WATER DEMAND FORECASTING: A FLEXIBLE APPROACH

by

Sina Shabani

B.Sc., American University of Sharjah, 2011
M.Sc., American University of Sharjah, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLEGE OF GRADUATE STUDIES

(Civil Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

September 2018

© Sina Shabani, 2018

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

WATER DEMAND FORECASTING: A FLEXIBLE APPROACH

submitted by Sina Shabani in partial fulfillment of the requirements of

the degree of Doctor of Philosophy

Dr. Gholamreza Naser, School of Engineering

Supervisor

Dr. Abbas Milani, School of Engineering

Supervisory Committee Member

Dr. Deborah Roberts, School of Engineering

Supervisory Committee Member

Dr. Jeff Curtis, School of Engineering

University Examiner

Dr. Yves Filion, School of Engineering and Applied Science, Queen's University

External Examiner

Abstract

Water distribution systems (WDS) operators would benefit greatly from educated estimates of water demand to help with the water scarcity threatening many regions worldwide. A wide range of stochastic and deterministic techniques have been proposed to model water demands in urban WDS. Every WDS needs a fair estimate of its future state for both development and operational purposes. Statistical models like time series forecasting and regression analysis have been widely used in the field of water demand forecasting. Recently, so-called artificial intelligence techniques became increasingly popular among scholars due to their high accuracy in prediction as well as not being limited to certain statistical assumptions. This research introduced gene expression programming (GEP), and support vector regression (SVM) and well-known artificial neural networks (ANN) as supervised learning algorithms for predictive analytics of water demand. K-means clustering as an unsupervised learning algorithm was tested to group the data based on a suitable distance metrics. The performance of the developed models was improved through phase space reconstruction of the time series data using optimum lag time determined by average mutual information. Monthly long-term water demand data of the City of Kelowna district (CKD) was used as the main case study throughout the research. Due to unavailability of water demand at finer short-term resolutions in Kelowna, the water demand data in the City of Milan was also used as a second case study for developing a framework of studying different temporal resolutions in the short-term analysis of water demand. Some scholars believe that predictive models are often wrong given the significant uncertainty in the conditions of underlying complex engineering systems. A novel technique based on data augmentation (data cropping and distorted data), and information theory were used to propose a flexible range of water demand (358 ML for upper bound and 335 ML for lower bound) for City of Kelowna which anticipates a wide range of uncertainties WDS.

Lay Summary

Natural water resources are becoming increasingly scarce. This scarcity has led water utilities to estimate the future demand for water for both operation and development purposes. Traditionally, a simple projection of historical water demand would have been sufficient for a good estimate of the future need for water utilities. However, recent rapid urbanizations and a gradual shift in climatic conditions have created a growing necessity to use sophisticated statistical models and machine learning algorithms for predictive analytics of the future state of water demand. The constant improvements in complex engineering systems and digitization of such facilities introduced a large amount of data to engineers which can be used by data scientists to propose highly accurate predictive models using machine learning algorithms. A novel technique is proposed using predictive analytics for a flexible forecast of water demand which can anticipate future changes in highly complex water distribution systems.

Preface

A version of chapter 2 has been published in the Journal of Procedia Engineering as a selected paper of WDSA conference proceeding. Shabani, S. and Naser, G., 2015. Dynamic nature of explanatory variables in water demand forecasting. *Procedia Engineering*, *119*, pp.781-787. I developed the technique and wrote the manuscript which was further edited by Dr. Bahman Naser.

A version of chapter 3 has been published in the book titled Water Stress in Plants by Intech. Shabani, S., Yousefi, P., Adamowski, J. and Naser, G., 2016. Intelligent soft computing models in water demand forecasting. In *Water Stress in Plants*. InTech. I defined the input variables based on the availability of data, developed the models, conducted the simulations, and wrote the manuscript which was edited by Dr. Bahman Naser. Dr. Jan Adamowski and Mr. Peyman Yousefi provided some constructive comments.

A version of chapter 4 has been published in the Journal of Procedia Engineering as a selected paper of the WDSA conference proceeding. Shabani, S., Yousefi, P. and Naser, G., 2017. Support vector machines in urban water demand forecasting using phase space reconstruction. *Procedia Eng*, *186*, pp.537-543. I developed the models, conducted the simulations, and wrote the manuscript which was edited by Dr. Bahman Naser and Mr. Peyman Yousefi provided some constructive comments.

A version of chapter 5 has been published in the Journal of Water. Shabani, S., Candelieri, A., Archetti, F. and Naser, G., 2018. Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand

Forecasts. *Water*, *10*(2), p.142. I developed the technique, simulated the models, and wrote the manuscript which was further edited by Dr. Bahman Naser. Dr. Antonio Candelieri and Dr. Francesco Archetti provided me with the hourly data on water demand of the City of Milan and helped me with the deployed unsupervised learning algorithm.

A version of chapter 6 has been submitted for publication in the Journal of Water (under review). I developed the technique, simulated the models, and wrote the manuscript which was further edited by Dr. Bahman Naser.

Publications from the research presented in this dissertation are listed below:

1. Shabani, S., Naser, G., 2018. Deep Learning Practices in Long-Term Water Demand Forecasting Models: A Flexible Approach. Submitted to *Water* journal is currently under peer review, Manuscript ID: water-335759.
2. Shabani, S., Candelieri, A., Archetti, F. and Naser, G., 2018. Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts. *Water*, *10*(2), p.142.
3. Shabani, S., Yousefi, P. and Naser, G., 2017. Support vector machines in urban water demand forecasting using phase space reconstruction. *Procedia Eng*, *186*, pp.537-543.
4. Shabani, S., Yousefi, P., Adamowski, J. and Naser, G., 2016. Intelligent Soft Computing Models in Water Demand Prediction. Published in the book entitled "Water Stress in Plants" edited by Ismail Md. Mofizur Rahman, Zinnat Ara Begum, and Hiroshi Hasegawa, *InTech Europe*, ISBN 978-953-51-4804-3.
5. Shabani, S. and Naser, G., 2015. Dynamic nature of explanatory variables in water demand forecasting. *Procedia Engineering*, *119*, pp.781-787.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Words are not enough to thank my supervisor Dr. Bahman Naser for his constant guidance and mentorship over the duration of this Ph.D. program. He was always available to help, guide, and hear me under different circumstances.

I would like to thank Dr. Deborah Roberts and Dr. Abbas Milani for serving as my Ph.D. committee members. Their constructive comments and insights helped me to represent my work in this dissertation better.

I would like to express my gratitude to my lovely parents and my fiancé. Their continuous support and endless love made me the person I am today.

Special thanks to my dear friend and roommate, Amin Bigdeli for his endless support. In the end, I would like to thank all the people I cannot think about without a smile on my face: Shahin, Saber, Ahmad, Ehsan, Bardia, Mohammad, Farhad, Ronak, Priscila, Behzad, Ali, Hamed, Peyman, and Alireza. You are always in my mind, and I am fortunate to have you all in my life.

Dedicated to my parents and my fiancé.

Chapter 1: Introduction

While water scarcity has become a key concern worldwide, it is particularly so in arid and semi-arid regions with limited potable water sources. In designing water distribution systems (WDS), engineers have typically used a "fixture unit" method, which considers the sum of fixture unit demands, facility types, and socio-economic factors to determine peak demand. However, this overestimates the actual peak demand by as much as 100% [1]. Due to these uncertainties, including those associated with demand, engineers often include large safety factors when designing WDS. Given that WDS rely mainly on regional energy and resources, an over-designed system can have environmental impacts that will appear in the region(s) beyond the jurisdictional boundaries of the system. While short-term demand forecasts are critical to a WDS' daily operations [2], long-term forecasts are required for future planning and management of the systems. In providing an accurate estimate of water demand, a robust demand-forecasting model assists engineers in designing a more environmentally sustainable WDS and in managing available water resources more efficiently. When coupled with a water demand management strategy, such models can help managers overcome operational problems (e.g., low pressure during peak demands) and issues related to asset management (e.g., non-replacement of assets or replacement by lower capacity assets reaching the end of their economic life). It has been estimated that a well-predicted monthly average demand might be up to 400% lower than peak demands that cause low pressure; however, a more realistic model can enhance resource management and efficient operating systems. This will eventually lead to significant savings for water and energy (for running pumps, treatment plants, etc.) in industries. The concept of flexible design for high value infrastructures/assets has received significant attention by engineers for the sake of durability or longer lifespan of the complex systems [3]. Many researchers also proposed flexibility as a remedy

to a high level of uncertainties in designing complex engineering designed systems [4], [5]. In fact, flexibility can build a platform to achieve the benefits of uncertainty, therefore being able to respond to future needs in an economically optimized manner. However, the concept of flexibility did not receive enough attention in the design of WDS. Thus, the prime objective of this research is to explore the flexibility of WDS. Even though many factors may affect the flexibility of WDS, this research only focused on the impact of water demand. In this regard, the present research created a forecasting model for the monthly average water demand. While the present research proposed a generic framework that could be easily adjusted for any specific WDS, the City of Kelowna (British Columbia, Canada) was employed as a test case. The City of Milan was used as another case study in chapter 5 since the finer resolution of time series data for water demand was not available for the City of Kelowna and multi-scale modeling was part of the overall objectives of this dissertation.

A secured and consistent supply of potable water to consumers is the primary goal behind a WDS design. This supply should meet the desired flow and pressure during operation. Basic design criteria for engineers to follow can be illustrated through a flowchart shown in Figure 1-1. It is known that valuable assets/infrastructures such as WDS should be designed for a long lifespan that can serve future demand. However, high levels of uncertainty are associated with the future of a complex system like WDS. Therefore, predicting the future always involves risks and uncertainties. Traditionally, engineers have been designing a *WDS* for a specific anticipated future with deterministic assumptions which is a projection of historical water demand trends and population. This approach is highly uncertain. Presumably, water demand is not a simply predictable quantity since it couples to highly complex weather, climate, ecological, sociological, and political dynamics. Therefore, the traditional approach can lead to over-sized/under-designed

systems. As shown in figure 1-1, water demand analysis is the backbone of WDS design. To insert the flexibility concept in *WDS* design, a flexible educated forecast of water demand is an essential need. Flexibility is the ability to handle unplanned eventualities. A flexible demand analysis is one that can propose a range of demand (instead of a deterministic approach) which considers the mentioned sources of uncertainties in their models. This research will focus on long-term water demand forecasting which can be inserted into a novel flexible demand analysis.

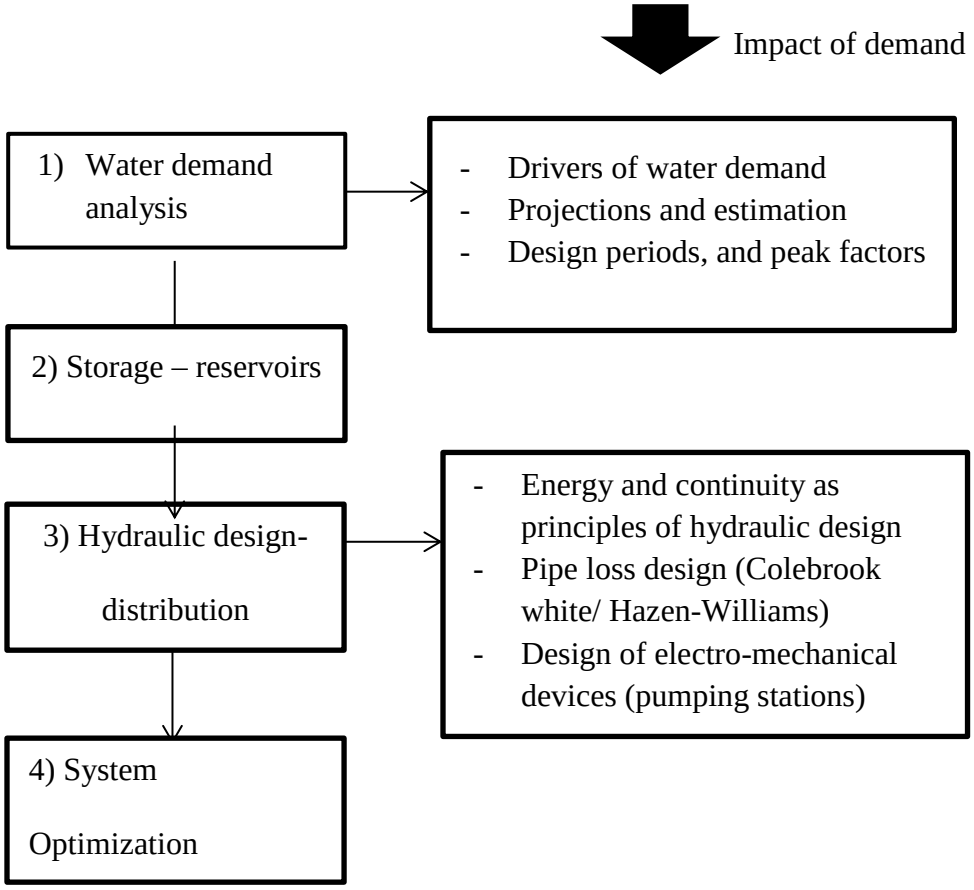


Figure 1-1 Basic design criteria WDS

Predictive modeling can be defined as the generation of accurate predictions using a set of mathematical tools such as statistical models or machine learning algorithms. However, there are two major concerns regarding the development of such models. The primary concern of predictive modeling practitioners is the high level of uncertainty associated with the initial conditions of the engineering systems in the prediction horizon. Our engineering world is subject to constant changes affecting the design, operation, and data acquisition for infrastructure. Therefore, this study tried to define a new general framework which can anticipate possible future changes in a predictive modeling tool. Another major concern would be a wide range of reasons that might fail a predictive model. The current practices in water demand forecasting models are dealing with insufficient pre-processing of the feeding data, inadequate model validation and comparison within a large set of developed models, unjustified future projections, and models which fail in flexibility, and are over-fitted to certain existing data. Recently, scholars and researchers have been trying different predictive modeling techniques which perform better than conventional methods. One should remember, predictive models, are highly dependent on the relevant information or data. A highly accurate model is not guaranteed to perform as good if the initial conditions are changed. Unfortunately, the current trend in scholarly papers is forcing models to have high accuracy. Consequently, interpreting them becomes a complex task since they ignore the application of such models and focus on the given case study. Long-term predictions can always be inaccurate given the high levels of uncertainties in complex engineering systems. Therefore, for the first time, this dissertation tried to define a framework in a flexible forecast model which gives a range of demand instead of a deterministic and uncertain single value.

1.1 Water Demand Forecasting

Sustainable design and operation of a WDS requires accurate knowledge of water demand that vary temporally and spatially. Water demand varies greatly both regionally and seasonally. Increasing urbanization and industrialization, as well as emerging issues such as shifting weather patterns and population growth, have significant impacts on water demand. The main components in demand prediction are the explanatory variables and time scales used. Selecting explanatory variables for a predictive model depends on the desired time scale and the availability of data. Simple models using very few explanatory variables have shown promising accuracy for short-term predictions [6], [7]. In general, the explanatory variables affecting water demand are of two types: weather (e.g., temperature, relative humidity, and rainfall) and socioeconomic (e.g., population and income). Weather conditions affect short-term prediction while their socioeconomic counterparts can affect long-term [8]–[10]. As has been highlighted by significant worldwide changes in climate, water availability is prone to great uncertainty [11]. Therefore, the impact of evolving climate conditions on long-term water demand predictions should receive greater attention. Furthermore, researchers who have considered weather conditions in short-term water demand prediction have established that it is not feasible to feed online automated WDS with real-time weather information [12]. As a result, limited studies have considered weather conditions in their demand forecasting models [13]–[15]. Table 1-1 summarizes the relevant literature; however, more recent studies are discussed in each chapter according to their corresponding topic. Temperature, precipitation, pan evaporation, and the number of days since the last rainfall were used in a forecasting model [16]. Another study used temperature, relative humidity, rainfall, wind speed, and air pressure as weather parameters in their hourly water demand model for Sao Paulo, Brazil [15]. Table 1-1 shows the previous researchers did not consider

socioeconomic and weather conditions simultaneously since their effects are dependent on the forecast horizon. Traditionally, WDS utilities have used historical patterns as explanatory variables in predicting future water demands. Scarce water reserves and the rapid increase in urbanization have raised awareness and led to the implementation of statistical approaches. Multiple linear regression (MLR) and time series were the most popular techniques used in the early stages of demand forecasting [9].While MLR has been widely used to better understand the determinants of water demand [17]–[21], its major drawback is the fact that it considers linear relationships among variables and water demand, while such relationships are often nonlinear necessarily. Time series have been introduced along with regression as methods for demand forecasting [16], [22]. Due to the common belief that they can deal with complex systems [23], artificial neural networks (ANNs) have been widely applied in water demand forecasting [2], [24]–[26]. Comparing regression, univariate time series, and ANN models, a research found ANN models drawing on standard rainfall and maximum temperature data could better predict weekly water demand [9]. Similarly, drawing on temperature and rainfall data in their forecasting models, researchers concluded that ANN models provided more reliable forecasts for peak weekly demand than time series and simple and multiple linear regressions [25]. Results of another study showed ANN models performed better for hourly forecasts, whereas regression models were more accurate in forecasting daily demand [26]. To improve the accuracy and robustness of demand forecasting models, hybrid models combining or modifying ANN, MLR, and time series techniques have been tested [27]–[30]. However, application of nonlinear regression in demand forecasting has remained limited to studies using support vector machines (SVM) [31]–[33] and training nonlinear relationships through linear regression models [9], [34]. Other computational techniques used in water demand forecasting are fuzzy approach [35] and agent-based approach [36].The hybrid

approaches include pattern recognition [37], neural-fuzzy [30], [38], [39], and model-tree approaches [40]. However, most of those studies focus on future demand forecasting without considering the uncertainties associated with the future changes of WDS' initial conditions. Monte Carlo simulation has also been studied to explore the uncertainties/risks (in water demand forecasting) associated with population growth, climate changes, and public behavior [41]. GEP was used as a new modeling technique (which results in an equation rather than a black-box model) and was compared with SVM and ANN as well-known black-box models in this research. If the dynamic time series nature of data is considered, it will facilitate decision making on the number/type of the parameters that should be studied in the modeling to have more accurate models.

Table 1-1 Relevant literature on water demand forecast models

#	Ref	Method	Determinant	Time Scale
1	[19]	Linear Regression	Seasonal dummies, derivatives of weather, price	Monthly demand
2	[20]	Linear regression	Density, building size, lot size, household size, income, price, temperature, rain, drought dummies	Bimonthly demand
3	[21]	Regression using Bayesian moment entropy	Population density	Annual demand
4	[16]	Decomposed daily demand followed by composite forecasts	Daily and hourly demands	Daily demand
5	[22]	Univariate time series	Delayed demand	Annual residential demand
6	[25]	Regression and ANN	Temperature, rainfall, and lags of peak demand	Peak weakly demand
7	[26]	ANN	Temperature, rainfall, and delayed demand	Daily Demand
8	[2]	Time series	Univariate demand series, the temperature in a multivariate model	Daily, weekly, monthly, annual
9	[29]	Time series and ANN	Delayed demands, temp, and rainfall	Weekly demand
11	[30]	Weighted average regression and ANN	Historical demand and time	Annual demand
12	[34]	Decomposed annual demand, Regression, and ANN	GDP, population, temperature, greenery coverage, delayed demand	Annual demand
13	[31]	Wavelet –denoising, and ANN	7-year long time series of demand	Monthly demand
14	[32]	SVM with RBF function is compared with ANN	Delayed demand, population	Daily demand
15	[33]	ANN, SVM, Monte Carlo	Rain, demand, wind speed, atmospheric pressure	Hourly demand

1.2 Research Objectives

The long-term objective of this research has been exploring the impacts of water demand on *WDS* flexibility. To arrive at such an approach, appropriate learning models and their tunings should have been investigated. This approach could ultimately lead to the development of a novel flexibility approach in designing and analysis of a WDS. The long-term objective will be accomplished via the following five short-term objectives:

Objective 1 – Investigate the existence of low dimensional chaos in the dynamic nature of the determinants used in the field of water demand forecasting through the following sub-objectives:

- to identify the optimum lag time of the time series data
- to perform phase space reconstruction on each time series data
- to use correlation dimension method through a range of embedding dimensions
- to investigate if using deterministic methods are possible in water demand forecasting models

Objective 2 – Investigate the performance of GEP as a new evolutionary and SVM as a machine learning technique in the field of water demand forecasting through the following sub-objectives:

- to define the input variables for multi-variate analysis
- to identify the optimum lag time for phase space reconstruction of the input variables
- to design the input for feeding the predictive models
- to use performance metrics for comparing the outcome of developed models

Objective 3 – Identify if the optimum lag time associated with the predicted value can be generalized to the input variables of predictive models through the following sub-objectives:

- to identify the optimum lag time of each input variables
- to conduct an input variable design based on individual optimum lag times of the determinants
- to check if the approach in previous objective was valid

Objective 4 – Investigate the outcome of coupling an unsupervised learning algorithm with the developed forecast models (supervised learning) as a two-stage learning process through the following sub-objectives:

- to use k-means clustering as an unsupervised learning method
- to investigate the performance of predictive models coupled with unsupervised learning

Objective 5 - Perform multi-scale modeling for a better understanding of the data acquisition frequency and the consumption behavior of water demand.

Objective 6 - Develop a concept of flexibility in the water demand analysis of a WDS through the following sub-objectives:

- to identify the impact of water demand on WDS's flexibility
- to investigate the flexibility of well-known predictive models
- to deploy the predictive models for proposing a range of demand
- to use deep learning practices for anticipating the future uncertainties

1.3 Data Collection and Analysis

Water demand depends on many factors including population, hydrologic, climatic, social, and economic development, consumers, and their behavioral patterns, etc. One of the objectives was to study the impacts of population, hydrologic and climatic factors on water demand. Thus, the relevant data of water demand, climatic information, and demographic variables required for the research were collected. The data with different temporal resolutions were synchronized. All data were collected in collaboration with the local and national offices. Water demand in a WDS is a dynamic variable, and its variations are affected by many environmental and non-environmental forcing that is nonlinear and interconnected. WDS, climatic, and social parameters were studied in detail. The WDS parameters include historical data about water consumption. Such data were collected through field measurement by the City of Kelowna. In practice, the numbers of field measurements are limited as they are highly costly. Moreover, such measurements are often affected by noises due to limitations of measuring devices. Information about water use was provided by the five water districts in Kelowna area including the City of Kelowna District, Glenmore Ellison Irrigation District, Black Mountain Irrigation District, Rutland Water District and South East Kelowna Irrigation District (http://www.kelowna.ca/CM/page130.aspx).

Explanatory variables studied in this research were precipitation, relative humidity, air temperature and as climatic information. The relevant data were collected from Canada Weather (http://climate.weather.gc.ca/). The City of Kelowna provided the weather data from their station located at 49°57’22” N and 119°22’40” W (http://climate.weather.gc.ca). Demographic variables include information about the population. Such data were collected from the City of Kelowna and the Statistics of Canada (http://www12.statcan.gc.ca/census-recensement/index-eng.cfm). While

Canada has the highest freshwater availability in the world (7% of total world's fresh water reported by Environment Canada), aging water supply and distribution systems and lack of new water resources has put water authorities under stress. According to Environment Canada, the threat to water availability ranges from low (for Quebec, British Columbia, northern and Atlantic Canada) to high (for the prairie region and the south-west part of Ontario). The threat for the Okanagan Valley (British Columbia) is medium; meaning that the water availability is a constraint on development and significant investment is required to meet the demand. This research focused on the City of Kelowna (British Columbia, Canada). The City has five water districts including the City of Kelowna District (CKD), Glenmore Ellison Irrigation District (GEID), Black Mountain Irrigation District (BMID), Rutland Water District (RWD) and the South East Kelowna Irrigation District (SEKID). This research will primarily employ CKD as the study area. The data sets were divided into two groups. The 1st group (training set) was used for model development. The 2nd group as the most recent data set (validation set) was used for the model-validation purpose. In practice, it is often very difficult (if not impossible) to conduct direct measurements of all the variables at fine temporal resolutions because either the required technology does not exist, or it is not economically feasible. As such, these variables are always measured with different temporal resolutions (e.g., one parameter is measured on daily-base, while another one is recorded weekly). Once relevant explanatory data were collected, models were designed in three different categories to evaluate the nature of data fed to developed models. 1) Demand-based data: water demand can be used as the main input to forecast models (as in practice). 2) Demand + Weather Data based: this category took climatic information into consideration; therefore, it was able to verify if such data is useful for long-term forecasts or not. 3) Demand + Weather + Population Data: socio-economic factors were added as another classification of data into proposed models.

1.4 Thesis Structure

This chapter provided a general introduction to the flexible approach of this dissertation followed by the relevant literature in the field of water demand forecasting. The following chapters provide a background on the specific research objective followed by the methods used to meet each objective (Figure 1-2). Chapter 2 to 6 address each research objectives mentioned in section 1.2. The evidence of low dimensional deterministic chaos was checked in chapter 2 using correlation dimension method which resulted in highly chaotic or noisy behavior of climatic variables in water demand forecasting. Therefore, GEP and SVM were selected as stochastic methods to forecast long term water demand in City of Kelowna (chapter 3). To arrive at this approach, phase space reconstruction of input variables using optimum lag time of water demand was performed to feed the learning algorithms. Finally, the superior models were identified by comparing their underlying performance indices. Chapter 4 validated if the approach in previous chapter in generalizing the optimum lag time of response variable is a good practice. The possibility of further improvement of GEP learning algorithm was investigated by coupling an unsupervised learning algorithm (K-means clustering) as well as multi-scale modeling in chapter 5. Using data augmentation techniques, the supervised learning algorithms were fed with a wide range of input designs to widen the range of forecasting as much as possible. This was performed to reach the flexible range of demand which anticipates future uncertainties dealing with *WDSs*. Finally, Chapter 7 provides the conclusive of remarks followed by suggestions for future research.

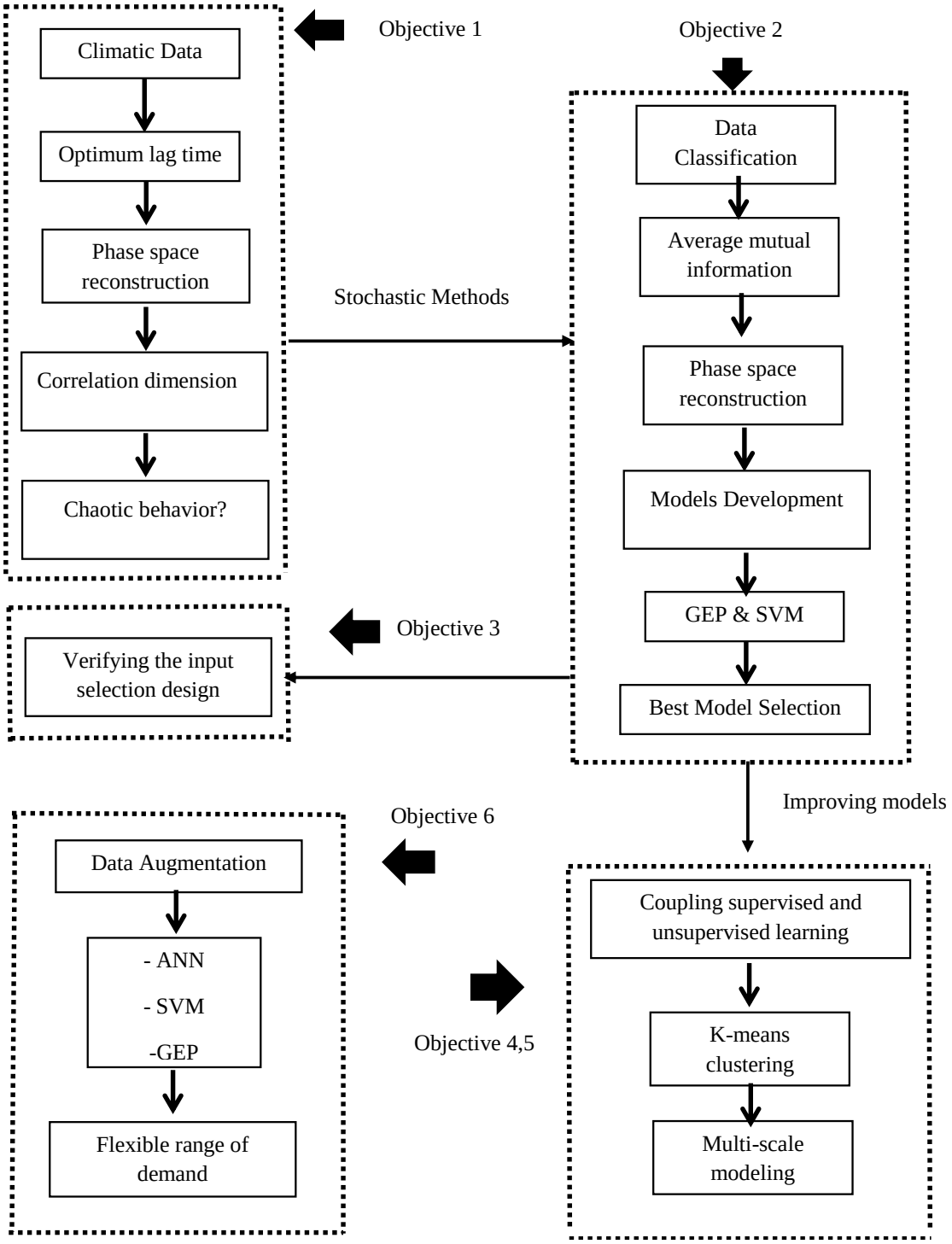


Figure 1-2 Flow diagram of thesis

Chapter 2: Dynamic Nature of Explanatory Variables in Demand Forecasts [1]

2.1 Overview

A wide range of explanatory variables have been used in water demand forecasting models. Dynamic nature of the data used in these modeling techniques are not similar. In fact, not enough attention has been paid to the periodic or chaotic nature of the time series deployed in water demand forecasting techniques. The purpose of this study was to investigate existence of low dimensional chaos in weather information variables used in demand forecasting models. The original objective of this chapter was to use a chaotic approach for water demand forecasting. For this purpose, one should always check if the explanatory variables can exhibit low dimensional chaos. Using the correlation dimension method to check the evidence of deterministic chaos, this research proved these explanatory variables can exhibit high dimensional chaos or colored noise in stochastic systems. It is important that climatic information was only used in this chapter since demographic variables are not associated with lag times in their dynamic nature.

2.2 Background

Water demand forecast models are mainly based on the available data for a set of explanatory variables without considering the dynamic nature of the data (i.e., black-box models). If the dynamic nature is considered, it will facilitate decision making on the number/type of the parameters that should be studied in the modeling. Moreover, it will improve the accuracy of the

[1] A version of this chapter has been published as a full paper: Shabani S, Naser G. Dynamic nature of explanatory variables in water demand forecasting. Procedia Engineering. 2015 Jan 1; 119:781-7.

results while making the model more computationally efficient. Many studies have considered weather conditions like temperature, rainfall, relative humidity, and wind speed in their demand forecasting models [13]–[15]. The existence of low dimensional chaos has been proved among hydrological factors [42], which are basically used as explanatory variables in water demand forecasting models. Therefore, the objective of this research was to investigate the possible use of chaotic approach in water demand forecasting models. The correlation dimension method was used to quantify low chaos among climatic variables of water demand forecasting models in this chapter. This method has been used in other fields like river engineering [43], [44]. The research outcome will assist the authorities with their planning, design, operation, asset management, and future financial planning and rate adjustment of their WDS. Indeed, an accurate water demand forecasting model will be a basis for any strategic decision making for selecting water resources, upgrading the available water resources, designing for the future water demand management options. As such, water resources are not exhausted and competing users can adequately access to those resources.

2.3 Methodology

Correlation dimension method has been widely used as a proof for evidence of chaos. This method can be defined in four steps which are explained later in detail:

1) The phase space reconstruction using Taken's theorem [45].
2) Calculation of Euclidian distances in between the points representing the reconstructed phase space in the 1st stage.
3) A plot which shows how Correlation function (Log C(r)) is changing over a certain range of *r*.

