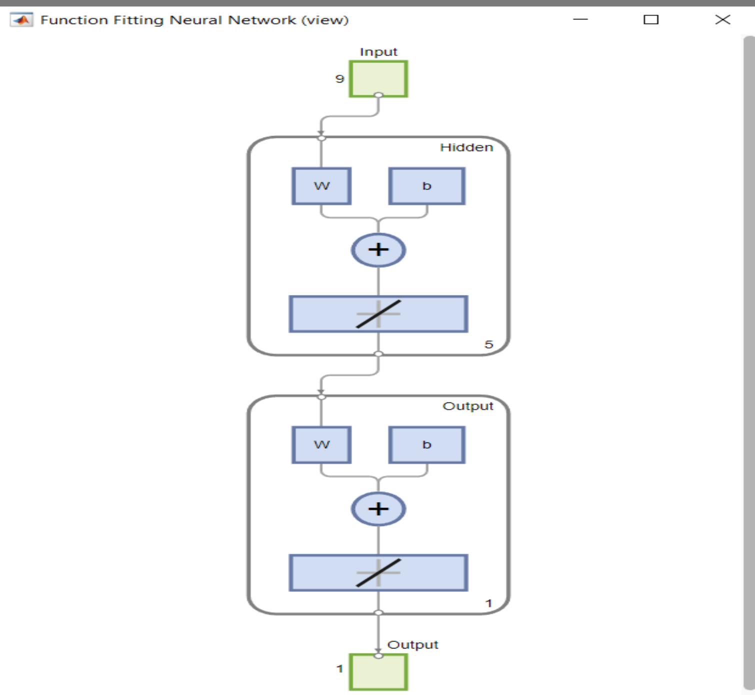

Figure 4-5 Input, hidden and output layer of ANN with five hidden nodes

The result indicating the input to hidden and hidden to output weights together with the bias weights at these layers is as shown in table 4.2 and 4.3. Figure 4.6 show the results obtained for this neural network architecture shown in Figure 4.5.

Table 4-3 the value of IW1, LW1 and bias

IW1									LW1	B1	B2
POP	PREC	AVE.TEMP	RH	H&F	CA	LS	IG	GR			
-0.2666	-0.5999	0.6396	0.8821	0.2294	0.4425	0.1660	-0.4540	-0.2670	-0.1770	0.3529	0.2669
0.4575	0.6912	-0.3999	0.4911	0.7409	0.8307	1.0554	-0.3774	0.8845	0.1781	0.6221	
-0.5666	0.1477	-0.8917	0.4330	-0.6162	-0.1067	-0.4281	0.2036	0.4854	-0.0894	-0.7544	
-0.8276	0.3774	-0.4100	-0.1227	0.4644	0.7005	0.1931	0.6087	-0.7443	-0.0320	0.1463	
-0.4665	1.0384	-0.4037	0.7028	-0.0649	-0.3526	-0.0402	0.7379	0.0294	-0.2760	0.8573	

Table 4.3 the value of LW1 and bias

LW1*IW1										B2+∑B1*LW1
Pop	PREC	AVE.TEMP	RH	H&F	CA	LS	IG	GR	B1*LW1	**0.1414**
0.0472	0.1062	-0.1132	-0.1562	-0.0406	-0.0783	-0.0294	0.0804	0.0473	-0.0625	
0.0815	0.1231	-0.0712	0.0875	0.1320	0.1480	0.1880	-0.0672	0.1576	0.1108	
0.0507	-0.0132	0.0797	-0.0387	0.0551	0.0095	0.0383	-0.0182	-0.0434	0.0675	
0.0265	-0.0121	0.0131	0.0039	-0.0149	-0.0224	-0.0062	-0.0195	0.0238	-0.0047	
0.1287	-0.2866	0.1114	-0.1940	0.0179	0.0973	0.0111	-0.2037	-0.0081	-0.2366	

0.3346	**-0.0825**	**0.0198**	**-0.2974**	**0.1495**	**0.1541**	**0.2018**	**-0.2282**	**0.1771**	-0.1255	

Based on these weights the weighted contribution of each input is as depicted in the following equation.

$$WC = -0.2282 * (IA) + 0.1771 * (EG) + 0.3346 * (POP) + 0.0198 * (TEMP) - 0.0825 * (PREC) - 0.2974 * (RH) + 0.2018 * (LS) + 0.1495 * (H\&F) + 0.1541 * (CA) + 0.1414$$

Note: - All the above model the result of water consumption value must be de-normalized because the input and the output value are under normalized conditions.

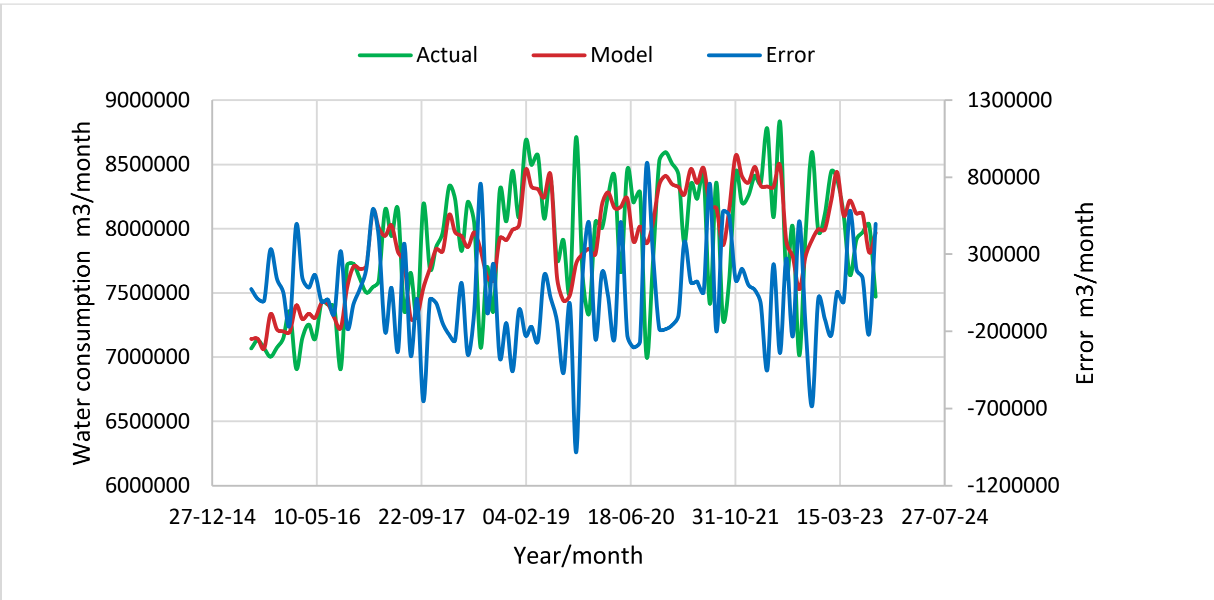

Figure 4-6 Analyzing and contrasting the ANN model's predicted and real water consumption with five hidden nodes (see Appendix 10 for tabular data)

4.4 Model Selection

The model with the lowest MSE and RMSE was looked for in order to identify the best model from the four models previously described.

Table 4-4 model training, testing and validation

Model	MSE (billion) (m^3/Month)			RMSE (million m^3/Month)			NRMSE		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
Linear regression	131.099	126.517	103.015	0.3621	0.3557	0.3210	0.04643	0.04337	0.04037
Nonlinear regression	125.2065	92.7321	88.982	0.3538	0.3045	0.2983	0.04537	0..03713	0.03752
ANN with one hidden	100.989	114.214	118.957	0.3178	0.3380	0.3449	0.04075	0.04121	0.04338

node									
ANN with five hidden nodes	95.4512	115.321	109.829	0.3089	0.3396	0.3315	0.03962	0.04141	0.04167

Table 4-5 MSE, RMSE, and NRMSE of water consumption for the developed models

Model	MSE (billon)	RMSE (million)	NRMSE (-)	MAPE (%)	Accuracy (%)
Linear regression GA model	126	0.355	0.0451	3.63	96.37
Nonlinear regression GA model	115	0.339	0.0430	3.40	96.60
ANN model with one hidden node	106	0.325	0.0413	3.17	96.83
ANN model with five hidden nodes	102	0.3195	0.0405	3.073	96.93

The difference between the modeled and observed values, or Y't - Yt, is how each of these models expresses the error. When evaluating model performance for a given sample data set, the model with the lowest NRMSE score is deemed to be the most accurate. The accuracy of the four created models is shown in Table 4.5.

When evaluating the four models' performances in relation to their NRMSE values. Better performance is indicated by a lower NRMSE. The least root mean square error (NRMSE) in this study is attributed to the ANN model with five hidden nodes, which is followed by the ANN model with one hidden node (0.0413), the nonlinear regression GA model (0.0430), and the linear regression GA model (0.0451). This implies that the most accurate representation of the underlying relationship between the variables is provided by the ANN model with five hidden nodes. It is the most accurate and has the least amount of mistake. We'll have more conversations based on this chosen paradigm.

4.5 Percentage contribution of the impact on Water consumption

To calculate the total weight by summing the absolute values coefficients of the input variable. Then, dividing each weight contribution (absolute value of the coefficient) by the total weight and multiplying by 100% will give result in the relative weight (percentages) for each input. Table 4.7 show these weight values for each input.

$$percentage\ contribution(\%) = \frac{Wi}{\sum Wi}$$

Where,

Wi= each weight value of the input and ∑Wi=the total weight of all factors

ANN with five hidden nodes percentage contribution shows population number have the highest weight percentage due to its larger coefficient. Relative humidity, Industrial activity, Number of livestock, economic growth and construction activity found to have significant contributions because their coefficients have relatively high absolute values.

Table 4-6 Percentage contribution of each factor by ANN with five hidden nodes

		ANN with five hidden nodes	
No,	Factors	weight	weight percentage
1	Population	0.334583	18.73041
2	Precipitation	-0.08252	4.618227
3	Average temperature	0.019794	1.108374
4	Relative humidity	-0.29744	16.64801
5	Holiday and festival	0.14951	8.368786
6	Construction growth	0.154077	8.626288
7	Number of livestock	0.201819	11.29646
8	Industrial activity	-0.22817	12.77429
9	Economic growth	0.177135	9.913793
10	Other factor	0.14138	7.915361

4.6 significance of both negative and positive weights corresponding real-world explanations

Based on the best model (ANN with five hidden nodes) for water consumption forecasting, the weights assigned to each independent variable hold significant meaning. These weights, both positive and negative, indicate the relative influence of each factor on water consumption.

Table 4-7 negative and positive weight meaning corresponding real-world

No,	Factors	Weight	Real-World Explanations
1	Population	0.334583	Population growth has a positive relationship with water consumption. More people simply require more water for domestic purposes (drinking, cooking, Bathing and hygiene, Sanitation, Laundry, Cleaning and Outdoor uses).
2	Precipitation	-0.08252	Precipitation has an inverse relationship with water consumption. When it rains more, people rely less on public water sources for watering plants and other outdoor chores.
3	Average temperature	0.019794	The average temperature and water usage are in a positive relationship. People often use more water in warmer climates to meet their needs for hydration, air conditioning, and sweating. They might also take more showers or baths.
4	Relative humidity	-0.29744	Water usage is inversely correlated with relative humidity. Because there is less evaporation under high humidity conditions, people might use less water for outdoor activities like irrigation.
5	Holiday and festival	0.14951	Holidays and festivals have a positive relationship with water consumption. During holidays and festivals, there may be increased water use for leisure activities, gatherings, and ornamental purposes.
6	Construction growth	0.154077	Water use is positively correlated with construction activities. Building operations need a large quantity of water.
7	Number of livestock	0.201819	Number of livestock has a positive relationship with water consumption. More livestock require more water for drinking and other purposes.

8	Industrial activity	-0.22817	Water consumption is inversely correlated with industrial activity. This might be the result of a change in the region's industries away from water-intensive ones or from sectors becoming more water-efficient. i.e. This could be explained by the fact that industrial operations employ water recycling techniques, which lower water consumption overall.
9	Economic growth	0.177135	Water consumption and economic growth are positively correlated. Population growth and general living standards are expected to rise with an expanding economy and might result in increased water use.

Policymakers and managers of water resources can learn a great deal about the complex relationships that influence Addis Ababa's water use by comprehending the weight values. Using this information, focused initiatives may be created to encourage water conservation, guarantee sustainable water management techniques, and increase climate change resilience.

4.7 Identification of driving factors

It is essential to get a precise and insightful analysis and explanation of the consumption patterns since knowledge of water consumption is essential for planning and managing water resources. The model is used to create the equation and determine the driving factors' weight values, which show how much each component contributes to the water consumption that is modeled. Combining these weight values offers a thorough knowledge of how the algorithm produces broad forecasts. The weight percentage contribution as shown in the above table. The best model for this research is ANN model with one hidden layer with five nodes. Using this model weight percentage contribution is as shown in table 4.7 and figure 4.7.

