INTEGRATED MANAGEMENT PLAN OF WATER DISTRIBUTION SYSTEMS: FORECASTING APPROACH

by

Peyman Yousefi

B.Sc., Islamic Azad University, 2009

M.Sc., University of Tabriz, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLAGE OF GRADUATES STUDIES

(Civil Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

January 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

INTEGRATED MANAGEMENT PLAN OF WATER DISTRIBUTION SYSTEMS: FORECASTING APPROACH

submitted by Peyman Yousefi in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Civil Engineering

Examining Committee:

Dr. Hadi Mohammadi, School of Engineering, UBC (Okanagan)
Supervisor

Dr. Gholamreza Naser, School of Engineering, Shippensburg University of Pennsylvania
Co-supervisor

Dr. Zheng Liu, School of Engineering, UBC (Okanagan)
Supervisory Committee Member

Dr. Sumi Siddiqua, School of Engineering, UBC (Okanagan)
Supervisory Committee Member

Dr. Jian Liu, School of Engineering, UBC (Okanagan)
University Examiner

Dr. Mohammad Reza Najafi, Department of Civil Engineering, Western University
External Examiner

Abstract

Having an integrated plan for the water distribution system (WDS) is a precondition to guarantee the safety of supply. Therefore, the development of such a plan can support the needs of WDS. Both integrated plans require highly accurate estimated values since they should not only cover failures in the system but also meet customers' demand. Relevant literature lists many mathematical and empirical models for modeling water consumption in the recent decade. Most of the methods in literature are on low-resolution monthly scales because of the limitations in data availability. Besides, increasing the resolution of data would decrease the errors of forecasting models. This study suggested techniques to model high resolution temporal scale and improvement of the models. To analyses the nature of data, average mutual information, phase space reconstruction (PSR), and correlation dimension; where all these methods provide identification of chaotic behavior. The results showed the existence of chaos in the data in different temporal scales with the value of 3.4. Then, non-linear local approximation method (NLA), artificial neural networks, genetic expression programming, and multiple linear regressions techniques are applied for short-term (daily) and mid-term (monthly) forecasting. NLA showed the best performance with the accuracy of 98% for the test period among all models. Then, the highly accurate model of each techniques, were selected to be combined with PSR and wavelet decomposition pre-processing to improve the accuracy of the models. The results presented 1.2% improvement by PSR techniques. For transition of different resolution, this study employed the continuous random cascade method to transfer the resolution. Random Cascade gave the daily value out of monthly scale by 90% accuracy. Finally, using expected consumption values resulted by forecasting models, an approach to make the leakage detection procedure resource-efficient is to estimate the timeline of leakage

event within the target district. The analysis of the value of consumption in two scenarios that were simulated in the target district related to consumption was provided using a detection classifier to investigate anomalies in the consumption patterns. This method detected the approximate time for the defined artificial leakage within the test case.

Lay Summary

Global water scarcity is becoming increasingly of concern due to climate change, urban development, population growth, industrial development, economic expansion, and the cost of drinking water. It is imperative that governments invest in integrated management plans that address the consequences of water problems such as scarcity of available water resources, enough allocation, and pipeline maintenance. Traditionally, the investigation of recorded data of previous consumption was enough to forecast the needs of the future. Climate change, urbanization, and socioeconomic variables highlight the need for accurate, sophisticated techniques that should be replaced with traditional methods. Machine learning and artificial intelligence techniques are the recently developed methods to cover the limitations of conventional techniques. This study proposed novel use of available techniques for reliable analysis and forecasting of water consumption. These models have the ability of anticipation for future changes within a complex water distribution system in order to pattern anomaly detection.

Preface

A version of chapter 2 has been accepted with minor revision for the publication as: Yousefi, P., Curtice, G., Naser, G., Mohammadi, H. 2019. Nonlinear Dynamic Modeling of Urban Water Consumption - A Chaotic Approach, Journal of Water MDPI. I developed the technique and wrote the manuscript. Which was further detailed by Greg Curtice and edited by Dr. G. Naser and Dr. H. Mohammadi.