4) A plot of correlation exponent (slope of the curve plotted in logarithmic scale) corresponding to embedding dimensions.

2.3.1 Phase Space Reconstruction

To identify the stochastic or dynamic nature of the data, a phase space is reconstructed using the collected data with a proper lag time (τ) and the minimum sufficient embedding dimension (m). Taken's theorem transforms a time-series data into the geometry of a single moving point along a trajectory, where each of its points corresponds to a data. This approach guarantees consideration of dynamics of the system in the m-dimensional space where state vectors $\vec{Y}_t$ can be represented through delay coordinates. Each one of these delay coordinates is a point in the reconstructed phase space.

$$Y_t = \left\{x_t, x_{t-\tau}, x_{t-2\tau}, .., x_{t-(m-1)\tau}\right\} \tag{2-1}$$

2.3.2 Lag Time

Literature lists three methods for estimating lag time, average mutual information (AMI), autocorrelation function (ACF), and correlation integral (CI) [46]–[48]. AMI has been the most popular method in the recent literature because ACF reflects only linear properties and CI requires a large set of data [44]. Average mutual information is calculated as:

$$I(\tau) = \sum_{X(i)X(i+\tau)} P(X(i), X(i+\tau)) \log_2 \left[\frac{P(X(i), X(i+\tau))}{P(X(i)P(X(i+\tau))} \right] \tag{2-2}$$

where join probability of two successive time series ($P(X(i), X(i+\tau)$) and product of their individual marginal probability were used to find the optimum lag time. This delay can contribute to the maximum information added on $X(i)$ by the successive time series $X(i+\tau)$. The prime objective of using this approach was to make sure these two time series were independent of each other to better

represent the dynamics of the system in the phase space. This method guaranteed they were not that independent resulting in any connections between them. In other words, a balanced independency was desired to help identifying an optimum delay time.

2.3.3 False Nearest Neighbors

False nearest neighbor method (FNN) has been widely used among researchers to find the minimum sufficient embedding dimension (m) before building the phase space [49]. This optimum embedding dimension can also be defined as a number of determinant variables to represent the dynamics of the underlying system in the phase space. The embedding dimension at which the percentage of FNN drops to zero is the desired value in this method. This optimum value represents a status in which the orbit systems of the attractors do not intersect with each other. Thus, FNN is used to examine if vectors are true neighbors.

Correlation Dimension

The correlation function and exponent are determined following the approach initially proposed by Grassberger and Procaccia [50] and later applied by others [42], [51]. For every embedding dimension, the correlation exponent will be found as the slope of the curve (plotted in logarithmic scale) for the correlation function corresponding to that embedding dimension. This curve indicates if chaos exists. In a stochastic scenario, any increase in embedding dimension would constantly increase the correlation exponent. For scenarios which exhibit low dimensional chaotic behavior, the correlation exponent initially increases by increasing the embedding dimension until it eventually remains unchanged. The correlation function is used as:

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i.j=1}^{N} H(r - |Y_i - Y_j|) \quad (2.3)$$

Equation (2.3) shows how correlation scales with *r* (radius of the sphere centered on the attractors). The advantage of this method is how trajectory $U = r - |Y_i - Y_j| \geq 0$ points are directly used without a need to partition the state of phase space. *C(r)* is calculated for a certain range of *r*. The important feature of this equation is the heavy side function *H(u).* This function works as an argument which considers 1 for positive and 0 for negative values of *U*.

$$C(r) \propto \alpha r^{D_2} \quad (2.4)$$

$$D_2 = \lim_{r\to 0} \frac{\ln C(r)}{\ln r} \quad (2.5)$$

As mentioned earlier, D_2 as correlation exponent is the slope of the curve plotted in logarithmic scale. If D_2 saturates, it is an indication of low dimensional chaos. Otherwise, the system can be considered as purely stochastic.

2.4 Study Area

The studied area in this research was the City of Kelowna (British Columbia, Canada). Weather indices used in this study were limited to daily records of average temperature, average relative humidity and average wind speed. These data were collected from the Environment Canada (http://kelowna.weatherstats.ca/). All data were obtained from the weather station A (Latitude: 49° 57' 13" N: Longitude: 119° 22' 29" W) located in the City of Kelowna's airport. Relevant literature suggests maximum temperature as a potentially better determinant in water demand forecasting

[9]. However, the mean temperature is a better determinant for the City of Kelowna based on the preliminary correlation analysis.

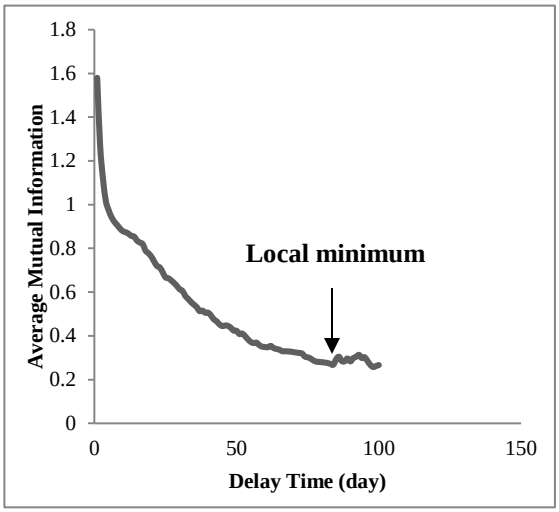
Table 2-1 Statistics of the studied variables

Statistic	Relative Humidity (%)	Average wind speed (km/h)	Average temperature (°C)
No. data	2197	2197	2197
Mean	67.43	7.91	8.46
Std. Deviation	16.26	3.58	8.77
Max	98.7	32	26.4
Min	28.2	0.33	-18.7

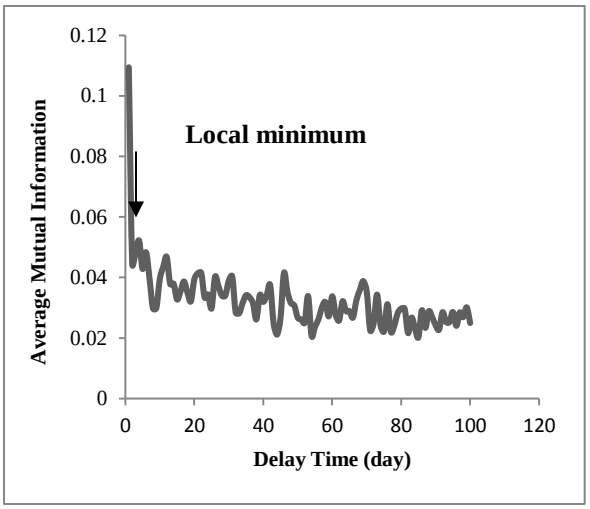
2.5 Results and Discussion

Phase space reconstruction as the first step of the correlation dimension method was accomplished using appropriate delay time and embedding dimension. Figure 2-1 shows the results of AMI method applied on a) average daily temperature, b) average daily wind speed, and c) average daily relative humidity. The first minimum value can be a good estimate of lag time. Therefore, 84 days has been used for daily temperature. Moreover, 2 days was the first local minimum for average daily wind speed and 10 days for daily average relative humidity. False nearest neighbors (%) is plotted versus the embedding dimension (figure 2-2) to find the optimal embedding dimension. Interestingly, the results of FNN were quite similar, both showing an embedding dimension of m=5 can be the minimum sufficient number of determinants in the reconstructed phase spaces. Figures in 2-3 show how correlation function C(r) is changing over the selected region of the radius. Using the optimal desired lag time and the embedding dimension, these graphs are plotted with different embedding dimensions to assess how the correlation exponent is changing. For average daily temperature, the correlation function is calculated for m=5-12 (figure 2.3a). This is

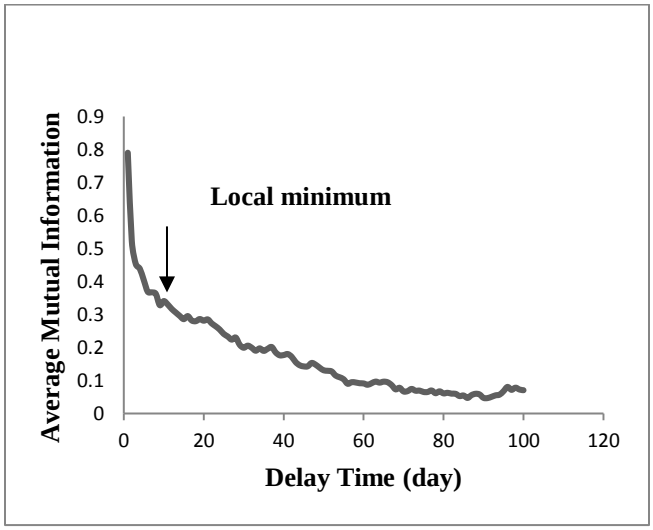
due to the higher duration of lag time (84 days) for this variable. Any dimensions larger than ten cannot give a good estimate of deterministic chaos as a much larger data set would be needed. However, for average daily wind speed with a lag time of 2 days, the correlation function is calculated for m= 5-18 (figure 2.3b). Figure in 2.3 show correlation exponent is increasing linearly with increase in embedding dimensions. Unlike the findings of previous studies in other engineering fields [42]–[44], this study showed some of the weather information might have stochastic nature in the whole system as correlation exponent is not reaching saturation in both cases.



a)



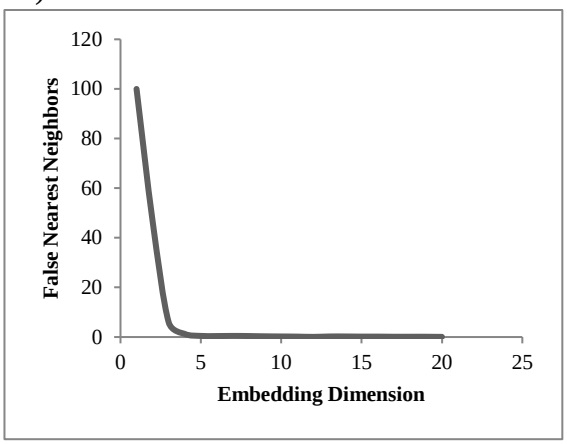
b)



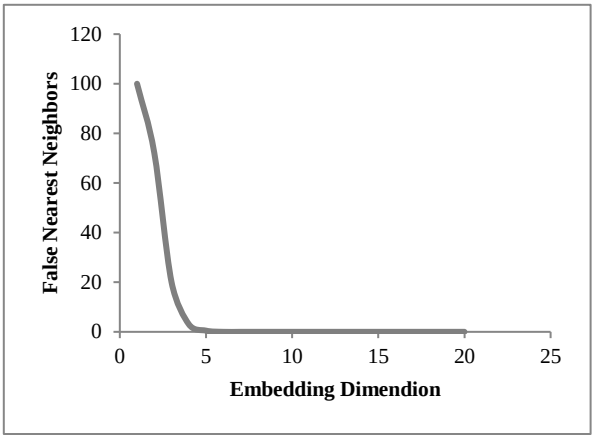
c)

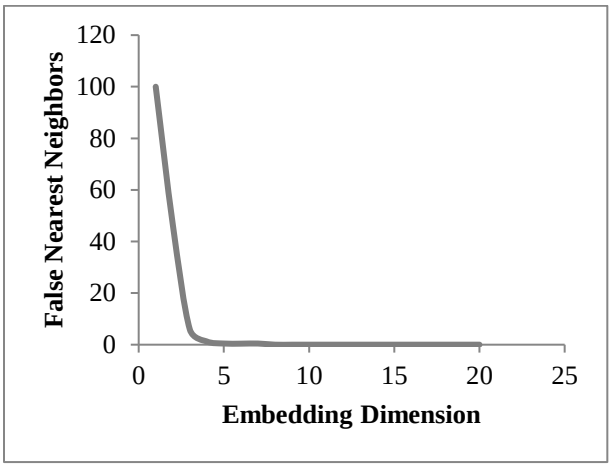
Figure 2-1 Average mutual information (a) average daily temperature (b) average daily wind speed (c) average daily relative humidity

a)



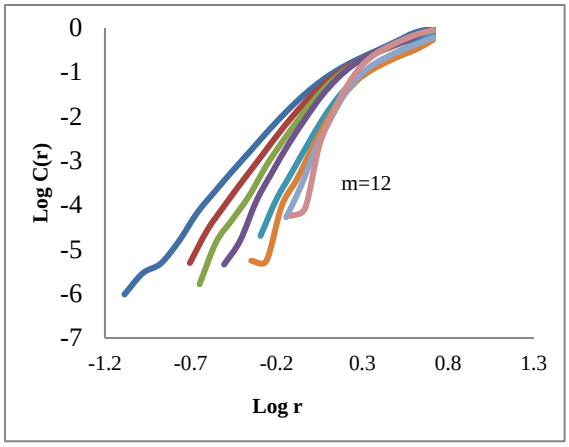
b)



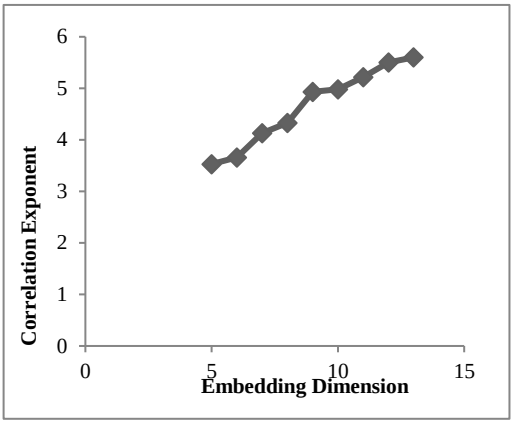


c)

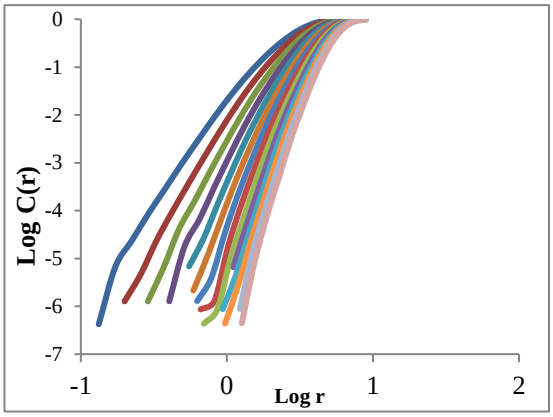
Figure 2-2 False nearest neighbors (a) average daily temperature (b) average daily wind speed (c) average daily relative humidity



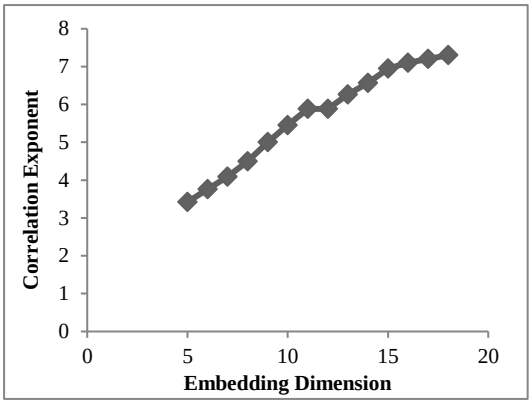
a)



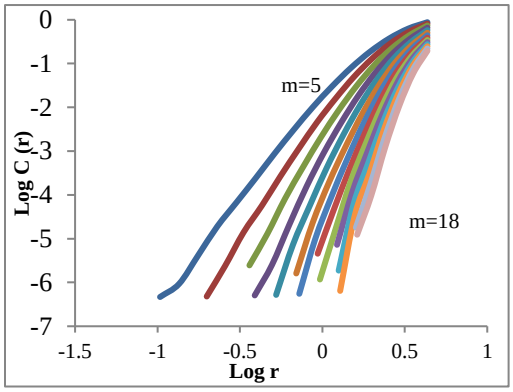
a)



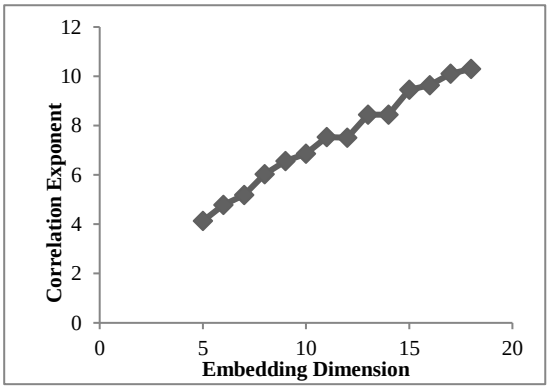
b)



b)



c)



c)

Figure 2-3 Correlation dimension and exponent (a) average daily temperature (b) average daily wind speed (c) average daily relative humidity

2.6 Summary

Chaos theory has attracted many researchers in the recent years as it could unfold the deterministic chaos within dynamic systems. Water demand forecasting can be related to a wide range of explanatory variables. Presumably, water demand is not a simply predictable quantity since it couples to highly complex weather, climate, ecological, sociological, and political dynamics. Thus, there is no reason to expect such a time series to exhibit low-dimensional chaos, although the models might. A dynamical model would predict future values of a single variable based on past values, and if it were nonlinear, it could be chaotic; this dissertation suggests this approach in water demand forecasting rather than looking for meaningful low dimensional deterministic chaos. Results of correlation dimension calculation in this study might be a typical representation of colored noise. As a practical matter, colored noise is indistinguishable from high dimensional chaos. Therefore more investigation is required to understand the dynamic nature behind this complex system.

Chapter 3: Water Demand Forecasting Using Phase Space Reconstruction to Feed SVM and GEP Models[2]

3.1 Overview

Previous chapter results showed a deterministic approach would not suit the climatic data available in this research because the nature of these explanatory variables exhibited either high dimensional chaos or colored noise. Therefore, other stochastic approaches should be used to forecast water demand from time series data. Performance of evolutionary techniques like Gene Expression Programming (GEP) is yet to be investigated in the field of water demand forecasting. Support Vector Machine (SVM) regression is an emerging machine learning algorithm used for forecast models in this field. This chapter proposed a new rationale and a novel technique in forecasting water demand using phase space reconstruction to feed the determinants of water demand with proper lag times followed by the development of GEP and SVM models. To have a better understanding of these determinants input factors were classified to 1) demand factor 2) demand and climatic factors, and 3) demand, climatic, and socio-economic factors. Different combinations of genetic operators were also used to investigate which mathematical operations can define correlations of water demand with its determinants. Three superior models of M1D3OP1, M2D3OP1, and M3D3OP2 (explained later in this chapter) were compared based on performance indices: R^2 (coefficient of determination), MAE (mean absolute error), RMSE (root mean square error), and E (Nash Sutcliff). Results showed that GEP models were highly sensitive to

[2] A version of this chapter has been published as a book chapter: Shabani S, Yousefi P, Adamowski J, Naser G. Intelligent soft computing models in water demand forecasting. InWater Stress in Plants 2016. InTech.

classification of data, genetic operators, and optimum lag time. However, M2D3K2 as the superior SVM model slightly outperformed GEP models using Polynomial kernel function. This research proved how phase space reconstruction could improve water demand forecasts using soft computing techniques.

3.2 Background

Considering climate conditions and population, the prime objective of this research was to develop a predictive model for monthly average water demand. While the research proposed a generic framework that can be easily adjusted for any case, the city of Kelowna (British Columbia, Canada) was employed as a test case. To study the dynamic nature of data, the present research proposed nonlinear modeling techniques, such as gene expression programming (GEP), support vector machines (SVM). These techniques are yet to be explored in depth by scholars in this field. Like genetic algorithm (GA) and genetic programming (GP), GEP is a genetic-based technique. Inspired by Darwin's theory of evolution, GEP was recently proposed in engineering disciplines to optimize the structure of input variables fed into predictive models [52]–[55]. The main difference among GA, GP and GEP is the nature of underlying genes. GA studies the genes as linear strings of fixed length, while they are nonlinear entities of different sizes and shapes in GP. GEP considers the genes as linear strings of fixed length expressed as nonlinear entities of different sizes and shapes. Being a self-learning algorithm, GEP has several advantages over conventional linear predictive models. GEP defines individual block structures (input variables, response, and function sets) and selects the optimized operating functions and multipliers through the process of learning algorithms. One research indicated a GEP-based model outperformed traditional linear models in the field of hydrology [54]. Since weather information is one of the major determinants

of water demand, this research employed GEP to develop an accurate predictive model. For the first time, this research proposed using GEP in water demand forecasting.

3.3 Methodology

3.3.1 Model Development

To determine water demand (*D*) in million liter (*ML*), this research used population (*P*) and hotel occupancy factor (*HOR*) as socio-economic parameters, and temperature (*T*) in °C, relative humidity (R_h) in percent, and rainfall (*R*) in mm as weather parameters. As they did not have the same order of magnitude, each parameter was normalized prior to models' developments using:

$$X = \frac{x - \mu}{\sigma} \tag{3.1}$$

where *X* is the standardized magnitude of parameter *x*, and μ and σ are the corresponding mean and standard deviation. Phase space reconstruction of each explanatory variable was used prior to GEP modelling to define the structure of the inputs of the models. This was done to identify the stochastic or deterministic nature of the collected data. For a given proper lag time, the phase space was built by using Taken's theorem [45] that transforms time-series data into the geometry of a single moving point along a trajectory, where each of its points corresponds to a data. Average mutual information (AMI) was used to determine the proper lag time of water demand for phase space reconstruction of all input factors. This has been done to attain a comprehensive understanding of the input factors, self-interaction of used variables, and finally assessing the use of lag times in demand forecasting models. A total number of twenty-seven models were created using 3 classifications (input's nature), 3 types of genetic operators, and 3 different lag times. Table 3-1 shows the structure of these models. M1, M2, M3 represent the classifications of the explanatory variables meaning (demand info based) for M1, (demand + climatic info based) for

M2, (demand + climatic + demographic info based) variables for M3. D1, D2, and D3 are used to label the number of months used as lag time in developing these models. OP1, OP2, and OP3 mean which genetic operators are used for creation of GEP models (Table 3-2).

Table 3-1 Structure of classified models (t refers to current month)

Classification	Model	Input Variables Combination
Demand Based	M1D1	$D(t\text{-}1)$
	M1D2	$D(t\text{-}1)$, $D(t\text{-}2)$
	M1D3	$D(t\text{-}1)$, $D(t\text{-}2)$, $D(t\text{-}3)$
Demand + Weather info Based	M2D1	$D(t\text{-}1)$, $T(t\text{-}1)$, $R(t\text{-}1)$, $R_h(t\text{-}1)$
	M2D2	$D(t\text{-}1)$, $D(t\text{-}2)$, $T(t\text{-}1)$, $T(t\text{-}2)$, $R(t\text{-}1)$, $R(t\text{-}2)$, $R_h(t\text{-}1)$, $R_h(t\text{-}2)$
	M2D3	$D(t\text{-}1)$, $D(t\text{-}2)$, $D(t\text{-}3)$, $T(t\text{-}1)$, $T(t\text{-}2)$, $T(t\text{-}3)$, $R(t\text{-}1)$, $R(t\text{-}2)$, $R(t\text{-}3)$, $R_h(t\text{-}1)$, $R_h(t\text{-}2)$, $R_h(t\text{-}3)$
Demand + Weather + Population info Based	M3D1	$D(t\text{-}1)$, $T(t\text{-}1)$, $R(t\text{-}1)$, $R_h(t\text{-}1)$, P, HOR
	M3D2	$D(t\text{-}1)$, $D(t\text{-}2)$, $T(t\text{-}1)$, $T(t\text{-}2)$, $R(t\text{-}1)$, $R(t\text{-}2)$, $R_h(t\text{-}1)$, $R_h(t\text{-}2)$, P, HOR
	M3D3	$D(t\text{-}1)$, $D(t\text{-}2)$, $D(t\text{-}3)$, $T(t\text{-}1)$, $T(t\text{-}2)$, $T(t\text{-}3)$, $R(t\text{-}1)$, $R(t\text{-}2)$, $R(t\text{-}3)$, $R_h(t\text{-}1)$, $R_h(t\text{-}2)$, $R_h(t\text{-}3)$, P, HOR

Table 3-2 Genetic operators

OP1	$\{+, -, x\}$
OP2	$\{+, -, x, x^2, x^3\}$
OP3	$\{+, -, x, x^2, x^3, \sqrt{}, e^x, \log, \ln\}$

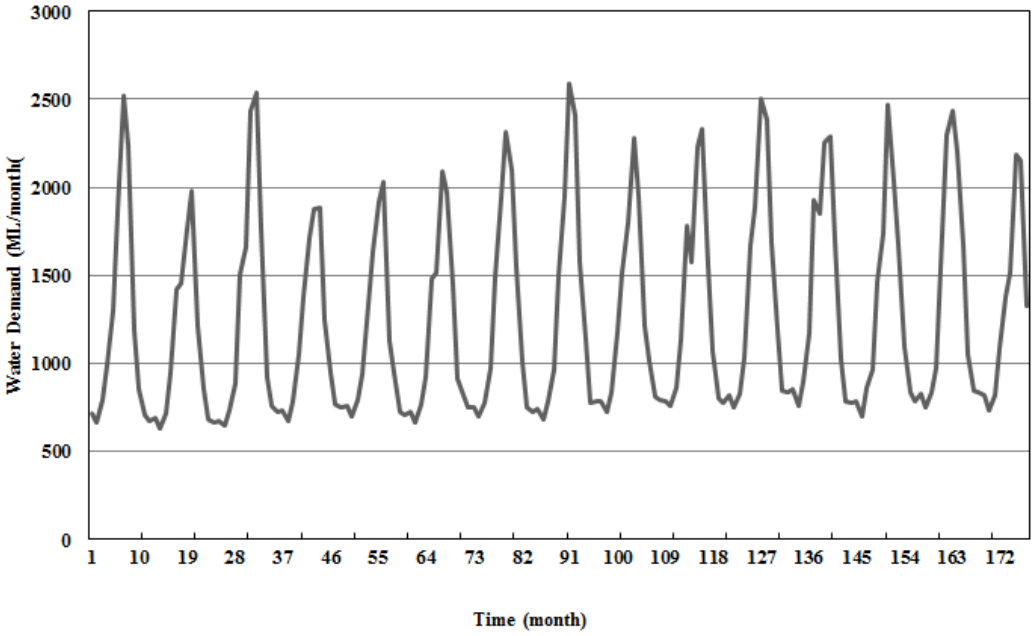


Figure 3-1 Time series of water demand in COK 1996- 2010

Data were used in partitions of 144 samples for training (1996 to 2007) and 35 for validation (2008 to 2010). Figure 3-1 shows a time series of water demand over this period. The figure shows that COK experiences a periodic cycle water demand, which is relatively in a regular pattern due to seasonal changes.

3.3.2 Gene Expression Programming

Ferreira [52] introduced a GEP model as one of the emerging soft computing techniques. The strategy used for the learning algorithms were the optimal evolution using the genetic operators. The general settings of the learning/training algorithms were 30 chromosomes, 8 head size, and 3 numbers of genes [52]. These settings defined the overall structure of the model. The selected head size determined how complex each parameter of the model was. Each head of the genes went under a set of different arrangements to model the feeding data. Selecting new random populations were followed by reproduction to reach the most suitable model with the optimized stopping condition. Models were developed based on 3 genes linked together by addition function. Number of genes

per chromosome was to specify the layers or blocks building the whole model. Although a large gene was useful; dividing the chromosomes into simpler units resulted in a more efficient and manageable learning process. Root mean square of error (RMSE) was used as fitness function to fit a curve to target values. The stopping condition was maximum fitness and correlation coefficient (R^2). Ten numerical constants were used as floating-point data in each gene. Appendix A shows the step by step approach used in GeneXprotools V 5.0 to develop GEP models.

3.3.3 Lag Time

As mentioned in chapter 2, average mutual information (equation 2-2) was used to determine the optimum lag time. A balanced independency was of desire to help identifying an optimum delay time through this method.

3.3.4 Support Vector Machines

Support vector machine method was initially proposed as a classification method in 1995 [56]. It was later modified as a regression technique. Using kernel functions, SVM regressors can account for nonlinearity in the systems. SVMs have been recently used in predictive models [57]; however, this new technique is yet to be deployed in water demand forecasting. This research applied SVMs in urban water demand forecasting along with pre-processed population based and climatic information.

In this method the input vectors are considered as supports forming the backbone of the whole model structure through a training process. If $\boldsymbol{N}$ samples of the population given by, $X \in R^m$, $Y \in R$. Where, $\boldsymbol{X}$ represents an input parameter $\{X_K, Y_K\}_{K=1}^{N}$ with m number of components along with $\boldsymbol{y}$ as its response output variable. A function or support vector machine estimator on a regression can be considered as:

$$f(x) = W.\phi(X) + b \tag{3.3}$$

Where *W* is a weight vectors, and *b* represents a bias. φ is used as a transfer function which exhibits nonlinear behavior, mapping the input vectors into a higher dimensional space. Furthermore, these mapped vectors can compromise complex nonlinear regression of the input space. In order to solve the previous function, Cortes and Vapnik [56] introduced the convex optimization problem with an insensitivity loss function as below:

Minimize *w*, *b*, ξ, ξ^*

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{N}\left(\xi_k + \xi_k^*\right)$$

Subject to:

$$\left\{\begin{array}{l} y_k - \mathbf{w}^T\phi(\mathbf{x}_k) - b \leqslant \varepsilon + \xi_k \\ \mathbf{w}^T\phi(\mathbf{x}_k) + b - y_k \leqslant \varepsilon + \xi_k^* \\ \xi_k,\ \xi_k^* \geqslant 0 \end{array}\right\} \quad k = 1,\ 2,\ \cdots,\ N \tag{3.4}$$

Where ξ and $\xi^*$are slack variables that penalize training errors by the loss function over the error tolerance *e*, and *C* is a positive tradeoff parameter that determines the degree of the empirical error in the optimization problem. Following the approach [58] and [59], Optimization of this proposed equation was done through Lagrangian multipliers and the Karush Kuhn-Tucker (KTT) conditions simultaneously. The structure of SVM model is shown in figure 3-2. Kernel functions are used to map the input vectors into higher dimensions in space. Linear (Lin), polynomial (Poly), radial basis (RBF), and sigmoid (Sig) are commonly used kernel functions in Literature. This study compared performance of RBF, Poly, and Lin kernel functions.

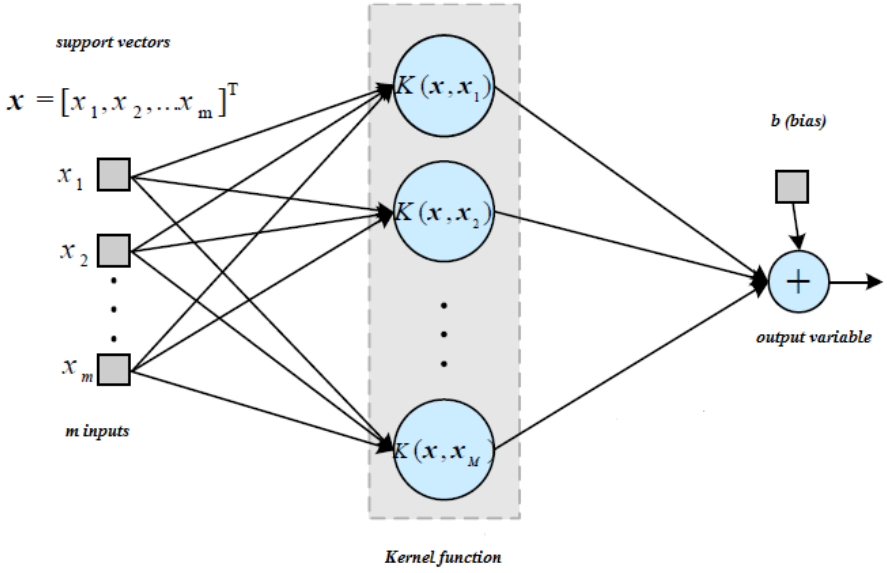


Figure 3-2 SVM model's structure

3.4 Results and Discussion

Prime objective of using phase space reconstruction was to find a proper lag time for evolving our models. To have a comprehensive understanding of the model performance, GEP models were defined by all lag times up to the optimum value determined for water demand of COK. AMI calculations of the water demand in COK resulted in a proper lag time of 3 months. Figure 3-3 shows the first local minimum point occurs at 3 months in which AMI was giving the optimum lag time for phase space reconstruction (τ = 0.6591 for 2 months, τ = 0.5073 for 3).