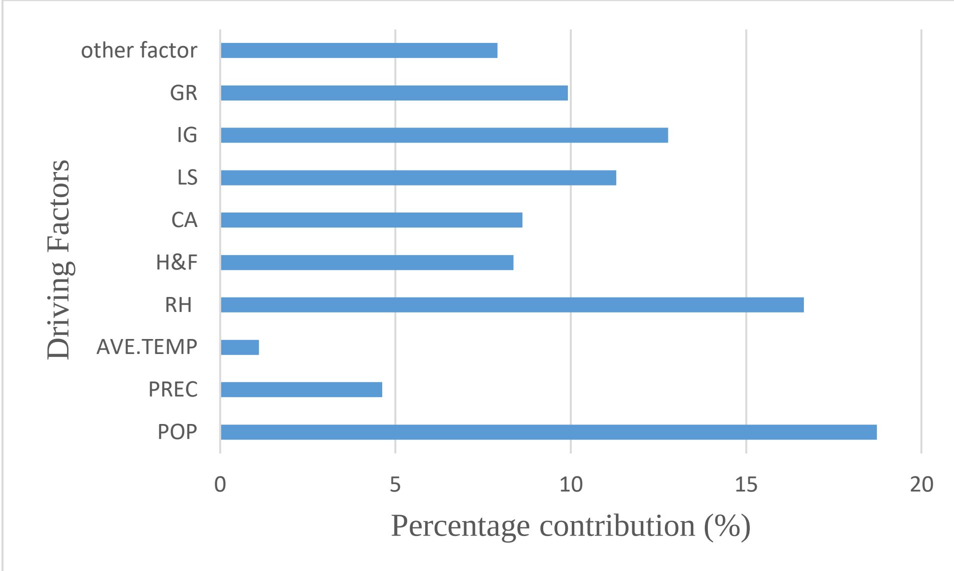

Figure 4-7 driving factors percentage contribution

The results show that Addis Ababa's population number has a significant effect on the amount of water used there. Factors relative contribution to water consumption in Addis Ababa in their order of significance are population, relative humidity, industrial activity, number of livestock, economic growth, construction activity, presence of holiday/festival, other factor, precipitation and average temperature. The result also shows that other causes for water consumption (not analyzed in this model) takes a toll of 8%. Factors that are related to population (number, temperature, relative humidity, economic growth, and Holidays) contributes to 55% of the total consumption. Thus, more than half of the consumption can be explained through issues related to population. These results show that demographic characteristics account for the majority of the variation in water consumption, with the key components being found to be congruent with the city's industrial structure. Furthermore, there are additional factors that are not directly linked to population, such as precipitation, construction activity, livestock, industrial activity, and other variables. These factors contribute to 45% of the overall impact, emphasizing the importance of conducting comprehensive research for accurate water consumption analysis and forecasting. Traditionally, population forecasting has been relied upon to determine future demands, but it is crucial to consider these other factors as well.

Therefore, to achieve sustainable water resource management, planners need to identify the main drivers of water consumption based on the current and future driving variables of the city, which requires a precise prediction of future water consumption.

4.8 Sensitive Analysis

Sensitivity analysis is a technique used to assess how much a model's output changes when you modify its input variables. It helps you understand which input variables have the most significant impact on the final outcome. It is also a valuable tool that offers a multifaceted approach to risk assessment, informed decision-making, and model improvement. It helps navigate uncertainties inherent in any model or system, leading to more reliable predictions and effective strategies.

To choose the optimal model in order to determine sensitivity analysis. The ANN model with five hidden nodes is the most suitable model for this investigation. This model is thus used to understand the sensitivity of each input parameters on the overall result of the model. A 5% increase or decrease of the coefficient of on driving factors were used to conduct the analysis. The code based on ANN model with five hidden nodes is presented in appendix 11.

Table 4-8 RMSE value, sensitive index and range of the Sensitive analyses with 5% increasing and decreasing

Factors	POP	PREC	AVE.TEMP	RH	H&F	CA	LS	IG	EG	True
RMSE 5% increase	320102	319621	319558.2	319413	319072	319511	320071	319424	319394	319000
Sensitive index	1.0035	1.00195	1.00175	1.0013	1.0002	1.0016	1.0034	1.0013	1.0012	1
RMSE 5% decrease	319274	319495	319532	319849	320105	319648	319140	319792	319763	319000
Sensitive index	1.0009	1.00155	1.001668	1.0027	1.0035	1.0020	1.0005	1.0025	1.0024	1
Range (Absolut value)	828	126	26.2	436	1033	137	931	368	369	

The absolute value of the difference between the two RMSE values (|5% increase-% 5 decrease|) is used in this research to establish the range value instead of the difference between the highest and lowest values. The highest range value in Figure 4.8 shows greater sensitivity to the other, whilst the shortest-range value denotes less sensitivity.

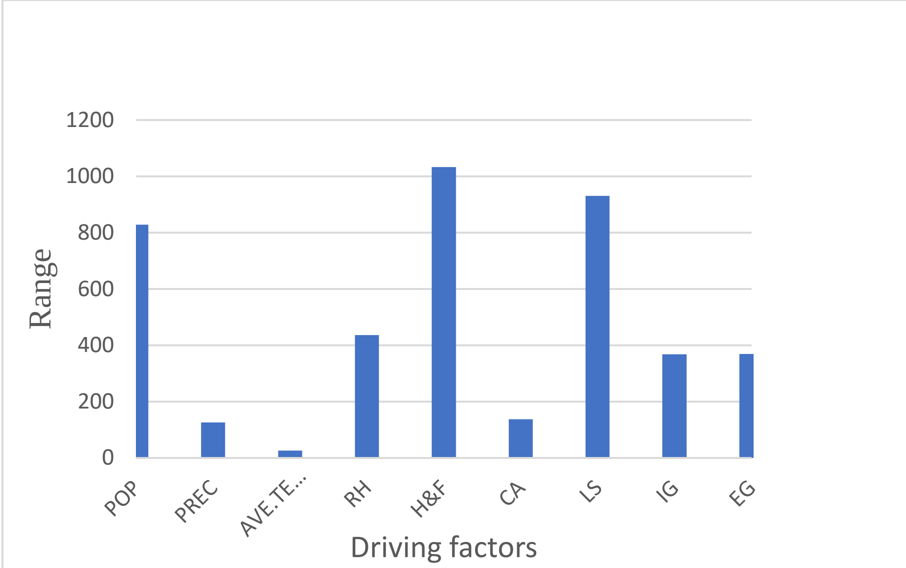

Figure 4-8 sensitive analysis with 5% increasing and decreasing on the weight of the factors

A significant increase in RMSE compared to the original value indicates high sensitivity to that particular input variable. This means small changes in its coefficient significantly affect the model's ability to match the reference data.

Conversely, a minimal change in RMSE suggests the model's output is not very sensitive to that input variable. Changing its coefficient by a small amount doesn't significantly impact the model's accuracy based on the reference data. Based on RMSE value with 5% increasing and decreasing on the weight of the factors to determine the Range. The largest range value shows the input factors is very sensitive but the reverse value shows less sensitive. For in this case holiday and festival, number livestock and population are very sensitive but average temperature is less sensitive to the output of the model.

By understanding how sensitive your model is to different input variables based on both coefficient changes, RMSE and range, one can gain valuable insights. This can help you determine which input variables require more precise data collection or highlight areas where the model might be less reliable due to high sensitivity to specific input variations.

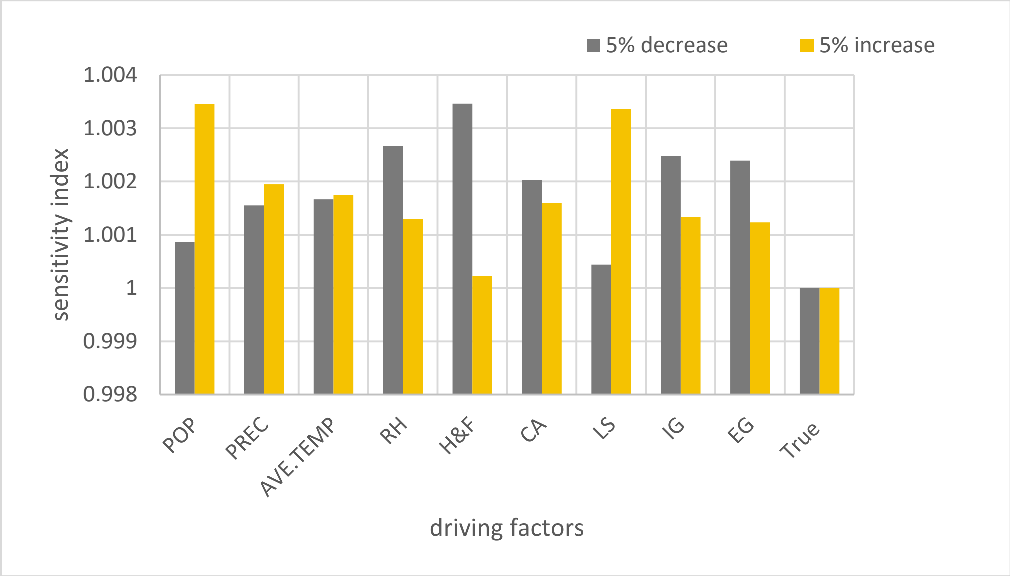

Figure 4-9 sensitive index with 5% increasing and decreasing on the weight of the driving factors

Based on the above figure show holyday and festival, population and number of livestock are more sensitive to other remain factors. So, by using three more sensitive input variable to develop the following model.

$$\mathrm{WC} = 0.3168 * POP + 0.1715 * H\&F + 0.2282 * LS + 0.0979$$

Table 4-9 model training, testing and validation

Model	MSE (billion) (m^3/Month)			RMSE (million m^3/Month)			NRMSE		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
ANN with five hidden nodes	98.9	26.8	33.9	0.3144	0.1637	0.1841	0.0587	00.1328	0.1513

Table 4-10 Accuracy of the calibrated of the model

Model	MSE (billon)	RMSE (million)	NRMSE (-)	MAPE (%)	Accuracy (%)
ANN model with five hidden nodes	161	0.4008	0.0509	4.0794	95.92

4.9 Future projections of water consumption

Water consumption forecasting has traditionally relied solely on population growth, neglecting other significant factors. This research proposes a more comprehensive approach that incorporates additional variables to improve forecasting accuracy. These variables include the number of livestock, average temperature, industrial activity and economic activity levels, the frequency of festivals and holidays, construction activity and relative humidity. By considering these additional influences more accurate prediction of water consumption is possible.

This study shows a solid grasp of the main factors influencing water use. One can improve the model's accuracy and prediction capacity by possibly adding the extra factors and concentrating on the data gathering and analysis for these components. In order to ensure sustainable water consumption and plan for future needs, this will be helpful for water resource management.

5. CONCLUSIONS AND RECOMMENDATION

5.1 CONCLUSIONS

Accurate water consumption forecasting and determination of the driving factors are crucial for water utilities, as it informs planning and operational decisions. Traditional models relying solely on population growth limited their effectiveness. This research proposes an enhanced forecasting model that incorporates additional variables like livestock, average temperature, industrial activity, economic growth, relative humidity, construction activity, holiday and festival. This broader approach provides a more comprehensive understanding of water usage patterns, as population alone cannot account for variations due to agriculture, climate, and industry.

GA Multiple Linear regression, GA non-linear regression, ANN model with one hidden node and ANN model with five hidden nodes models were developed to forecast monthly water consumption and determination of its driving factors in Addis Ababa city. The outcomes show the dominance of the ANN model with five hidden nodes, achieving a simulation accuracy of 96.93% compared to 96.83%, 96.60 and 96.37% for ANN model with one hidden node, nonlinear and linear GA model respectively. This highlights the effectiveness of the ANN model with five hidden nodes approach in water consumption modeling and determination of the driving factors.

Based on the best model selection (ANN model with five hidden nodes) Addis Ababa water consumption forecasting by considering the driving factors: population, average temperature, construction activity, relative humidity, economic development, agricultural activity, industrial development, holiday/festival, and precipitation was enhanced. Overall, the enhanced model, considering a wider range of influencing factors, offers more accurate future water consumption estimations. Based on sensitive analysis holyday and festival, population and number of livestock are more sensitive. By providing the knowledge required for efficient planning and management of water resources, this enables decision-makers to adopt sustainable and effective methods that adapt to changing circumstances.

5.2 RECOMMENDATION

Preceding analysis informed the ANN with five hidden nodes appears to be the best recommendation for water consumption prediction and determination of driving factors. However, here are some additional recommendations incorporating the use of Genetic Algorithms (GAs) and considering potential additional data:

Hybrid Approach with Feature Selection:

- Utilize a Genetic Algorithm (GA) for feature selection. This involves the GA searching for the optimal subset of input variables that contribute most significantly to water consumption prediction and driving factors determination.
- Train an ANN model with the selected features identified by the GA. This can potentially improve prediction accuracy and reduce model complexity compared to using all initial input variables.