A version of chapter 3 has been published as: Yousefi, P., Naser, G., Mohammadi, H. 2018. Surface Water Quality Model: Impacts of Influential Variables. Journal of Water Resources Planning and Management 144 (5), 04018015. I developed the methodology and wrote the manuscript and edited by Dr. G. Naser. In addition, another publication as; Yousefi, P., Shabani, S., Mohammadi, H., Naser, G. 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering 186, 544-550. I developed the technique and wrote the manuscript. S. Shabani worked on the draft and edited by Dr. Naser. I have cooperated as the co-Author for the published book titled water stress in plants by Intech. Shabani, S., Yousefi, P., Adamowski, J., Naser, G. 2016. Intelligent soft computing models in water demand forecasting. In water stress and plants. Intech. I defined combination of input variables for the models and cooperated in preparing the manuscript.

A version of chapter 4 published in book titled Wavelet Theory and Its Applications by Intech. Yousefi, P., Naser, G., Mohammadi, H. 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications, Intech. I developed the technique and wrote the manuscript and edited by Dr. G. Naser and Dr. H. Mohammadi. In addition, another publication as; Yousefi,

P., Shabani, S., Mohammadi, H., Naser, G. 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering 186, 544-550. I developed the technique and wrote the manuscript. S. Shabani worked on the draft and edited by Dr. Naser. In addition, another publication as; Yousefi, P., Naser, G., Mohammadi, H. 2018. Hybrid Wavelet and Local Approximation Method for Urban Water Demand Forecasting – Chaotic Approach. Published in WDSA / CCWI Joint Conference 2018. I developed the technique and wrote the manuscript and edited by Dr. G. Naser and Dr. H. Mohammadi.

A version of chapter 5 published in book titled Wavelet Theory and Its Applications by Intech. Yousefi, P., Naser, G., Mohammadi, H. 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications, Intech. I developed the technique and wrote the manuscript and edited by Dr. G. Naser and Dr. H. Mohammadi. In addition, another publication as; Yousefi, P., Naser, G., Mohammadi, H. 2018. Estimating High Resolution Temporal Scale of Water Demand Time Series – Disaggregation Approach (Case Study). HIC 2018. 13th International Conference on Hydroinformatics 3, 2408-2416. I developed the technique and wrote the manuscript and edited by Dr. G. Naser and Dr. H. Mohammadi.

A version of chapter 6 prepared to be submitted in Journal of Water Resources Planning and Management. Yousefi, P., Curtice, G., Naser, G., Mohammadi, H. 2019. Leakage detection in water distribution system by signal analytics approach. I developed the technique and wrote the manuscript. Which was further detailed by Greg Curtice and edited by Dr. G. Naser and Dr. H. Mohammadi.

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

1. Yousefi, P., Courtice, G., Naser, G., Mohammadi, H. 2019. Dynamic Modeling of Urban Water Consumption – A Chaotic Approach. Accepted with minor revision for publication in the Journal of Water.

2. Yousefi, P., Naser, G., Mohammadi, H. 2018. Surface Water Quality Model: Impacts of Influential Variables. Journal of Water Resources Planning and Management, ASCE 2018 Feb 22;144(5):04018015.
3. Yousefi, P., Shabani, S., Mohammadi, H., Naser, G. 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering. 186:544-50.
4. Yousefi, P., Naser, G., Mohammadi, H. 2018. Application of wavelet decomposition and phase space reconstruction in urban water consumption forecasting– Chaotic approach (Case study). Application of Wavelet Algorithm, InTech.
5. Yousefi, P., Naser, G., Mohammadi, H. 2018. Estimating High Resolution Temporal Scale of Water Demand Time Series – Disaggregation Approach (Case Study). 13th International Conference on Hydroinformatics, Italy 2018.
6. Yousefi, P., Naser, G., Mohammadi, H. 2018. Hybrid Wavelet and Local Approximation Method for Urban Water Demand Forecasting – Chaos Approach. 1st International Conference on WDSA / CCWI, Canada.
7. Yousefi, P., Curtice, G., Naser, G., Mohammadi, H. 2019. Leakage detection in water distribution system by signal analytics approach. To be submitted in Journal of Water Resources and Planning Management.