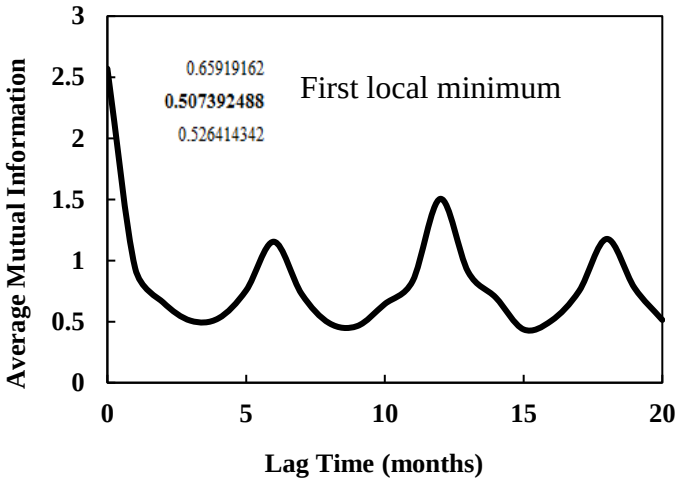


Figure 3-3 AMI for water demand.

Figures 3-4a, 3-4b, and 3-4c show the phase space diagrams of water demand for τ =1, 2, and 3 months, respectively. Each one of these figures represent the state of WDS demand at the given time. The evolution of phase space in this time series was given by a reconstructing a pseudo phase space in which the demand of COK, which was a nonlinear system, was considered by its self-interaction using AMI [44]. Figure 4c is more bilateral (if a 45-degree line is drawn and either sides are compared) and has a regular pattern in comparison with the other two previous states of phase space. Therefore, Figure 3-4 comes in agreement with the estimated lag time of τ = 3 months, since the phase space diagrams at different states proved lag time of 3 months was the optimum. Prior to analysis of GEP models, a correlation table between the explanatory variables and water demand provided a better understanding of how to define the input factors (Table 3-3). The correlations were in order of 0.92, 0.84, -0.83, 0.11, and -0.01 for (*D-T*), (*D-HOR*), (*D-*R_h), (*D-P*), (*D-R*) respectively. Interestingly, outdoor water use was highly correlated to temperature and hotel occupancy rate in COK, which describes the periodic cycle of demand due to seasonal changes. This research, however, employed all input factors in evolving the GEP models.

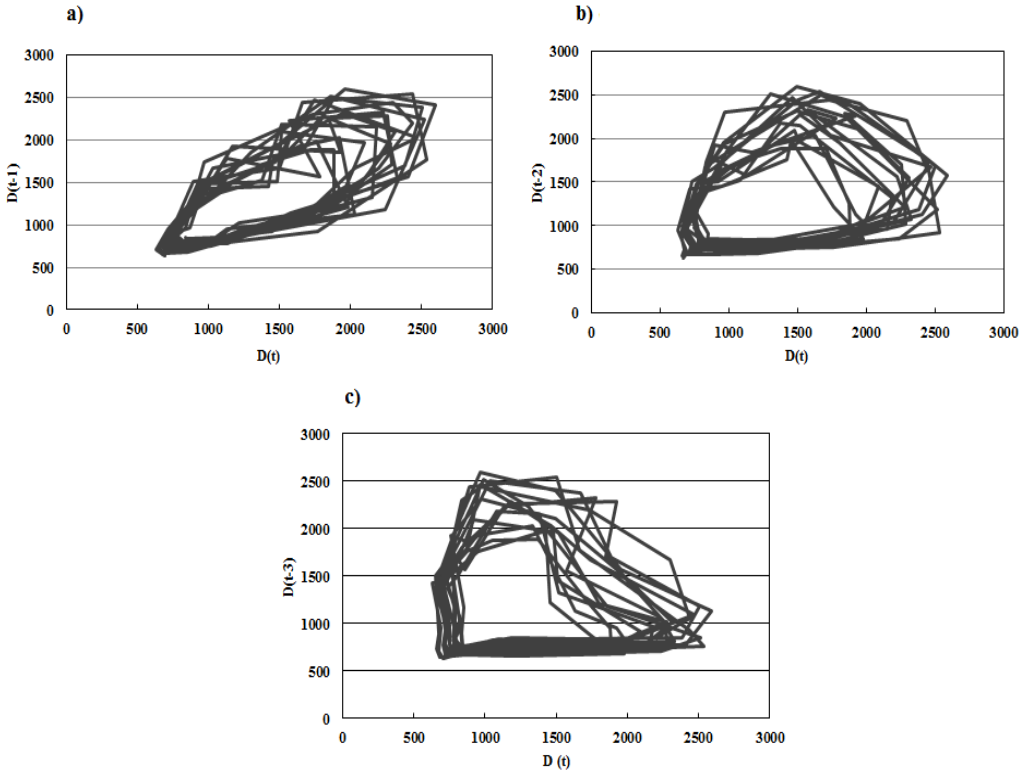


Figure 3-4 Phase space diagrams lag times (1-3 months)

Table 3-3 Correlation between studied factors

	D	***T***	***R***	R_h	***P***	***HOR***
D	1.00	0.92	-0.01	-0.83	0.11	0.84
T	0.92	1.00	0.10	-0.89	0.00	0.92
R	-0.01	0.10	1.00	-0.05	-0.26	0.11
R_h	-0.83	-0.89	-0.05	1.00	0.02	-0.84
P	0.11	0.00	-0.26	0.02	1.00	-0.09
HOR	0.84	0.92	0.11	-0.85	-0.09	1.00

Table 3-4 shows a total of 27 GEP models developed in this research. Three superior models were highlighted in each category or classification of determinants. Interestingly, the lag time of 3 months outperformed other combinations in all different classifications which show the importance of using phase space construction in studying complex systems. This proves how

proper lag time determined by *AMI* can significantly improve the performance of the forecasting models. Different genetic operators were also used to understand which mathematical operations better define the nature of these determinants. It is known that the first operator {+, -, x} showed better performance in the first two classifications for demand based and demand plus climatic info based categories. The second operator (OP2) {+, -, x, x^2, x^3} outperformed other operators in the third classification (demand + socio-economic +climatic information) of input parameters in which socio-economic factors were included. It is interesting that using more complex mathematical operations OP3 {+, -, x, x^2, x^3, $\sqrt{}$, e^x, log, ln} reduced the quality of the model's performance consistently. This showed that water demand forecasting could be reasonably explained in models using basic mathematical operations despite its complexity in nature. Three performance indices of MAE, RMSE, and R^2 were used to investigate the sensitivity of the models to determinant's classification, genetic operators, and lag time. Comparing the superior models in each category [a) M1D3OP1, b) M2D3OP1, and c) M3D3OP2] through these indices could not give us a solid reason to pick one over the others. This is due to the very close testing MAE of 0.304, 0.3035, and 0.291, respectively. RMSE values were also close to each other as 0.3984, 0.3664, and 0.3660 for corresponding superior models. Finally, R^2 values showed better performance of M2 and M3, however, R^2 was not enough to judge the performance of the models (Figure 3-5).

Table 3-4 Performance of GEP models

Model ID*	Training			Testing		
	MAE	RMSE	R^2	MAE	RMSE	R^2
$M_1D_1OP_1$	0.4687	0.6974	0.6284	0.4833	0.6067	0.6343
$M_1D_1OP_2$	0.4718	0.6100	0.6252	0.4849	0.6120	0.6300
$M_1D_1OP_3$	0.4672	0.6118	0.6235	0.4800	0.6112	0.6281
$M_1D_2OP_1$	0.3552	0.4721	0.7754	0.378	0.4607	0.7892
$M_1D_2OP_2$	0.3574	0.4721	0.7756	0.3794	0.4608	0.7892
$M_1D_2OP_3$	0.3008	0.4049	0.8481	0.4188	0.5188	0.8346
$M_1D_3OP_1$	**0.3229**	**0.4317**	**0.8156**	**0.3040**	**0.3984**	**0.8452**
$M_1D_3OP_2$	0.2858	0.3641	0.8691	0.3488	0.3106	0.8452
$M_1D_3OP_3$	0.3545	0.4647	0.7849	0.3637	0.4548	0.8029
$M_2D_1OP_1$	0.3777	0.4790	0.7735	0.4529	0.5296	0.7552
$M_2D_1OP_2$	0.3955	0.4933	0.7560	0.4423	0.5169	0.7546
$M_2D_1OP_3$	0.3914	0.4893	0.7903	0.4596	0.5488	0.7643
$M_2D_2OP_1$	0.2463	0.3359	0.8867	0.3015	0.3981	0.8426
$M_2D_2OP_2$	0.3236	0.4022	0.8438	0.3455	0.4176	0.8473
$M_2D_2OP_3$	0.3580	0.4450	0.8048	0.3987	0.4798	0.8077
M2D3OP1	***0.2957***	***0.3758***	***0.8623***	***0.3035***	***0.3664***	***0.8945***
$M_2D_3OP_2$	0.3619	0.4445	0.8085	0.3893	0.4649	0.8139
$M_2D_3OP_3$	0.3033	0.4184	0.8502	0.3339	0.4562	0.8260
$M_3D_1OP_1$	0.2776	0.3810	0.8542	0.4201	0.5869	0.7087
$M_3D_1OP_2$	0.3474	0.4194	0.8237	0.4154	0.5348	0.7919
$M_3D_1OP_3$	0.2780	0.3601	0.8861	0.3933	0.5410	0.7714
$M_3D_2OP_1$	0.2875	0.3694	0.8778	0.4987	0.6332	0.6999
$M_3D_2OP_2$	0.3514	0.4543	0.8147	0.5694	0.6959	0.7027
$M_3D_2OP_3$	0.3944	0.2205	0.7827	0.5219	0.6408	0.7401
$M_3D_3OP_1$	0.3213	0.3961	0.8609	0.5624	0.6556	0.6839
$M_3D_3OP_2$	**0.2483**	**0.3230**	**0.9005**	**0.2910**	**0.3660**	**0.8882**
$M_3D_3OP_3$	0.3907	0.4801	0.7800	0.3655	0.4582	0.8236

* M_1, Demand; M_2, Demand + Climactic; M_3, Demand + Climactic + Socioeconomic; D_1, $\boldsymbol{\tau}$ (lag) = 1 month; D_2, $\boldsymbol{\tau}$ = 2 months; D_3, $\boldsymbol{\tau}$ = 3 months; OP_1, {+, -, x}; OP_2, {+, -, x, x2, x3}; OP_3, {+, -, x, x2, x3, √, ex, log, ln}; R^2, coefficient of determination; MAE, mean absolute error; RMSE, root mean square error.

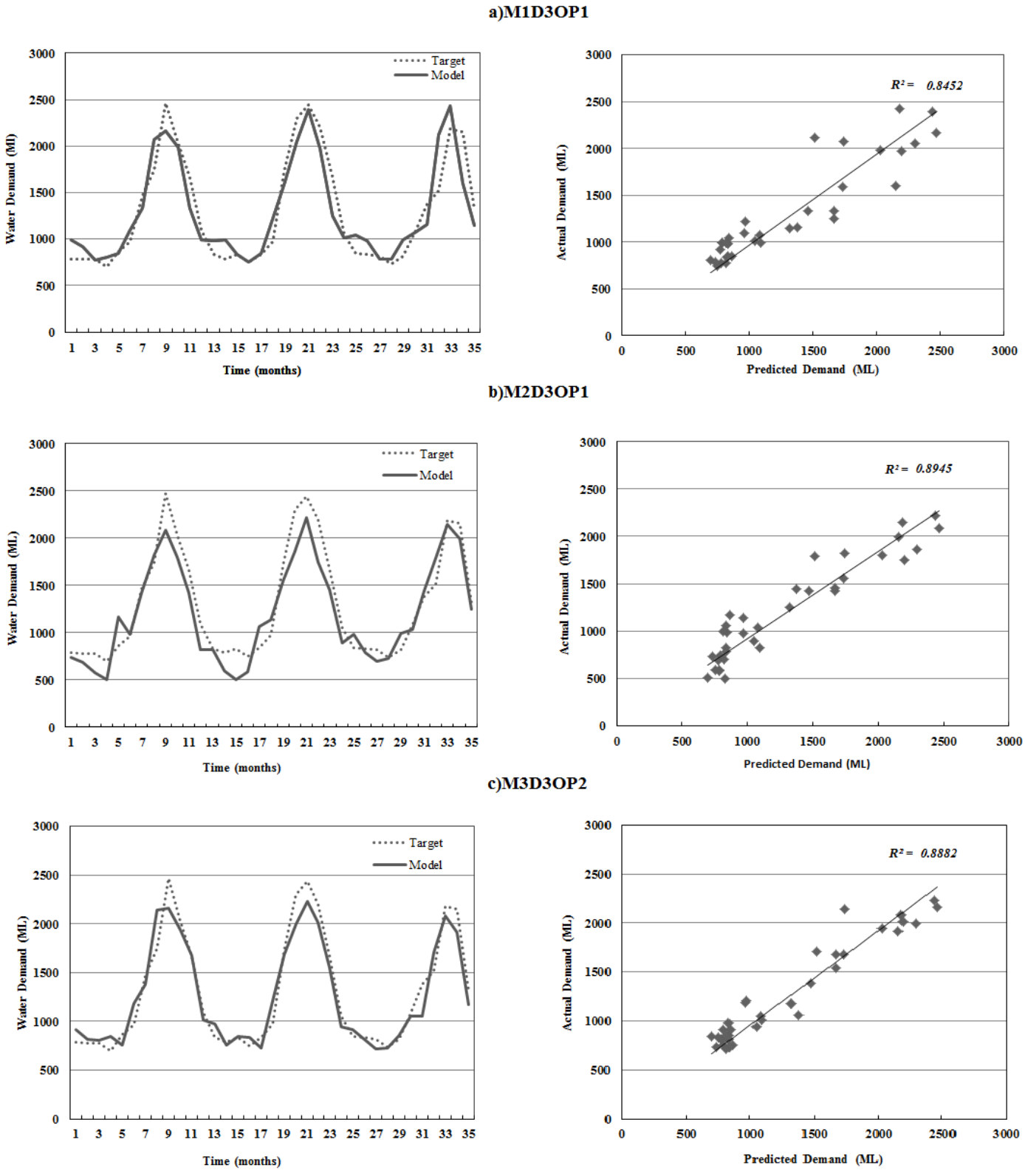


Figure 3-5 Superior GEP models a) M1D3OP1 b) M2D3OP1 c) M3D3OP2

Therefore, accumulative water demand was calculated in testing and target/actual values. M1 and M3 were closer to target accumulative demand as shown in Figure 3-6. This shows M2 was not the best model due to the error it exhibited when comparing with accumulative demand. This revealed M1 and M3 as the best possible models in this research. To pick one of them, accumulative (target – model output) was plotted in Figure 3-7. This figure helps to pick M3 as the best due to lower fluctuations of errors and following a consistent pattern in most of the graphs' domain. This might be since combining socio-economic factors with demand + climatic info might have resulted in a smoother model, which lowered the error associated with the other two classifications.

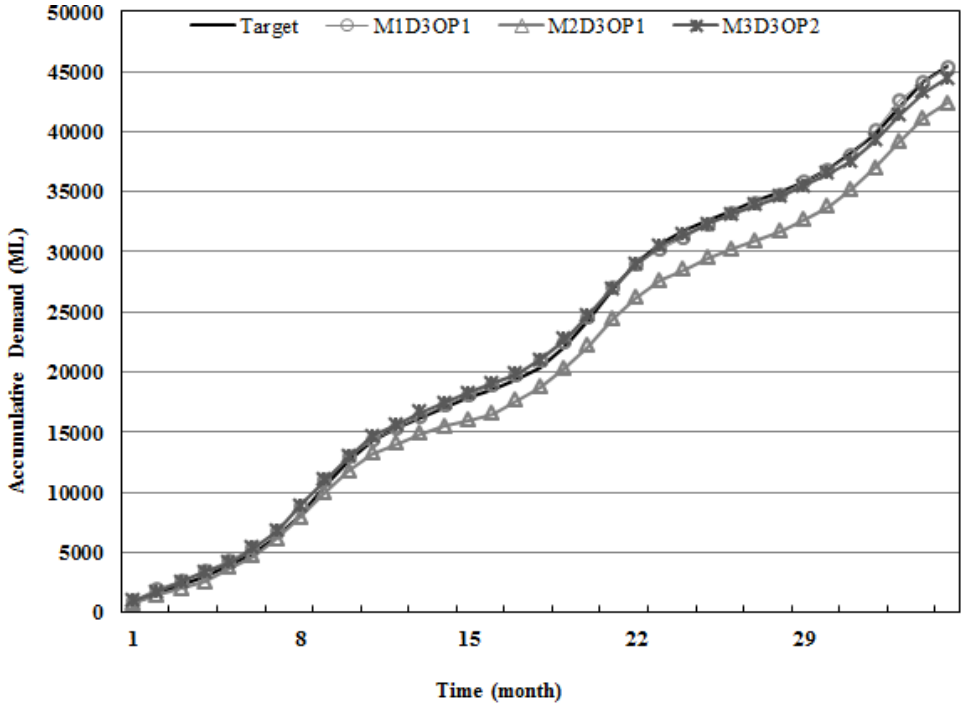


Figure 3-6 Accumulative demand with time

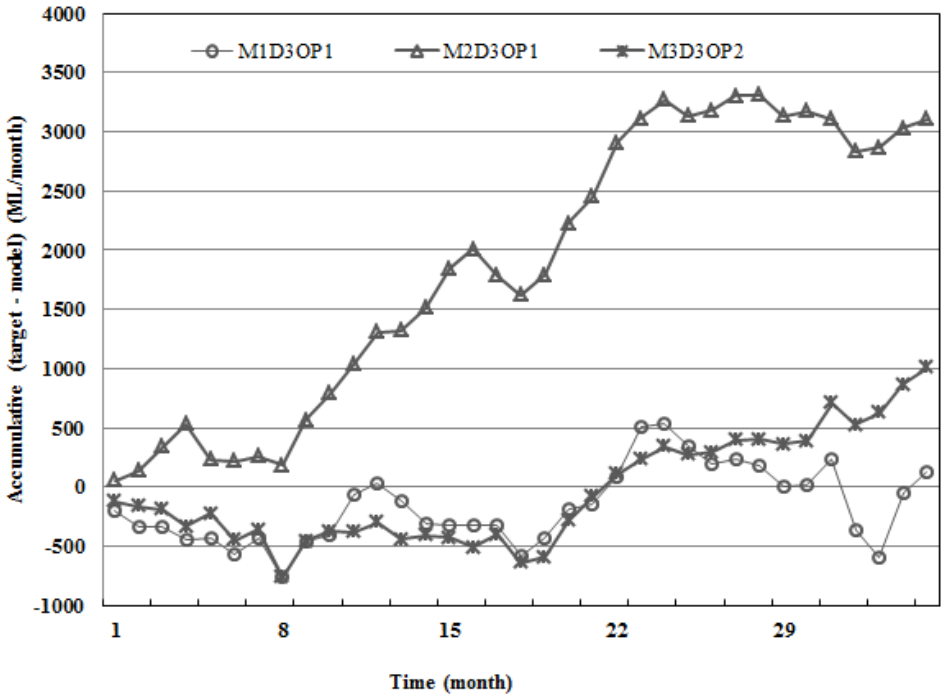


Figure 3-7 Accumulative (target-model) with time

The superior GEP models from each classification were compared with SVM models using three different kernel functions (RBF, Poly, and Lin). Since genetic operators are not part of this approach, target models are defined without OP1-3. Table 3-5 shows the training and testing performance indices for developed SVM models using three kernel functions. Results show that Poly kernel functions outperformed other kernel functions in this study. Interestingly Lin kernels performed poorly which shows the nature of input parameters cannot be considered using these functions. M2D3 is selected as the superior SVM model to be compared with GEP models (Figure 3-8).

Table 3-5 Performance of SVM models.

Model ID*	Training			Testing		
	R^2	*RMSE*	*E*	R^2	*RMSE*	*E*
M_1D_3*RBF*	0.9545	0.2123	0.9546	0.8397	0.4051	0.8387
M_2D_3*RBF*	0.9856	0.1201	0.9855	0.8701	0.3678	0.867
M_3D_3*RBF*	0.9416	0.2407	0.9415	0.9258	0.3014	0.9107
M_1D_3*Poly*	0.9308	0.2618	0.9309	0.8206	0.4278	0.8201
M_2D_3*Poly*	0.9372	0.2497	0.9371	0.9343	0.2593	0.9339
M_3D_3*Poly*	0.9428	0.239	0.9424	0.9279	0.3002	0.9114
M_1D_3*Lin*	0.7864	0.4602	0.7864	0.7945	0.4592	0.7927
M_2D_3*Lin*	0.8894	0.3311	0.8894	0.8977	0.323	0.8974
M_3D_3*Lin*	0.9093	0.2998	0.9004	0.9084	0.3344	0.8901

* M_1, Demand; M_2, Demand + Climactic; M_3, Demand + Climactic + Socioeconomic; D_1, $\boldsymbol{\tau}$ (lag) = 1 month; D_2, $\boldsymbol{\tau}$ = 2 months; D_3, $\boldsymbol{\tau}$ = 3 months; RBF, Poly, Lin R^2, coefficient of determination; *RMSE*, root mean square error; *E*, Nash-Sutcliffe coefficient.

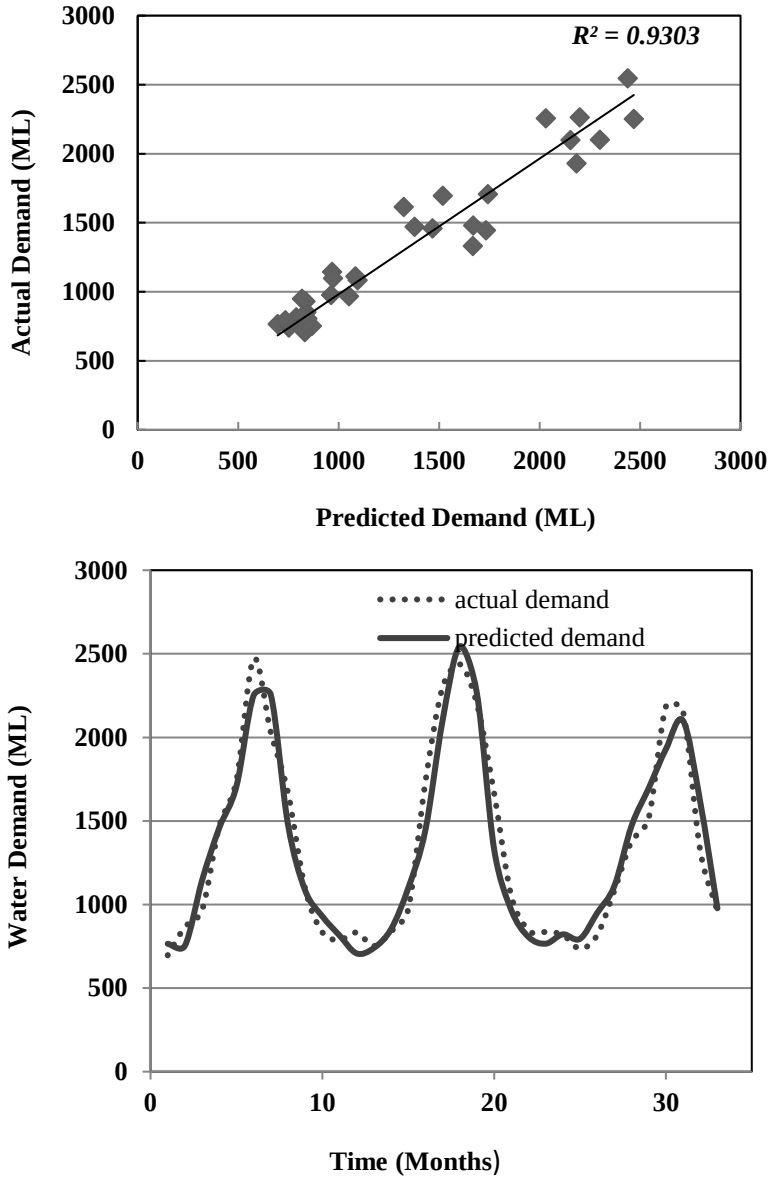


Figure 3-8 Superior SVM model

3.5 Summary

A wide range of modeling techniques has been proposed by researchers over the recent years, trying to improve the models' accuracy in prediction. However, this research defined a new rationale for modelling water demand which opts genetic expression programming along with phase space reconstruction. Proper lag time determined by AMI method defined the structure of the explanatory variables deployed in models. The outcome of this research proved GEP models are highly sensitive to classification of input factors, proper lag time, and selection of genetic operators. In general, soft computing techniques like GEP should receive more attention in forecasting behaviors of complex systems like WDS because of their high accuracy and capability of showing internal relationships between explanatory variables. These models can offer valuable information to WDS operators and designers to deploy optimum determinants in their forecast models since they are not built based on a black-box approach like artificial intelligence models or other conventional modelling approaches like linear regression which cannot account for nonlinearity of the determinants. The three selected superior GEP models proposed in this research were compared using different performance indices, however differentiating between them were not possible due to close values of indices, therefore a third GEP model was selected due to lower error when accumulation of demand is under consideration and it had less fluctuation in comparison with the first model. However, these models were slightly outperformed by the superior SVM model with better performance indices. This shows both GEP and SVM can be emerging techniques in water demand forecasting with can account for nonlinearity of the input parameters. The outcome of this research can be used both regionally for Kelowna and on national level as educated guess or forecast of water demand can significantly help decision making of governments in dealing with water availability dilemma.

Chapter 4: Optimum Lag Time Determination in Analysis of Water Demand Using SVM [3]

4.1 Overview

The optimum lag time associated with time series data of water demand with monthly resolution was used to perform the phase space reconstruction of the input variables in chapter 3. This approach raises the question of how appropriate it is to generalize this optimum lag, using it for all explanatory variables while defining inputs feeding the predictive models. Since using AMI is relatively new in phase space reconstruction of multi-variate analysis of water demand, there is a need to explore this concept further in detail. The prime objective of this chapter was investigating transformation of deployed time series of input data through phase space reconstruction to allow for use of different lag times to be explored in the final output of the model. This was done to confirm if the approach in chapter 3 was a reasonable practice. This research applied SVMs in urban water demand forecasting coupled with pre-processed water demand and climatic information. Results of this study showed optimum lag time of the input variables can significantly improve the performance of SVM models if the optimum lag time of the response variable is used in input design.

[3] A version of this chapter has been published as a whole paper: Shabani S, Yousefi P, Naser G. Support vector machines in urban water demand forecasting using phase space reconstruction. Procedia Engineering. 2017 Jan 1; 186:537-43.

4.2 Methodology

The main purpose of this approach was to determine the impact of lag time on support vector machines using phase space reconstruction. AMI (average mutual information) is opted for determination of optimum lag time. This technique is selected rather than auto-correlation function (ACF) and correlation integral (CI) as these alternatives require large sets of data or fail to exhibit nonlinearity of the models [44].

4.2.1 Phase Space Reconstruction: proper lag time

Equation 4.1 can determine the appropriate lag time between two independent time series. This approach uses the joint probability of sequential time series which succeed one another by an increment (equal to 1 unit of the lag time). Moreover, the product of their marginal probability is also utilized to determine the optimum lag time. Like Shannon's entropy, this technique can be a good estimate of how entropy levels can change the dynamics of these deployed time series in forecasting models. An optimum lag time is used to make sure sufficient information is added on the balanced independent time series which can magnify the behavior of these time series in a desired phase space. Figure 4-1 shows how the first local minimum of these graphs estimated the optimum lag time. In this case, optimum lag time was selected as 1 month for precipitation. On the other hand, temperature and water demand showed an optimum lag time of 3 months.

$$I_\tau = \sum_{i=1}^{i=n} P(X_i, X_{i+\tau}) \cdot \log_2 \frac{P(X_i, X_{i+\tau})}{P(X_i) \cdot P(X_{i+\tau})} \tag{4.1}$$

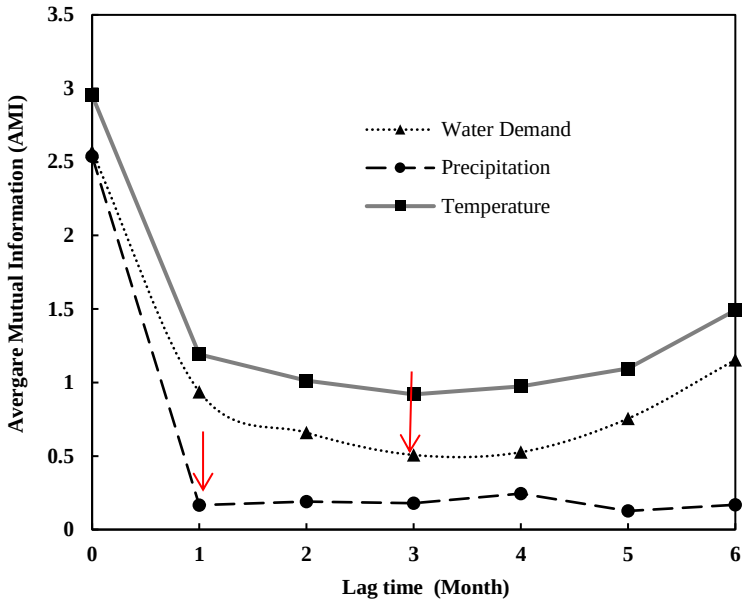


Figure 4-1 Average mutual information.

4.2.2 Model Design

Phase space reconstruction improved the performance of the GEP and SVM models proposed in previous chapter. However, the impact of feeding models explicitly with individual optimum lag times remains to be explored. Therefore, this research considered 4 different design combinations as shown in table 4-1. The first model labelled as τD^{opt} only considers the optimum lag time of demand (3 months). Followed by τT^{opt} and τR^{opt} which consider their corresponding individual optimum lag times of 3 and 1 accordingly. Finally, the last designed combination considers all possible lag times from 1 to the optimum lag time of the demand as the focus of this study. This is to investigate which design combination can result in better performance of models.

Table 4-1 Design combinations

*Model ID**	Input variables
τD^{opt}	$D_{t-3}, T_{t-1}, P_{t-1}$
τT^{opt}	$D_{t-1}, T_{t-3}, P_{t-1}$
τP^{opt}	$D_{t-1}, T_{t-1}, P_{t-1}$
$\tau_{1,2,3}$	$D_{t-1}, D_{t-2}, D_{t-3}, T_{t-1}, T_{t-2}, T_{t-3}, P_{t-1}, P_{t-2}, P_{t-3}$

* *t* refers to current month; τ represents the lag, *D is* demand in *Ml*; *T* is temperature *°C*; *P is* total precipitation in *mm*.

4.2.3 Results and Discussion

As mentioned earlier in this chapter, 4 combinations were used to assess the impact of lag time in the inputs of demand forecasting models. Table 4-2 compares the performance of the SVM models with their corresponding design. The models are compared based on coefficient of determination (R^2) and Root mean square of error (RMSE). Since R^2 is not sensitive to data point's outliers [60], *RMSE* was used to have a better error analysis of the developed models. R^2 ranges between 0-1, with R^2=1 being the best model which can replicate the field data explained by developed models. Its complementary index, RMSE, gives an estimate of prediction error, with values close to 0 being the best possible outcome. Results showed optimum lag time of demand (τD^{opt}), can lead to promising results. Using demand's lag time explicitly could slightly outperform the second model using temperature's optimum lag time (τT^{opt}). However, explicit use of optimum lag time for precipitation (τP^{opt}), performed significantly poorly in comparison with the other two design combinations. In the end, the final model which considered all possible lag times of 1-3 months for explanatory variables outperformed other model designs which comes in agreement with the outcome of chapter 3 [61]. Figure 4-2 illustrates a visual comparison of the superior model in this

research with actual water demand. It shows SVM models are highly sensitive to application of phase space reconstruction.