Non-linear Regression with GA:

- While your analysis suggests ANN outperforms linear regression, consider exploring non-linear regression models like Support Vector Regression (SV R) or Gaussian Process Regression (GPR) in conjunction with GA. Train these non-linear models with the full set of input variables and compare their performance to the ANN model. They might capture complex non-linear relationships that linear regression misses.

Additional Data Exploration:

- Investigate the potential impact of including additional relevant data:
  - **Time-series data:** Incorporate historical water consumption data to capture temporal patterns and trends.
  - **Socio-economic factors:** Consider factors like income levels, water pricing, and population density.
  - **Urbanization:** The percentage of people living in cities relative to the overall population

- **Climate data:** Include variables like wind speed, sunshine hours, and evapotranspiration.
- **Tourism data:** number of tourists increase the number of populations
- Analyze the impact of these additional data points on both the GA and ANN

- To determine future water demand based on consumption value, it recommended to adjust the consumption value by multiplying it with several factors. These factors account for various influences that can impact future water needs.

REFERENCE

Ababa, A. (2007). "Climate change national adaptation programme of action (Napa) of Ethiopia." National Meteorological Services Agency, Ministry of Water Resources, Federal Democratic Republic of Ethiopia, Addis Ababa.

Altunkaynak, A., M. Özger and M. Çakmakci (2005). "Water consumption prediction of Istanbul city by using fuzzy logic approach." Water resources management **19**: 641-654.

Baeck, T., D. Fogel and Z. Michalewicz (2018). A history of evolutionary computation. Evolutionary Computation 1, CRC Press**:** 78-96.

Balakrishnan, S. and R. Weil (1996). "Neurocontrol: A literature survey." Mathematical and Computer Modelling **23**(1-2): 101-117.

Basheer, I. A. and M. Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application." Journal of microbiological methods **43**(1): 3-31.

Baxter, C., S. Stanley, Q. Zhang and D. W. Smith (2002). "Developing artificial neural network models of water treatment processes: a guide for utilities." Journal of Environmental Engineering and Science **1**(3): 201-211.

Billings, R. B. and C. V. Jones (2011). Forecasting urban water demand, American Water Works Association.

Bowden, G. J., G. C. Dandy and H. R. Maier (2005). "Input determination for neural network models in water resources applications. Part 1—background and methodology." Journal of Hydrology **301**(1-4): 75-92.

Bowden, G. J., H. R. Maier and G. C. Dandy (2002). "Optimal division of data for neural network models in water resources applications." Water resources research **38**(2): 2-1-2-11.

Bozokalfa, G. (2005). " ANN" artifical neural networks and fuzzy logic models for cooling load prediction, Izmir Institute of Technology (Turkey).

Cabral, M., D. Loureiro, A. Mamade and D. Covas (2014). "Water demand projection in distribution systems using a novel scenario planning approach." Procedia Engineering **89**: 950-957.

Donkor, E. A., T. A. Mazzuchi, R. Soyer and J. Alan Roberson (2014). "Urban water demand forecasting: review of methods and models." Journal of Water Resources Planning and Management **140**(2): 146-159.

Fenta, T. M. (2014). "Demands for urban public transportation in Addis Ababa." Journal of intelligent transportation and urban planning **2**(3): 81-88.

Fernando, T., H. Maier, G. Dandy and R. May (2005). Efficient selection of inputs for artificial neural network models. Proc. of MODSIM 2005 International Congress on Modelling and Simulation: Modelling and Simulation Society of Australia and New Zealand, Citeseer.

Fuangkhon, P. (2014). "An incremental learning preprocessor for feed-forward neural network." Artificial Intelligence Review **41**(2): 183-210.

Garg, S. K. (2005). Soil Mechanics and Foundation Engineering: For Civil Engineering Degree Students; AMIE (section B) Exams-New Scheme; GATE Exams; UPSC and Other State Service Competitions; and for Professional Field Engineers, Khanna.

Gündüz, A. Y. and B. Akkoyunlu (2020). "Effectiveness of gamification in flipped learning." Sage Open **10**(4): 2158244020979837.

Hartley, J. A. (1995). A neural network and rule based system application in water demand forecasting, Brunel University School of Engineering and Design PhD Theses.

Haykin, S. (2009). Neural networks and learning machines, 3/E, Pearson Education India.

Herrera, M., L. Torgo, J. Izquierdo and R. Pérez-García (2010). "Predictive models for forecasting hourly urban water demand." Journal of hydrology **387**(1-2): 141-150.

Huang, H., Z. Zhang, Z. Lin and S. Liu (2022). "Hourly water demand forecasting using a hybrid model based on mind evolutionary algorithm." Water Supply **22**(1): 917-927.

Hunt, K. J., D. Sbarbaro, R. Żbikowski and P. J. Gawthrop (1992). "Neural networks for control systems—a survey." Automatica **28**(6): 1083-1112.

Loucks, D. P. and E. Van Beek (2017). Water resource systems planning and management: An introduction to methods, models, and applications, Springer.

Maier, H., N. Morgan and C. Chow (2004). "Use of artificial neural networks for."

Maier, H. R. and G. C. Dandy (2000). "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications." Environmental modelling & software **15**(1): 101-124.

Mitchell, M. (1995). Genetic algorithms: An overview. Complex., Citeseer.

Mohri, M., A. Rostamizadeh and A. Talwalkar (2018). Foundations of machine learning, MIT press.

Ravn, O., L. Hansen, N. Poulsen, M. Norgaard and M. Noergaard (2000). Neural networks for modelling and control of dynamic systems, Springer.

Rodriguez, M. J. and J.-B. Sérodes (2004). "Application of back-propagation neural network modeling for free residual chlorine, total trihalomethanes and trihalomethanes speciation." Journal of Environmental Engineering and Science **3**(Supplement 1): S25-S34.

Tolon, U. W. (2008). "Comparision of urban upgrading projects on development cooperation in Ethiopia."

Appendix 1: determination of outlier by using XLSTAT

XLSTAT 2023.1.6.1410 - Dixon test for outliers

Significance level (%): 5

Alternative hypothesis: Two-sided

Iterations: Maximum=1

p-value: Monte Carlo method

Number of simulations: 1000000

Maximum time (s): 180

Seed (random numbers): 123456789

Table 1 Summary statistics:

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Consumption	99	0	99	4895988.000	9611861.000	7830543.525	663119.126

Dixon test for outliers / Two-tailed test:

R22(Observed value) =0.34

R22 (Critical value) = 0.283

p-value (Two-tailed) =0.010

Alpha= 0.05

The p-value has been computed using 1000000 Monte Carlo simulations. 99% confidence interval on the p-value:

Test interpretation:

H0: There is no outlier in the data

Ha: The minimum or maximum value is an outlier

As the computed p-value is less than the significance level alpha=0.05, one cannot reject the alternative hypothesis, Ha. The risk to reject the alternative hypothesis Ha

Table 2 Z-value of water consumption with outlier

Observation	Value	Z-score	Observation	Value	Z-score	Observation	Value	Z-score
Obs1	**4895988**	**-4.425**	Obs34	7826970	-0.005	Obs67	**9611861**	**2.686**
Obs2	6738029	-1.648	Obs35	8208224	0.570	Obs68	8416673	0.884

Obs3	**5990320**	**-2.775**	Obs36	8033656	0.306	Obs69	7884647	0.082
Obs4	7003118	-1.248	Obs37	7076139	-1.138	Obs70	8345317	0.776
Obs5	6775683	-1.591	Obs38	7698564	-0.199	Obs71	8232874	0.607
Obs6	7147025	-1.031	Obs39	7364790	-0.702	Obs72	8415169	0.882
Obs7	7352301	-0.721	Obs40	8300142	0.708	Obs73	7415000	-0.627
Obs8	6910522	-1.387	Obs41	8057189	0.342	Obs74	8358479	0.796
Obs9	7146994	-1.031	Obs42	8450074	0.934	Obs75	7296807	-0.805
Obs10	7253931	-0.870	Obs43	8088169	0.389	Obs76	7596651	-0.353
Obs11	7140756	-1.040	Obs44	8681775	1.284	Obs77	8433124	0.909
Obs12	7427536	-0.608	Obs45	8496602	1.004	Obs78	8203007	0.562
Obs13	**6217731**	**-2.432**	Obs46	8569319	1.114	Obs79	8256512	0.642
Obs14	7388208	-0.667	Obs47	8078743	0.374	Obs80	8412426	0.877
Obs15	6909406	-1.389	Obs48	8397558	0.855	Obs81	8353262	0.788
Obs16	7713993	-0.176	Obs49	7758647	-0.108	Obs82	8779371	1.431
Obs17	7727011	-0.156	Obs50	7910023	0.120	Obs83	8089063	0.390
Obs18	7611241	-0.331	Obs51	7507056	-0.488	Obs84	8830469	1.508
Obs19	7502639	-0.494	Obs52	8712274	1.330	Obs85	7629820	-0.303
Obs20	7545231	-0.430	Obs53	7629318	-0.303	Obs86	8012902	0.275
Obs21	7592636	-0.359	Obs54	7337532	-0.743	Obs87	7013975	-1.231
Obs22	8147592	0.478	Obs55	8051808	0.334	Obs88	7942867	0.169
Obs23	7942723	0.169	Obs56	8003864	0.261	Obs89	8596591	1.155
Obs24	8150153	0.482	Obs57	8250575	0.633	Obs90	7982098	0.229
Obs25	7361087	-0.708	Obs58	8408200	0.871	Obs91	8111402	0.424
Obs26	7656120	-0.263	Obs59	7660307	-0.257	Obs92	8445107	0.927
Obs27	7315411	-0.777	Obs60	8457027	0.945	Obs93	8382996	0.833
Obs28	8192470	0.546	Obs61	8203648	0.563	Obs94	8108950	0.420
Obs29	7685774	-0.218	Obs62	8281121	0.679	Obs95	7638935	-0.289
Obs30	7858423	0.042	Obs63	7004501	-1.246	Obs96	7912454	0.124

Obs31	7971852	0.213	Obs64	7719573	-0.167	Obs97	6835784	-1.500
Obs32	8330614	0.754	Obs65	8526112	1.049	Obs98	8030272	0.301
Obs33	8231172	0.604	Obs66	8595465	1.154	Obs99	7470289	-0.543
Values displayed in bold are outliers								

value of relative humidity with outlier

XLSTAT 2023.1.6.1410 - Dixon test for outliers -
Significance level (%): 5
Alternative hypothesis: Two-sided
Iterations: Maximum=1
p-value: Monte Carlo method
Number of simulations: 1000000
Maximum time (s): 180
Seed (random numbers): 123456789
Summary statistics

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean
RH	99	0	99	30.297	79.923	55.703

Dixon test for outliers / Two-tailed test:

R22(Observed value) =0.099

R22 (Critical value) = 0.283

p-value (Two-tailed) =0.869

Alpha= 0.05

Test interpretation:
H0: There is no outlier in the data
Ha: The two lowest or greatest values are outliers

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0

Table 3 Z-value of relative humidity data with outlier

Observation	Value	Z-score	Observation	Value	Z-score	Observation	Value	Z-score
Obs1	47.761	-0.620	Obs34	52.090	-0.282	Obs67	43.961	-0.916
Obs2	47.761	-0.620	Obs35	61.677	0.466	Obs68	42.097	-1.062
Obs3	47.761	-0.620	Obs36	56.948	0.097	Obs69	45.271	-0.814

Obs4	47.761	-0.620	Obs37	67.052	0.885	Obs70	36.871	-1.469
Obs5	47.761	-0.620	Obs38	77.690	1.715	Obs71	41.097	-1.140
Obs6	47.761	-0.620	Obs39	77.226	1.679	Obs72	51.303	-0.343
Obs7	52.895	-0.219	Obs40	61.910	0.484	Obs73	59.026	0.259
Obs8	52.895	-0.219	Obs41	54.439	-0.099	Obs74	59.026	0.259
Obs9	52.895	-0.219	Obs42	52.587	-0.243	Obs75	77.045	1.665
Obs10	52.895	-0.219	Obs43	45.242	-0.816	Obs76	66.710	0.859
Obs11	52.895	-0.219	Obs44	37.897	-1.389	Obs77	31.613	-1.879
Obs12	52.895	-0.219	Obs45	38.935	-1.308	Obs78	43.097	-0.984
Obs13	66.555	0.847	Obs46	45.490	-0.797	Obs79	48.310	-0.577
Obs14	76.968	1.659	Obs47	53.761	-0.151	Obs80	48.961	-0.526
Obs15	76.987	1.661	Obs48	44.019	-0.912	Obs81	38.548	-1.338
Obs16	70.194	1.131	Obs49	68.742	1.017	Obs82	43.677	-0.938
Obs17	46.858	-0.690	Obs50	76.819	1.647	Obs83	48.987	-0.524
Obs18	48.555	-0.558	Obs51	75.710	1.561	Obs84	46.587	-0.711
Obs19	46.806	-0.694	Obs52	60.355	0.363	Obs85	67.787	0.943
Obs20	35.974	-1.539	Obs53	54.948	-0.059	Obs86	79.923	1.890
Obs21	35.974	-1.539	Obs54	56.232	0.041	Obs87	76.858	1.650
Obs22	47.465	-0.643	Obs55	54.800	-0.070	Obs88	70.910	1.186
Obs23	45.471	-0.798	Obs56	48.477	-0.564	Obs89	60.994	0.413
Obs24	63.787	0.631	Obs57	39.910	-1.232	Obs90	50.090	-0.438
Obs25	58.110	0.188	Obs58	52.561	-0.245	Obs91	51.058	-0.362
Obs26	78.516	1.780	Obs59	55.671	-0.002	Obs92	**30.297**	**-1.982**
Obs27	76.639	1.633	Obs60	60.045	0.339	Obs93	35.084	-1.609
Obs28	72.865	1.339	Obs61	66.200	0.819	Obs94	64.129	0.657
Obs29	58.316	0.204	Obs62	69.019	1.039	Obs95	64.129	0.657
Obs30	50.213	-0.428	Obs63	74.645	1.478	Obs96	69.858	1.104
Obs31	46.439	-0.723	Obs64	70.110	1.124	Obs97	67.523	0.922