And, the related publications that I cooperated as the co-author:

1. Shabani, S., Yousefi, P., Naser, G. 2017. Support vector machines in urban water demand forecasting using phase space reconstruction. Procedia Engineering,186:537-43.
2. Shabani, S., Yousefi, P., Adamowski, J., Naser, G. 2016. Intelligent soft computing models in water demand forecasting. In Water Stress in Plants, InTech.

Table of Contents

List of Tables

List of Figures

List of Symbols

D_t : Vector of the consumption data at t time

$P(d_t)$: Marginal probabilities for measurements and joint probability

τ : Lag time

$C_m(r)$: Correlation integral value

X_i : The i^{th} state vector

C_e : Correlation exponent

m_{opt} : Optimum embedding dimension

X_j : Vector of dimension

w_{ij} : Weigh in i and j dimension

$\emptyset$: Transfer function

θ : Threshold limit

λ_{max} : Largest Lyapunov exponent

$\left|u_{Y_{n0}}\right|$: The number of neighbors to point Y_{n0}

Imp_j : The relative importance

β : Exponent of power spectrum

List of Abbreviations

ACF: Auto Correlation Function

AI: Artificial Intelligence

AMI: Average Mutual Information

ANN: Artificial Neural Networks

ARIMA: Auto Regressive Integrated Moving Average

ARIMAX: Auto Regressive Integrated Moving Average with explanatory variable

CE: Correlation Exponent

CM: Correlation Dimension

db: Daubechies

GA: Genetic Algorithm

GEP: Gene Expression Programming

GP: Genetic programming

IVS: Input Variable Selection

MLP: MultiLayer Perceptron

MLR: Multiple Linear Regression

NLA: Non-linear Local Approximation

PSR: Phase Space Reconstruction

SCADA: Supervisory Control and Data Acquisition

SLP: Single Layer Perceptron

SVM: Support Vector Machines

TF: Transfer Function

WDS: Water Distribution Systems

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Hadi Mohammadi for his continuous support of my Ph.D. study.

I would like to thank my co-supervisor, Dr. Bahman Naser for his guidance, mentorship and above all his friendship since I have worked under his supervision.

Besides my supervisors, I would like to thank my thesis committee, Dr. Zheng Liu and Dr. Sumi Siddiqua for their insightful comments and encouragement.

I am grateful to my lovely parents, who have provided me through moral and emotional support in my life.

Also, I would like to thank City of Kelowna Utility to provide me with the requested information for my research.

Last but by no means least, I would thank to Dr. Yousef Hassanzadeh and Dr. Mohammad A. Ghorbani for their supports, and to my friends, Dr. B. Mohajer, Dr. V. Pourmostaghimi, Dr. M. K. Brar, Mrs. H. Asadi, Mr. M. Amirrahmani, Mr. H. Ghannad and Mr. A. Yaghouti. Thanks for all your help and encouragement!

Dedicated to my parents.

Chapter 1: Introduction

Climate change significantly affects the water availability all over the globe. This effect plays a crucial role in arid and semiarid regions. On the other hand, urban development, population growth, industrial development, and economic expansion also critically increase the water scarcity concerns worldwide. The availability of fresh potable water has huge impacts on society. Therefore, the authorities must be prepared in advance for relevant consequences. Moreover, water conflicts are more severe than any other natural resources scarcity, which are prevalent around the world (e.g., oil lands), especially in arid and semiarid regions that are suffering from water resources limitation. Moreover, such conflicts are becoming more prevalent and severe due to rapid increases in demand, while reliable sources of fresh water remain unchanged (if not depleted). Table 1.1 shows the approximate number of water-related conflicts around the world chronologically (data extracted from the Pacific Institute at www.worldwater.org). The Table indicates a large increase in total water related conflicts from 20 and 170 in the 19th and 20th centuries to 476 within the last two decades (https://www.worldwater.org/water-conflict/). Therefore, it is crucial to have an integrated plan to manage the allocations and consumption efficiently, and above all to have a long/short-term plan that is based on the future information of the system.