Table 4-2 Performance of SVM models

Model ID	Training		Testing	
	R^2	RMSE	R^2	RMSE
τD^{opt}	0.9258	0.3778	0.9434	0.3551
τT^{opt}	0.946	0.324	0.9521	0.314
τP^{opt}	0.8476	0.5301	0.7374	0.6807
$\tau_{1,2,3}$	0.9582	0.2863	0.9662	0.2615

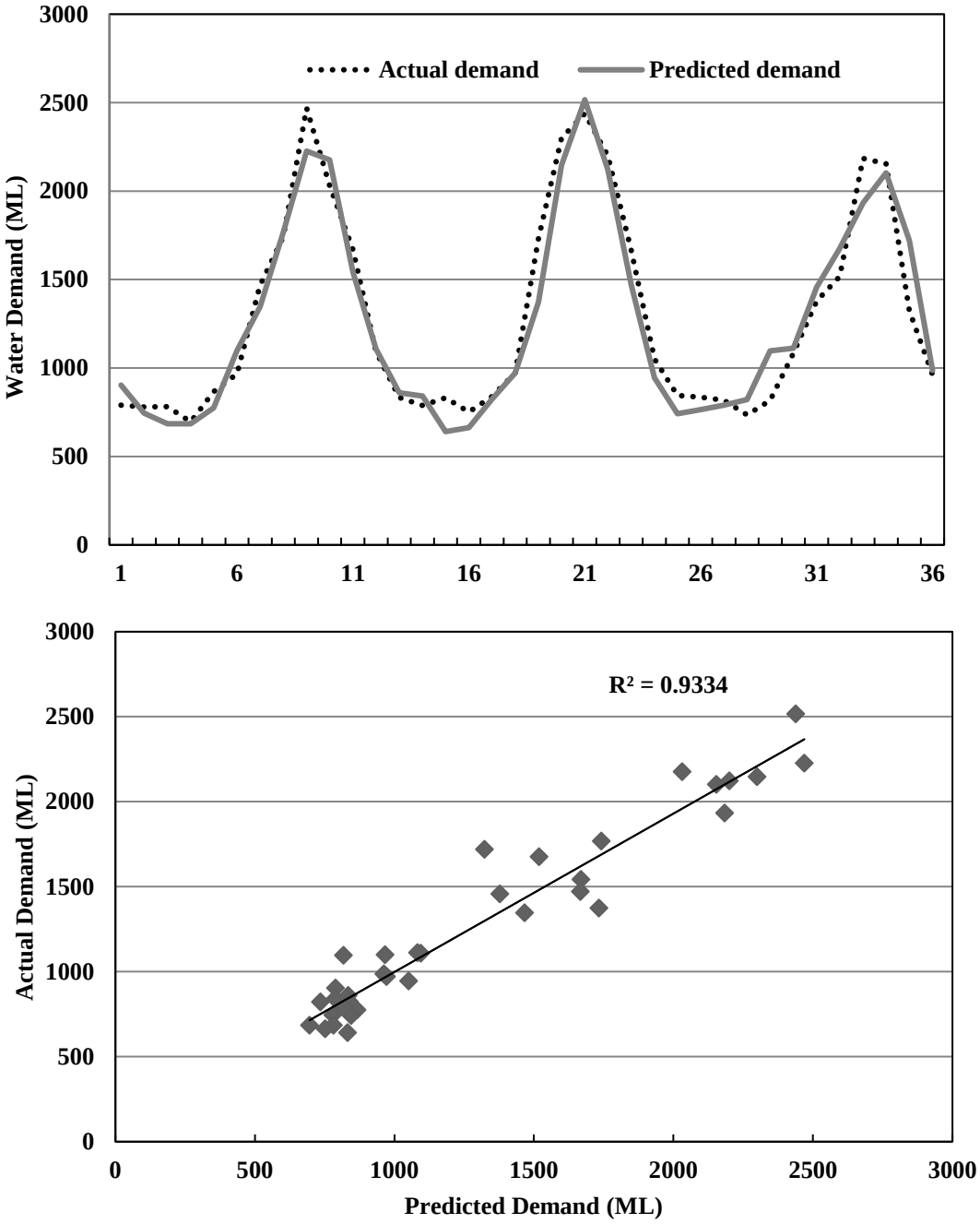


Figure 4-2 Actual demand Vs predicted demand by model τ 1,2,3

4.3 Summary

Time series of weather information and water demand are considered as input variables of water demand forecasting models. Presumably, a given lag time can be associated with such time series that can better explain their natural cycles or behaviour. This research used AMI as a well-known technique to determine the appropriate lag time for water demand, temperature, and rainfall as explanatory variables in the developed models. Following chapter 3[61] , this work focused the optimum lag time of these variables separately. Followed by, a model which considered all given lags up to the optimum value. Results showed that a model which uses extra valuable information as independent time series can outperform models which target individual optimum lag time of these variables. Support vector machines showed their high level of sensitivity to phase space reconstruction of input variables. In general, this research could highlight the importance of design combination of inputs fed into demand forecasting models by employing the concept of optimum lag time determined by average mutual information.

Chapter 5: Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts[4]

5.1 Overview

This chapter proposes a new general approach in short-term water demand forecasting based on a two-stage learning process that couples time-series clustering with gene expression programming (GEP). Since short-term water demand data of Kelowna was not available, this approach was tested on the real-life water demand data of the city of Milan, in Italy as a framework which can be applied to any data. Moreover, multi-scale modeling using a series of head-time (multiple resolutions) was deployed to investigate the optimum temporal resolution under study. Multi-scale modeling was performed based on rearranging hourly based patterns of water demand into 3, 6, 12, and 24 h lead times to investigate the impact of data acquisition frequency in short-term water demand forecast models. Results showed that GEP should receive more attention among the emerging nonlinear modelling techniques if coupled with unsupervised learning algorithms in detailed spherical k-means.

5.2 Background

Water demand forecasting is a key predictive analytic among researchers in the field of water resource management. Due to a shortage of potable water, a lack of access to a mitigation of water resources in semi-arid and arid regions, climate change, and rapid worldwide urbanization [62], public awareness and concerned water authorities have led researchers to come up with novel

[4] A version of this chapter has been published as a journal paper: Shabani S, Candelieri A, Archetti F, Naser G. Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts. Water. 2018 Feb 2;10(2):142.

techniques in this endeavor. Consumer satisfaction is the prime objective of water utility operators. However, population growth has caused infrastructures dealing with a significant amount of stress to meet sustainable states of engineered systems. A key parameter for water distribution systems (WDS) operators in decision making for pumping schedules, storage, treatment, and distribution of water is an accurate and reliable forecast of short-term water demand [63]. Such a forecast can certainly aid the operators with an optimized water supply system that takes a fairly accurate demand of water into account for the near future [2]. Over recent years, there has been a boom in data analytics [64], which has brought to the attention of researchers and engineers that traditional approaches are not enough in designing a WDS for its future state. Predictive models are becoming increasingly useful, since more data are available now than in the past. This popularity is highlighted in the field of water demand even more due to a lack of records on the usage of water in the past. Indeed, the water utilities' archives are relatively poor regarding water demand data with higher sampling rates. Unfortunately, most utilities have readings of water use every few months. Therefore, short-term forecasts of water demand are usually based on the experience of the WDS operators in situations in which SCADA systems are not yet deployed. On the other hand, residential water use is different than actual production due to the high percentage (10–50%) of water loss or leakage through WDS [65]. Indeed, there is a need for a comprehensive assessment of temporal resolutions through time-scale modelling, which can assess different time-scales for short term water demand.

A survey of scholarly research articles in the field of water demand forecasting reveals the novelty of these studies was due to the consideration of temporal resolutions, the type/number of input variables fed into models, and modelling approaches. There is a common convention among water demand forecast modelers that short term forecasts are those targeting temporal resolutions hourly,

daily, or weekly that are used for operational purposes of WDS [2]. Furthermore, other researchers considered temperature, precipitation, and humidity in their analysis [12], [25], [61], [66]–[68]. Most of the models in the literature are data-driven techniques defined around using water demand with a lead time to predict the future demand. This research should be categorized under short-term forecasts due to the temporal resolutions selected. Timescale modelling proposed in this research was aimed to help the decision makers tackle the data acquisition challenges of water utilities in their operations. Not all the water utilities use supervisory control and data acquisition (SCADA) systems to keep track of water use. Thus, this research targeted short term demand forecasting in a long-time span (1 year) for the sake of operation, as well as management. Researchers have used multiple approaches for predictive analytics in the field of water demand forecasts, from very simple projections to the sophisticated data-driven and machine learning models used these days. Early stages of this field show linear regression was adopted by some researchers [18]–[20]; however, due to nonlinear trends in the characteristics of water demand, time series analysis or using the periodic pattern of data has been used by others [12], [28]. Most of recent research uses data-driven techniques with learning algorithms for predictive analysis. Artificial neural networks (ANN) are probably the most popular ones [69]–[71], also most studies are done with slight changes in such models when pre-processing the data or changes in the structures of the defined ANN models. One study combined ANN with the wavelet bootstrapping machine learning approach as a hybrid model to improve performance of the models by pre-processing the data [72]. In another study, performance of ANN was improved through a hybrid approach using the Fourier time series to model the water demand forecast [73]. Another recent study proposes the coupling of the kernel partial least squares-autoregressive moving average with wavelet transformation as a hybrid approach for modeling annual urban water demand [74]. On

the other hand, support vector regression/machine (SVR/SVM)-based models have become increasingly popular recently [32], [33], [75], [76].

Inspired by Darwin's theory of evolution, Ferreira introduced genetic expression programing (GEP), which brings the optimum selection of input variables in regressions/function findings [52]. The *GEP* is used in many engineering disciplines [55], [77], [78], and its operating functions are subjected to a vigorous learning process to find the optimum ones to use in the gene structures. As a tool to perform data-driven or self-learning techniques, GEP has some advantages over the conventional predictive models. GEP defines individual block structures (input variables, response, and function sets) and selects the optimized operating functions and multipliers through the process of learning algorithms. Furthermore, GEP has a built-in sensitivity analysis that selects the most important variables. The ability of GEP to propose a function/equation at the end of analysis is also unique, while most other data-driven techniques are considered as a black-box model that fails to provide a mathematical function. Therefore, this research investigated the performance of GEP in water demand forecast models for short-term time-scales, which has not yet been explored in this field. The increasing use of time series, also due to the adoption of high-frequency sensors and devices, has initiated many research and development attempts in time series data mining. Time series clustering is only a part of the effort in time series data mining research, but it has always aroused great research interest. The data mining approaches with regard to time-series data are often categorized into pattern recognition and clustering, classification, rule discovery, and summarization [79]. This chapter proposed a coupled deployment of a two–stage learning process, with clustering as unsupervised learning followed by GEP as supervised learning process to forecast short-term water demand. The main outcome of this study was:

- Evaluation of GEP as an alternative to other black-box models used in the literature that have not been explored by other researchers in the field of short-term water demand forecasting
- Investigation of coupling time series clustering with GEP in short-term water demand forecasts to reduce the adverse effect of seasonality and holidays/working days on performance of proposed forecast models
- Proposing a suitable sampling frequency for WDS operators through a time-scale modeling process

5.3 Model Development

The input design of the proposed models is based on reaching a broad understanding of the nature of input factors in the data-driven model, the self-interaction of the water demand, and the use of appropriate lag times in demand forecasting models (Figure 5-1). This is labelled as $K_aHT_bOP_c$, in which $a \in [1, 2, 3, 4, 5, 6]$ number of clusters, $b \in [1, 2, 3, 4, 5]$ number of headlines, and $c \in [1, 2, 3]$ mathematical operators.

A total of 90 models (6 clusters × 5 head-times × 3 operators) were created. Through a two-stage spherical k-means clustering, data were divided into six different groups. Five different head-times were used to perform a time-scale modeling to obtain the optimum temporal resolution (1, 3, 6, 12, 24 h). Three types of mathematical operators [OP_1, {+, −, ×}; OP_2, {+, −, ×, ×2, ×3}; OP_3, {+, −, ×, ×2, ×3, √, ex, log, ln}] were used in the developing of the *GEP* models. All of these 90 models were used in partitions of 80% for train and 20% for test sets.

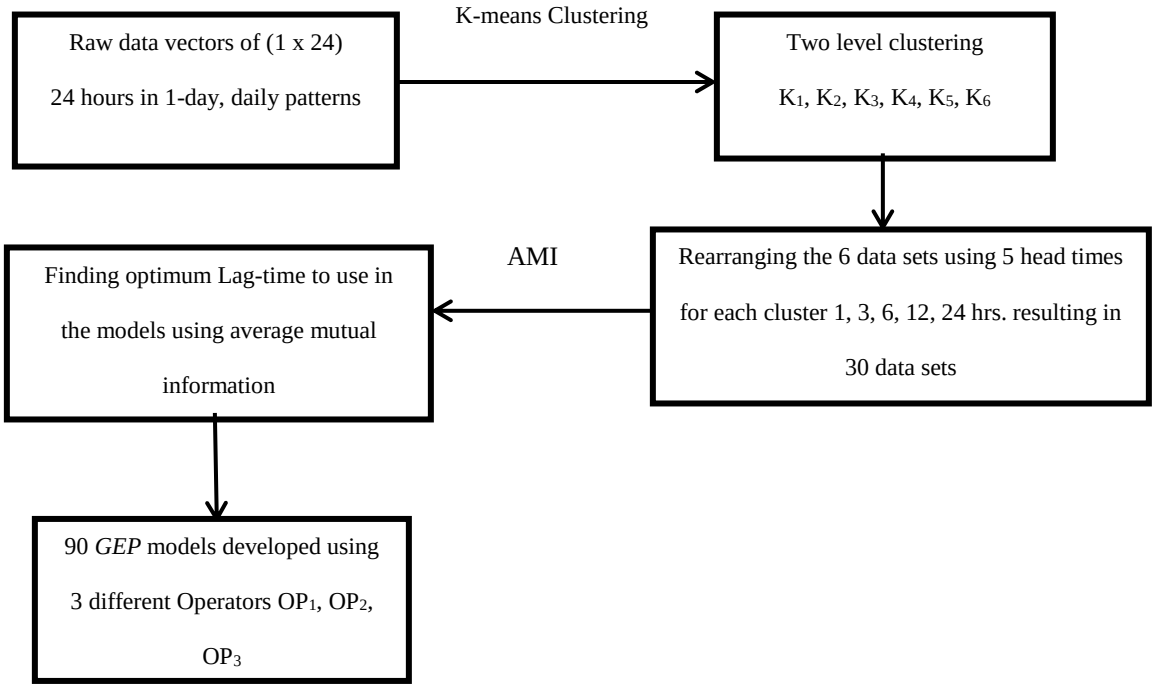


Figure 5-1 Schematic of the proposed approach

5.3.1 Unsupervised Learning: K-Means Clusters

The same data were clustered in a recent study in [80], in which SVM regression was used as forecasting mechanism; therefore, further details on the clustering stage can be found in the related paper. The clustering procedure is briefly summarized here. The time-series used all have the same length: a (1 × 24) vector representing daily water demand. The following picture shows an example of a typical daily water demand pattern selected randomly for illustration.

As Figure 5-2 shows, the nature of water demand is bound to the temporal shifts, which are due to the habits of consumers in a 24 *h* time window. Therefore, triangle similarity (also known as cosine similarity) is used in the clustering algorithm, since it can effectively deal with "similarity in time" (temporal alignment of peaks and bursts) [80]. More specifically, two time series are considered similar in time when they vary in a similar way on each time step. Other possible similarities for

comparing time series are “similarity in shape”, based on the occurrence of trends at different times or similar subpatterns, and “similarity in change”, which identifies two series as similar according to their variations from time step to time step. As water consumption behaviors can be associated with the recurrent occurrence of peaks and bursts at some hours of the day, the triangle similarity, which is a similarity in time measure, was used to compare and cluster water demand time series. Triangle similarity is the cosine value of the angle between two vectors and is computed as:

$$s(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|} \tag{5.1}$$

It is equal to 1 if x and y have the same orientation, 0 if they are orthogonal, and −1 if they have opposite orientations, independently of their magnitude. Since the components of the urban water demand vectors are not negative, triangle similarity varies in [0; 1].

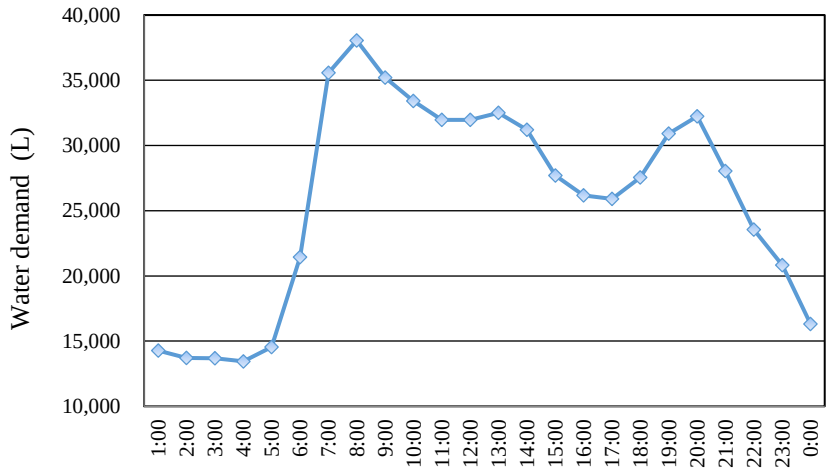


Figure 5-2 An example of time series data representing hourly water consumption in a day (L/h)

The main objective of the clustering phase was to group the days in clusters that can consider the seasonality in water consumption behaviors. To reach the evidence of this seasonality, a two-level

clustering was performed to group the months in the first stage, followed by clustering of the days with the "month clusters".

The unsupervised learning algorithm used in this chapter was K-means, available as "skmeans" among R packages. The following equation shows how the cosine similarity concept is implemented through this algorithm. Calinski-Harabasz and Silhouette [34] proposed methods to measure the validity of the numbers of selected clusters. Therefore, their methods were used to find the optimum number of K in this study. More details on this approach are explained in [80].

$$d(x,y) = 1 - s(x,y) = 1 - \frac{\langle x,y \rangle}{\|x\|\|y\|} \tag{5.2}$$

5.3.2 Average Mutual Information

Performing data-driven models requires selection of an appropriate lag time, as it improves the computational efficiency by adding more valuable information for training, as inputs feeding the GEP models. Traditionally, lag times have been used up to the point where using more lag times would not result in a model's improvement. Some methods like autocorrelation function (ACF) or correlation integral (CI) methods could bring an educated guess about which lag time should be used in the development of such models [46]–[48]. However, we have used average mutual information (AMI), which does not require large data sets, unlike the ACF and CI methods; it has also been widely used in the field of hydro-informatics in recent years [44]. The following equation (5-3) is the method for computing AMI for each one of the data sets designed as input variables of the GEP models. It simply uses the joint probability of two successive time series, as well as the marginal probability of them. It should be noted that it is similar to Shannon's entropy; therefore, it shows less entropy will be in the selection of the optimum lag time based on AMI.

$$I_\tau = \sum_{i=1}^{i=n} P(X_i, X_{i+\tau}) \cdot \log_2 \frac{P(X_i, X_{i+\tau})}{P(X_i) \cdot P(X_{i+\tau})} \quad (5.3)$$

In this equation, the joint probability of two successive time series $P(X(i), X(i+\tau))$ and the product of their individual marginal probability are used to find the appropriate lag time. This delay can contribute to the maximum valuable information added on $X(i)$ by the successive time series $X(i+\tau)$.

5.3.3 Gene Expression Programming

GEP's learning process begins with the random generation of chromosomes for the given raw data/population. The generated populations work with two entities: chromosomes and expression trees. Environment selection or the fitness criteria will evaluate which of the offspring solutions can outclass the others. This repetitive process will eventually deliver a good candidate to be selected. In this study, the general settings of the learning/training algorithms were 30 chromosomes, 8 head size, and 3 numbers of genes as suggested by default conditions of the GenExprotools program. Selection of the head size determined the complexity of each one of the parameters. Each head of the genes was exposed to a variety of arrangements before feeding data into the models. Reproduction of the randomly generated populations could reach the superior model with the optimized stopping condition. Figure A1 (Appendix A) shows the expression or tree diagram of the proposed model. As shown, the model was based on 3 genes (sub-expression tree diagrams) linked together by the addition function. The number of genes used in a chromosome characterized the different layers/blocks building the whole structure. Using a very big gene could result in more accurate models; however, in this study chromosomes are used in smaller units for simpler computation as a limited number of generations were used. The last part of *GEP* modeling is the selection of stopping condition-function. Root mean square of error

(RMSE) was used as the fitness function to fit a curve to target values. The stopping condition was 3000 generations for all the runs or models to have a fair comparative assessment between all input designs.

Logical sequence of steps in function finding through nonlinear regression of GEP is explained below:

- Creating a random initial population of chromosomes
- Expressing chromosomes in a tree diagram with subsets
- Comparing the new offspring solutions based on fitness criteria
- Keeping the best solution, followed by reproduction.

5.4 Study Area and Data Collection

The proposed approach mentioned in the previous section was applied to the urban water demand in Milan, Italy, recorded between 1 October 2012 and 30 September 2013. The WDS in Milan serves drinking water to more than 5000 buildings in this city, which serves a population of approximately 1 million people. Other features of this WDS are listed below:

- 149,639 junctions
- 118,950 pipes
- 26 pumping stations
- 501 wells and well pumps
- 33 storage tanks
- 95 booster pumps

- 36,295 valves
- 602 check valves
- Total base demand 7.5 ± 4.2 (m3/s)

Samples of water demand are recorded through a Supervisory Control and Data Acquisition system (SCADA). The quantity of water pumped into the network by each one of the pumping stations in the city is collected with a sampling rate of 1 sample/min. Collected data are then sent to the centralized SCADA, which sum measures over the different stations to provide the overall water demand of the city. Moreover, SCADA allows for modifying the time scale; more specifically, data used in this study were retrieved from SCADA according to an hourly resolution. The multi-scale analysis of the data was performed by scaling the recorded data to head-time bases of hourly, 3 h, 6 h, 12 h, and 24 h. The scaled data were then prearranged into a time-series dataset $D = \{x_1, x_2, \ldots, x_n\}$ consisting of n vectors, one for each day in the observation period, in which each vector x_i is a set of 24 ordered values for hourly, followed by 8, 4, 2, and 1 for the rest of the scaled data, which are under study for the i-th day.

5.5 Results

The main purpose of clustering in data mining is to illustrate the typical patterns of the trend in residential use of water that is inferred from recurrent peak/burst hours depending only on consumers' habits. This assignment is done without using any information about the data in the learning process (time of the day or week and working/non-working days). Since the trend of water consumption follows a specific pattern shown earlier in this paper, "cosine similarity", known as triangle similarity, was opted for as a similarity index in the clustering process (spherical k-means).

Results of the time-series data clustering (i.e., a two-level clustering approach) allow us to identify 6 typical daily urban water demand patterns (i.e., consumption behaviors), in which the number of clusters was defined according to the best values of the Silhouette and Calinski-Harabasz indices (respectively, 0.74 and 97.87, averaged on the two-level clustering) obtained by varying the possible number of clusters from 2 to 24. More details regarding the detailed results of clustering can be found in [33]. The following figure 5-3 shows a calendar with the cluster assignment for every day of the observation period. This kind of visualization makes seasonality and cyclic behaviors more evident.

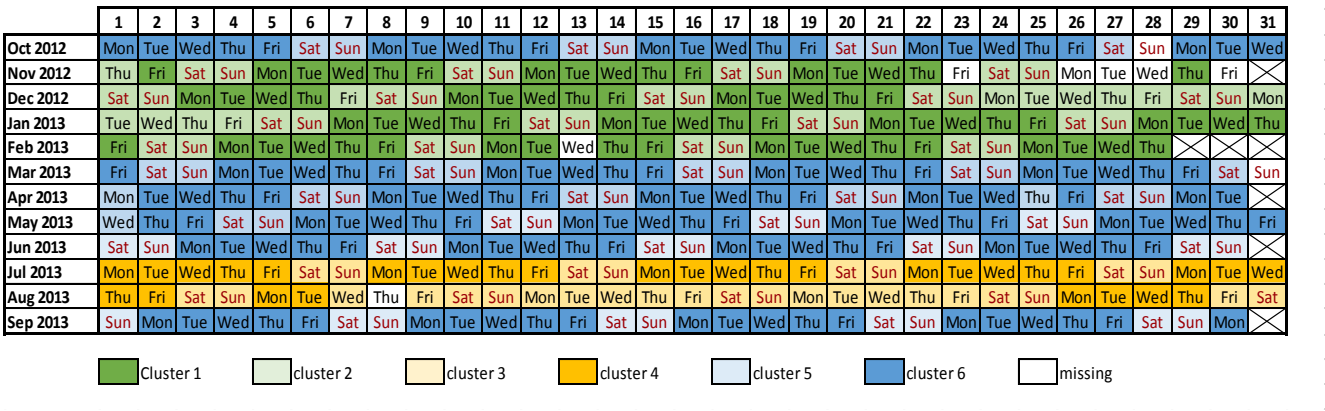


Figure 5-3 Cluster assignments

Looking into the 3 clusters of the first stage, one can identify these clusters as (1) Months of November, December, January, February, and March, which correspond to Fall and Winter; (2) Months of April, May, June and September, and October; and (3) July and August, which correspond to the period of largest consumption during summer holidays. The second level of clustering is to target the working/non-working days, which shows how holidays can affect the consumption behavior of water demand.

Centroids of the 6 prototypes of daily water demand are shown in figure 5-4; looking into this figure; one can easily recognize the differences in the peaks of the mornings and evenings that

differentiate these 6 clustered prototypes. To be more precise, we can capture a meaningful definition that shows the peak in the mornings of holidays and weekends is always delayed by approximately 1 h compared to that of working days for each period of the year.

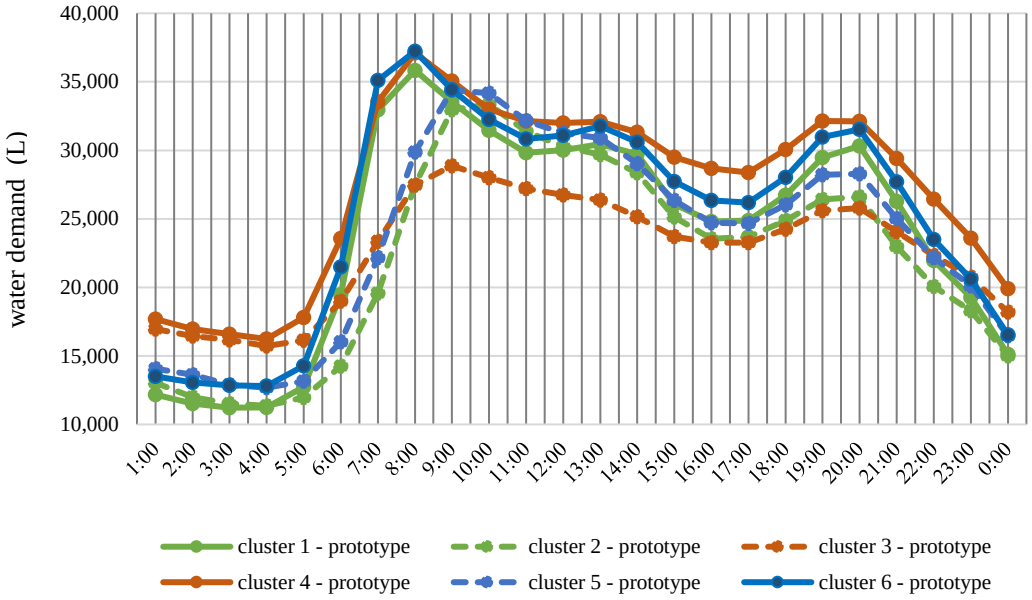


Figure 5-4 The k = 6 typical water demand patterns identified through the two-level clustering procedure

Results of the AMI code applied to the rescaled time-series for each head-time under study are shown in Figure 5-5. In this bar chart, the X-axis represents all 30 models (6 clusters × 5 head-times) labeled accordingly. Y-axis is the computed AMI value, which is discussed earlier in this paper through equation 2. These values were used to define the data-driven GEP models in this study. It is important to note that the AMI values make sense in that they are always at their maximum value when the head-time is 1 h, except for K_4, in which the maximum is in head-time 3. This exception is because the pattern of consumer behaviors shifts temporally. Moreover, more data is acquired when time-scale is on an hourly basis.

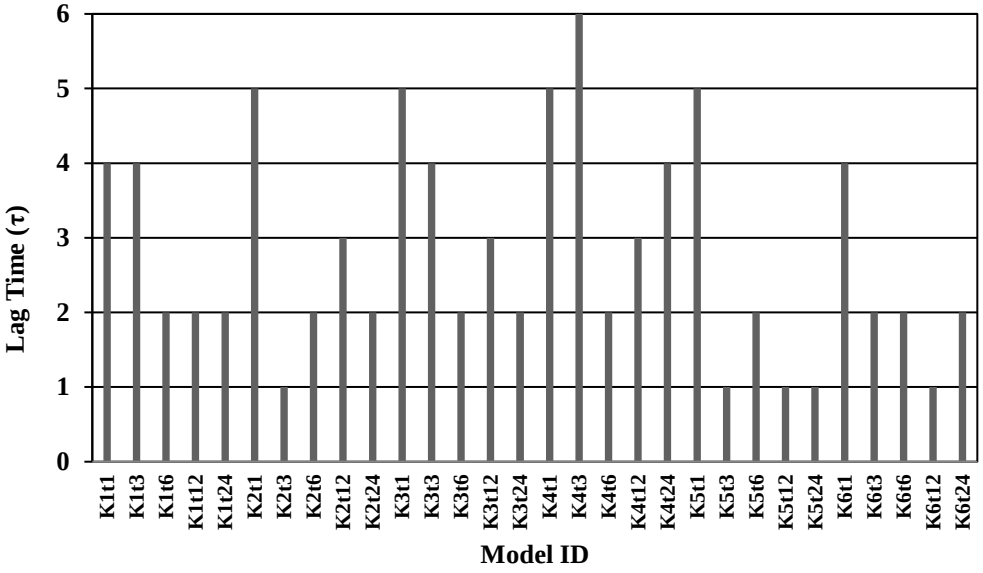


Figure 5-5 Average mutual information for all input designs k1-6 clusters t1,3,6,12,24 head times

Table 5-1 shows the performance of GEP models in the testing period separated by the performance indices used (MAE, RMSE, R^2, and MAPE). It is important to note that these data are averaged on each cluster since the overall performance is what matters to predictive modelers. However, the detailed results are shown in Table A1 (Appendix A) for further investigation of these performances. Results showed the hourly-based models could significantly outperform the other sampling rate/frequencies. It was expected that hourly-based models could outperform other resolutions because data-driven methods perform better with larger datasets. It is known that aggregating the data temporally can improve forecasting due to the temporal averaging which reduces the noise; however, since averaging was not the focus of this study, higher head times did not improve the models. Another valid reason is "oversampling", which can increase the resolution of the data; consequently, there would be a better definition of the trend in the time series. Moreover, it is known that temporal aggregations can lead to loss of valuable information about

the primary data process. MAE measures the mean absolute error between prediction and actual values. The GEP operators performed similarly, as MAE values are very close to each other (MAE of 0.2319 ± 0.0391 being the best using 2nd GEP operator Table 5-1a). The same story was repeated for RMSE, as 2nd GEP operator outperformed other models with RMSE of 0.3048 ± 0.0632 Table 5-1b. The third GEP operator was the best among all R^2, with the highest value of 0.8962 ± 0.0569 averaged on clusters. Interestingly, the performance of 6-h head-time models was significantly better than 3-h models. This improved performance might be because the behavior of consumer's demand or the temporal shifts of the consumption can be better defined with a frequency of 1 sample per 6 h rather than 3 h. Optimum resolution of the data might be better represented using this time-scale. Moreover, it is known that the performance of these models deteriorated with less frequent data for 12 and 24 h due to much smaller data sets within these time-scales. Results showed that the mathematical operations used in GEP operators do not play a very significant role; however, since these data are not linear, the second {+, −, ×, ×2, ×3} and third operators {+, −, ×, ×2, ×3, √, ex, log, ln} consistently outperformed the first one {+, −, ×} in all simulations. MAPE (minimum absolute percentage error) is used just for comparing testing sets since it provides a relative magnitude regarding percent values. Like other performances indices, the 1 h head time models could outperform others with a value of 0.8850 ± 0.0089 averaged on clusters.