Obs32	47.010	-0.678	Obs65	48.729	-0.544	Obs98	79.484	1.855
Obs33	47.581	-0.634	Obs66	38.439	-1.347	Obs99	74.677	1.480
Values displayed in bold are outliers								

Average temperature of Summary statistics

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Ave.temp	99	0	99	13.120	19.757	16.457	1.262

Dixon test for outliers / Two-tailed test:

R22(Observed value) =0.188

R22 (Critical value) = 0.283

The p-value has been computed using 1e+06 Monte Carlo simulations.
99% confidence interval on the p-value:
] 0.351, 0.354 [

Test interpretation:
H0: There is no outlier in the data
Ha: The two lowest or greatest values are outliers

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

p-value (Two-tailed) =0.352

Alpha= 0.05

Table 4 Z-value of average temperature data with outlier

Observation	Value	Z-score	Observation	Value	Z-score	Observation	Value	Z-score
Obs1	17.068	0.484	Obs34	17.925	1.163	Obs67	15.585	-0.691
Obs2	16.966	0.403	Obs35	17.792	1.058	Obs68	16.320	-0.108
Obs3	15.980	-0.378	Obs36	18.245	1.417	Obs69	15.321	-0.900
Obs4	16.068	-0.308	Obs37	16.402	-0.043	Obs70	18.349	1.499
Obs5	17.388	0.738	Obs38	14.647	-1.434	Obs71	17.842	1.098

Obs6	16.136	-0.254	Obs39	15.050	-1.114	Obs72	16.595	0.109
Obs7	15.401	-0.837	Obs40	15.813	-0.510	Obs73	16.178	-0.220
Obs8	16.007	-0.357	Obs41	16.241	-0.171	Obs74	14.263	-1.738
Obs9	16.883	0.338	Obs42	15.203	-0.993	Obs75	14.878	-1.251
Obs10	**19.738**	**2.600**	Obs43	15.285	-0.928	Obs76	15.737	-0.570
Obs11	18.240	1.413	Obs44	15.638	-0.648	Obs77	16.145	-0.247
Obs12	18.098	1.301	Obs45	16.149	-0.243	Obs78	15.452	-0.796
Obs13	16.790	0.264	Obs46	**19.757**	**2.615**	Obs79	16.794	0.267
Obs14	16.217	-0.190	Obs47	18.456	1.585	Obs80	16.292	-0.131
Obs15	15.366	-0.864	Obs48	18.700	1.777	Obs81	15.954	-0.398
Obs16	16.016	-0.349	Obs49	16.770	0.249	Obs82	18.178	1.364
Obs17	16.758	0.239	Obs50	15.898	-0.443	Obs83	17.745	1.021
Obs18	15.924	-0.422	Obs51	15.615	-0.667	Obs84	18.614	1.709
Obs19	15.916	-0.428	Obs52	15.421	-0.821	Obs85	14.211	-1.779
Obs20	15.887	-0.452	Obs53	16.273	-0.145	Obs86	**13.120**	**-2.644**
Obs21	15.645	-0.643	Obs54	16.064	-0.311	Obs87	16.099	-0.283
Obs22	18.246	1.418	Obs55	15.717	-0.586	Obs88	15.704	-0.596
Obs23	17.715	0.997	Obs56	16.654	0.156	Obs89	15.953	-0.399
Obs24	17.439	0.778	Obs57	16.757	0.238	Obs90	15.435	-0.810
Obs25	17.426	0.768	Obs58	**19.214**	**2.185**	Obs91	16.394	-0.049
Obs26	16.148	-0.244	Obs59	18.427	1.562	Obs92	16.204	-0.200
Obs27	15.503	-0.755	Obs60	18.784	1.844	Obs93	15.903	-0.438
Obs28	15.202	-0.994	Obs61	16.027	-0.340	Obs94	17.552	0.868
Obs29	16.317	-0.111	Obs62	14.958	-1.187	Obs95	17.714	0.996
Obs30	15.075	-1.094	Obs63	15.037	-1.125	Obs96	17.758	1.031
Obs31	14.532	-1.525	Obs64	15.227	-0.974	Obs97	17.255	0.632
Obs32	15.585	-0.691	Obs65	15.975	-0.381	Obs98	17.516	0.839
Obs33	15.676	-0.619	Obs66	15.152	-1.034	Obs99	17.516	0.839

Values displayed in bold are outliers

Precipitation data with outlier

Significance level (%): 5
Alternative hypothesis: Two-sided
Iterations: Maximum=1
p-value: Monte Carlo method
Number of simulations: 1000000
Maximum time (s): 180
Seed (random numbers): 123456789
Dixon test for outliers / Two-tailed test:

R22(Observed value) =0.009

R22 (Critical value) = 0.283

p-value (Two-tailed) =0.012

Summary statistics

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
PREC	99	0	99	0.000	11.673	3.288	3.598

The p-value has been computed using 1000000 Monte Carlo simulations. 99% confidence interval on the p-value:

Test interpretation:

H0: There is no outlier in the data

Ha: The minimum or maximum value is an outlier

As the computed p-value is less than the significance level alpha=0.05, one cannot reject the alternative hypothesis, Ha. The risk to reject the alternative hypothesis

Table 5 Z-value of precipitation data with outlier

Observation	Value	Z-score	Observation	Value	Z-score	Observation	Value	Z-score
Obs1	4.666	0.383	Obs34	0.753	-0.704	Obs67	0.126	-0.879
Obs2	8.265	1.383	Obs35	2.522	-0.213	Obs68	0.000	-0.914
Obs3	8.818	1.537	Obs36	2.688	-0.167	Obs69	1.190	-0.583
Obs4	2.808	-0.134	Obs37	7.554	1.186	Obs70	0.164	-0.868

Obs5	0.090	-0.889	Obs38	**11.166**	**2.189**	Obs71	3.569	0.078
Obs6	0.389	-0.806	Obs39	**11.570**	**2.302**	Obs72	2.552	-0.205
Obs7	0.013	-0.910	Obs40	1.573	-0.477	Obs73	3.566	0.077
Obs8	0.839	-0.681	Obs41	1.083	-0.613	Obs74	7.076	1.053
Obs9	0.754	-0.704	Obs42	0.407	-0.801	Obs75	8.514	1.452
Obs10	2.138	-0.320	Obs43	0.000	-0.914	Obs76	6.297	0.836
Obs11	4.531	0.345	Obs44	0.000	-0.914	Obs77	2.651	-0.177
Obs12	4.474	0.329	Obs45	0.254	-0.843	Obs78	0.023	-0.908
Obs13	4.477	0.330	Obs46	0.925	-0.657	Obs79	0.000	-0.914
Obs14	6.973	1.024	Obs47	3.012	-0.077	Obs80	0.194	-0.860
Obs15	7.684	1.222	Obs48	3.287	0.000	Obs81	0.027	-0.906
Obs16	4.385	0.305	Obs49	5.376	0.580	Obs82	1.335	-0.543
Obs17	0.720	-0.714	Obs50	**10.958**	**2.131**	Obs83	0.998	-0.636
Obs18	0.184	-0.863	Obs51	8.537	1.459	Obs84	0.000	-0.914
Obs19	0.012	-0.911	Obs52	8.009	1.312	Obs85	4.560	0.354
Obs20	0.000	-0.914	Obs53	0.446	-0.790	Obs86	**11.584**	**2.305**
Obs21	0.767	-0.701	Obs54	0.556	-0.759	Obs87	10.108	1.895
Obs22	1.231	-0.572	Obs55	0.323	-0.824	Obs88	5.432	0.596
Obs23	0.902	-0.663	Obs56	0.019	-0.908	Obs89	1.106	-0.607
Obs24	3.784	0.138	Obs57	0.026	-0.906	Obs90	0.129	-0.878
Obs25	2.378	-0.253	Obs58	2.328	-0.267	Obs91	0.000	-0.914
Obs26	10.198	1.920	Obs59	3.602	0.087	Obs92	0.162	-0.869
Obs27	9.799	1.810	Obs60	3.681	0.109	Obs93	0.256	-0.843
Obs28	9.674	1.775	Obs61	5.113	0.507	Obs94	4.380	0.303
Obs29	0.453	-0.788	Obs62	**11.541**	**2.293**	Obs95	4.306	0.283
Obs30	0.005	-0.912	Obs63	**11.673**	**2.330**	Obs96	6.505	0.894
Obs31	0.000	-0.914	Obs64	7.229	1.095	Obs97	0.760	-0.702
Obs32	0.000	-0.914	Obs65	0.682	-0.724	Obs98	7.412	1.146

Obs33	0.940	-0.653	Obs66	0.012	-0.911	Obs99	1.287	-0.556
Values displayed in bold are outliers								

Appendix 2: Input data without outlier

Table 6 Input data without outlier

ID	Date	POP (Number)	PREC (mm)	Ave.Temp (° c)	RH (%)	H&F (Number)	CA (%)	LS (Number)	IG (%)	GR (%)	Cons (m3/mont
0	42156	3871000	4.665591	17.06806	47.76129	0	-0.02837	1306892	1.566134	0.759153	7067265
1	42186	3871000	8.264516	16.96581	47.76129	1	-0.02837	1306892	1.566134	0.759153	7131412
2	42217	3871000	8.818065	15.97984	47.76129	0	-0.02837	1306892	1.566134	0.759153	7067265
3	42248	3871000	2.807527	16.06774	47.76129	3	-0.02837	1306892	1.566134	0.759153	7003118
4	42278	3871000	0.090323	17.38806	47.76129	0	-0.02837	1306892	1.566134	0.759153	7075072
5	42309	3871000	0.38871	16.13613	47.76129	0	-0.02837	1306892	1.566134	0.759153	7147025
6	42339	3871000	0.012903	15.40073	52.89516	1	-0.02837	1306892	1.566134	0.759153	7352301
7	42370	4040000	0.839247	16.00661	52.89516	2	-0.02445	1389816	1.566134	0.759153	6910522
8	42401	4040000	0.754409	16.88266	52.89516	0	-0.02445	1389816	1.566134	0.759153	7146994
9	42430	4040000	2.137634	17.56133	52.89516	1	-0.02445	1389816	1.566134	0.759153	7253931
10	42461	4040000	4.530645	18.24	52.89516	1	-0.02445	1389816	1.566134	0.759153	7140756
11	42491	4040000	4.473656	18.09839	52.89516	3	-0.02445	1389816	1.566134	0.759153	7427536
12	42522	4040000	4.476882	16.78968	66.55484	0	-0.02445	1389816	-0.5143	0.573243	7407872
13	42552	4040000	6.973118	16.2171	76.96774	1	-0.02445	1389816	-0.5143	0.573243	7388208
14	42583	4040000	7.684409	15.36581	76.9871	0	-0.02445	1389816	-0.5143	0.573243	6909406
15	42614	4040000	4.385484	16.01613	70.19355	3	-0.02445	1389816	-0.5143	0.573243	7713993
16	42644	4040000	0.72043	16.75839	46.85806	0	-0.02445	1389816	-0.5143	0.573243	7727011
17	42675	4040000	0.184409	15.92387	48.55484	0	-0.02445	1389816	-0.5143	0.573243	7611241
18	42705	4040000	0.011828	15.91645	46.80645	0	-0.02445	1389816	-0.5143	0.573243	7502639
19	42736	4216000	0	15.88677	35.97419	2	-0.01822	1489622	-0.5143	0.573243	7545231
20	42767	4216000	0.767204	15.64452	35.97419	0	-0.01822	1489622	-0.5143	0.573243	7592636
21	42795	4216000	1.230645	18.24581	47.46452	1	-0.01822	1489622	-0.5143	0.573243	8147592
22	42826	4216000	0.901613	17.71452	45.47097	2	-0.01822	1489622	-0.5143	0.573243	7942723
23	42856	4216000	3.783871	17.43871	63.7871	3	-0.01822	1489622	-0.5143	0.573243	8150153
24	42887	4216000	2.377957	17.42613	58.10968	1	-0.01822	1489622	0.040742	0.674413	7361087
25	42917	4216000	10.19839	16.14806	78.51613	0	-0.01822	1489622	0.040742	0.674413	7656120
26	42948	4216000	9.799462	15.50323	76.63871	0	-0.01822	1489622	0.040742	0.674413	7315411
27	42979	4216000	9.673656	15.20226	72.86452	3	-0.01822	1489622	0.040742	0.674413	8192470
28	43009	4216000	0.452688	16.3171	58.31613	0	-0.01822	1489622	0.040742	0.674413	7685774
29	43040	4216000	0.005376	15.07548	50.2129	1	-0.01822	1489622	0.040742	0.674413	7858423
30	43070	4216000	0	14.53226	46.43871	0	-0.01822	1489622	0.040742	0.674413	7971852
31	43101	4400000	0	15.58516	47.00968	2	-0.0163	1589422	0.040742	0.674413	8330614