Table 1.1 The summary of water conflicts around the world before 0 BC to present.

Dates	Years	Conflicts			Total Conflicts	Conflict/Year
		Casualty	Trigger	Weapon		
Before 0 BC	NA	4	0	23	27	NA
0 - 1799	1800	1	2	9	12	0.007
1800 - 1899	100	8	8	4	20	0.200
1900 - 1999	100	78	41	51	170	1.7000
2000 - 2019	19	232	188	56	476	23.800

The reliability of water distribution systems (WDS) can be improved through accurate simulation of hydraulic conditions in pipeline systems based on future values of water consumption. In other words, water consumption forecasting provides suppliers with the necessary future demand information to ensure consumption needs can be met [1,2]. Water consumption forecasting is a dynamic process as forecasting are essential for optimum operation, as well as sustainable growth and development of urban water supply systems [3].

Although recent studies have improved the understanding of the nonlinearity and complexity of water consumption variables, still more research is required to highlight the impact of water consumption variables on the accuracy of consumption models to forecast the future values. Accurate estimation and forecasting data are influenced by availability of high-resolution temporal scale datasets. Additionally, availability of other influential factors, such as holidays, humidity, peak hour consumption, air temperature, population growth, and consumers' income, are important to accurately forecast the consumption data. Moreover, forecasting the consumption in short-, mid-, and long-term time (e.g. less than a week, a week to a month, a month to a year or more) play a crucial role in making the operation of WDS more efficient by an integrated plan. Various factors are essential in developing an integrated plan including optimized pumping, pipeline maintenance, minimizing energy and water supply cost, improving system reliability, and the quality of allocated water [4–6].

Over the past three decades, two groups of deterministic and probabilistic methods have been proposed to forecast urban water consumption. The deterministic approach is solely based on input variables and their initial conditions, whereas, a probabilistic model relies on modeling uncertainties and randomness of input variables. Researchers have confirmed the superiority of nonlinear deterministic approaches over probabilistic ones [7,8]. Individuals' consumption habits

are important factors in modeling and forecasting the future consumption. However, there is a lack of detailed research about the importance of the individuals' consumption habits in the forecasting models.

Consumer habits are not commonly considered in the total recorded consumption that is used for modeling. Nonetheless, an individual's habits influence the consumption, and thus, the future values in the system will be highly influenced by individual consumer's habits, resulting in a highly complex system. In this complex system, "the behavior of the integral part of the system is simple, but the behavior of the overall system is complex" [9]. Given the significant challenges and complexity of probabilistic methods and knowing the fact that deterministic methods can provide a useful approximation to their probabilistic counterparts, this research has focused on a deterministic approach to forecast the short- and the mid-term water consumption.

Majority of well-known methods need to use high-resolution temporal scales of data, for reliable estimation of water consumption. Nevertheless, only the well-equipped supervisory control and data acquisition (SCADA) systems can record such a high-resolution data (e.g., high-resolution temporal scale, proper time series of pressure values in pipelines, etc.). Therefore, development of a method to transfer the low-resolution scales (e.g., annual and monthly values) to a high-resolution scale (e.g., hourly or daily events) is indispensable. A reliable consumption management plan for an urban area requires understanding of the information about the behavior of data series and the consumption pattern in different temporal resolutions [10,11]. The high-resolution temporal scale of data series gives the information that is required to understand the behavior of the dataset in more detail. Having the details of the dataset offers models with higher accuracy to estimate the goal values. Accurate estimation of the consumption values plays a crucial role in developing maintenance plans in WDS. For example, estimation of the future values of