$K_3HT_1OP_1$ is further investigated as one of the superior models to show how GEP models could perform when hourly data is acquired. Table A1 (Appendix A) shows $K_3HT_1OP_1$ is selected with the highest values of R^2, 0.95, and 0.93 for training and testing data set accordingly. A model that is neither over nor under-trained should have similar performances in test and train periods, an important point overlooked by many scholars in our area. On the other hand, MAE and RMSE

values were also the lowest compared to other models with 0.19 and 0.24 for the training set, followed by 0.18 and 0.23 for testing set. Figure A2 illustrates how close the prediction and actual demand are for this particular model. MAE and RMSE should have the same unit as the estimated quantity (water demand (L)); however, all values were scaled between 0 and 1 for fair comparative assessment between the head times under study.

Table 5-1 Performance indices averaged clusters: (a) MAE, (b) RMSE, (c) R2, and (d) MAPE%

(a). *MAE* * (Mean ± Standard Deviation)					
Operator	**Head Time = 1**	**Head Time = 3**	**Head Time = 6**	**Head Time = 12**	**Head Time = 24**
GEP-operator_1	0.240 ± 0.0388	0.7497 ± 0.4449	0.5943 ± 0.3548	0.5399 ± 0.2932	0.6163 ± 0.3039
GEP-operator_2	0.2319 ± 0.0391	0.7448 ± 0.3885	0.4386 ± 0.1096	0.5061 ± 0.1660	0.6242 ± 0.3952
GEP-operator_3	0.2387 ± 0.0444	0.6031 ± 0.2228	0.4854 ± 0.2857	0.5276 ± 0.2211	0.6433 ± 0.3808
(b). *RMSE* * (Mean ± Standard Deviation)					
Operator	**Head Time = 1**	**Head Time = 3**	**Head Time = 6**	**Head Time = 12**	**Head Time = 24**
GEP-operator_1	0.3087 ± 0.0563	0.7861 ± 0.3763	0.7731 ± 0.5048	0.6896 ± 0.3230	0.8272 ± 0.5226
GEP-operator_2	0.3048 ± 0.0632	0.8595 ± 0.3398	0.5274 ± 0.1483	0.6215 ± 0.2052	0.8401 ± 0.5591
GEP-operator_3	0.3116 ± 0.0842	0.7627 ± 0.3338	0.6104 ± 0.3661	0.6718 ± 0.2800	0.8149 ± 0.5710
(c). R^2 (Mean ± Standard Deviation)					
Operator	**Head Time = 1**	**Head Time = 3**	**Head Time = 6**	**Head Time = 12**	**Head Time = 24**
GEP-operator_1	0.8900 ± 0.0498	0.4455 ± 0.1681	0.6221 ± 0.2732	0.4227 ± 0.3275	0.3174 ± 0.2151
GEP-operator_2	0.8906 ± 0.0491	0.4332 ± 0.1593	0.6776 ± 0.2314	0.5131 ± 0.2665	0.2229 ± 0.2288
GEP-operator_3	0.8962 ± 0.0569	0.5077 ± 0.2248	0.6551 ± 0.3585	0.4395 ± 0.3204	0.2727 ± 0.2255
(d). *MAPE* % (Mean ± Standard Deviation)					
Operator	**Head Time = 1**	**Head Time = 3**	**Head Time = 6**	**Head Time = 12**	**Head Time = 24**
GEP-operator_1	0.900 ± 0.0113	1.2067 ± 0.0080	2.1450± 0.0133	1.4267 ± 0.0098	2.0667 ± 0.0106
GEP-operator_2	0.9400 ± 0.0110	1.4367 ±0.0113	2.3933± 0.0090	1.5950 ± 0.0012	2.0683 ± 0.0101
GEP-operator_3	0.8850 ± 0.0089	1.2033 ± 0.0115	2.1867 ±0.0084	1.4667 ± 0.0091	2.0917 ± 0.0010

5.6 Summary

The prime objective of this paper was to propose a coupled deployment of supervised and unsupervised learning in short-term water demand forecast models. In the proposed two-stage approach, spherical k-means clustering was used to organize daily water demand patterns into six different clusters based on the computed Silhouette and Calinski-Harabasz indices. Gene

expression programming was further used, as our supervised learning part of the approach, to model these six clusters separately. The measurement of errors in this paper was done using four performance indices on both training and testing data sets. MAE, RMSE, R^2, and MAPE are the common methods of error measurement in the field of hydro-informatics, and they are widely adopted by researchers. Results of this study showed GEP should receive more attention in this area due to the highly accurate predictions it can provide while coupled with unsupervised learning algorithms. It is not a black-box model like most of the proposed ANN or SVM models; therefore, meaningful internal relations within the input water demand will be provided, as well as a mathematical equation (through nonlinear regression) to be used by operators of WDS (Appendix A, Figure A1 and Python code in appendix detailed equation). The proposed approach could have a profound impact on the operations of water utilities, as well as on managerial decisions. The frequency of the collected data is a major decision that is used to plan for the next hour, week, month or even year. The seasonality of water demand patterns is not unestablished; however, the two-level clustering provided (in a completely unsupervised and data-driven paradigm) six groups of data that are not usually used in classified data of predictive models in this field. The positive impacts of this approach are a better understanding of how one can utilize the time series of water demand in short-term forecasting through a completely data-driven technique with a fair understanding of how to opt for the suitable temporal resolution, the proper lag time of the feeding inputs to the models, and an equation/function that can help the operators to use a wide range of models based on their desired duration or the time of year.

Chapter 6: Data Augmentation Practices in Long-Term Water Demand Forecasting Models: A Flexible Approach[5]

6.1 Overview

An extensive number of predictive analytics methods have been explored by researchers in water demand forecasting on short, intermediate, and long-term evolution of this valuable natural resource. However, the inevitable uncertainty dealing with the future state of water demand driving factors and associated risks of black-box models has highlighted the growing necessity for considering the sources of these uncertainties in the design process of water distribution systems (*WDS*). Demand analysis has been mainly dependent on such forecast models. This research proposed a novel method for inserting flexibility concept into demand analysis of *WDS*. The applied predictive analytics of water demand consisted of average monthly values of both hydrological factors (water demand, temperature, humidity, and rainfall) – and demographic variables (population and hotel occupancy factors) for 15 years (1996-2010) in Kelowna, Canada. The bounds and limits of well-known hydroinformatics modeling techniques, artificial neural networks (*ANN*), support vector machines (*SVM*), and gene expression programming (*GEP*) were computationally tested using fundamental deep learning practices like data augmentation (data cropping and distorted data), and information theory. The results indicated that *SVM* could be considered as the most flexible algorithm in hydroinformatics since it is less sensetive to data augmentation practices. Moreover, the prime objective of this chapter was met by proposing a

[5] A version of this chapter has been submitted to *Water* journal and is currently under peer review, Manuscript ID: water-335759

flexible range of demand (358 *ML* for upper bound and 335 *ML* for lower bound) which takes many sources of uncertainties into account.

6.2 Background

Once a cheap commodity for development, water is now considered as blue gold for managers of water utilities and governments due to the increased awareness after recent global urbanization (mostly agricultural). Therefore, the growing necessity for accurate water demand forecasts has led researchers to try a wide range of predictive analytics techniques in this endeavor. However, before designing of WDS, the predicted values through such forecast models are subject to change by a safety factor defined by municipalities. These current safety factors are highly uncertain based on engineering intuition. Traditionally, engineers have been designing WDSs for a specific anticipated future with deterministic assumptions which is a projection of historical water demand trends and population. Therefore, the traditional approach can lead to unenviable over-sized/under-designed systems. Valuable assets/infrastructures such as WDS should be designed for a long lifespan that can serve future demands. The critical part in long-term decision making of planning for developments in water systems is the fact that the bases for these decisions are highly subjected to future changes. Therefore, an attempt to anticipate these uncertainties is very much needed in WDS design. For a better understanding of the dilemma, one can refer to this interesting quote - "We need to recognize the limits to human foresight. We need to recognize that forecasts are always wrong and that our future is inevitably uncertain. We thus need to look at a wide range of possible futures and design our projects to deal effectively with these scenarios." [5]. Basic design criteria for engineers to follow can be illustrated through a flowchart shown in Figure 1-1 (chapter 1). As shown in this figure, water demand analysis is the backbone of WDS design. To insert the flexibility concept in *WDS* design, a flexible educated forecast of water demand is an essential

need. Flexibility can be defined as the ability to handle eventualities and unplanned happenings. Flexible demand analysis is one that can propose a range of demand (instead of a deterministic approach) which considers the sources of uncertainties in their models.

Long-term water demand forecasting models in practice can be categorized into five main approaches [81]: 1) Temporal extrapolation models, 2) Unit fixture method, 3) End-use models 4) Estimations based on projecting land use, and 5) Multivariate statistical models. First, temporal extrapolation methods are historically, the projection of water demand for a deterministic future using time-series analysis [22], [82], [83]. The major drawback of these models is its assumption that considers initial conditions would not change under the lifetime of the design period. Second, restricted to a number of users, unit fixture unit method is often used by WDS designers to determine peak demands based on the type of habitats (per consumers, per employee, per student, per unit of the occupants). This would result in a significant overestimation of actual demand [1]. Third, end-use models which consider the facility type of the users and the reason for consumption [84]–[87]. These models are mainly used for residential urban water forecast models; however, there would usually be a need for a multivariate statistical model to accompany the end-use models for an effective long-term forecast, for instance: ANN model [88]. Fourth, geographic information systems (GIS) has been enabled water authorities to use real-time urbanization information to determine future water demand. However, this method needs very detailed documentation of municipality information coupled with GIS maps [81]. Finally, one could claim that multivariate statistical models and emerging soft computing techniques in hydroinformatics are taking over traditional methods, especially among researchers in this field.

WDS is a very complex dynamic infrastructure driven by many socio-economic, political, and hydrological factors; therefore, there has always been a huge interest among researchers to come

up with sophisticated statistical models to propose more accurate water demand forecast models. Multivariate linear regression models have been used in the early stages of water demand forecasting models [19], [20]. The linearity assumption behind this technique is a major disadvantage, especially in water demand which has highly non-linear trends in its global drivers. Therefore, researchers had to come up with more sophisticated models able to deal with high level of complexity and nonlinearity within data.

An emerging field has been introduced as hydroinformatics which can be considered as a subdivision of informatics using water data along with artificial intelligence techniques such as ANN, SVM, and GEP. ANN is a global approximation method which performs well in forecasting complex engineering systems' future states [23]. ANN is the most popular method in water demand forecast models [2], [34], [69], [89]. Although ANN is considered as the superior technique in forecast models, it is a black-box model which fails to understand the internal relationships between the determinants of the response variable. SVR/SVM algorithms have also been emerging as one strong alternative to *ANN* models for predictive analytics of water demand over the recent years[32], [61], [90]–[92]. Being originally used as a classifier, *SVR* models use only a subset of the training examples for prediction. GEP models have also been used recently in hydrology [53]–[55], and in water demand forecast models (chapters 3,4) [61], [75]. All these studies have neglected the concept of flexibility in water demand design in their forecasts. Figure 6-1 shows a concept of flexible demand forecasting as defined in this chapter. Long-term demand forecasts use average weekly, monthly, or yearly values. However, this figure shows there is an upper and lower bound for yearly water demand over a certain horizon. A flexible long-term forecast model must be able to consider the sources of uncertainties to propose a range of demand or upper/lower bound as well as an average to be used for design purposes.

Flexibility theory has been the focus of many scholars in different engineering fields. But, not many people defined it for WDS designs. Shah et al. [93] characterize flexibility as 'the ability of a system to respond to potential internal or external changes affecting its value delivery, in a timely and cost-effective manner.' Another definition by Fricke and Schulz [3] is system's ability to be changed easily. External changes have to be implemented to cope with changing environments. One study on the flexibility of *WDS* design is defined through a multi-objective approach by Monto Carlo sampling technique followed by decision trees in multiple interventions considering minimization of cost and maximizing resilience index [94]. In another study, flexibility in the capacity of delivery was solved through a multi-objective stochastic optimization problem [95]. There is a lack of clear definition for flexibility in complex systems like WDS, especially for demand design. Therefore, this research will only try to use this concept in demand assessment before hydraulic design of WDS to propose an efficient long-term demand forecasting of water demand that brings an educated guess by considering climate change, and other global drivers of water demand as major sources of uncertainties. A novel flexible demand analysis for WDS was defined in this chapter. This flexible system is defined to fulfill the hydraulic design conditions for both upper and lower band of the proposed range of water demand. Secured and consistent supply of potable water to consumers is the primary goal behind a WDS design. This supply should meet the sufficient desired flow and pressure during operation.

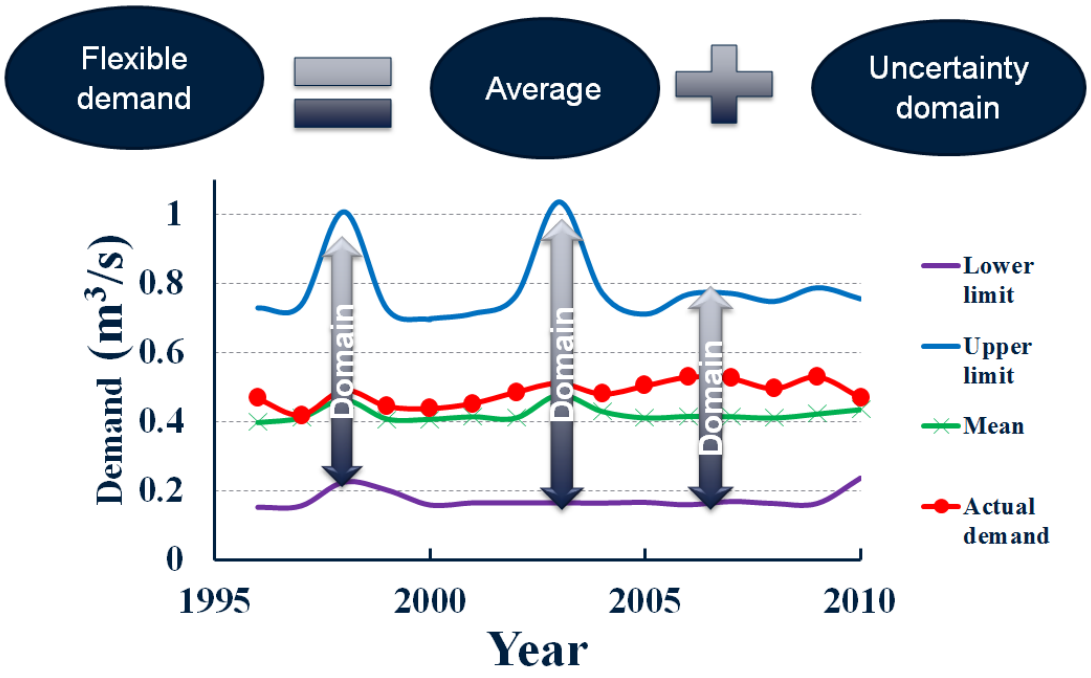


Figure 6-1 Concept of flexibility in water demand

6.3 Research Methodology

6.3.1 Input Variable Design

This study used population (*P*) and hotel occupancy factor (*HOR*) as demographic variables (as Kelowna is considered a hot spot for tourism in North America, and temperature (*T*) in °C, relative humidity (*RH*) in percentage, and rainfall (*R*) in millimetre as hydrological variables to predict water demand (*D*) in millions of litters (*ML*). Average mutual information was used to determine the appropriate lag time (3 months) for feeding the models as shown in the previous chapter. Therefore, the resulting input variable design considered $D_{t-1}, D_{t-2}, D_{t-3}, T_{t-1}, T_{t-2}, T_{t-3}, R_{t-1}, R_{t-2}, R_{t-3}$ $RH_{t-1}, RH_{t-2}, RH_{t-3}, P, HOR$ as inputs to predict (*D*) at present time as a dynamic model. The analysis of water demand was further

subjected to some fundamental data augmentation methods borrowed from deep learning practices to determine the bounds and limits of the modelling techniques.

6.3.2 Data augmentation practices

Constant improvement of engineering structures and digitization of the society has provided sufficient big data for applications of deep learning. Andrew NG [96] introduced data augmentation practices for improvement of deep learning in computer vision. Random cropping of data, mirroring, distortion are major methods of data augmentation. Although these methods are applied on pixels of image data in computer vision, they can be applied in time-series data if deep learning of these models is of interest. Data augmentation is mainly creating new sets of training example, therefore shuffling training examples have also been used based on information theory which says "Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred" [97]. Since reaching an uncertainty domain was the prime objective of this paper, time-series data of input variables have gone through this deep learning practice to create a range of forecasted demand.

- **Clustering (K-means):** 6 prototypes or unique clusters of annual water consumption were the outcomes of an unsupervised learning algorithm (spherical k-means clustering using Cosine similarity as distance metrics) in this study. The same algorithm in the previous work of the authors of this paper was used to perform clustering [98]. To simulate data cropping on the time-series, the 6 unique clusters were used only in test period leaving the training procedure with the other 5 clusters.
- **Shuffling training examples:** 5 new randomly shuffled training sets were generated using random permutation algorithm in MATLAB. The important aspect of this shuffling was

ensuring that feature vectors come in agreement with their proper labels after this random permutation.

- **Extreme climatic conditions:** following the idea of data cropping, the highest and lowest 20% of extreme weather conditions were cropped out of the time-series of the reconstructed phase space of the model and placed into test period. This cropping has been done to make sure the uncertainty sources of sudden changes in weather conditions are considered in this novel approach.
- **Distorted data:** blurring of images or changes in the color combinations of *RGB* (red, green blue) has been used in computer vision. Having a time-series data, the addition of white Gaussian noise on the raw data seemed a reasonable practice. However, in this approach models were only developed as completely data-driven depending only on water demand. The *AWGN* algorithm in MATLAB was used to perform data distortion technique.

6.3.3 Artificial Neural Networks

ANN is a powerful data-driven technique, which employs the structure of a human's brain to compute complex relationships. It is shaped based on an interconnected network of the three layers of input, hidden, and output used in ANN. In a typical 3-layer feed forward ANN (Figure 6-2), the input layer contained the independent variables under the study. The hidden layer contained several neurons simulating algebraic functions, which are set with defined boundary values for training [99]. Feedforward back propagation was used as the most popular learning algorithm to distribute the error within the layers of ANN. The input information was processed through signals passing the neurons. Once output was calculated, the difference between the calculated and actual target was used for backward signals. This training process was continued until certain conditions

(error or epochs as cycles) were satisfied. The number of neurons was changed from 1 to 10 neurons with a single hidden layer resulting in 10 different models. There was a trade for having either high R^2 value or flexible model since a high number of neurons result in over-fitting of data and longer simulation time whereas low number of neurons results in a less rigid algorithm [100]. Hyperbolic tangent sigmoid was used as a transfer function to modify the signals connecting neurons to other layers, setting modified weights. Moreover, Levenverg-Marquardt algorithm was used for minimizing the nonlinear function in feed forward backpropagation process.

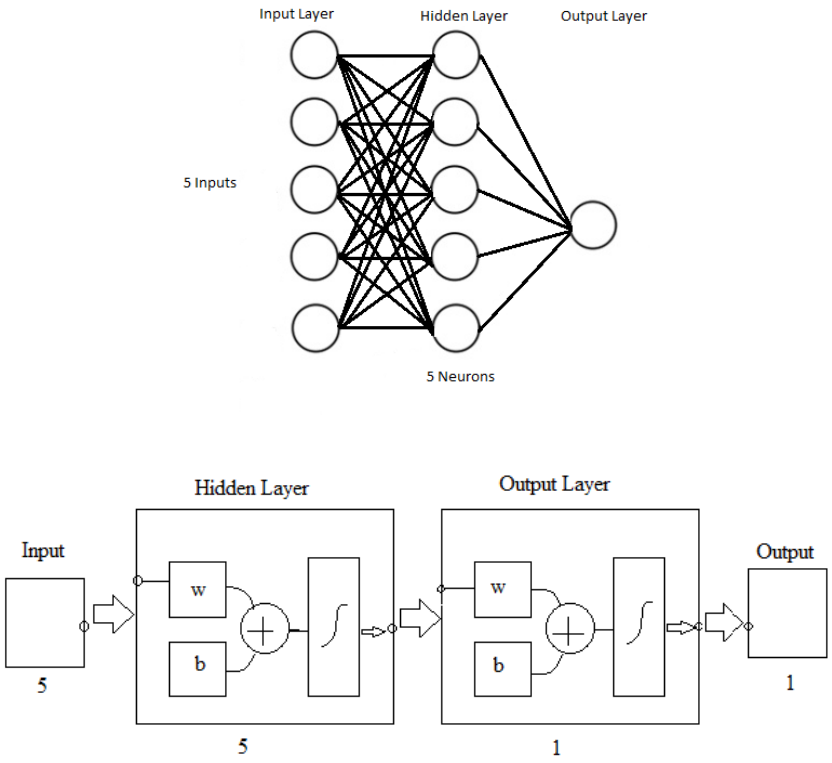


Figure 6-2 Example of a neural network model structure

6.4 Results and Discussion

The results of pre-processing methods based on deep learning practices are reported in this section as well as the final decision on the flexible range of water demand based on the modeling techniques used in this study.

6.4.1 Unique clusters in testing

One of the ideas in this study was using unique prototypes of yearly water demand as an unknown future or test period for the developed models. Yearly water demand pattern with the monthly resolution was studied for the observation period that is 15 years (1996-2010). Spherical k-means (i.e., k-means using cosine distance) was applied for k ranging from 2 to 10, and the two following validity measures were used to identify the best value of k: Calinski-Harabasz (*CH*) and Silhouette (*Sil*). Calinski-Harabasz is an internal clustering criterion. There is no "acceptable" cut-off value to use; the common indication is that the higher the value, the "better" is the solution. A solution (i.e., k value) giving a peak or at least abrupt elbow in the line-plot of *CH* values should be chosen. If, on the contrary, the line-plot is smooth - horizontal or ascending or descending – just like the preliminary results of this study, then there is no reason to prefer one solution to others. Thus, the Silhouette is the only index which supports us – in this case – in defining the most suitable value of k. K=6 offers the highest Silhouette value (averaged on the clusters) (figure 6-3). The result of the clustering process is summarized in figure 6-4 which reports the 6 prototypes identified (i.e., centroids of the 6 clusters), representing 6 different typical yearly water demand patterns. The final results of cluster assignments are further shown in Table 6-1.

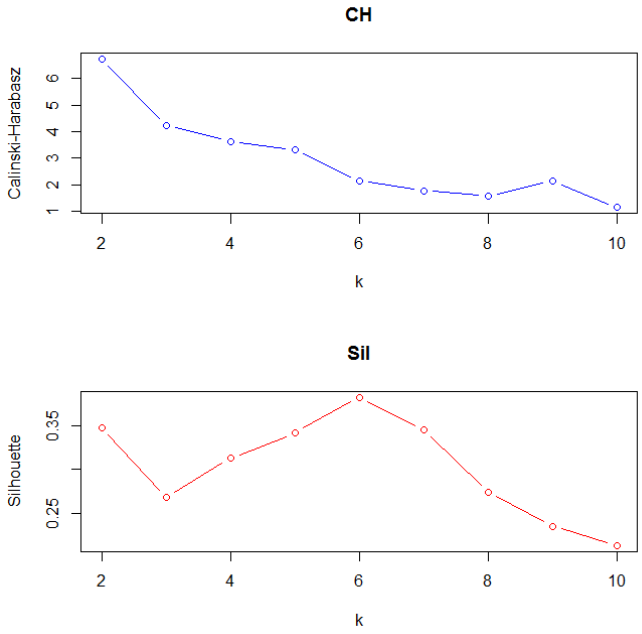


Figure 6-3 Calinsky and Sillouhte factors values for K= 2-10

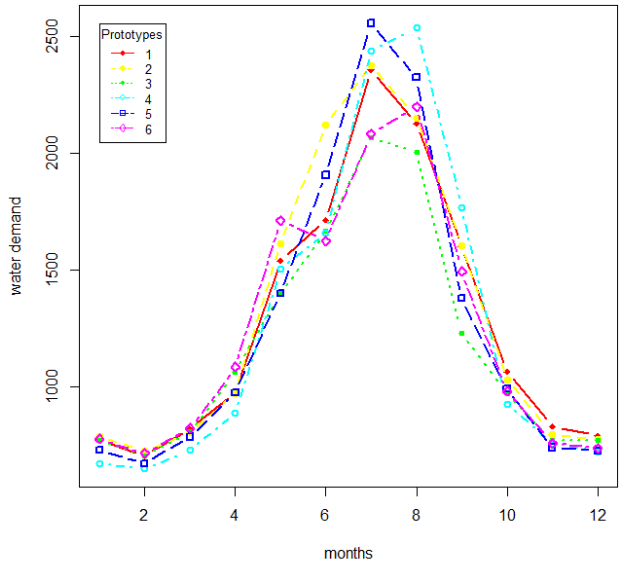


Figure 6-4 Six unique clusters or prototypes identified for daily yearly water demand

Table 6-1 Cluster assignments to different years

year\month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1996	720	666	793	990	1303	1857	2517	2241	1185	852	705	674
1997	693	631	715	951	1424	1455	1762	1982	1216	856	683	662
1998	675	651	731	890	1503	1661	2439	2536	1769	926	757	723
1999	733	674	787	1050	1376	1712	1880	1883	1251	952	769	750
2000	756	697	790	943	1333	1634	1923	2032	1127	948	725	710
2001	725	665	769	923	1482	1513	2095	1968	1449	918	818	752
2002	750	698	773	976	1493	1945	2314	2104	1546	1012	748	723
2003	743	678	781	965	1498	1957	2594	2407	1573	1133	777	782
2004	785	721	840	1184	1517	1803	2280	1948	1214	1020	811	794
2005	789	758	859	1134	1782	1572	2230	2329	1556	1067	806	775
2006	821	753	831	1034	1669	1886	2505	2382	1690	1186	845	836
2007	853	761	899	1170	1928	1847	2258	2288	1708	1024	789	779
2008	781	696	866	967	1467	1742	2469	2031	1669	1094	834	788
2009	832	752	839	972	1733	2300	2438	2200	1667	1051	845	835
2010	819	735	817	1083	1378	1518	2184	2154	1323	962	794	836

Performance of GEP, SVM, and ANN models in test periods versus the actual water demand are compared through figures (6-5 to 6-10). The Green dashed lines representing the SVM models show how this prediction technique was close to actual water demand almost in all clusters. Comparing GEP and ANN these models based on performance indices R^2 and RMSE would give a fair comparative assessment of them. Figure 6-11 is a bar chart showing the average of these indices to compare their performance both in train and test periods. SVM had the best average R^2 in training with a high value of 96%. ANN and GEP also were great in training close to 90% R^2. However, in the test period SVM had the lowest value of R^2= 40% compared to GEP and ANN which had similar performances close to 84%. Since R^2 is not sufficient for a comparative assessment (lack of capability to reflect outliers), RMSE was used as a performance index as well. SVM models outperformed GEP and ANN with an average RMSE value of 179 ML (lowest

compared to ANN and GEP) in the test period. However, figure 6-11 shows RMSE values are close for three of them in the training period.