32	43132	4400000	0.939785	15.67581	47.58065	0	-0.0163	1589422	0.040742	0.674413	8231172
33	43160	4400000	0.753226	17.92484	52.09032	0	-0.0163	1589422	0.040742	0.674413	7826970
34	43191	4400000	2.521505	17.79194	61.67742	1	-0.0163	1589422	0.040742	0.674413	8208224
35	43221	4400000	2.687634	18.24548	56.94839	2	-0.0163	1589422	0.040742	0.674413	8033656
36	43252	4400000	7.554301	16.40194	67.05161	3	-0.0163	1589422	0.723042	0.971385	7076139
37	43282	4400000	4.56371	14.64694	77.69032	1	-0.0163	1589422	0.723042	0.971385	7698564
38	43313	4400000	3.068414	15.05032	77.22581	0	-0.0163	1589422	0.723042	0.971385	7364790
39	43344	4400000	1.573118	15.81258	61.90968	2	-0.0163	1589422	0.723042	0.971385	8300142
40	43374	4400000	1.082796	16.24129	54.43871	0	-0.0163	1589422	0.723042	0.971385	8057189
41	43405	4400000	0.406989	15.20323	52.5871	1	-0.0163	1589422	0.723042	0.971385	8450074
42	43435	4400000	0	15.28548	45.24194	0	-0.0163	1589422	0.723042	0.971385	8088169
43	43466	4592000	0	15.63839	37.89677	2	-0.00314	1689214	0.723042	0.971385	8681775
44	43497	4592000	0.254301	16.14935	38.93548	0	-0.00314	1689214	0.723042	0.971385	8496602
45	43525	4592000	0.925269	17.3029	45.49032	1	-0.00314	1689214	0.723042	0.971385	8569319
46	43556	4592000	3.011828	18.45645	53.76129	2	-0.00314	1689214	0.723042	0.971385	8078743
47	43586	4592000	3.286559	18.69968	44.01935	3	-0.00314	1689214	0.723042	0.971385	8397558
48	43617	4592000	5.375806	16.77032	68.74194	1	-0.00314	1689214	1.229956	0.549737	7758647
49	43647	4592000	6.956452	15.89806	76.81935	0	-0.00314	1689214	1.229956	0.549737	7910023
50	43678	4592000	8.537097	15.61452	75.70968	1	-0.00314	1689214	1.229956	0.549737	7507056
51	43709	4592000	8.00914	15.42065	60.35484	2	-0.00314	1689214	1.229956	0.549737	8712274
52	43739	4592000	0.446237	16.27323	54.94839	0	-0.00314	1689214	1.229956	0.549737	7629318
53	43770	4592000	0.556452	16.06355	56.23226	1	-0.00314	1689214	1.229956	0.549737	7337532
54	43800	4592000	0.322581	15.71677	54.8	0	-0.00314	1689214	1.229956	0.549737	8051808
55	43831	4592000	0.019355	16.65355	48.47742	2	0.036861	1689214	1.229956	0.549737	8003864
56	43862	4794000	0.026344	16.75742	39.90968	0	0.036861	1722998	1.229956	0.549737	8250575
57	43891	4794000	2.328495	17.59242	52.56129	1	0.036861	1722998	1.229956	0.549737	8408200
58	43922	4794000	3.602151	18.42742	55.67097	2	0.036861	1722998	1.229956	0.549737	7660307
59	43952	4794000	3.680645	18.78419	60.04516	4	0.036861	1722998	1.229956	0.549737	8457027
60	43983	4794000	5.112903	16.0271	66.2	0	0.036861	1722998	1.229956	0.573243	8203648
61	44013	4794000	5.641935	14.95839	69.01935	1	0.036861	1722998	0.705319	0.573243	8281121
62	44044	4794000	6.170968	15.03677	74.64516	0	0.036861	1722998	0.705319	0.573243	7004501
63	44075	4794000	7.229032	15.2271	70.10968	2	0.036861	1722998	0.705319	0.573243	7719573
64	44105	4794000	0.68172	15.97548	48.72903	1	0.036861	1722998	0.705319	0.573243	8526112
65	44136	4794000	0.011828	15.15194	38.43871	0	0.036861	1722998	0.705319	0.573243	8595465
66	44166	4794000	0.126452	15.58452	43.96129	0	0.036861	1722998	0.705319	0.573243	8506069
67	44197	4794000	0	16.32	42.09677	2	-0.00033	1722998	0.705319	0.573243	8416673
68	44228	5006000	1.190323	15.32065	45.27097	0	-0.00033	1766063	0.705319	0.573243	7884647
69	44256	5006000	0.163978	18.34903	36.87097	1	-0.00033	1766063	0.705319	0.573243	8345317
70	44287	5006000	3.568817	17.84202	41.09677	1	-0.00033	1766063	0.705319	0.573243	8232874
71	44317	5006000	2.551742	16.59476	51.30323	5	-0.00033	1766063	0.705319	0.573243	8415169
72	44348	5006000	3.566452	16.17839	59.02581	0	-0.00033	1766063	0.581066	0.706861	7415000

73	44378	5006000	7.076344	14.26339	59.02581	1	-0.00033	1766063	0.581066	0.706861	8358479
74	44409	5006000	8.513978	14.87839	77.04516	0	-0.00033	1766063	0.581066	0.706861	7296807
75	44440	5006000	6.297419	15.73742	66.70968	2	-0.00033	1766063	0.581066	0.706861	7596651
76	44470	5006000	2.651075	16.14516	31.6129	1	-0.00033	1766063	0.581066	0.706861	8433124
77	44501	5006000	0.022581	15.45161	43.09677	0	-0.00033	1766063	0.581066	0.706861	8203007
78	44531	5006000	0	16.79409	48.30968	0	-0.00033	1766063	0.581066	0.706861	8256512
79	44562	5006000	0.193548	16.29194	48.96129	2	0.004472	1766063	0.581066	0.706861	8412426
80	44593	5228000	0.026613	15.95403	38.54839	0	0.004472	1476070	0.581066	0.706861	8353262
81	44621	5228000	1.335484	18.17823	43.67742	1	0.004472	1476070	0.581066	0.706861	8779371
82	44652	5228000	0.998387	17.74516	48.9871	2	0.004472	1476070	0.581066	0.706861	8089063
83	44682	5228000	0	18.61371	46.5871	4	0.004472	1476070	0.581066	0.706861	8830469
84	44713	5228000	4.560215	14.21129	67.7871	0	0.004472	1476070	0.581066	0.706861	7629820
85	44743	5228000	7.334301	15.15516	79.92258	1	0.004472	1476070	0.581066	0.706861	8012902
86	44774	5228000	10.10839	16.09903	76.85806	0	0.004472	1476070	0.581066	0.341599	7013975
87	44805	5228000	5.432258	15.70387	70.90968	2	0.004472	1476070	0.581066	0.341599	7942867
88	44835	5228000	1.105645	15.95262	60.99355	1	0.004472	1476070	0.581066	0.341599	8596591
89	44866	5228000	0.129032	15.43492	50.09032	0	0.004472	1476070	0.581066	0.341599	7982098
90	44896	5228000	0	16.39435	51.05806	0	0.004472	1476070	0.581066	0.341599	8111402
91	44927	5228000	0.162366	16.20403	43.07097	2	0.009346	1476070	0.581066	0.341599	8445107
92	44958	5461000	0.255914	15.90323	35.08387	0	0.009346	1653139	0.581066	0.341599	8382996
93	44986	5461000	4.37957	17.55161	64.12903	1	0.009346	1653139	0.581066	0.341599	8108950
94	45017	5461000	4.306452	17.71371	64.12903	3	0.009346	1653139	0.581066	0.338757	7638935
95	45047	5461000	6.505376	17.75806	69.85806	3	0.009346	1653139	0.581066	0.338757	7912454
96	45078	5461000	0.760484	17.25484	67.52258	1	0.009346	1653139	0.581066	0.338757	7971363
97	45108	5461000	7.411794	17.51613	79.48387	0	0.009346	1653139	0.581066	0.338757	8030272
98	45139	5461000	1.286559	17.51613	74.67742	0	0.009346	1653139	0.581066	0.338757	7470289

Appendix 3: multiple Linear Regression GA model Code

```
clc
clear
%This demonstration illustrates how a function fitting genetic algorithm with
linear regression model
%The Problem: Predict water consumption and driving factors determination in
the case of Addis Ababa city. To build genetic algorithm for the predication
of water consumption and driving factor for the future
%The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. Average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. Economic growth (%)
 %10. Water consumption data (m3/month)
 %Input data
```

```
ainp1=xlsread('Input','NND');
% The following specifies input and output variables
data=ainp1;
n=size(data,1);
idx=randperm(n);
trian_idx=idx(1:round(0.7*n));
val_idx=idx(round(0.7*n)+1:round(0.85*n));
test_idx=idx(round(0.85*n)+1:end);
trian_data=data(trian_idx,:);
val_data=data(val_idx,:);
test_data=data(test_idx,:);
modinp=trian_data(:,3:11);
[ndata,nc]=size(modinp);
out1=trian_data(:,12);
% Normalization of input and output data
inmax=max(modinp); % the highest values across all data
inmin=min(modinp); % the lowest numbers over the whole
lfmax=1; % the highest number that should be given to the logistic activation
function
lfmin=-1; % the lowest number that should be entered for the logistic
activation function
indnorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1:ndata
    for j=1:nc
        indnorm(i,j)=lfmin+(lfmax-lfmin)*(modinp(i,j)-inmin(j))/...
            (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*(out1(i,1)-min(out1))/...
            (max(out1)-min(out1));
end
%% multiple linear regression
obf=@(x)sqrt(mean(((x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,
3)+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm
(:,7)+ x(1,8)*indnorm(:,8)+ x(1,9)*indnorm(:,9)+x(1,10))-outdnorm(:,1)).^2));
Ai1=diag(-1*ones(1,10)); bi1=zeros(10,1);% constraint that makes all the nine
variables to be > 1
Ai1(10,10)=0;%and the last variable to be any number%(@consan,12);
Ai1(1,1)=-1;
Ai1(2,2)=1;
Ai1(4,4)=1;
if  Ai1(4,4) > Ai1(1,1)
  disp('Absolute value of The element at (1,1) is greater than the element at
(4,4).');
else
  disp('Absolute value of The element at (1,1) is less than or equal to the
element at (4,4).');
end
[x,fval1,exitflag1,output1,population1]= ga(obf,10,Ai1,bi1,[],[]);
xmodr=
x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)+x(1,4)*indnorm(:,
4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1,9)*indnorm(:,9)+x(1,10);
figure
plot(trian_data(:,2),outdnorm, '-b',trian_data(:,2),xmodr,'-r')
title('Water Consumption')
```

```
xlabel('Year/Month')
ylabel('Normalized Consumption (m3)')
legend('Measured','Modelled')
disp('Optimized coefficients:');
disp(x);
dnorm=min(out1)+ (max(out1)-min(out1))*(xmodr(:,1)-lfmin)/(lfmax-lfmin);
%% Fitting Error Histogram
error = out1 -dnorm; % 'dnorm is Forecasted water consumption
ploterrhist(error)
figure
plot(trian_data(:,2),out1, '-b', trian_data(:,2),dnorm,'-
r',trian_data(:,2),error,'g')
title('Water Consumption')
xlabel('Year/Month')
ylabel('Consumption (m3)')
legend('Measured','Modelled','error')
%validation
inputval=val_data(:,3:11);
[ndata,nc]=size(inputval);
outval=val_data(:,12);
% Normalization of validation data
inmax=max(inputval);
inmin=min(inputval);
1fmax=1;
1fmin=-1;
indnormv=zeros(ndata,nc);%normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
        indnormv(I,j)=1fmin+(1fmax-1fmin)* inputval (I,j)-inmin(j))/...
                              (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*(outval(i,1)-min(outval))/...
            (max(outval)-min(outval));
end
modelval= x(1,1)*indnormv(:,1)+x(1,2)*indnormv(:,2)+ x(1,3)*indnormv(:,3)...
+x(1,4)*indnormv(:,4)+x(1,5)*indnormv(:,5)+x(1,6)*indnormv(:,6)+x(1,7)indnorm
v(:,7)+ x(1,8)*indnormv(:,8)+ x(1,9)*indnormv(:,9)+ x(1,10);
outvalN=min(outval)+ (max(outval)-min(outval))*(modelval(:,1)-lfmin)/(lfmax-
lfmin);
errorva=outval-outvalN;
plotregression(outval,outvalN)
%test
inputtest=test_data(:,3:11);
[ndata,nc]=size(inputtest);
outtest=test_data(:,12);
% Normalization of test data
inmax=max(inputtest);
inmin=min(inputtest);
1fmax=1;
1fmin=-1;
indnormT=zeros(ndata,nc); %normalized input data
outdnormT=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
        indnormv(I,j)=1fmin+(1fmax-1fmin)* inputtest (I,j)-inmin(j))/...
```

```
                                   (inmax(j)-inmin(j));
    end
    outdnormT (i,1)=lfmin+(lfmax-lfmin)*( outtest (i,1)-min(outtest)/...
            (max(outtest)-min(outtest));
end
modeltest= x(1,1)*indnormT(:,1)+x(1,2)*indnormT(:,2)+ x(1,3)*indnormT(:,3)...

+x(1,4)*indnormT(:,4)+x(1,5)*indnormT(:,5)+x(1,6)*indnormT(:,6)+x(1,7)*indnor
mT(:,7)....
+x(1,8)*indnormT(:,8)+ x(1,9)*indnormT(:,9)+x(1,10);
modeldet=min(outtest)+ (max(outtest)-min(outtest))*(modeltest(:,1)-
lfmin)/(lfmax-lfmin);
errortest=outtest-modeldet;
```