consumption in a specific time, will give the minimum requirements of WDS to achieve the consumers satisfaction. Furthermore, these plans help municipal authorities to overcome the probable failures such as bursts, and leakages within the system to aid in WDS's maintenance [12,13]. These integrated plans for the efficient WDS utilize interpolation and extrapolation modeling methods based on the SCADA records. Therefore, high-resolution temporal scale records have a linear relationship with the efficiency of the WDSs plans. This research proposes a new technique to transform temporal scales in reliable approximate accuracy. The proposed technique allows for the transformation of data from a low-resolution temporal scale to a high-resolution set.

Online monitoring requires the water consumption indicators to be constantly observed. Consequently, it can be resource-demanding and impractical in certain cases. Mathematical modeling, in contrast, is less demanding and often the best tool. However, the models must be calibrated by independent reliable sets of field data in order to guarantee their accuracy. Thus, the availability of high-resolution dataset is essential for accuracy of the model. Further, modeling can be computationally demanding and inefficient depending upon the type and number of variables under consideration. A comprehensive model is often mathematically complex and requires a rich knowledge that may be unavailable particularly in small water districts [14]. Using their local knowledge, such districts prefer employing a reliable predictive model that only requires basic and commonly available data resources. It can be shown that, increasing the number of input variables in a model does not necessarily improve its accuracy [15]. Therefore, application of input variable selection techniques, in combination with pre-processing techniques, will improve the accuracy of models while reducing the complexities due to high number of input variables. Given the significant challenges and complexity of probabilistic methods and the fact that pre-processing

methods can provide a useful approximation to their probabilistic counterparts, this research also proposes the application of pre-processing and input variable selection methods to make the forecasting models more efficient, improve the accuracy, and reduce the complexity. The research will focus on the dynamic characteristics of water demand in the city of Kelowna (British Columbia, Canada) as a test case.

1.1 Water Consumption Modeling and Forecasting

Information about future water consumption helps the authorities to develop an integrated and efficient plan to mitigate (if not eliminate) the water-related stresses for the suppliers and consumers. Recent evidence shows an exponential growth in peak and average consumptions, which overburden the water storage systems [2,16]. Therefore, efficient demand management and forecasting are required to solve this problem. Accurate estimation of drinking water demand and availability of resources is considered as one of the solutions to reduce the water stresses [17–21]. Figure 1.1. shows the approximate number of publications related to water consumption modeling and forecasting.

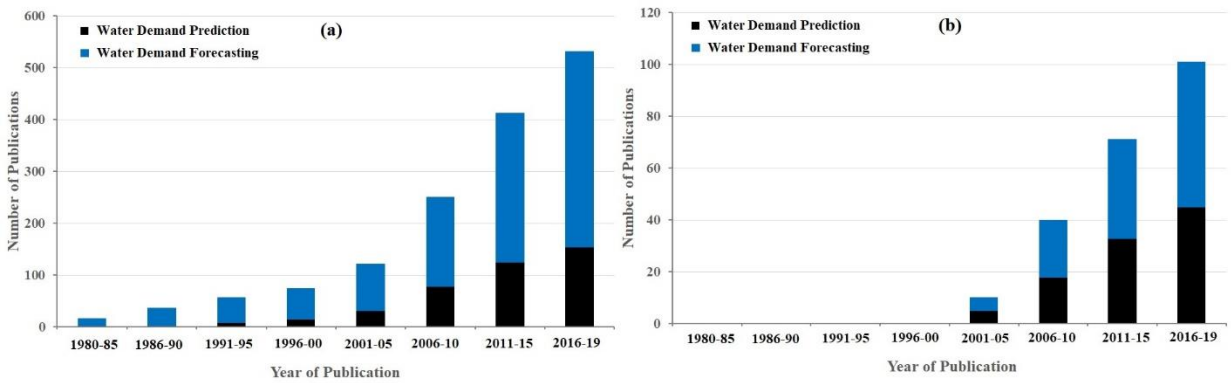


Figure 1.1 Published papers with titles including (a) Water Demand Prediction/Forecasting (b) Water Consumption Prediction/Forecasting based on Google Scholar and Web Science data base.