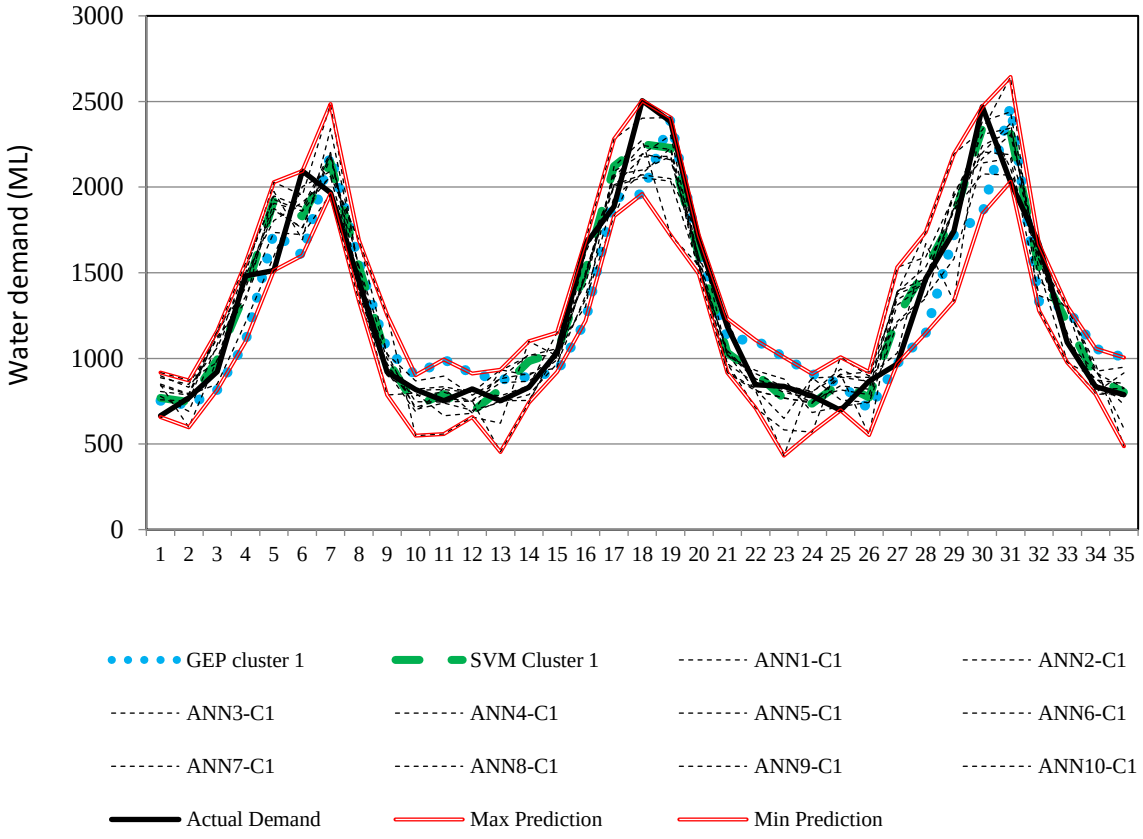


Figure 6-5 Cluster 1 in test period

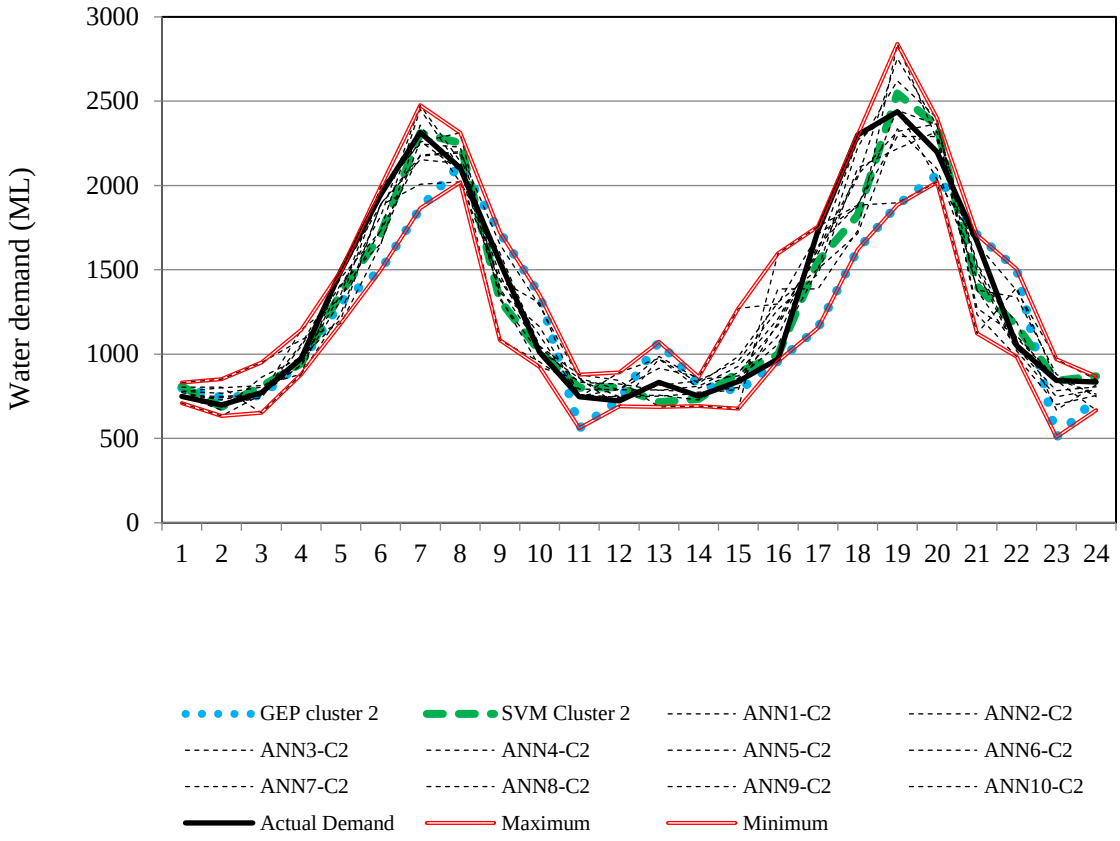


Figure 6-6 Cluster 2 in the test period

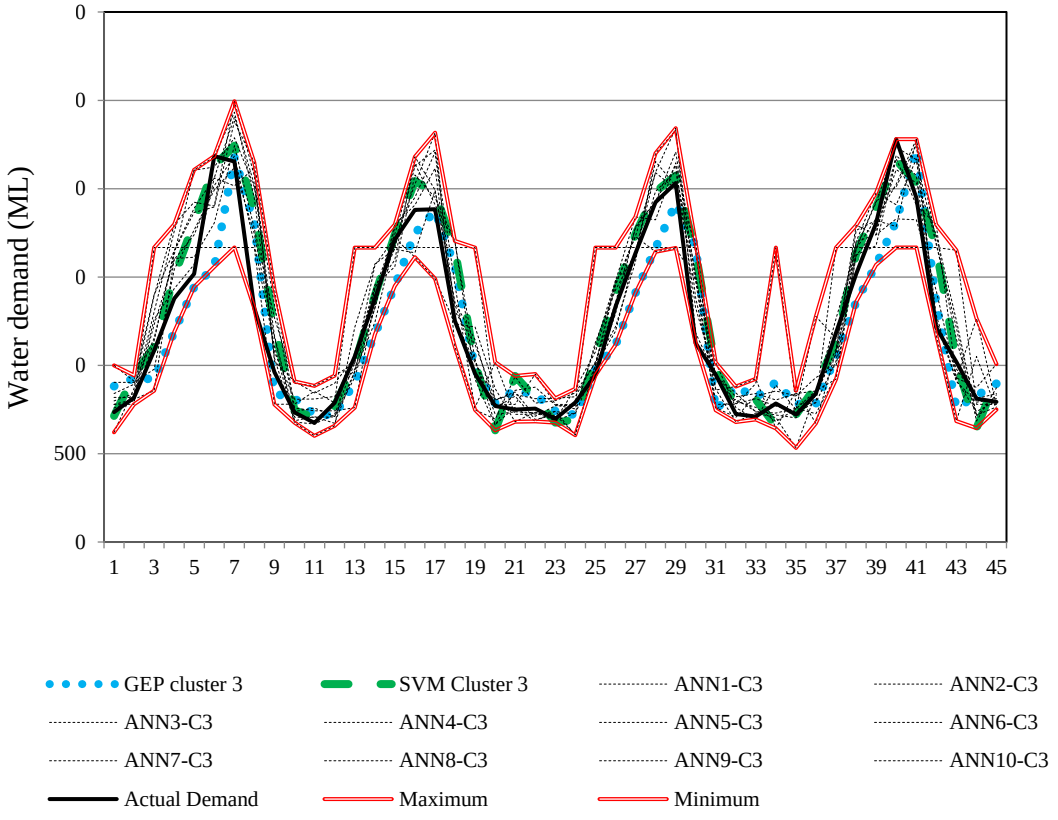


Figure 6-7 Cluster 3 in the test period

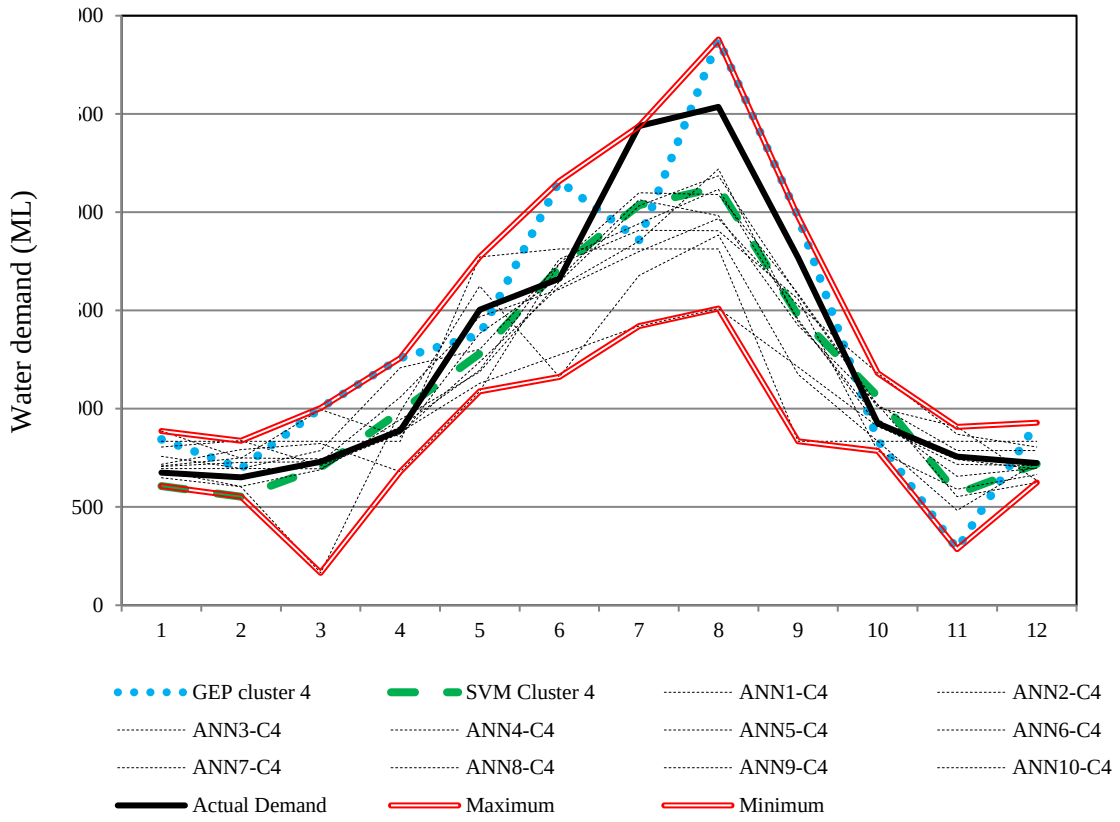


Figure 6-8 Cluster 4 in the test period

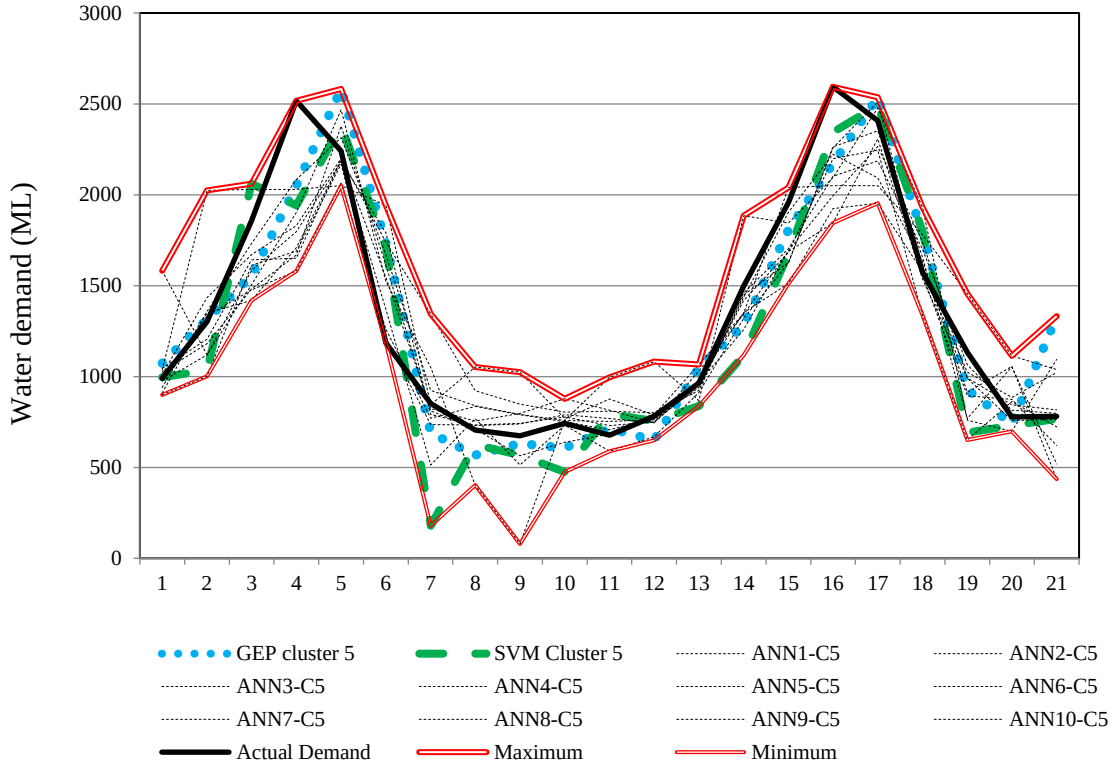


Figure 6-9 Cluster 5 in the test period

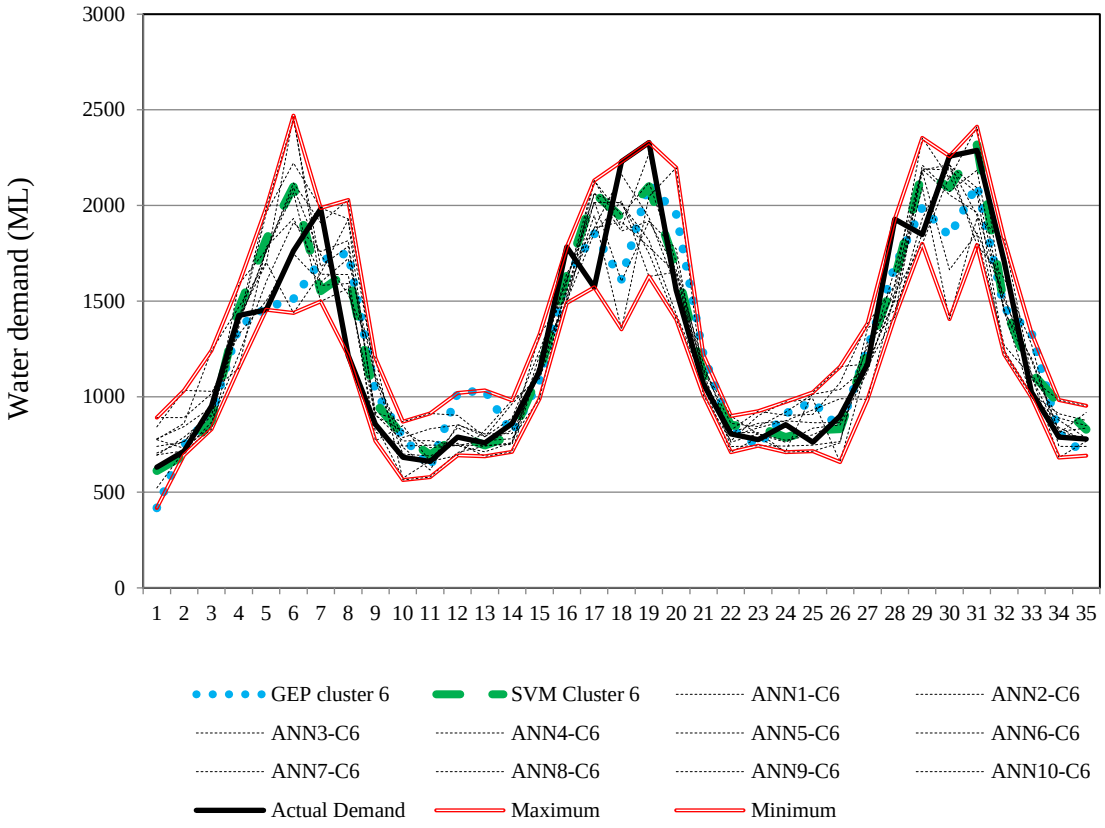


Figure 6-10 Cluster 6 in the test period

Figure 6-11 Comparing models based on unique clusters in the test period

6.4.2 Shuffled training examples

A simple random permutation was used to shuffle each row of training examples in MATLAB based on information theory. This random shuffling was performed to check what would be the response of predictive analysis if sudden changes happen in initial conditions. Figure 6-12 shows the 5 GEP models in dashed blue lines, 1 SVM model in dashed green, and 50 ANN models in dashed black lines compared to solid black showing the actual demand. This figure shows how SVM is close to actual demand. The reason SVM is only represented as one model is the fact that this technique gave similar results both in train and test periods regardless of how training examples are ordered. Contrary to the performance of ANN and GEP, the underlying SVM algorithm used for regression always converges to a subset or same solution of training examples

[101]. Therefore, the *SVM* algorithm is not dependent on the order of training examples. Figure 6-13 shows the comparison of the deployed models based on RMSE and R^2. In the training period, *SVM* had the highest average (R^2 = 94%) value followed by ANN and GEP with 90%, and 78%. SVM was also the best with lowest average RMSE of 132 ML followed by GEP and ANN with 155 and 207 ML. In the test period, SVM also had the highest R^2 of 92% followed by 81% and 78% for ANN and GEP respectively. Comparing them based on RMSE in test period also suggests that SVM was the best with lowest average RMSE of 167 ML followed by GEP and ANN with RMSE values of 300 and 321 *ML* which are relatively poor. Thus, SVM had the best performance among other modeling technique which can compensate sudden initial changes which is the criteria of the flexibility concept in this study.

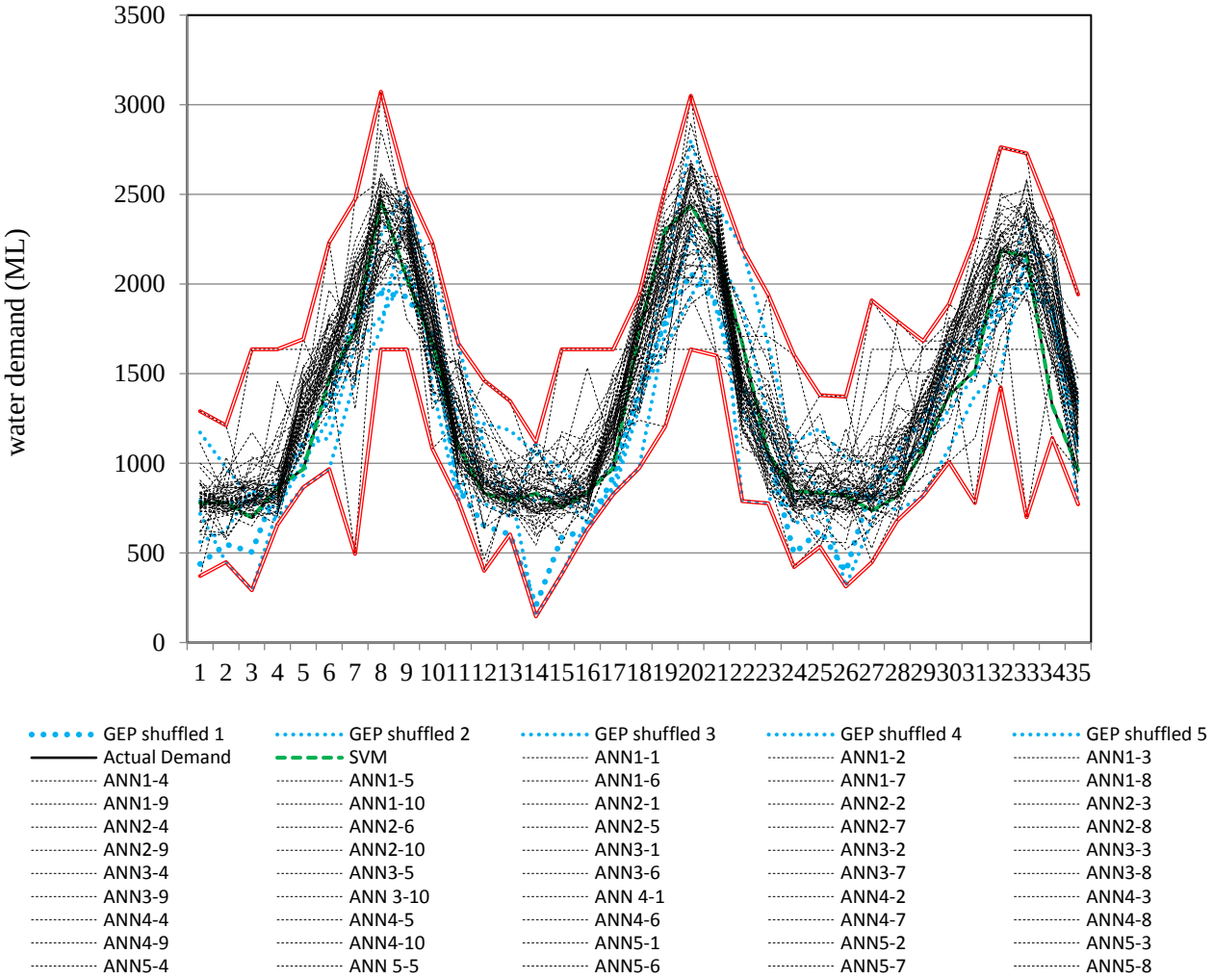


Figure 6-12 Shuffled training examples

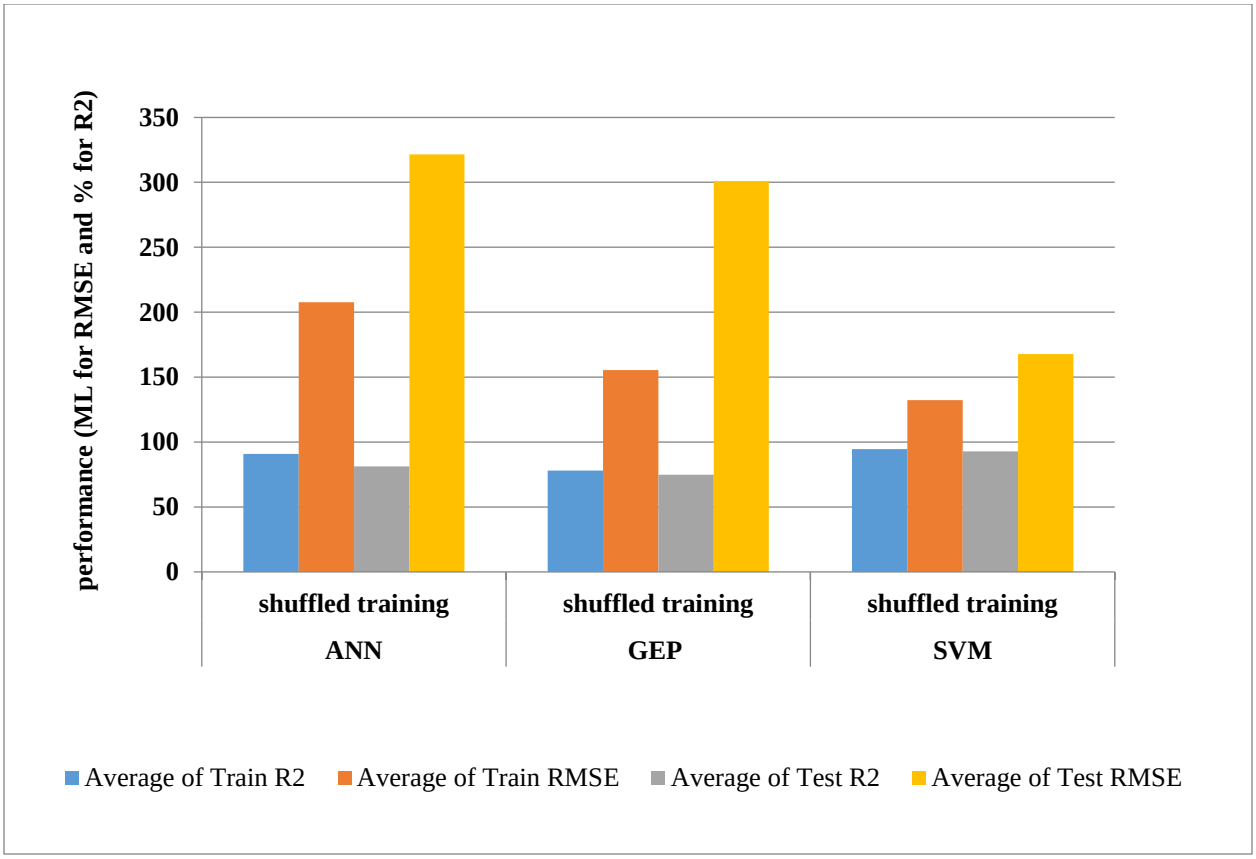


Figure 6-13 Comparing models based on shuffling training examples

6.4.3 Extreme climatic conditions

One of the methods of deep learning is to use data cropping in computer vision. This cropping allows the models to learn the bounds and limits of a training set of data. A deeply learned model would be able to predict the continues labelled output even if there is important information missing (copped data). Extreme climatic conditions were used in this study to simulate data cropping. As 80/20% data partitioning was applied to all models in train/test period, the 20 % extreme climatic conditions were cut and placed in test period for all time-series under study. Figure 6-14 shows the comparison between the deployed models when upper 20% of water demand placed in the test period as unknown future. This figure shows almost all techniques

underestimated the actual water demand. However, SVM and GEP performed much better than *ANN* models. On the other hand, lower 20% water demand in the test period showed better performance of these models (figure 6-15), although overestimation of actual demand can be easily noticed through this observation. The hottest days were used in the test set as upper 20% temperature. Figure 6-16 shows GEP was better in this extreme condition as blue is closer to actual demand visually. However, SVM shows a high fluctuation and poor performance in this circumstance. A different scenario was observed for the coldest days as GEP showed the poorest performance by highly overestimating most of the predictions (figure 6-17). SVM and ANN showed much closer prediction compared to GEP. Figure 6-18 and 6-19 show extreme conditions with rainfall could be easily taken by all models deployed, as predictions are close to actual demand based on this observation. To compare all these models, Figure 6-20 can be the fair method of comparison based on average R^2 and RMSE values. In training, SVM was the best with an R^2 of 96% followed by 89% for ANN and 65% for GEP. Comparing *RMSE* values in training shows all of them roughly similar performances with the best one as 148 ML for SVM, and 158 ML and 163 ML for ANN and GEP respectively. A similar comparison in testing showed all these models are really poor in prediction when extreme weather conditions are cropped out of the original training set. However, SVM was still the best with 40% R^2 which is still extremely poor. Surprisingly, SVM was significantly better in prediction based on RMSE, with an average RMSE of 179 ML followed by 272 and 568 ML for ANN and GEP respectively.

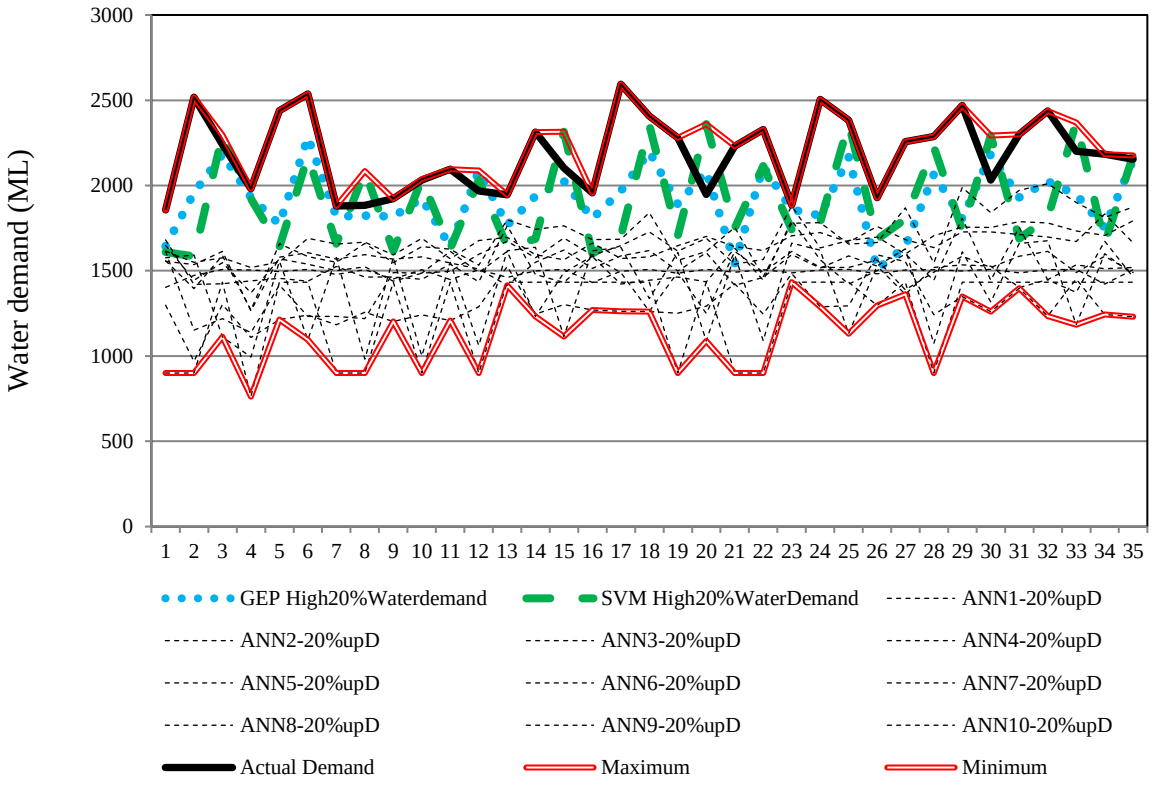


Figure 6-14 Upper 20% demand in the test period (36 months)

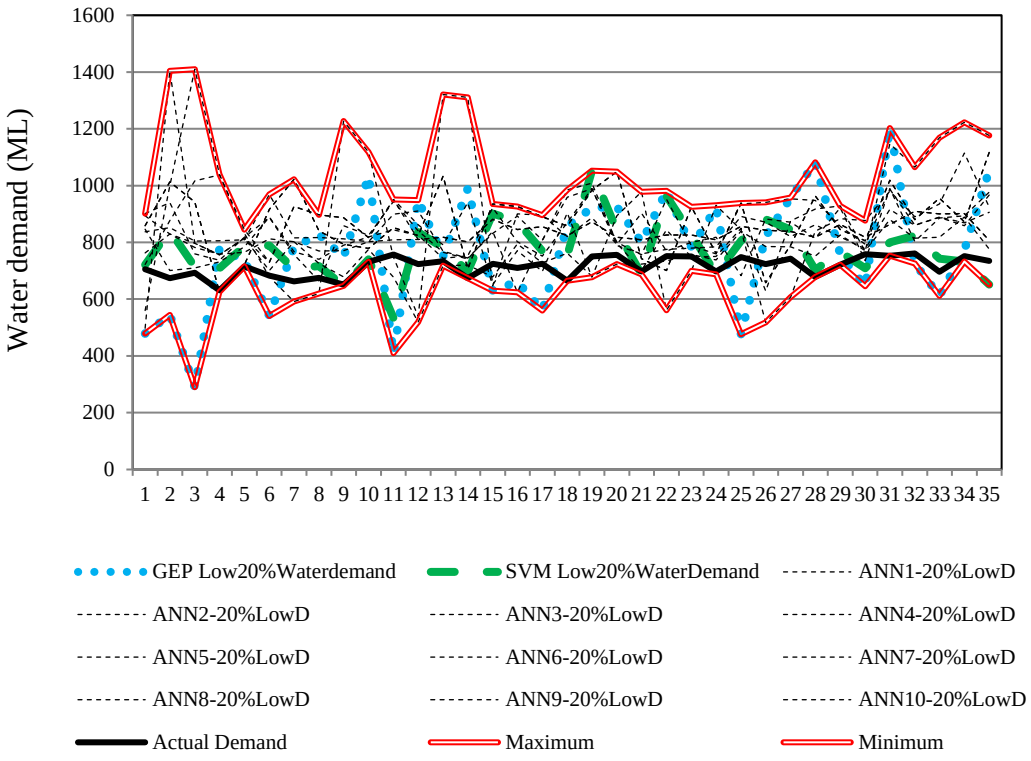


Figure 6-15 Lower 20% demand in the test period (36 months)

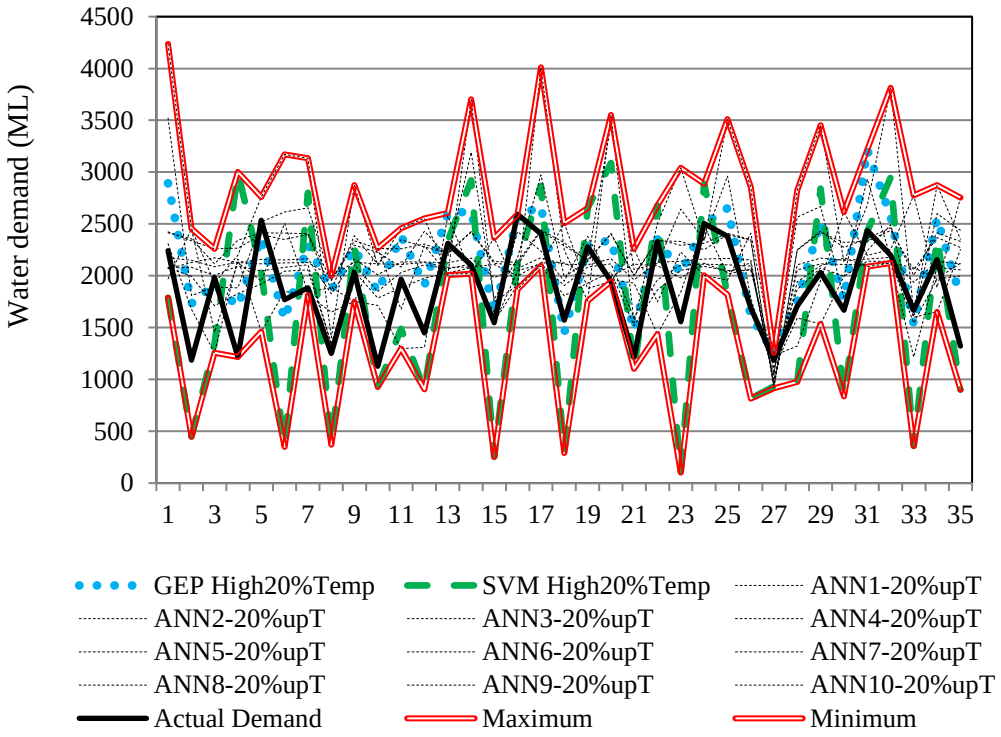


Figure 6-16 Upper 20% temperature in the test period (36 months)