5.3 Appendix 4: Water usage simulation results and real inputs for the GA linear regression model

Table 7 Water usage inputs and outputs for the GA linear regression model from 2015 to 2023 (M3/month) that are calculated and real.

Year	Actual	Model	Error	Year	Actual	Model	Error
6/1/2015	7067265	7174357	107092.2	7/1/2019	7507056	7714462	207406.2
7/1/2015	7131412	7219079	87667.73	8/1/2019	8712274	7841976	-870298
8/1/2015	7067265	7042309	-24956	9/1/2019	7629318	7941549	312230.6
9/1/2015	7003118	7619354	616235.9	10/1/2019	7337532	8032837	695304.7
10/1/2015	7075072	7482868	407796.8	11/1/2019	8051808	7944915	-106893
11/1/2015	7147025	7304645	157620.3	12/1/2019	8003864	8227260	223395.7
12/1/2015	7352301	7595035	242734.3	1/1/2020	8250575	8194844	-55730.7
1/1/2016	6910522	7703854	793331.6	2/1/2020	8408200	8135350	-272850
2/1/2016	7146994	7500828	353834.2	3/1/2020	7660307	8114010	453703.4
3/1/2016	7253931	7558685	304754	4/1/2020	8457027	8355043	-101984

4/1/2016	7140756	7475960	335203.9	5/1/2020	8203648	7843838	-359810
5/1/2016	7427536	7944372	516836.1	6/1/2020	8281121	7906724	-374397
6/1/2016	7407872	7198114	-209758	7/1/2020	7004501	7745660	741159.2
7/1/2016	7388208	7183813	-204395	8/1/2020	7719573	7948479	228905.7
8/1/2016	6909406	7016173	106766.9	9/1/2020	8526112	8223713	-302399
9/1/2016	7713993	7487911	-226082	10/1/2020	8595465	8209227	-386238
10/1/2016	7727011	7464406	-262605	11/1/2020	8506069	8169536	-336533
11/1/2016	7611241	7469011	-142230	12/1/2020	8416673	8383393	-33279.7
12/1/2016	7502639	7486904	-15735	1/1/2021	7884647	8238953	354305.8
1/1/2017	7545231	7943468	398237.1	2/1/2021	8345317	8447777	102460.5
2/1/2017	7592636	7706011	113375	3/1/2021	8232874	8296226	63351.77
3/1/2017	8147592	7726273	-421319	4/1/2021	8415169	8677645	262476.4
4/1/2017	7942723	7854096	-88626.7	5/1/2021	7415000	8120779	705778.8
5/1/2017	8150153	7731733	-418420	6/1/2021	8358479	7915418	-443061
6/1/2017	7361087	7603741	242653.6	7/1/2021	7296807	7818501	521693.6
7/1/2017	7656120	7127885	-528235	8/1/2021	7596651	8096037	499386.1
8/1/2017	7315411	7167223	-148188	9/1/2021	8433124	8442491	9367.339
9/1/2017	8192470	7495047	-697423	10/1/2021	8203007	8033897	-169110
10/1/2017	7685774	7461916	-223858	11/1/2021	8256512	8322192	65680.36
11/1/2017	7858423	7926854	68430.95	12/1/2021	8412426	8769805	357378.8
12/1/2017	7971852	7701136	-270716	1/1/2022	8353262	8201375	-151887
1/1/2018	8330614	8086106	-244508	2/1/2022	8779371	8459736	-319635
2/1/2018	8231172	7839877	-391295	3/1/2022	8089063	8586938	497874.8
3/1/2018	7826970	7824380	-2590.05	4/1/2022	8830469	8576423	-254046
4/1/2018	8208224	7711959	-496265	5/1/2022	7629820	7607041	-22778.7
5/1/2018	8033656	7933210	-100446	6/1/2022	8012902	7764233	-248669
6/1/2018	7076139	8143334	1067195	7/1/2022	7013975	7432033	418058.4
7/1/2018	7698564	7733234	34669.71	8/1/2022	7942867	7846404	-96463

8/1/2018	7364790	7686456	321665.6	9/1/2022	8596591	8148183	-448408
9/1/2018	8300142	8054888	-245254	10/1/2022	7982098	7965210	-16887.7
10/1/2018	8057189	7914978	-142211	11/1/2022	8111402	8115484	4081.72
11/1/2018	8450074	8192525	-257549	12/1/2022	8445107	8225348	-219759
12/1/2018	8088169	7776088	-312081	1/1/2023	8382996	8027283	-355713
1/1/2019	8681775	8455687	-226088	2/1/2023	8108950	8104042	-4907.54
2/1/2019	8496602	8232255	-264347	3/1/2023	7638935	8315302	676367.2
3/1/2019	8569319	8273067	-296252	4/1/2023	7912454	8198584	286130.1
4/1/2019	8078743	8251917	173173.7	5/1/2023	7971363	8021707	50343.87
5/1/2019	8397558	8413258	15700.17	6/1/2023	8030272	7786719	-243553
6/1/2019	7758647	7897361	138713.7	7/1/2023	7470289	8037192	566903.2
				8/1/2023	7470289	8037192	566903.2

Appendix 5: Non-régression GA model Code

```
clc
clear
% nonlinear regression GA model code
% The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. Average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
 %10. Water consumption data (m3/month)
 % Input data
ainp1=xlsread('Input','NND');
%dividing data
data=ainp1;
n=size(data,1);
idx=randperm(n);
train_idx=(idx(1: round(0.7*n)));
val_idx = (idx(1: round(0.7*n)+1:round(0.85*n));
test_idx =idx(round (0.85*n)+1:end);
trian_data=data (train_idx, :);
val_data=data (val_idx, :);
test_data=data (test_idx, :);
modinp=trian_data(:3:11);
[ndata,nc]=size(modinp);
out1 data =trian_data(:,12);
% Normalization trian data
```

```
inmax=max(modinp);
inmin=min(modinp);
1fmax=1;
1fmin=-1;
indorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1);%normalized output data
for i=1: ndata
    for j=1:nc
indnorm(I,j)=1fmin+(1fmax-1fmin)* modinp(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*(out1(i,1)-min(out1))/...
            (max(out1)-min(out1));
End

%% nonlinear regression
obf=@(x)sqrt(mean(((x(1,1)*(indnorm(:,1).^x(1,10))+x(1,2)*(indnorm(:,2).^x(1,
11))+x(1,3)*(indnorm(:,3).^x(1,12))+x(1,4)*((1,15))+x(1,7)*(indnorm(:,7).^x(1
,16))+x(1,8)*(indnorm(:,8).^x(1,17))+x(1,9)*(indnorm(:,9).^x(1,18))+x(1,19))-
outdnorm(:,1)).^2));
Ai1=diag(-1*ones(1,19)); bi1=zeros(19,1);   %
Ai1(19,19)=0
Ai1(2,2)=1;
Ai1(4,4)=1
thresh=1;
if  Ai1(4,4) > Ai1(1,1)
  disp('Absolute value of the element at(1,1) is greater than  the element at
(4,4).');
else
  disp(Absolute value of the element at(1,1) is less than or equal to  the
element at  (4,4).');
end
[x,fval1,exitflag1,output1,population1]= ga(obf,19,Ai1,bi1,[],[]);
xmodr=x(1,1)*(indnorm(:,1).^x(1,10))+x(1,2)*(indnorm(:,2).^x(1,11))+x(1,3)*(i
ndnorm(:,3).^x(1,12))+x(1,4)*(indnorm(:,4).^x(1,13))+x(1,5)*(indnorm(:,5).^x(
1,14)+x(1,6)*(indnorm(:,6).^x(1,15))+x(1,7)*(indnorm(:,7).^x(1,16))+x(1,8)*(i
ndnorm(:,8).^x(1,17))+x(1,9)*(indnorm(:,9).^x(1,18))+x(1,19);
figure
plot(trian_data(:,2),outdnorm, '-b',trian_data(:,2),xmodr,'-r')
title('Water Consumption')
xlabel('Year/Month')
ylabel('Normalized Consumption (m3)')
legend('Measured','Modelled')
disp('Optimized coefficients:');
disp(x);
%` dnorm is the reverse of normalization
dnorm=min(out1)+ (max(out1)-min(out1))*(xmodr(:,1)-lfmin)/(lfmax-lfmin);
% Fitting Error Histogram
error = out1 -dnorm;% `dnorm is Forecasted water consumption
ploterrhist(error)
plotregression(out1,dnorm);
figure
plot(trian_data(:,2),out1, '-b', trian_data(:,2),dnorm,'-
r',trian_data(:,2),error,'g')
title('Water Consumption')
xlabel('Year/Month')
```

```
ylabel('Consumption (m3)')
legend('Measured','Modelled','error')
%validation
inputval=val_data(:,3:11);
[ndata,nc]=size(inputval);
outval=val_data(:,12);
% Normalization of validation data
inmax=max(inputval);
inmin=min(inputval);
1fmax=1;
1fmin=-1; %
indormv=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
indnormv(i,j)=1fmin+(1fmax-1fmin)*inputv(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*(outval(i,1)-min(outval))/...
            (max(outval)-min(outval));
end
modelval=x(1,1)*(indnormv(:,1).^x(1,10))+x(1,2)*( indnormv(:,2).^x(1,11))+x(1
,3)*( indnormv(:,3).^x(1,12))+x(1,4)*( indnormv(:,4).^x(1,13))+x(1,5)*( indno
rmv(:,5).^x(1,14)+x(1,6)*( indnormv(:,6).^x(1,15))+x(1,7)*( indnormv(:,7).^x(
1,16))+x(1,8)*( indnormv(:,8).^x(1,17))+x(1,9)*( indnormv(:,9).^x(1,18))+x(1,1
9);
outvalN=min(outval)+ (max(outval)-min(outval))*(modelval(:,1)-lfmin)/(lfmax-
lfmin);
errorva=outval-outvalN;
%test
inputtest=test_data(:,3:11);
[ndata,nc]=size(inputval);
outtest=test_data(:,12);
% Normalization of test data
inmax=max(inputtest);
inmin=min(inputtest);
1fmax=1;
1fmin=-1;
indormT=zeros(ndata,nc); %normalized input data
outdnormT=zeros(ndata,1);%normalized output data
for i=1: ndata
    for j=1:nc
indnormT(i,j)=1fmin+(1fmax-1fmin)* inputtes (i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnormT(i,1)=lfmin+(lfmax-lfmin)*(outtest(i,1)-min(outtest))/...
            (max(outtest)-min(outtest));
end
    outdnormT(i,1)=lfmin+(lfmax-lfmin)*(outtest(i,1)-min(outtest))/...
            (max(outtest)-min(outtest));
end
modeltest=x(1,1)*( indnormT(:,1).^x(1,10))+x(1,2)*( indnormT(:,2).^x(1,11))+x
(1,3)*( indnormT(:,3).^x(1,12))+x(1,4)*( indnormT(:,4).^x(1,13))+x(1,5)*( ind
normT(:,5).^x(1,14)+x(1,6)*( indnormT(:,6).^x(1,15))+x(1,7)*( indnormT(:,7).^
x(1,16))+x(1,8)*( indnormT(:,8).^x(1,17))+x(1,9)*( indnormT(:,9).^x(1,18))+x(1
,19);
```

```
modeldet=min(outtest)+ (max(outtest)-
min(outtest))*(modeltest(:,1)lfmin)/(lfmax-lfmin);
errortest=outtest-modeldet;
```