To collect the information shown in Figure 1.1, “Publish and Perish V6.4” is used to approximate the number of studies with the different keywords in their titles. Since, studies in the area used both forecasting and prediction for consumption and demand, different keywords are considered to extract the information related to the studies. Figure 1.1 shows the number of publications with the keywords including “water demand forecast”, “water demand prediction”, “water consumption forecast” and “water consumption prediction” up to date based on “Cross Ref” and “Google Scholar”. It should be noticed that “Web of Science” is not considered as a source in this stage. As it is shown in Figure 1.1, the number of studies has been increasing with a considerable rate. However, there is not a significant difference between consumption and demand in the literature, to the best knowledge of the author “Consumption” is suitable for the related titles. Since the forecasting models in the literature are developed based on the past recorded data of the consumption values, the mentioned data are registered based on allocation and consumption while the term “Demand” is related to an estimated based model created from statistics and population information. Therefore, this study considers the keyword of “Consumption Forecasting” to present the methodologies and results. Moreover, the rate of increase in the number of publications shows the availability of dataset in the area. As it is shown in Figure 1.1, water consumption forecasting has become an interesting field of study to researchers after the year 2000. One reason for that is the availability of data from well-equipped WDS such as the counters and gages that register the value of consumption within a period. This evidence can be another reason why the term “Consumption” is a better choice rather than “Demand”. As the dataset used for the studies are more relevant to the numbers given by equipment that records and registers the value of allocations and consumptions. Table 1.2 shows the statistics about the publications that are shown in Figure 1.1, in a specific period.

Table 1.2 Statistical information of publications in the field of water demand/consumption prediction/forecasting from 1980 to the present.

	Water Demand Prediction	Water Demand Forecasting	Water consumption Prediction	Water consumption Forecasting
Publication Year	1980-2019	1980-2019	2001-2019	1992-2019
Citation Years	39	39	18	27
Papers	154	377	45	56
Citations	983	4929	409	228
Cites/year	25.2	126.3	22.72	8.44
Cites/paper	6.18	13.07	8.52	3.62
Authors/paper	2.72	2.73	2.71	2.76
h-index	13	34	8	8
g-index	28	63	20	12
hI, norm	10	23	6	4
hI, annual	0.26	0.59	0.33	0.15

The variables affecting the consumption values can be categorized into two groups: climatic (e.g. temperature, relative humidity, rainfall, etc.) and socioeconomic (e.g. population and income) [13,22]. Commonly climatic variables are used in short- and mid-term forecasting, while socioeconomic variables are used for long-term demand predictions [23,24]. Common climatic factors considered in the literature are temperature, precipitation, and previously recorded demands [25–27]. While the literature provides a rich list of studies for short- and mid-term demand forecasting, a limited number of researches have focused on the impact of climatic factors on demand forecasting [28–30].

Literature enlists the implementation of different statistical and probabilistic approaches for consumption forecasting. Conventional techniques were reported prevalent for a better understanding of chosen variables in the modeling of water consumption [31–33]. This type of

modelling considers certain linear relationships among the functional variables and the value of the water consumption, though the observations indicate that these relationships exhibit nonlinear behaviors. The literature is mostly categorized into physical-based and black-box models. Physical-based models approximate the general internal sub-process and physical mechanism by fundamental laws of mass, energy, and momentum. Black-box models implement artificial intelligence, fuzzy based and nonlinear deterministic methods (e.g., artificial neural networks, heuristic algorithms such as genetic programming, support vector machine, nonlinear local approximation, etc.) to ascertain the relationship between the input and output variables.