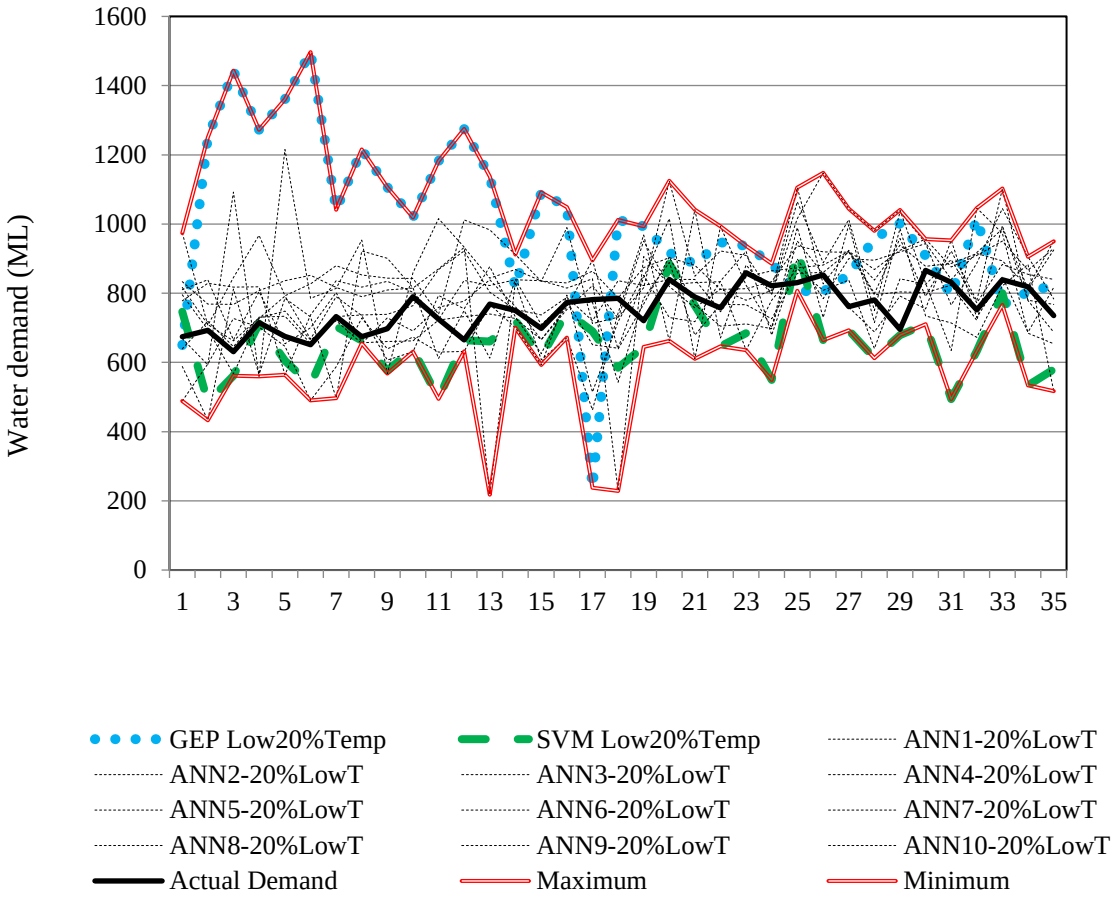


Figure 6-17 Lower 20% temperature in the test period (36 months)

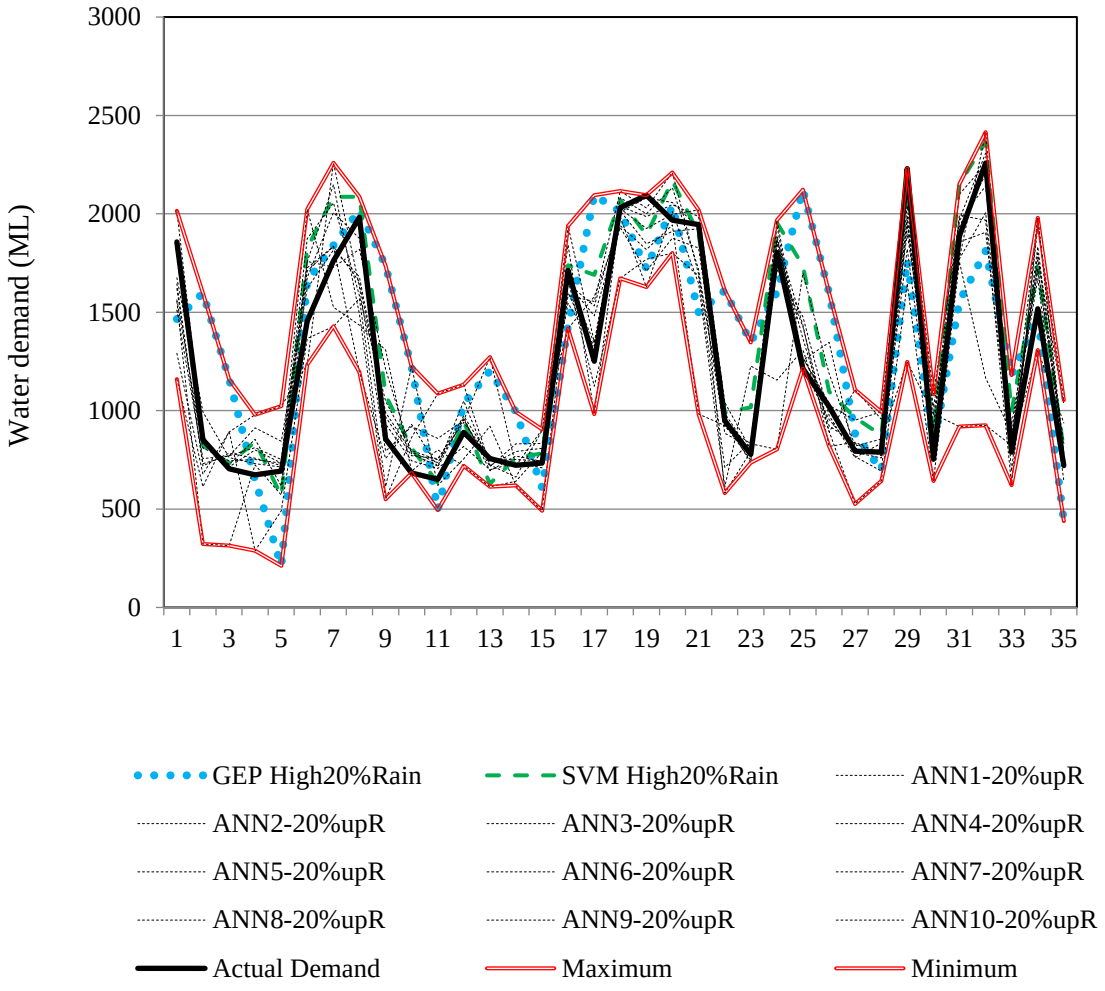


Figure 6-18 Upper 20% rainfall in the test period (36 months)

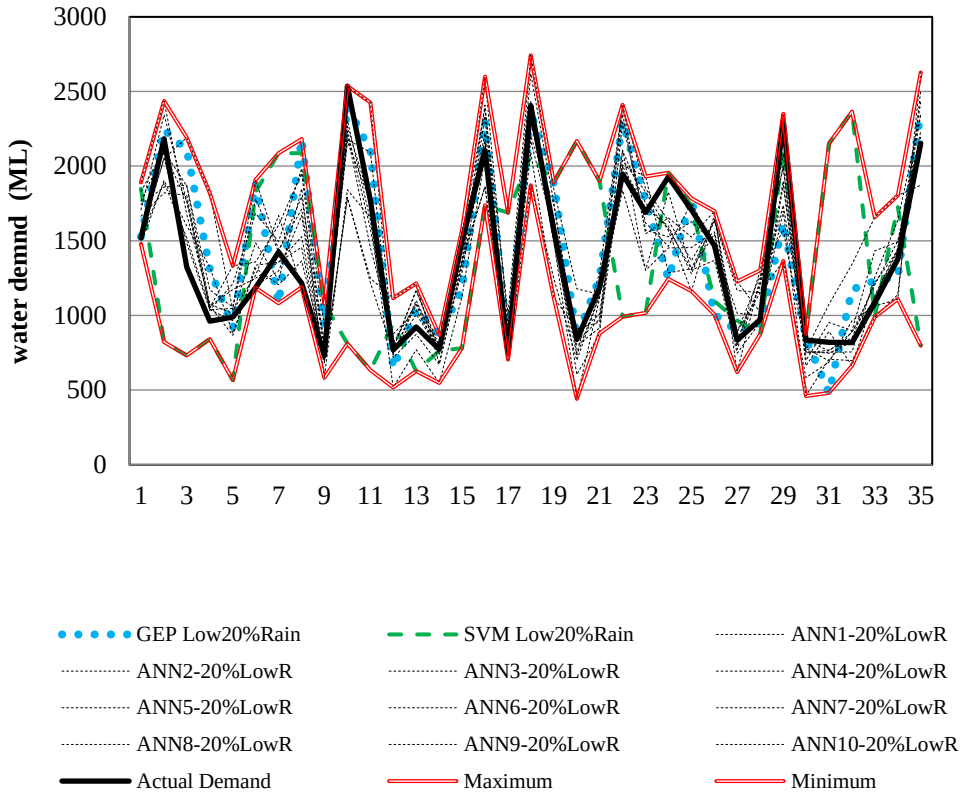


Figure 6-19 Lower 20% rainfall in the test period (36 months)

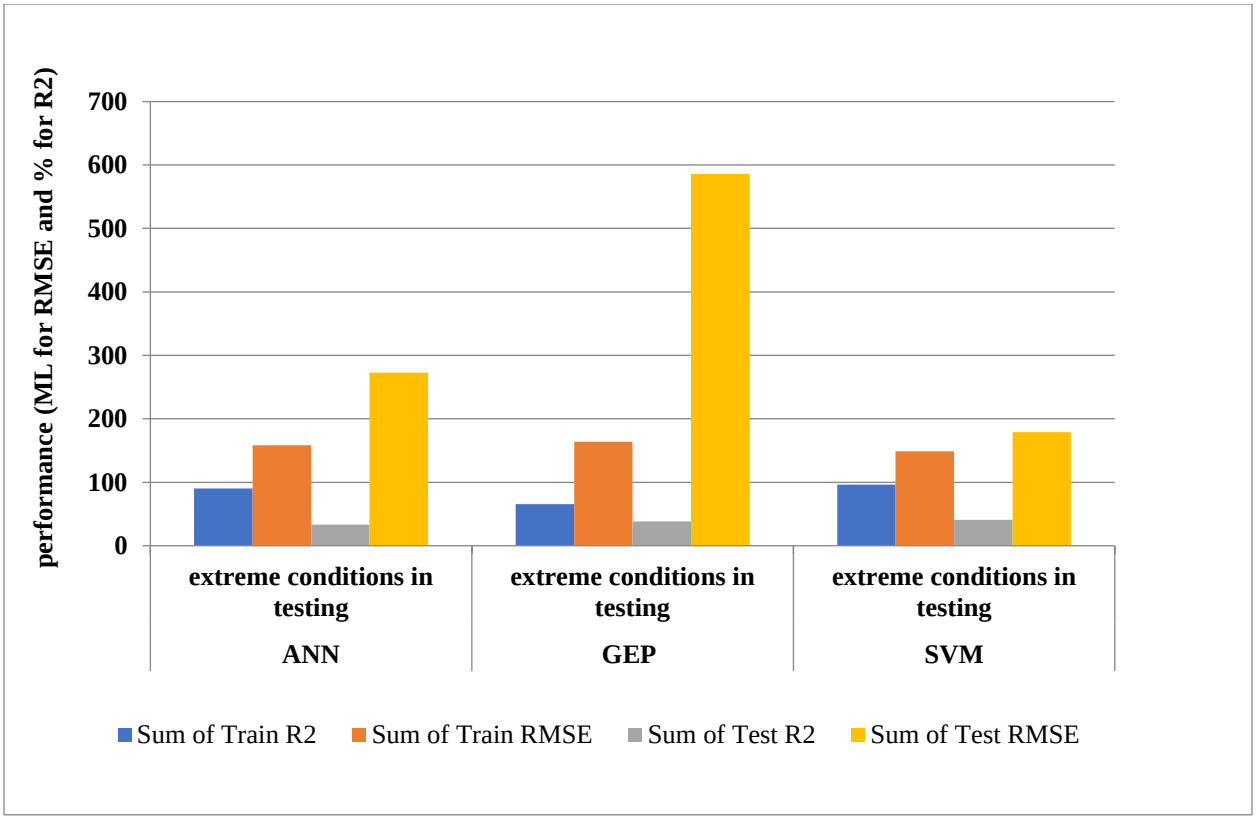


Figure 6-20 Comparing models based on an extreme condition in the test period

6.4.4 Distorted data- added white Gaussian noise

Using distorted data is also another method for training deep learning models. In the context of computer vision, images are blurred, or the combination of RGB (red, green, blue) is modified (for example, an image with the more blueish combination). This would make the model better understand the cases in which the image is not perfectly conveying the same pattern learned by model. To apply the same concept in this study, the time series of water demand has been modified by adding white Gaussian noise to its original time series. Moreover, a completely data-driven model was designed based on the modified time series as its main input. The results showed all these models would be extremely poor if data is modified with added white Gaussian noise. Figure

6-21 shows the performance of these models in the test period. Actual performance of these models can be better illustrated through figure 6-22 by comparing them based on the performance indices in both train and test sets. The average R^2 of the training data set was 40% for ANN followed by 36% and 30% for GEP and SVM respectively. However, in testing, ANN was the highest with R^2 of 36% followed by 34% and 23% for SVM and GEP. RMSE values of training showed ANN should be considered the best model in training with a significantly lower value of 105 ML compared to 491 and 575 ML for GEP and SVM respectively. However, all models had poor performance in test period with GEP giving the average RMSE value of 384 ML followed by 458 and 570 ML for SVM and ANN. After all, one could claim that the ANN is less sensitive to distorted data in training. But all models performed poorly in testing.

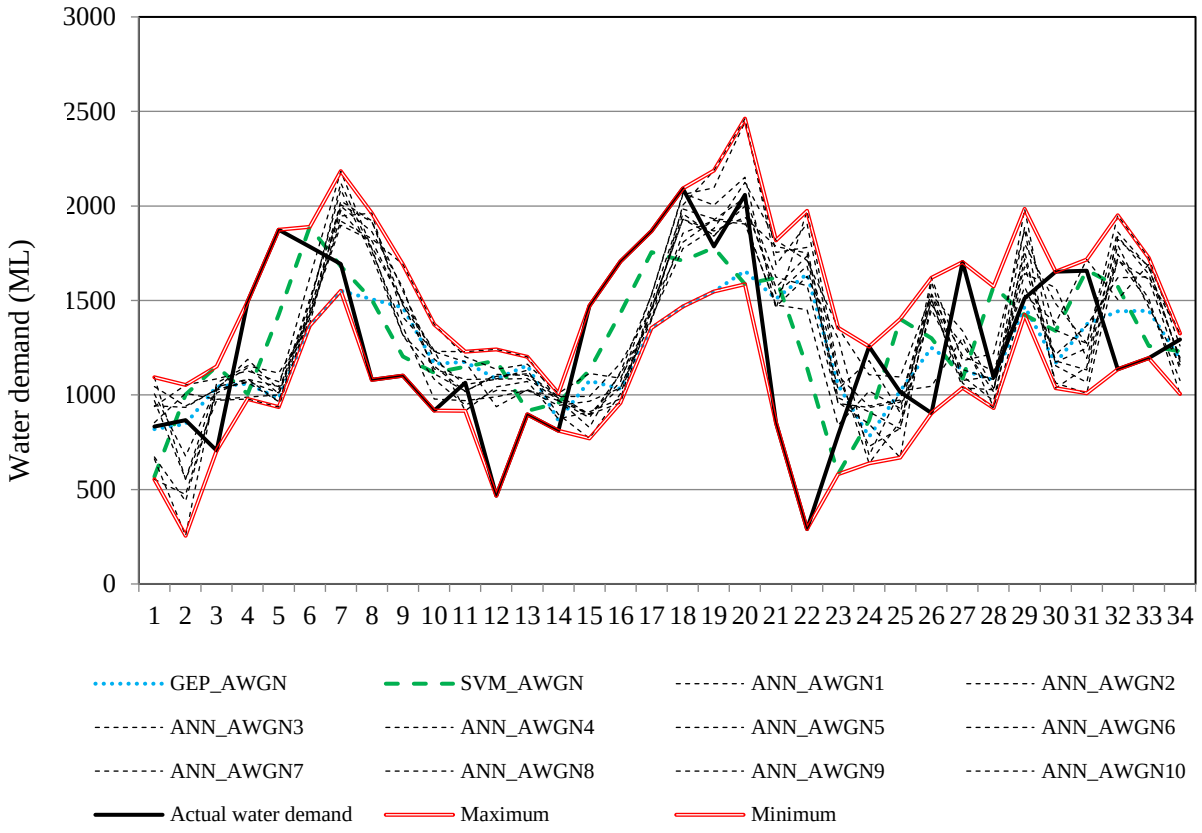


Figure 6-21 Added white Gaussian noise (distorted data) (36 months)

Figure 6-22 Comparing models based on distorted data

This section tried to answer the question about which one of the deployed modeling techniques in Hydroinformatics would be less sensitive to sudden changes, in other words, which one would be a more flexible tool in predictive analytics of water demand. Comparing these models based on four different criteria of distorted data, cropped data, shuffled training, and using unique prototypes in test period showed overall SVM could be considered the most flexible one in shuffled training examples. ANN was the superior model regarding distorted data in training. However, all models were poor in testing. SVM was more flexible in data cropping/ using extreme conditions in the test set too. Also, when unique prototypes of yearly demand were used, SVM could outperform other modeling types in being more flexible to take sudden changes in pattern or behavior of consumers in water demand. (Figure 6-23 and Table 6-2)

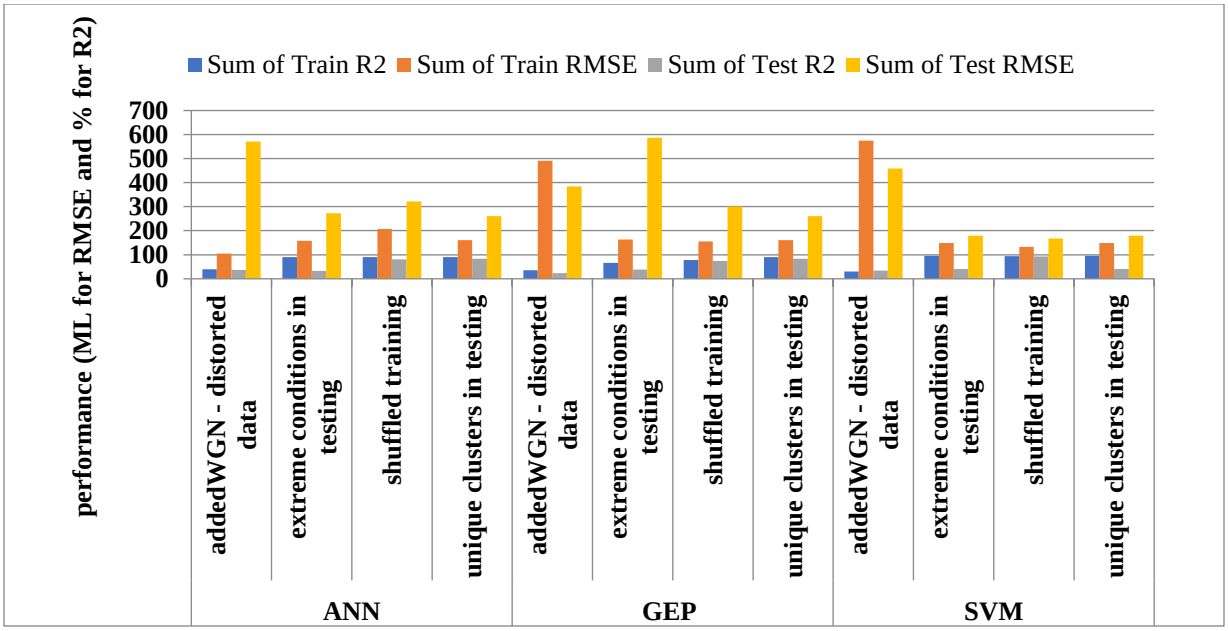


Figure 6-23 Overall comparison of all models

Table 6-2 performance indices of models averaged on all data augmentation practices

Model	Input design	Train R^2	Train *RMSE*	Test R^2	Test *RMSE*
GEP	shuffled training	78.03	155.38	74.75	300.84
SVM	shuffled training	94.49	132.28	92.96	167.78
ANN	shuffled training	90.78	207.67	81.17	321.67
GEP	addedWGN - distorted data	36.18	491.37	23.54	383.85
SVM	addedWGN - distorted data	30.28	574.92	34.18	458.41
ANN	addedWGN - distorted data	40.02	104.97	36.37	570.62
GEP	extreme conditions in testing	65.68	163.91	38.47	586.21
SVM	extreme conditions in testing	96.05	148.63	40.81	179.05
ANN	extreme conditions in testing	89.98	158.29	33.08	272.65
GEP	unique clusters in testing	90.19	160.17	84.15	260.58
SVM	unique clusters in testing	96.05	148.63	40.81	179.05
ANN	unique clusters in testing	90.19	160.17	84.15	260.58

6.4.5 Flexible range of demand

All these 212 numerical of models in chapter 6 were intended to propose a range of demand fulfilling the concept of flexibility defined in this project. Sources of uncertainty in water demand

could change the initial conditions of WDS. Therefore, the input design of these models was modified to make sure some possible sudden changes are considered in the modeling phase. Table 3 shows the values which are the consequences of all these numerical experiments based on common deep learning practices. Average values of (Max prediction - Actual) as well as (Actual – Min perdition) are presented through the 212 runs of numerical experiments designed in this study. Finally, this novel approach would propose an average of 358 ML for upper bound, and 335 *ML* for lower bound of water demand in the case of the City of Kelowna (Table 6-3). It is extremely surprising that after 212 numerical experiments applying different practices of deep learning, the upper and lower bound came very close to each other with only (358-335=23 ML) difference between upper and lower bounds, which could be neglected in this scale.

Table 6-3 Upper/lower bound for flexibility concept

Data Augmentation	**average (Max prediction - Actual) *ML***	**average (Actual - Min prediction) *ML***
Shuffled training	363.79	294.14
Cluster 1 test	206.87	228.48
Cluster 2 test	187.87	244.72
Cluster 3 test	340.87	187.34
Cluster 4 test	206.02	97.82
Cluster 5 test	549.80	345.07
Cluster 6 test	244.56	217.77
Upper20%demand test	41.72	1052.38
Lower20%demand test	330.86	93.20
Upper20%Temp test	986.36	598.85
Lower20%Temp test	328.79	181.42
Upper20%rainfall test	372.32	386.92
Lower20%rainfall test	491.62	475.75
Distorted data	363.79	294.14

6.5 Summary

Flexible demand design for complex infrastructures such as WDS could be very challenging in real engineering practice. High levels of uncertainties associated with the initial conditions of WDS, along with the fact that most of the hydrological and socio-economic factors are subjected to future changes could add further complexity to this approach. This research utilized fundamental practices in deep learning to deploy the bounds and extremes of well-known modeling techniques in hydroinformatics for proposing a range of future demand by considering climatic and socio-economic factors. Unlike the common practice which uses a safety factor (usually >1.5) defined by municipalities for demand design before hydraulic design of WDS, this research tried considering climate change, as well as two demographic variables (population and hotel occupancy rate) to evaluate main uncertainties of initial conditions of WDS. Implementing distorted data by adding white Gaussian noise, data cropping by using extreme climatic conditions in testing, shuffling training examples to evaluate entropy in modeling, and using unique prototypes of consumer behavior in yearly water demand, this research tried many possible sources of uncertainties to widen the range of the proposed demand. Therefore, it would be logical to claim using results of this approach can be a safe, as well as an educated guess to propose upper and lower bound of water demand.

Chapter 7: Conclusions and future work

7.1 Conclusions

To improve the current practices in long-term water demand forecasting, this research used a flexible approach in predictive analytics of water demand. The research outcomes are listed below in detail:

1. The dynamic nature of explanatory variables used in predictive models can provide valuable information to the modeler before selection of predictive analytics method and define input variables. The second chapter of this dissertation deployed phase space reconstruction of time series data for climatic information in the City of Kelowna using average mutual information. The applied correlation dimension method showed all the studied variables are good examples of noisy data and the level of chaos exhibited by them makes it in appropriate to use deterministic methods.
2. Hydro-informatics is a new branch of informatics defined around computational hydraulics and other subjects related to hydrology. GEP and SVM as main soft computing approaches used in this field have been used to develop models for long-term forecasting of water demand for the city of Kelowna (1996-2010) using average monthly data of water demand as response variables. An input variable design paradigm has been defined to check the performance of 3 input design: (water demand based), (water demand + climatic information based), and (water demand + climatic information + demographic information based) model. This was coupled with average mutual information of water demand for an optimum lag selection of all independent variables for phase space reconstruction. Results showed both deployed modeling techniques had significantly accurate predictions. However, SVM performed slightly better than GEP.

3. In chapter 3, the phase space reconstruction of all input variables was conducted using the optimum lag time of average monthly water demand time series (3 months). One could question how the models would perform if the optimum lag time of each explanatory is used to reconstruct the phase space before feeding the models. Chapter 4 answered this question through different input designs considering individual optimum lag times separately and compared it with the method in chapter 3. The results could validate the input design developed in chapter 3 where the optimum lag time of the response variable, water demand, was used to perform phase space reconstruction of all input variables.
4. A novel two-stage learning approach is proposed in chapter 5. K-means clustering was used as an unsupervised learning algorithm to group the given daily data vectors (24, 1) with cosine similarity as distance metrics. This was followed by deploying GEP as a supervised learning evolutionary technique to predict the water demand in a wide range of different time-scales. Time-scale modeling of water demand can aid the water utilities with a better understanding of the consumption behavior in WDS. Not all WDSs are using a real-time SCADA system that provides water demand data in finer resolutions depending on the frequency of their data acquisition system. Results of the chapter proved the efficiency of labeling data through unsupervised learning in improving the performance of predictive models. Due to the behavior of daily water consumption, a cosine similarity distance metrics was a suitable selection to find the distances between corresponding daily vectors. When coupled with a supervised evolutionary algorithm like GEP, predictive models can be highly accurate among their cluster assignments.
5. Chapter 6 of this dissertation proposed a novel approach to insert flexibility concept in demand design of WDS design. The supervised learning algorithms used in chapter 3

(SVM and GEP) were deployed along with ANN as conventional machine learning algorithm in water demand forecasting models to generate a range of demand based on different sources of uncertainties. Basic deep learning practices in data augmentation were used to widen the range of the forecasted water demand as much as possible. The outcome of this chapter was a range of demand which anticipates future changes in initial conditions of WDS. The obtained results proved SVM could be considered as the most flexible technique among the supervised learning algorithms used in this dissertation. Moreover, this approach resulted in a flexible range of predicted demand (358 ML for upper bound and 335 ML for lower bound) which considered many sources of uncertainties into account.

A demand design based on the proposed approach would result in a sustainable system which can foresee possible unexpected changes in initial conditions of WDS. Eventually, WDS operators would not be facing issues or problems associated with pumping schedules in times of high/low demand, and rehabilitation/changing trenching of pipelines due to serving more consumers in case of rapid urbanization. Consequently, this design would be neither over nor under-designed.

7.2 Main Contributions to Knowledge

- Verifying that the climatic data used for the forecasting models does not exhibit low dimensional chaotic properties.
- Merging climatic and demographic variables for long term water demand forecasting models to account for climate change.

- Exploring the use of GEP as new approach in this field and comparing it with SVM as an emerging modeling technique.
- Using AMI for phase space reconstruction and input variable design of water demand forecasting models.
- Proposing a new two-stage (supervised/unsupervised) learning approach to water demand forecasting models by coupling GEP with K-means clustering using triangle similarity.
- Developing a new concept of flexible water demand analysis using data augmentation techniques borrowed from deep learning practices.

7.3 Limitations

- The main weakness of the developed models is out-of-sample predictive analytics of the underlying determinants. Global climate change models or future projection of demographic variables should be used to deploy these time series data for such forecast models.
- Finding relevant data for input variable design is a challenging process in data science projects. This is even more difficult in small cities like Kelowna where not enough data is available to modelers for water demand forecasting.
- Deep learning of the algorithms is possible only in ANN as going deeper in layers of SVM algorithm is a limitation. Also, the concept of GEP in evolutionary programming is not as computationally efficient as the other two methods.
- Digitization of the society and infrastructures are introducing big data to predictive analytics practitioners. Currently, online learning is possible through platforms like TensorFlow which allows developing deep learning models in ANN. Also, they are far

more computationally efficient compared to other models as GPU (graphical processing unit) versions of these platforms perform much faster than CPU (central processing unit). This means, ANN can outperform other learning algorithms used in this dissertation if big data is of interest.

7.4 Future Work

New research ideas could be developed based on the findings of this dissertation. The following suggestions can be made:

- A deep understanding of the problem context coupled with engineering intuition is needed to lay a solid foundation for defining a predictive model's input variables. One major limitation associated with data intensive research projects is the data frame needed to feed predictive models. Input variable design is the backbone of predictive models; however, not always relevant data are available to modelers.
- This research focused on the available climatic information and demographic variables in the City of Kelowna. Land use, price of water, demographic variables like ratio of male/female, and income of consumers can be used in future research for long-term forecasting of water demand. This research focused on a flexible approach to anticipate uncertainties of *WDS* initial conditions in predictive models. All the supervised/unsupervised learning algorithms and data augmentation practices were used as tools to meet the prime objective of this research. If accuracy of these models is of interest, one could opt for further analysis of each algorithm through tuning the parameters, emerging deep learning practices, data mining, and pre-processing of the data.

- The distance metrics used in the clustering or data mining of this research was cosine similarity/distance between the vectors representing the consumption behavior in water demand. However, further research can be done using other distance metrices.
- It is important to notice this dissertation just proposed a new concept of flexibility, to implement this flexible range of demand, WDS designers might need to use alternate pipelines, or use different pumping schedules which can be interesting topics for future research ideas in design aspects of WDS.
- The engineering and cost implications of incorporating the flexible demand analysis approach into WDS design can develop interesting future research ideas following this dissertation. Implications for pipe, pump, tank size, system operation, capital cost, operational use, energy use, system's reliability, and system's resilience can be explored following the proposed approach.

Bibliography

[1] M. Blokker, I. Vloerbergh, and S. Buchberger, "Estimating peak water demands in hydraulic systems II-future trends," in *WDSA 2012: 14th Water Distribution Systems Analysis Conference*, 2012, p. 1138.

[2] M. Ghiassi, D. K. Zimbra, and H. Saidane, "Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model," *J. Water Resour. Plan. Manag.*, vol. 134, no. 2, pp. 138–146, 2008.

[3] E. Fricke and A. P. Schulz, "Design for changeability (DfC): Principles to enable changes in systems throughout their entire lifecycle," *Syst. Eng.*, vol. 8, no. 4, pp. 342–359, 2005.

[4] D. A. E. H. Astings, J. H. Saleh, and D. E. Hastings, "EXTRACTING THE ESSENCE OF FLEXIBILITY IN S SYSTEM DESIGN Extracting the Essence of Flexibility in System Design," 2001.