Appendix 6: Actual inputs and Forecasting of water consumption by GA non-linear regression

Table 8 Actual inputs and Forecasting of water consumption (m3/month) by GA non-linear regression Model from 2015 to 2023 (M3/month)

Year	Actual	Model	Error	Year	Actual	Model	Error
6/1/2015	7067265	7110736	43471.14	7/1/2019	7910023	7666156	-243867
7/1/2015	7131412	7175820	44408.21	8/1/2019	7507056	7790579	283523.1
8/1/2015	7067265	6997104	-70160.9	9/1/2019	8712274	7801868	-910406
9/1/2015	7003118	7343164	340045.9	10/1/2019	7629318	7977262	347944.1
10/1/2015	7075072	7434245	359173.4	11/1/2019	7337532	8125038	787506
11/1/2015	7147025	7360578	213552.8	12/1/2019	8051808	7995432	-56376.1
12/1/2015	7352301	7642746	290445.5	1/1/2020	8003864	8472740	468876
1/1/2016	6910522	7753175	842653.1	2/1/2020	8250575	8359975	109400.3
2/1/2016	7146994	7600753	453758.6	3/1/2020	8408200	8260921	-147279
3/1/2016	7253931	7665062	411130.7	4/1/2020	7660307	8205684	545377.3
4/1/2016	7140756	7565883	425127.3	5/1/2020	8457027	8203181	-253846
5/1/2016	7427536	7567580	140044	6/1/2020	8203648	7991587	-212061
6/1/2016	7407872	7357416	-50456.3	7/1/2020	8281121	8133890	-147231
7/1/2016	7388208	7446190	57981.81	8/1/2020	7004501	7955858	951356.7
8/1/2016	6909406	7261731	352325.5	9/1/2020	7719573	8092666	373093.3
9/1/2016	7713993	7521848	-192145	10/1/2020	8526112	8382652	-143460
10/1/2016	7727011	7560169	-166842	11/1/2020	8595465	8377428	-218037
11/1/2016	7611241	7641577	30335.77	12/1/2020	8506069	8313975	-192094
12/1/2016	7502639	7723916	221277.4	1/1/2021	8416673	8404699	-11973.8
1/1/2017	7545231	8082252	537021.5	2/1/2021	7884647	8043774	159126.6
2/1/2017	7592636	7698904	106267.5	3/1/2021	8345317	8347714	2396.547
3/1/2017	8147592	7827632	-319960	4/1/2021	8232874	8087831	-145043

4/1/2017	7942723	7854637	-88086.1	5/1/2021	8415169	8129370	-285799
5/1/2017	8150153	7694350	-455803	6/1/2021	7415000	7953721	538721.4
6/1/2017	7361087	7781552	420465.5	7/1/2021	8358479	7989834	-368645
7/1/2017	7656120	7382841	-273279	8/1/2021	7296807	7804092	507284.9
8/1/2017	7315411	7388569	73157.97	9/1/2021	7596651	8027228	430576.7
9/1/2017	8192470	7553158	-639312	10/1/2021	8433124	8156488	-276636
10/1/2017	7685774	7765652	79878.5	11/1/2021	8203007	8274116	71108.58
11/1/2017	7858423	8068711	210288.1	12/1/2021	8256512	8338991	82478.63
12/1/2017	7971852	7935758	-36094.2	1/1/2022	8412426	8399804	-12621.8
1/1/2018	8330614	8178666	-151948	2/1/2022	8353262	8360681	7419.151
2/1/2018	8231172	7779793	-451379	3/1/2022	8779371	8325526	-453845
3/1/2018	7826970	7803857	-23112.5	4/1/2022	8089063	8351770	262707.3
4/1/2018	8208224	7846587	-361637	5/1/2022	8830469	8594223	-236246
5/1/2018	8033656	7839405	-194251	6/1/2022	7629820	7947976	318156.5
6/1/2018	7076139	7729462	653323.1	7/1/2022	8012902	8085725	72822.59
7/1/2018	7698564	7805064	106500.3	8/1/2022	7013975	7648348	634373
8/1/2018	7364790	7701705	336915.5	9/1/2022	7942867	7924234	-18632.9
9/1/2018	8300142	7947513	-352629	10/1/2022	8596591	8122260	-474331
10/1/2018	8057189	7821708	-235481	11/1/2022	7982098	8092629	110530.9
11/1/2018	8450074	8055553	-394521	12/1/2022	8111402	8210296	98894.16
12/1/2018	8088169	8066205	-21964.3	1/1/2023	8445107	8277736	-167371
1/1/2019	8681775	8403825	-277950	2/1/2023	8382996	8170682	-212314
2/1/2019	8496602	8095855	-400747	3/1/2023	8108950	8071595	-37355
3/1/2019	8569319	8174153	-395166	4/1/2023	7638935	7988320	349384.7
4/1/2019	8078743	8049990	-28753.2	5/1/2023	7912454	7921628	9174.43
5/1/2019	8397558	8038688	-358870	6/1/2023	7971363	8180105	208742
6/1/2019	7758647	7876169	117521.8	7/1/2023	8030272	7734372	-295900
				8/1/2023	7470289	7970963	500674.4

Appendix 7: ANN Code

```
clc
clear
% ANN code
% The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
 %10. Water consumption data (m3/month)
% Input data
ainp1=xlsread('Input','NND');
% The following specifies input and output variables
inputvariable=ainp1(:,3:11);
inputvariable = inputvariable';
water consumption=ainp1(:,12);
water consumption=water consumption';
numNodesLayers=5;
net=fitnet(numNodesLayer);
% View(net)
% Describe  transfer function
net.layer{1}.transferFcn='purelin'; % hidden layer 1 transfer function
net.layer{2}.transferFcn='purelin'; % hidden layer 2 transfer function
%Configure network
net =train(net,inputvariable,waterconsumption);
view(net)
%network training
net.trainFcn='trainlm';
net.performFcn= 'mse' ;
[net,tr)=train(net,inputvariable,waterconsumption );
% nntraintool nntraintool
% creat network 'traingdm'
% net.trainparam.swow=50;
% net.trainparam.epochs=500;  % Number of training times
% net.trainpram.goal=1e-5; % accuracy of training
% net. trainparam.1r=01; % learning rate
% net=train (net, inputvariable,water consumption);
% weight values visualization
IW1=cell2mat(net.IW);   % this is the weights connecting input layer to the
first hidden layer.numlayers-by-numinputs cell array of input weight values
LW1=cell2mat(net.LW);   %% this is the weights connecting first hidden  layer
to the   next hidden layer.numlayers-by-numinputs cell array of input weight
values
b=cell2mat(net.b);  % numLayers -by-1 cell array of bias values
b1=net.b{1};
b1=net.b{1};
Y = sim(net,input variable);     %Y=net(input variable);
Y=Y';
plot(ainp1(:,2),ainp1(:,12), '-b', ainp1(:,2),Y,'-r')
legend('Measured','Modelled')
```

```
title('ANN result')
xlabel('Month')
ylabel('Consumption m3/month')

% determination of the model value by using ANN model equation with one hidden node
clc
clear
% The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
 %10. Water consumption data (m3/month)
 % Input data
ainp1=xlsread('Input','NND');
modinp=ainp1(:,3:11);
[ndata,nc]=size(modinp);
out1=ainp1(:,12);
% Normalization of input and output data
inmax=max(modinp);
inmin=min(modinp);
1fmax=1;
1fmin=-1;
indnorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
indnorm(i,j)=1fmin+(1fmax-1fmin)* modinp(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*( out1 (i,1)-min(out1))/...
             (max(out1)-min(out1));
end
x=xlsread('Input','xvalue1');
y_linear=x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)+x(1,4)*indnorm(:
,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)+x(1,8)*indnorm(:,8)+x
(1,9)*indnorm(:,9)+0.156596407;
dnorm=min(out1)+ (max(out1)-min(out1))*(y_linear(:,1)-lfmin)/(lfmax-lfmin);
```

Appendix 8: Actual inputs and forecasting water consumption by ANN Model with one hidden node

Table 9 Actual inputs and forecasting water consumption (m3/month) for ANN Model with one hidden layer from 2015 to 2023 (M3/month)

Year	Actual	Model	Error	Year	Actual	Model	Error
6/1/2015	7067265	7148671	81406.39	7/1/2019	7910023	7382015	-528008

7/1/2015	7131412	7247718	116306.2	8/1/2019	7507056	7496698	-10358
8/1/2015	7067265	7135721	68456.01	9/1/2019	8712274	7792730	-919544
9/1/2015	7003118	7459866	456747.7	10/1/2019	7629318	7666721	37402.8
10/1/2015	7075072	7157370	82298.03	11/1/2019	7337532	7754157	416624.9
11/1/2015	7147025	7149001	1975.752	12/1/2019	8051808	7665226	-386582
12/1/2015	7352301	7185786	-166515	1/1/2020	8003864	8127120	123256
1/1/2016	6910522	7447502	536980.2	2/1/2020	8250575	8145358	-105217
2/1/2016	7146994	7243294	96300.34	3/1/2020	8408200	8094408	-313792
3/1/2016	7253931	7350520	96589.15	4/1/2020	7660307	8163991	503684.3
4/1/2016	7140756	7351332	210576.1	5/1/2020	8457027	8321454	-135573
5/1/2016	7427536	7560401	132865	6/1/2020	8203648	7824812	-378836
6/1/2016	7407872	7331068	-76804	7/1/2020	8281121	7992484	-288637
7/1/2016	7388208	7298842	-89365.8	8/1/2020	7004501	7817087	812585.8
8/1/2016	6909406	7187227	277821.4	9/1/2020	7719573	8083213	363640
9/1/2016	7713993	7595732	-118261	10/1/2020	8526112	8259280	-266832
10/1/2016	7727011	7582061	-144950	11/1/2020	8595465	8278466	-316999
11/1/2016	7611241	7556387	-54854.2	12/1/2020	8506069	8212147	-293922
12/1/2016	7502639	7578402	75763.23	1/1/2021	8416673	8295062	-121611
1/1/2017	7545231	8101839	556608.1	2/1/2021	7884647	8169215	284568
2/1/2017	7592636	7889302	296666.1	3/1/2021	8345317	8399633	54316.25
3/1/2017	8147592	7866713	-280879	4/1/2021	8232874	8338740	105865.8
4/1/2017	7942723	7993636	50913.46	5/1/2021	8415169	8624761	209592
5/1/2017	8150153	7864137	-286016	6/1/2021	7415000	8135218	720217.7
6/1/2017	7361087	7699269	338182.3	7/1/2021	8358479	8222906	-135573
7/1/2017	7656120	7320259	-335861	8/1/2021	7296807	7894977	598170.3
8/1/2017	7315411	7340175	24763.8	9/1/2021	7596651	8242470	645818.7
9/1/2017	8192470	7700358	-492112	10/1/2021	8433124	8583255	150131.2
10/1/2017	7685774	7587531	-98242.9	11/1/2021	8203007	8334491	131484.1

11/1/2017	7858423	7786343	-72080	12/1/2021	8256512	8278001	21489.3
12/1/2017	7971852	7725049	-246803	1/1/2022	8412426	8496327	83900.96
1/1/2018	8330614	8098957	-231657	2/1/2022	8353262	8325976	-27286.5
2/1/2018	8231172	7881155	-350017	3/1/2022	8779371	8379120	-400251
3/1/2018	7826970	7839422	12451.89	4/1/2022	8089063	8415571	326508.3
4/1/2018	8208224	7821346	-386878	5/1/2022	8830469	8662354	-168115
5/1/2018	8033656	7987913	-45742.8	6/1/2022	7629820	7943573	313753.1
6/1/2018	7076139	8056825	980686.1	7/1/2022	8012902	7899063	-113839
7/1/2018	7698564	7707464	8899.916	8/1/2022	7013975	7531819	517843.6
8/1/2018	7364790	7613053	248263.5	9/1/2022	7942867	7820220	-122647
9/1/2018	8300142	8021014	-279128	10/1/2022	8596591	7846865	-749726
10/1/2018	8057189	7907761	-149428	11/1/2022	7982098	7876083	-106015
11/1/2018	8450074	8030209	-419865	12/1/2022	8111402	7870279	-241123
12/1/2018	8088169	8018011	-70157.6	1/1/2023	8445107	8198679	-246428
1/1/2019	8681775	8536994	-144781	2/1/2023	8382996	8318789	-64207
2/1/2019	8496602	8317020	-179582	3/1/2023	8108950	8065831	-43119.4
3/1/2019	8569319	8346523	-222796	4/1/2023	7638935	8274495	635559.5
4/1/2019	8078743	8352554	273810.7	5/1/2023	7912454	8200101	287646.9
5/1/2019	8397558	8580165	182607.4	6/1/2023	7971363	8024539	53176.16
6/1/2019	7758647	7595553	-163094	7/1/2023	8030272	7762341	-267931
				8/1/2023	7470289	7831231	360941.8