Recent studies on physical-based and black-box models include conventional regression models [34], artificial neural networks (ANN) [16,25,35,36], feedforward neural networks (FNN) [26,37], general regression neural networks (GRNNs) [38], deep belief neural network (DBNN) [39], support vector machines (SVMs) [22,40–43], gene expression programming (GEP) [44,45], adaptive neural fuzzy inference system (ANFIS) [46], Fourier analysis [4], hybrid models (e.g. combined wavelet) [27,47,48], fuzzy regression [49], fuzzy cognitive map learning method [50,51], epidemiology-based forecasting framework [52], temporal disaggregation [53], harmonic analysis [54], and wavelet de-noising [55]. Table 1.3 briefly shows the most recent published related works in the area with higher publication indices in comparison with the relevant researches in the field.

Table 1.3 shows that the temporal scales for the forecasting horizon is mostly considered as a short/mid-term time period. While other hydro-informatics studies (e.g. rainfall runoff, sedimentation, climate, etc.) are equally spread over short/long-term horizon, the complexity of consumption data caused weaker accuracy in long-term. Therefore, most studies calibrated forecasting models in short-term time period, whereas the models estimated acceptable fitness

values. The question is “how we can investigate the complexity of the data before calibrating the models?” and “how we can improve the accuracy of the models in long-term period?”. This research studies the possible answers for the mentioned questions.

Table 1.3 Noted relevant literature chronologically in water consumption forecasting methods and periods.

Method of Purpose	Determinants	Temporal Scale	Horizon Category	Reference
DEMC	Historical Data	Hourly	Short-Term	[56]
ARIMA, ANN	Historical Data	Yearly	Long-Term	[57]
ELM, ANN	Historical Data	Daily	Short-Term	[58]
Hybrid Model	Historical Data	Monthly	Short-Term	[59]
Moving Window	Historical Data	Daily, Weekly	Short-Term	[60]
MCP	Historical Data	Hourly, Daily	Short-Term	[60]
Support Vector Regression	Historical Data	Monthly	Short-Term	[61]
ANN, SVR, ELM, MLR	Historical Data	1-,3-Day	Short-Term	[62]
SVM, Clustering	Historical Data	Daily/Hourly	Short-Term	[63]
Chaos, ANN	Historical Data	Daily	Short-Term	[64]
ANN, SVM	Historical Data	Daily	Short-Term	[39]
Tree based methods	Historical Data	Monthly	Short-Term	[65]
Trend and Harmonic Analysis	Historical Data	Hourly	Short-Term	[54]
ANN	Historical Data	Daily	Short-Term	[66]
Review Article				
Review Article				
Review Article				
Review Article				

1.2 Problem Statement

Growth in peak water demand, which overburdens the urban water resources, requires an efficient management plan. This important challenge motivated the application of soft-computing techniques into urban water demand forecasting methods in order to develop methods that are applicable in solving the forecasting problems. Many techniques have been proposed to forecast water demand under differing time scales. However, there has been a limited investigation on performance comparisons among the models, to assist in selecting the best model under various conditions [67]. Moreover, non-stationary, non-linear, and inherent stochasticity of water demand data make the forecasting problems more challenging in this field [68]. Donkor et al. has reported that periodicity and forecasting horizon influence the performance of the methods in short- and long-term forecasting, such as artificial neural networks and econometric models, respectively [69]. Therefore, understanding the forecasting horizon, which is dependent on the dataset considered, will help to categorize the performance of the developed models. It can be considered as a classification for short-, mid-, and long-term forecasting models, which help the operators to select the appropriate model for implementation, considering the available dataset [20,25,70]. Forecasting horizon can improve the reliability of demand-forecasting methods, through identifying the most useful method for demand forecasting under specified time frames. In general, however, there is currently no acknowledgement of a time frame for these forecasting horizons [69]. This study focuses on the forecasting horizon for the considered models to inform performance under various periodicity (i.e. daily, 2-day, 4-day, weekly, bi-weekly, and monthly) and lead time.