[5] R. de Neufville and S. Scholtes, "Flexibility in engineering design (engineering systems)," p. 293, 2011.

[6] P. Troy and D. Holloway, "The use of residential water consumption as an urban planning tool: a pilot study in Adelaide," *J. Environ. Plan. Manag.*, vol. 47, no. 1, pp. 97–114, 2004.

[7] J. Koo, M. Yu, S. Kim, M. Shim, and A. Koizumi, "Estimating regional water demand in Seoul, South Korea, using principal component and cluster analysis," *Water Sci. Technol. Water Supply*, vol. 5, no. March 2005, pp. 1–7, 2005.

[8] S.-P. Miaou, "A Class of Time Series Urban Water Demand Models With Nonlinear Climatic Effects," *Water Resour. Res.*, vol. 26, no. 2, pp. 169–178, 1990.

[9] A. Jain, A. K. Varshney, and U. C. Joshi, "Short-term water demand forecast modeling at IIT Kanpur using artificial neural networks," *Water Resour. Manag.*, vol. 15, no. 5, pp. 299–

321, 2001.

[10] S. Gato-Trinidad, N. Jayasuriya, and P. Roberts, “Understanding urban residential end uses of water,” *Water Sci. Technol.*, vol. 64, no. 1, pp. 36–42, 2011.

[11] L. Beck and T. Bernauer, “How will combined changes in water demand and climate affect water availability in the Zambezi river basin?,” *Glob. Environ. Chang.*, vol. 21, no. 3, pp. 1061–1072, 2011.

[12] M. Bakker, J. H. G. Vreeburg, K. M. van Schagen, and L. C. Rietveld, “A fully adaptive forecasting model for short-term drinking water demand,” *Environ. Model. Softw.*, vol. 48, pp. 141–151, 2013.

[13] S. L. Zhou, T. A. McMahon, A. Walton, and J. Lewis, “Forecasting daily urban water demand: A case study of Melbourne,” *J. Hydrol.*, vol. 236, no. 3–4, pp. 153–164, 2000.

[14] A. Mukhopadhyay, A. Akber, and E. Al-Awadi, “Analysis of freshwater consumption patterns in the private residences of Kuwait,” *Urban Water*, vol. 3, no. 1–2, pp. 53–62, 2001.

[15] C. C. Dos Santos and A. J. Pereira Filho, “Water Demand Forecasting Model for the Metropolitan Area of São Paulo, Brazil,” *Water Resour. Manag.*, vol. 28, no. 13, pp. 4401–4414, 2014.

[16] S. L. Zhou, T. A. McMahon, A. Walton, and J. Lewis, “Forecasting operational demand for an urban water supply zone,” *J. Hydrol.*, vol. 259, no. 1–4, pp. 189–202, 2002.

[17] R. Anderson, T. Miller, and M. Washburn, “Water savings from lawn watering restrictions during a drought year, fort collins, Colorado,” *Water Resour. Bull.*, vol. 16, no. 4, pp. 642–625, 1980.

[18] D. Maidment and E. Parzen, “Monthly water use and its relationship to climatic variables in Texas,” *Water Resour.*, vol. 19, no. 8, pp. 409–418, 1984.

[19] L. Brekke, M. Larsen, M. Ausburb, and L. Takaichi, “Suburban Water Demand Modeling Using Stepwise Regression,” *J.Awwa*, vol. 94, no. 10, pp. 65–75, 2002.

[20] A. S. Polebitski, R. N. Palmer, M. ASCE, and P. Waddell, “Evaluating Water Demands Under Climate Change and Transitions in the Urban Environment,” *J. Water Resour. Plan. Manag.*, vol. 137, no. May-June, pp. 249–257, 2011.

[21] S. J. Lee, E. A. Wentz, and P. Gober, “Space-time forecasting using soft geostatistics: A case study in forecasting municipal water demand for Phoenix, Arizona,” *Stoch. Environ. Res. Risk Assess.*, vol. 24, no. 2, pp. 283–295, 2010.

[22] J. M. Alhumoud, “Freshwater consumption in Kuwait: Analysis and forecasting,” *J. Water Supply Res. Technol. - AQUA*, vol. 57, no. 4, pp. 279–288, 2008.

[23] R. Vemuri and R. Rogers, *Artifical neural networks, forecasting time series*. 1994.

[24] V. Crommelynck, C. Duquesne, M. Mercier, and C. Miniussi, “Daily and hourly water consumption forecasting tools using neural networks,” in *AWWA’s annual computer speciality conference*, 1992, pp. 665–676.

[25] J. Bougadis, K. Adamowski, and R. Diduch, “Short-term municipal water demand forecasting,” *Hydrol. Process.*, vol. 19, no. 1, pp. 137–148, 2005.

[26] L. Jentgen, H. Kidder, R. Hill, and S. Conrad, “Energy management strategies use short-term water consumption forecasting to ...,” *Water*, no. June, pp. 86–94, 2007.

[27] A. H. Aly and N. Wanakule, “Short-Term Forecasting for Urban Water Consumption,” *J. Water Resour. Plan. Manag.*, vol. 130, no. 5, pp. 405–410, 2004.

[28] S. Alvisi, M. Franchini, and A. Marinelli, “A short-term, pattern-based model for water-demand forecasting,” *J. Hydroinformatics*, vol. 9, no. 1, p. 39, 2007.

[29] X. Wang, Y. Sun, L. Song, and C. Mei, “An eco-environmental water demand based model

for optimizing water resources using hybrid genetic simulated annealing algorithms. Part I. Model development," *J. Environ. Manage.*, vol. 90, no. 8, pp. 2628–2635, 2009.

[30] W. Li and Z. Huicheng, "Urban water demand forecasting based on HP filter and fuzzy neural network," *J. Hydroinformatics*, vol. 12, no. 2, p. 172, 2010.

[31] I. S. Msiza, F. V Nelwamondo, and T. Marwala, "Water demand prediction using artificial neural networks and support vector regression," *J. Comput.*, vol. 3, no. 11, pp. 1–8, 2008.

[32] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-García, "Predictive models for forecasting hourly urban water demand," *J. Hydrol.*, vol. 387, no. 1–2, pp. 141–150, 2010.

[33] B. M. Brentan, E. Luvizotto, M. Herrera, J. Izquierdo, and R. Pérez-García, "Hybrid regression model for near real-time urban water demand forecasting," *J. Comput. Appl. Math.*, vol. 309, pp. 532–541, 2017.

[34] J. Adamowski, H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada," *Water Resour. Res.*, vol. 48, no. 1, pp. 1–14, 2012.

[35] A. Altunkaynak, M. Özger, and M. Çakmakci, "Water consumption prediction of Istanbul City by using a fuzzy logic approach," *Water Resour. Manag.*, vol. 19, no. 5, pp. 641–654, 2005.

[36] I. N. Athanasiadis, A. K. Mentes, P. A. Mitkas, and Y. A. Mylopoulos, "A Hybrid Agent-Based Model for Estimating Residential Water Demand," *Simulation*, vol. 81, no. 3, pp. 175–187, 2005.

[37] B. L. Shvartser, U. Shamir, M. ASCE, and M. Feldman, "Pattern Recognition Approach,"

Water Resour., vol. 119, no. 6, pp. 155–169, 1994.

[38] I. Pulido-Calvo and J. C. Gutiérrez-Estrada, "Improved irrigation water demand forecasting using a soft-computing hybrid model," *Biosyst. Eng.*, vol. 102, no. 2, pp. 202–218, 2009.

[39] M. A. Yurdusev, M. Firat, M. Mermer, and M. E. Turan, "Water use prediction by radial and feed-forward neural nets," *Proc. Inst. Civ. Eng. Manag.*, vol. 162, no. 3, pp. 179–188, 2009.

[40] D. P. Solomatine and Y. Xue, "M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China," *J. Hydrol. Eng.*, vol. 9, no. 6, pp. 491–501, 2004.

[41] K. B. Khatri and K. Vairavamoorthy, "Water Demand Forecasting for the City of the Future against the Uncertainties and the Global Change Pressures: Case of Birmingham," *World Environ. Water Resour. Congr. 2009*, vol. 41036, no. May 2009, pp. 1–15, 2009.

[42] B. Sivakumar, "Chaos theory in hydrology: important issues and interpretations\r," *Am. J. Respir. Crit. Care Med.*, vol. 227, no. 4, pp. 1–20, 2000.

[43] M. A. Ghorbani, O. Kisi, and M. Aalinezhad, "A probe into the chaotic nature of daily streamflow time series by correlation dimension and largest Lyapunov methods," *Appl. Math. Model.*, vol. 34, no. 12, pp. 4050–4057, 2010.

[44] R. Khatibi, B. Sivakumar, M. A. Ghorbani, O. Kisi, K. Koçak, and D. Farsadi Zadeh, "Investigating chaos in river stage and discharge time series," *J. Hydrol.*, vol. 414–415, pp. 108–117, 2012.

[45] F. Takens, "Detecting strange attractors in turbulence."

[46] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[47] J. Holzfuss and G. Mayer-Kress, “An Approach To Error-Estimation in The Application of Dimension Algorithms,” 1985.

[48] R. Hegger, H. Kantz, and T. Schreiber, “Practical implementation of nonlinear time series methods: The TISEAN package,” *Chaos*, vol. 9, no. 2, pp. 413–435, 1999.

[49] M. Kennel, R. Brown, and H. Abarbanel, “Determining embedding dimension for phase-space reconstruction using a geometrical construction,” *Phys. Rev. A*, vol. 134, no. 1, pp. 5–6, 1992.

[50] P. Grassberger and I. Procaccia, “Characterization of strange attractors,” *Phys. Rev. Lett.*, vol. 50, no. 5, pp. 346–349, 1983.

[51] J. Theiler, “Spurious dimension from correlation algorithms applied to limited time-series data,” *Phys. Rev. A*, vol. 34, no. 3, pp. 2427–2432, 1986.

[52] C. Ferreira, “Gene Expression Programming: a New Adaptive Algorithm for Solving Problems,” pp. 1–22, 2001.

[53] O. Kisi and J. Shiri, “Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy Conjunction Models,” *Water Resour. Manag.*, vol. 25, no. 13, pp. 3135–3152, 2011.

[54] J. Shiri *et al.*, “Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran,” *J. Hydrol.*, vol. 508, pp. 1–11, 2014.

[55] J. Shiri, P. Marti, and V. P. Singh, “Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning,” *Hydrol. Process.*, vol. 28, no. 3, pp. 1215–1225, 2014.

[56] C. Cortes and V. Vapnik, “Support Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp.

273–297, 1995.

[57] M. A. Mohandes, T. O. Halawani, S. Rehman, and A. A. Hussain, “Support vector machines for wind speed prediction,” *Renew. Energy*, vol. 29, no. 6, pp. 939–947, 2004.

[58] H. Yoon, S. C. Jun, Y. Hyun, G. O. Bae, and K. K. Lee, “A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer,” *J. Hydrol.*, vol. 396, no. 1–2, pp. 128–138, 2011.

[59] A. A. Jafarzadeh, M. Pal, M. Servati, M. H. FazeliFard, and M. A. Ghorbani, “Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction,” *Int. J. Environ. Sci. Technol.*, vol. 13, no. 1, pp. 87–96, 2016.

[60] D. R. Legates and G. J. McCabe Jr., “Evaluating the Use of ‘Goodness of Fit’ Measures in Hydrologic and Hydroclimatic Model Validation,” *Water Resour. Res.*, vol. 35, no. 1, pp. 233–241, 2005.

[61] S. Shabani, P. Yousefi, J. Adamowski, and G. Naser, “Intelligent Soft Computing Models in Water Demand Forecasting,” in *Water Stress in Plants*, InTech, 2016.

[62] P. H. Glecik, *The World’s Water Volume 7: Report on FreshWater*. 2011.

[63] T. G. Mamo, I. Juran, and I. Shahrour, “Urban Water Demand Forecasting Using the Stochastic Nature of Short Term Historical Water Demand and Supply Pattern Urban Water Demand Forecasting Using the Stochastic Nature of Short Term Historical Water Demand and supply Pattern,” no. SEPTEMBER 2013, pp. 1–10, 2015.

[64] L. Cao, “Data Science,” *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–42, 2017.

[65] A. Gupta, S. Mishra, N. Bokde, and K. Kulat, “Need of smart water systems in India,” *Int. J. Appl. Eng. Res.*, vol. 11, no. 4, pp. 2216–2223, 2016.

[66] A. Jain and L. E. Ormsbee, “Short-term water demand forecast modeling techniques -

Conventional methods versus AI," *J. / Am. Water Work. Assoc.*, vol. 94, no. 7, pp. 64–72, 2002.

[67] J. F. Adamowski, "Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks," *J. Water Resour. Plan. Manag.*, vol. 134, no. 2, pp. 119–128, 2008.

[68] S. Gato, N. Jayasuriya, and P. Roberts, "Temperature and rainfall thresholds for base use urban water demand modeling," *J. Hydrol.*, vol. 337, no. 3–4, pp. 364–376, 2007.

[69] M. A. Al-Zahrani and A. Abo-Monasar, "Urban residential water demand prediction based on artificial neural networks and time series models," *Water Resour. Manag.*, vol. 29, no. 10, pp. 3651–3662, 2015.

[70] M. K. Tiwari and J. Adamowski, "Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models," *Water Resour. Res.*, vol. 49, no. 10, pp. 6486–6507, 2013.

[71] M. Firat, M. A. Yurdusev, and M. E. Turan, "Evaluation of artificial neural network techniques for municipal water consumption modeling," *Water Resour. Manag.*, vol. 23, no. 4, pp. 617–632, 2009.

[72] M. K. Tiwari and J. F. Adamowski, "Medium-Term Urban Water Demand Forecasting with Limited Data Using an Ensemble Wavelet – Bootstrap Machine-Learning Approach," *J. Water Resour. Plan. Manag.*, vol. 141, no. 2001, pp. 1–12, 2015.

[73] F. K. Odan, L. Fernanda, and R. Reis, "Hybrid Water Demand Forecasting Model Associating Artificial Neural Network with Fourier Series," vol. 138, no. JUNE, pp. 245–256, 2012.

[74] L. Huang, C. Zhang, Y. Peng, and H. Zhou, "Application of a Combination Model Based on Wavelet Transform and KPLS-ARMA for Urban Annual Water Demand Forecasting,"

J. Water Resour. Plan. Manag., vol. 140, no. 8, p. 04014013, 2014.

[75] S. Shabani, P. Yousefi, and G. Naser, "Support Vector Machines in Urban Water Demand Forecasting Using Phase Space Reconstruction," *Procedia Eng.*, vol. 186, pp. 537–543, 2017.

[76] M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski, and A. Pandey, "Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5267–5276, 2014.

[77] A. K. Fernando, A. Y. Shamseldin, and R. J. Abrahart, "Use of Gene Expression Programming for Multimodel Combination of Rainfall-Runoff Models," vol. 17, no. September, pp. 975–985, 2012.

[78] R. Stull, "Wet-bulb temperature from relative humidity and air temperature," *J. Appl. Meteorol. Climatol.*, vol. 50, no. 11, pp. 2267–2269, 2011.

[79] T. C. Fu, "A review on time series data mining," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, 2011.

[80] A. Candelieri, "Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection," *Water*, vol. 9, no. 3, p. 224, 2017.

[81] J.-D. Rinaudo, "Long-Term Water Demand Forecasting," *Underst. Manag. Urban Water Transition. Glob. Issues Water Policy*, pp. 239–268, 2015.

[82] E. A. Donkor, S. M. Asce, T. A. Mazzuchi, R. Soyer, and J. A. Roberson, "Urban Water Demand Forecasting : Review of Methods and Models," vol. 140, no. February, pp. 146–159, 2014.

[83] B. Billings and C. V. Jones, *Forecasting urban water demand*. American Water Work Association, 2011.

[84] E. Blokker, “Simulating residential water demand with a stochastic end-use model,” *J. Water …*, vol. 137, no. February, pp. 19–26, 2009.

[85] E. R. Levin, W. O. Maddaus, N. M. Sandkulla, and H. Pohl, “Forecasting wholesale demand and conservation savings,” *J. / Am. Water Work. Assoc.*, vol. 98, no. 2, pp. 102–111, 2006.

[86] H. E. Jacobs and J. Haarhoff, “Application of a Residential End-Use Model for Estimating cold water and hot water demand, waste water flow and salinity,” *Water SA*, vol. 30, no. 3, pp. 305–316, 2004.

[87] P. W. Mayer *et al.*, “Residential End Uses of Water,” *Aquacraft, Inc. Water Eng. Manag.*, p. 310, 1999.

[88] C. Bennett, R. A. Stewart, and C. D. Beal, “ANN-based residential water end-use demand forecasting model,” *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1014–1023, 2013.

[89] J. Adamowski and C. Karapataki, “Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms,” *J. Hydrol. Eng.*, vol. 15, no. 10, pp. 729–743, 2010.

[90] S. Mouatadid and J. Adamowski, “Using extreme learning machines for short-term urban water demand forecasting,” *Urban Water J.*, vol. 14, no. 6, pp. 630–638, 2017.

[91] Y. Bai, P. Wang, C. Li, J. Xie, and Y. Wang, “A multi-scale relevance vector regression approach for daily urban water demand forecasting,” *J. Hydrol.*, vol. 517, pp. 236–245, 2014.

[92] L. Zhang, X. Chen, B. Liu, and Z. Wang, “SVM model of wate demand prediction based on AGA,” *J. China Hydrol.*, vol. 28, no. 1, pp. 38–42, 2008.

[93] N. B. Shah, J. Wilds, L. Viscito, A. M. Ross, and D. E. Hastings, “Quantifying Flexibility for Architecting Changeable Systems,” *6th Conf. Syst. Eng. Res.*, pp. 1–13, 2008.

[94] I. Basupi and Z. Kapelan, “Flexible Water Distribution System Design Under Uncertainty,” *J. Water Resour. Plan. Manag.*, vol. 141, no. 4, pp. 786–797, 2015.

[95] Y. Zhou and T. Hu, “Flexible design of delivery capacity in urban water distribution system,” in *Management and Service Science*, 2009, pp. 1–4.

[96] A. NG, “Deep learning specialization,” 2017. [Online]. Available: www.deeplearning.ai.

[97] R. Shukla, “How to train your deep neural network,” 2017. [Online]. Available: http://rishy.github.io.

[98] S. Shabani, A. Candelieri, F. Archetti, and G. Naser, “Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts,” *Water*, vol. 10, no. 2, p. 142, 2018.

[99] G. Deyfruz *et al.*, *Réseaux de neurones: méthodologie et applications*. .

[100] B. N. Karunanithi, W. J. Grenney, D. Whitley, and K. Bovee, “Neural Networks For River Flow Prediction,” vol. 8, no. 2, pp. 201–220, 1994.

[101] N. Cristianini, C. Campbell, and J. Shawe-Taylor, “Dynamically adapting kernels in support vector machines,” *Adv. neural Inf. Process. Syst.*, pp. 204–210, 1999.

Appendix A

Instruction on using GeneXproTools V 5.0:

1) Loading data in software environment is done by connecting your directory as database to read in data.

2) Partitioning data sets into training and validation/test sets (80-20% for instance).

3) There is an option to use random shuffling or going with actual order of data. Since time series data is used in this thesis, random shuffling was not selected.

4) Choosing the function sets can be done by selecting user defined functions or built-in mathematical functions available in the software environment. The functions used in this thesis were based on previous literature (3 different operators).

5) Selecting the model architecture is the next step were the candidate solutions are encoded as chromosomes (linear strings of randomly populated numbers). This architecture requires defining the bounds of a gene. Head/tail, and random constants sizes which are defined based on the software default recommended values. The linking function is the last step where we set it as addition.

6) Selecting fitness function is just like cost function in machine learning language. Here RMSE was selected as fitness function or selection environment of the genes.

7) Exploring the learning algorithms by running the models.

Table A1. Detailed performance of all 90 GEP models in both testing and training periods

Model ID	Train			Test			Model ID	Train			Test		
	MAE	*RMSE*	R^2	*MAE*	*RMSE*	R^2		*MAE*	*RMSE*	R^2	*MAE*	*RMSE*	R^2
K1HT1OP1	0.27	0.36	0.88	0.27	0.33	0.91	K4HT1OP1	0.20	0.29	0.92	0.25	0.35	0.84
K1HT1OP2	0.31	0.42	0.84	0.29	0.38	0.86	K4HT1OP2	0.23	0.32	0.90	0.25	0.36	0.82
K1HT1OP3	0.35	0.52	0.75	0.32	0.48	0.78	K4HT1OP3	0.31	0.41	0.84	0.25	0.31	0.90
K1HT3OP1	0.50	0.64	0.64	0.53	0.40	0.65	K4HT3OP1	0.47	0.55	0.70	1.63	1.48	0.38
K1HT3OP2	0.54	0.68	0.67	0.55	0.72	0.62	K4HT3OP2	0.61	0.73	0.49	1.52	1.53	0.32
K1HT3OP3	0.41	0.51	0.74	0.44	0.52	0.75	K4HT3OP3	0.46	0.56	0.69	1.01	1.41	0.24
K1HT6OP1	0.23	0.30	0.91	0.29	0.36	0.89	K4HT6OP1	0.42	0.49	0.76	1.28	1.75	0.13
K1HT6OP2	0.32	0.40	0.84	0.28	0.33	0.91	K4HT6OP2	0.38	0.46	0.81	0.58	0.76	0.28
K1HT6OP3	0.17	0.23	0.95	0.22	0.28	0.93	K4HT6OP3	0.27	0.35	0.88	1.01	1.28	0.07
K1HT12OP1	0.30	0.42	0.82	0.22	0.33	0.88	K4HT12OP1	0.30	0.44	0.80	1.02	1.18	0.09
K1HT12OP2	0.34	0.49	0.77	0.29	0.38	0.84	K4HT12OP2	0.35	0.48	0.76	0.56	0.62	0.55
K1HT12OP3	0.31	0.45	0.79	0.28	0.37	0.84	K4HT12OP3	0.35	0.50	0.75	0.83	1.03	0.15
K1HT24OP1	0.54	0.73	0.47	0.55	0.66	0.38	K4HT24OP1	0.40	0.52	0.72	1.12	1.82	0.26
K1HT24OP2	0.51	0.72	0.48	0.54	0.66	0.25	K4HT24OP2	0.41	0.48	0.78	1.35	1.88	0.14
K1HT24OP3	0.56	0.75	0.43	0.71	0.54	0.11	K4HT24OP3	0.36	0.44	0.80	1.30	1.88	0.27
K2HT1OP1	0.20	0.29	0.93	0.22	0.27	0.94	K5HT1OP1	0.28	0.35	0.89	0.23	0.28	0.90
K2HT1OP2	0.18	0.27	0.93	0.19	0.24	0.95	K5HT1OP2	0.24	0.31	0.91	0.20	0.24	0.92
K2HT1OP3	0.22	0.30	0.91	0.22	0.28	0.93	K5HT1OP3	0.22	0.30	0.92	0.22	0.28	0.92
K2HT3OP1	0.70	0.82	0.32	0.77	0.87	0.30	K5HT3OP1	0.71	0.83	0.30	0.62	0.74	0.25
K2HT3OP2	0.69	0.81	0.34	0.74	0.86	0.33	K5HT3OP2	0.71	0.84	0.32	0.63	0.75	0.26
K2HT3OP3	0.63	0.75	0.43	0.67	0.79	0.43	K5HT3OP3	0.71	0.83	0.30	0.62	0.73	0.27
K2HT6OP1	0.43	0.51	0.74	0.45	0.55	0.72	K5HT6OP1	0.49	0.60	0.64	0.52	0.59	0.63
K2HT6OP2	0.34	0.45	0.80	0.35	0.43	0.87	K5HT6OP2	0.45	0.55	0.71	0.44	0.51	0.66
K2HT6OP3	0.36	0.45	0.80	0.34	0.42	0.89	K5HT6OP3	0.34	0.48	0.77	0.34	0.46	0.80
K2HT12OP1	0.40	0.53	0.72	0.34	0.45	0.78	K5HT12OP1	0.70	0.78	0.35	0.64	0.80	0.35
K2HT12OP2	0.43	0.56	0.69	0.45	0.54	0.80	K5HT12OP2	0.70	0.78	0.35	0.64	0.80	0.35
K2HT12OP3	0.45	0.59	0.65	0.40	0.46	0.81	K5HT12OP3	0.68	0.72	0.38	0.58	0.75	0.44
K2HT24OP1	0.57	0.82	0.34	0.50	0.63	0.13	K5HT24OP1	0.63	0.81	0.33	0.81	0.95	0.08
K2HT24OP2	0.58	0.80	0.34	0.41	0.52	0.12	K5HT24OP2	0.64	0.83	0.29	0.78	1.05	0.06
K2HT24OP3	0.58	0.80	0.34	0.41	0.52	0.13	K5HT24OP3	0.67	0.79	0.36	0.80	1.05	0.08
K3HT1OP1	0.19	0.24	0.95	0.18	0.23	0.93	K6HT1OP1	0.33	0.43	0.81	0.30	0.38	0.82
K3HT1OP2	0.21	0.28	0.93	0.21	0.27	0.92	K6HT1OP2	0.29	0.39	0.85	0.25	0.34	0.86
K3HT1OP3	0.19	0.25	0.94	0.20	0.25	0.93	K6HT1OP3	0.26	0.36	0.88	0.22	0.28	0.91
K3HT3OP1	0.60	0.73	0.46	0.55	0.67	0.45	K6HT3OP1	0.42	0.57	0.67	0.41	0.55	0.64
K3HT3OP2	0.60	0.75	0.46	0.56	0.69	0.45	K6HT3OP2	0.45	0.58	0.70	0.48	0.60	0.63
K3HT3OP3	0.47	0.56	0.74	0.43	0.53	0.73	K6HT3OP3	0.45	0.59	0.66	0.46	0.59	0.61
K3HT6OP1	0.71	0.87	0.33	0.62	0.87	0.52	K6HT6OP1	0.33	0.43	0.81	0.40	0.52	0.84
K3HT6OP2	0.58	0.69	0.54	0.50	0.62	0.58	K6HT6OP2	0.45	0.55	0.70	0.47	0.51	0.77
K3HT6OP3	0.58	0.73	0.50	0.61	0.78	0.35	K6HT6OP3	0.38	0.48	0.79	0.40	0.45	0.89
K3HT12OP1	0.48	0.59	0.65	0.36	0.49	0.24	K6HT12OP1	0.68	0.92	0.14	0.66	0.89	0.19
K3HT12OP2	0.41	0.50	0.75	0.38	0.47	0.37	K6HT12OP2	0.73	0.94	0.12	0.73	0.92	0.17
K3HT12OP3	0.40	0.50	0.75	0.35	0.47	0.27	K6HT12OP3	0.65	0.89	0.22	0.73	0.95	0.12
K3HT24OP1	0.42	0.52	0.73	0.46	0.54	0.37	K6HT24OP1	0.30	0.44	0.80	0.26	0.36	0.68
K3HT24OP2	0.53	0.63	0.63	0.40	0.56	0.10	K6HT24OP2	0.32	0.46	0.78	0.26	0.37	0.67
K3HT24OP3	0.42	0.52	0.72	0.37	0.51	0.37	K6HT24OP3	0.31	0.45	0.80	0.28	0.39	0.67

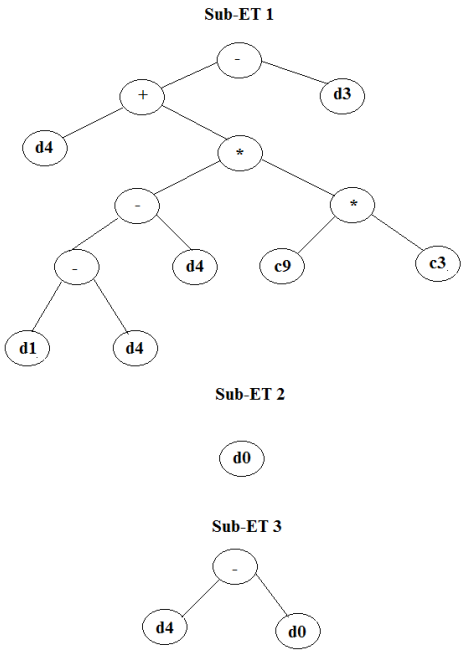


Figure A1- 3 genes (sub-expression tree diagram) linked together by addition function

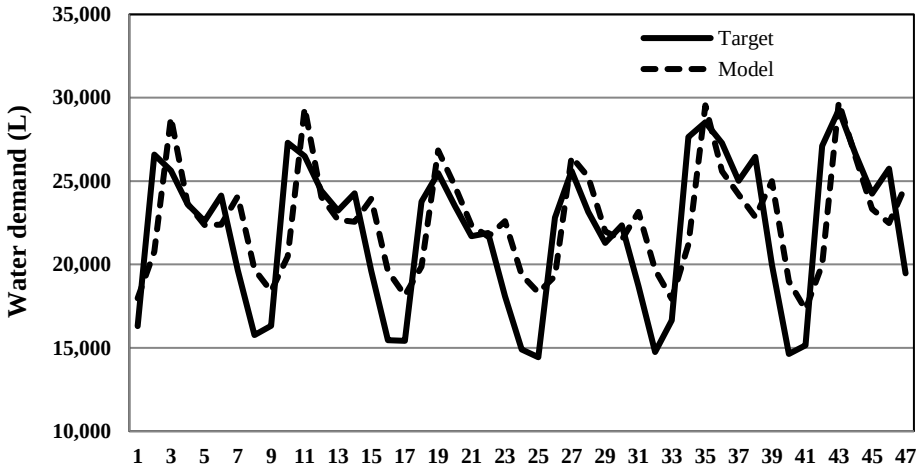


Figure A2- Target vs Model's prediction for the selected GEP model in chapter 6

Python Code for the selected GEP model:

Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts
Supplementary Materials: The following are available online at www.mdpi.com/link
Python Code for the Selected GEP Model

```
# Regression model generated by GeneXproTools 5.0 on 10/23/2017 12:24:57 PM
# GEP File: C:\Users\shabani\Desktop\Clustered + GEP Paper\ALL 90 models\K3HT1OP1.gep
# Training Records: 572
# Validation Records: 143
# Fitness Function: RMSE
# Training Fitness: 803.686978350874
# Training R-square: 0.940232394017262
# Validation Fitness: 809.891789934083
# Validation R-square: 0.931516795833398
#-------------------------------------------------------------------------
from math import *
def gepModel(d):
	G1C9 = -3.26853790974151e-02
	G1C3 = -3.18216498306223
	Standardize(d)
	y = 0.0
	y = ((d[4]+(((d[1]-d[4])-d[4])*(G1C9*G1C3)))-d[3])
	y = y + d[0]
	y = y + (d[4]-d[0])
	y = Reverse_Standardization(y)
return y
def Standardize(input):
	AVERAGE_0 = 22956.3688811189
	STDEV_0 = 5003.62843504768
	input[0] = (input[0] - AVERAGE_0) / STDEV_0
	AVERAGE_1 = 22964.636013986
	STDEV_1 = 4999.08753180465
	input[1] = (input[1] - AVERAGE_1) / STDEV_1
	AVERAGE_3 = 22978.497027972
	STDEV_3 = 4986.44287669382
	input[3] = (input[3] - AVERAGE_3) / STDEV_3
	AVERAGE_4 = 22980.8622377622
	STDEV_4 = 4983.37909553497
	input[4] = (input[4] - AVERAGE_4) / STDEV_4
def Reverse_Standardization(modelOutput):
	# Model standardization
	MODEL_AVERAGE = -1.03668530636382E-15
	MODEL_STDEV = 0.971247647537788
	modelOutput = (modelOutput - MODEL_AVERAGE)/MODEL_STDEV
```

```
# Reverse standardization
TARGET_AVERAGE = 22978.8772727273
TARGET_STDEV = 4985.99234334088
return modelOutput * TARGET_STDEV + TARGET_AVERAGE
```