Appendix 9: determination the model value for ANN model with five hidden nodes

```
clc
clear
% ANN model with five hidden nodes code
 %1. population (number)
 %2. Rainfall (mm)
 %3. average temperature {o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
```

```
 %10. Water consumption data (m3/month)
 % Input data
ainp1=xlsread('Input','NND');
modinp=ainp1(:,3:11);
[ndata,nc]=size(modinp);
out1=ainp1(:,12);
% Normalization of input and output data
inmax=max(modinp);
inmin=min(modinp);
1fmax=1;
1fmin=-1
indnorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
indnorm(i,j)=1fmin+(1fmax-1fmin)* modinp(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*( out1 (i,1)-min(out1))/...
             (max(out1)-min(out1));
end
x=xlsread('Input','xvalue5');
y_linear=x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)+x(1,4)*indnorm(:
,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)+x(1,8)*indnorm(:,8)+x
(1,9)*indnorm(:,9)+0.141379715;
dnorm=min(out1)+ (max(out1)-min(out1))*(y_linear(:,1)-lfmin)/(lfmax-lfmin);
```

Appendix 10: Actual inputs and forecasting water consumption for ANN Model with five hidden nodes

Table 10 Actual inputs and forecasting water consumption (m3/month) for ANN Model with five hidden layers from 2015 to 2023 (M3/month)

Year	Actual	Model	Error	Year	Actual	Model	Error
6/1/2015	7067265	7141899	74634.63	7/1/2019	7910023	7439823	-470200
7/1/2015	7131412	7142550	11138.14	8/1/2019	7507056	7483464	-23592
8/1/2015	7067265	7068303	1038.134	9/1/2019	8712274	7729115	-983159
9/1/2015	7003118	7334794	331676.4	10/1/2019	7629318	7802824	173506.2
10/1/2015	7075072	7215680	140608.4	11/1/2019	7337532	7841626	504093.7
11/1/2015	7147025	7200631	53606.37	12/1/2019	8051808	7801874	-249934
12/1/2015	7352301	7197079	-155222	1/1/2020	8003864	8185553	181688.8
1/1/2016	6910522	7402817	492295	2/1/2020	8250575	8282942	32367.02
2/1/2016	7146994	7296533	149539.3	3/1/2020	8408200	8161903	-246297
3/1/2016	7253931	7338119	84187.89	4/1/2020	7660307	8169712	509404.7

4/1/2016	7140756	7306564	165808.5	5/1/2020	8457027	8234609	-222418
5/1/2016	7427536	7421160	-6376.17	6/1/2020	8203648	7899491	-304157
6/1/2016	7407872	7414668	6796.359	7/1/2020	8281121	8017016	-264105
7/1/2016	7388208	7305387	-82820.8	8/1/2020	7004501	7885460	880958.8
8/1/2016	6909406	7229579	320173.4	9/1/2020	7719573	8039127	319554.1
9/1/2016	7713993	7538950	-175043	10/1/2020	8526112	8342565	-183547
10/1/2016	7727011	7705768	-21242.9	11/1/2020	8595465	8410398	-185067
11/1/2016	7611241	7687092	75850.86	12/1/2020	8506069	8346893	-159176
12/1/2016	7502639	7710392	207753.3	1/1/2021	8416673	8323128	-93545.2
1/1/2017	7545231	8137037	591806.1	2/1/2021	7884647	8265948	381300.7
2/1/2017	7592636	8008210	415573.9	3/1/2021	8345317	8463880	118563.2
3/1/2017	8147592	7944176	-203416	4/1/2021	8232874	8356756	123882.5
4/1/2017	7942723	8025895	83172.13	5/1/2021	8415169	8471250	56080.75
5/1/2017	8150153	7819604	-330549	6/1/2021	7415000	8171508	756508
6/1/2017	7361087	7731092	370005	7/1/2021	8358479	8158469	-200010
7/1/2017	7656120	7300095	-356025	8/1/2021	7296807	7870664	573856.8
8/1/2017	7315411	7323140	7728.821	9/1/2021	7596651	8149394	552743.4
9/1/2017	8192470	7539564	-652906	10/1/2021	8433124	8567137	134012.8
10/1/2017	7685774	7691912	6137.689	11/1/2021	8203007	8408956	205949.2
11/1/2017	7858423	7841827	-16596.2	12/1/2021	8256512	8358814	102301.5
12/1/2017	7971852	7824590	-147262	1/1/2022	8412426	8480617	68190.8
1/1/2018	8330614	8108830	-221784	2/1/2022	8353262	8333560	-19702
2/1/2018	8231172	7973335	-257837	3/1/2022	8779371	8328488	-450883
3/1/2018	7826970	7941597	114626.9	4/1/2022	8089063	8324769	235706
4/1/2018	8208224	7857055	-351169	5/1/2022	8830469	8490784	-339685
5/1/2018	8033656	7971621	-62034.8	6/1/2022	7629820	7902769	272948.8
6/1/2018	7076139	7834571	758431.9	7/1/2022	8012902	7781404	-231498
7/1/2018	7698564	7625745	-72818.6	8/1/2022	7013975	7528461	514486.1

8/1/2018	7364790	7600394	235603.7	9/1/2022	7942867	7783106	-159761
9/1/2018	8300142	7926018	-374124	10/1/2022	8596591	7912270	-684321
10/1/2018	8057189	7910682	-146507	11/1/2022	7982098	7994663	12565.22
11/1/2018	8450074	7991899	-458175	12/1/2022	8111402	7993201	-118201
12/1/2018	8088169	8028342	-59826.5	1/1/2023	8445107	8220570	-224537
1/1/2019	8681775	8454657	-227118	2/1/2023	8382996	8439897	56900.74
2/1/2019	8496602	8327781	-168821	3/1/2023	8108950	8103409	-5540.9
3/1/2019	8569319	8306858	-262461	4/1/2023	7638935	8219252	580317.2
4/1/2019	8078743	8243632	164889.4	5/1/2023	7912454	8117679	205224.6
5/1/2019	8397558	8414053	16494.94	6/1/2023	7971363	8115531	144168.2
6/1/2019	7758647	7624628	-134019	7/1/2023	8030272	7815395	-214877
				8/1/2023	7470289	7967457	497168.3

Appendix 11: Sensitive analysis

```
clc
clear
%by using the best model selected to determine sensitive analysis. For this research the best model is ANN model with five hidden nodes.
The equation of the model is:
```

$$WC = -0.2282 * (IA) + 0.1771 * (EG) + 0.3346 * (POP) + 0.0198 * (TEMP) - 0.0825 * (PREC) - 0.2974 * (RH) + 0.2018 * (LS) + 0.1495 * (H\&F) + 0.1541 * (CA) + 0.1414$$

```
% 5 % increasing on the coefficient of the input variable
By using the above equation determine sensitive analysis by using RMSE
% The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
 %10. Water consumption data (m3/month)
ainp1=xlsread('Input','NND');
modinp=ainp1(:,3:11);
[ndata,nc]=size(modinp);
out1=ainp1(:,12);
% Normalization of input and output data
```

```
inmax=max(modinp);
inmin=min(modinp);
1fmax=1;
1fmin=-1;
indnorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
indnorm(i,j)=1fmin+(1fmax-1fmin)* modinp(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*( out1 (i,1)-min(out1))/...
            (max(out1)-min(out1));
end
x=xlsread('Input','xvalue5');
y_linear1= 1.05*x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm1=min(out1)+ (max(out1)-min(out1))*(y_linear1(:,1)-lfmin)/(lfmax-lfmin);
y_linear2= x(1,1)*indnorm(:,1)+1.05*x(1,2)*indnorm(:,2)+ x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;

dnorm2=min(out1)+ (max(out1)-min(out1))*(y_linear2(:,1)-lfmin)/(lfmax-lfmin);

y_linear3=x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+1.05*x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm3=min(out1)+ (max(out1)-min(out1))*( y_linear3 (:,1)-lfmin)/(lfmax-lfmin);
y_linear4= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+1.05*x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm4=min(out1)+ (max(out1)-min(out1))*( y_linear4 (:,1)-lfmin)/(lfmax-lfmin);
y_linear5
= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+1.05*x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm5=min(out1)+ (max(out1)-min(out1))*(y_linear5(:,1)-lfmin)/(lfmax-lfmin);
y_linear6= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+1.05*x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm6=min(out1)+ (max(out1)-min(out1))*(y_linear6(:,1)-lfmin)/(lfmax-lfmin);
y_linear7= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+1.05*x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm7=min(out1)+ (max(out1)-min(out1))*(y_linear7(:,1)-lfmin)/(lfmax-lfmin);
y_linear8= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+1.05*x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm8=min(out1)+ (max(out1)-min(out1))*(y_linear8(:,1)-lfmin)/(lfmax-lfmin);
y_linear9= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+1.05* x(1.9)*indnorm(:,9)+0.1414;
dnorm9=min(out1)+ (max(out1)-min(out1))*(y_linear9(:,1)-lfmin)/(lfmax-lfmin);
```

```
% 5 % decreasing on the coefficient of the input variable
clc
clear
```

$$WC = -0.2282*(IA) + 0.1771*(EG) + 0.3346*(POP) + 0.0198*(TEMP) - 0.0825 * (PREC) - 0.2974*(RH) + 0.2018*(LS) + 0.1495*(H\&F) + 0.1541*(CA) + 0.1414$$

```
By using the above equation determine sensitive analysis by using RMSE
% The input data of the model
 %1. population (number)
 %2. Rainfall (mm)
 %3. average temperature (^{o}C)
 %4. Relative Humidity (%)
 %5. Holiday and festival (number)
 %6. construction activity (%)
 %7. number of livestock (number)
 %8. industrial activity (%)
 %9. economic growth (%)
 %10. Water consumption data (m3/month)
ainp1=xlsread('Input','NND');
modinp=ainp1(:,3:11);
[ndata,nc]=size(modinp);
out1=ainp1(:,12);
Normalization of input and output data
inmax=max(modinp);
inmin=min(modinp);
1fmax=1;
1fmin=-1;
indnorm=zeros(ndata,nc); %normalized input data
outdnorm=zeros(ndata,1); %normalized output data
for i=1: ndata
    for j=1:nc
indnorm(i,j)=1fmin+(1fmax-1fmin)* modinp(i,j)-inmin(j))/...
       (inmax(j)-inmin(j));
    end
    outdnorm(i,1)=lfmin+(lfmax-lfmin)*( out1 (i,1)-min(out1))/...
            (max(out1)-min(out1));
end
x=xlsread('Input','xvalue5');
y_linear1= 0.95*x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm1=min(out1)+ (max(out1)-min(out1))*(y_linear1(:,1)-lfmin)/(lfmax-lfmin);
y_linear2= x(1,1)*indnorm(:,1)+0.95*x(1,2)*indnorm(:,2)+ x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm2=min(out1)+ (max(out1)-min(out1))*(y_linear2(:,1)-lfmin)/(lfmax-lfmin);
y_linear3=x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+0.95*x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm3=min(out1)+ (max(out1)-min(out1))*( y_linear3 (:,1)-lfmin)/(lfmax-lfmin);
```

```
y_linear4= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+0.95*x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm4=min(out1)+ (max(out1)-min(out1))*( y_linear4 (:,1)-lfmin)/(lfmax-lfmin);
y_linear5= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+0.95*x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm5=min(out1)+ (max(out1)-min(out1))*(y_linear5(:,1)-lfmin)/(lfmax-lfmin);
y_linear6= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+0.95*x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm6=min(out1)+ (max(out1)-min(out1))*(y_linear6(:,1)-lfmin)/(lfmax-lfmin);
y_linear7= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+0.95*x(1,7)*indnorm(:,7)
+x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm7=min(out1)+ (max(out1)-min(out1))*(y_linear7(:,1)-lfmin)/(lfmax-lfmin);
y_linear8= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+0.95*x(1,8)*indnorm(:,8)+ x(1.9)*indnorm(:,9)+0.1414;
dnorm8=min(out1)+ (max(out1)-min(out1))*(y_linear8(:,1)-lfmin)/(lfmax-lfmin);
y_linear9= x(1,1)*indnorm(:,1)+x(1,2)*indnorm(:,2)+x(1,3)*indnorm(:,3)...
+x(1,4)*indnorm(:,4)+x(1,5)*indnorm(:,5)+x(1,6)*indnorm(:,6)+x(1,7)*indnorm(:,7)....
+x(1,8)*indnorm(:,8)+0.95* x(1.9)*indnorm(:,9)+0.1414;
dnorm9=min(out1)+ (max(out1)-min(out1))*(y_linear9(:,1)-lfmin)/(lfmax-lfmin);
```