Based on previous studies, the common influential variables in water demand forecasting include; observed consumption values (e.g. historical recorded consumption values over various time

period), climatic variables (e.g. temperature, humidity, and levels of snow and rainfall) and socioeconomic variables (e.g. population growth rate, and the economic factors such as income and water cost) [67]. Although these variables are used in the literature, there is limited study in using socioeconomic variables in long-term forecasting studies. Moreover, extrapolation models in water consumption forecasting are based on the recorded data and its error terms [69]. The limitation of these models is based on their dependence of past trends that will likely be observed in the future, but do not consider the important roles the socioeconomic variables and the consumer habits can play in influencing the future consumption values. Recently-published studies investigating these factors are limited and require 1) accurate estimation and forecasting of water consumption, considering different periodicity and forecasting horizon to classify the models based on their performance, and 2) determination of the degree of nonlinearity among the influential variables to consider exogenous variables (e.g., socioeconomic). Therefore, such a complex system requires a method which considers the mentioned variables by investigating the dataset. Additionally, the application of the methods in different periodicity and lead time should be clarified regarding the models in their forecasting horizon.

1.3 Research Objectives

The overall objective of this research is to detect anomalies in consumption pattern caused by leakages and failures within the target district in WDS. To reach the goal, a reliable estimation of expected values of the water consumption in future is crucial. It will assist water authorities detecting likely anomalies in consumption values within the system (with the reasons of leakage, pipe breaks, etc.). The long-term objective will be achieved through the following phsases, as indicated below:

1) Phase I – data collection and analysis: This step will investigate the accuracy of the dataset that is collected from municipalities and provide techniques to generate the possible missing data.
2) Phase II – modeling the future value of water consumption: The aim is to develop novel models to forecast short- and long-term water consumption.
3) Phase III – accuracy improvement of the selected models: This step aims to improve the accuracy of selected forecasting models. Combination of wavelet decomposition and input variable selection techniques will be applied to pre-process the input variables.
4) Phase IV – anomaly detection in WDS: This step applies the concept of failures detection to interperet the anomalies of consumption information within the system.

The research applies the City of Kelowna as the test case.

1.4 Thesis Structure

Figure 1.2 depicts the structure of this dissertation including the following five chapters:

1) Chapter 1 provided an overall introduction about the importance of water resources. Also, introduction about the impact of forecasting models and their roles in investigation of water consumption management plans in the field of urban water management.
2) Chapter 2 investigated the reliability of the recorded data as the case test. Phase space reconstruction and dynamic investigation of explanatory variables and classification of the dataset were explained in this chapter. The existence of chaos and forecasting horizon for the test case was investigated.
3) Chapter 3 discussed four methods were applied to define forecasting models. These include non-linear local approximation (NLA), multilayer perceptron artificial neural network (MLP-ANN), gene expression programming (GEP) and multiple linear regression (MLR)

methods. The performance of the models was investigated with and without PSR in different lead time.

4) Chapter 4 applied pre-processing methods to improve the performance of selected models in previous chapter. Wavelet decomposition and PSR are the techniques that were applied for accuracy improvement of the selected models.

5) Chapter 5 evaluated the performance of disaggregation techniques in transferring the temporal scales from low resolution to high resolution temporal scale. Non-linear deterministic and random cascade models were used to transition. Moreover, the application of disaggregation methods in accuracy improvement of the models were investigated.

6) Chapter 6 employed the models output in anomaly detection of water consumption values within WDS and failure detection. Two scenarios were defined based on artificial leakage. Selected forecasting model was used to estimate the expected value of consumption for short-term future. Then, wavelet analysis was employed to detect the anomalies caused by different scenarios.

7) Chapter 7 summarized the results of the research as conclusion and suggest future work along with this study.

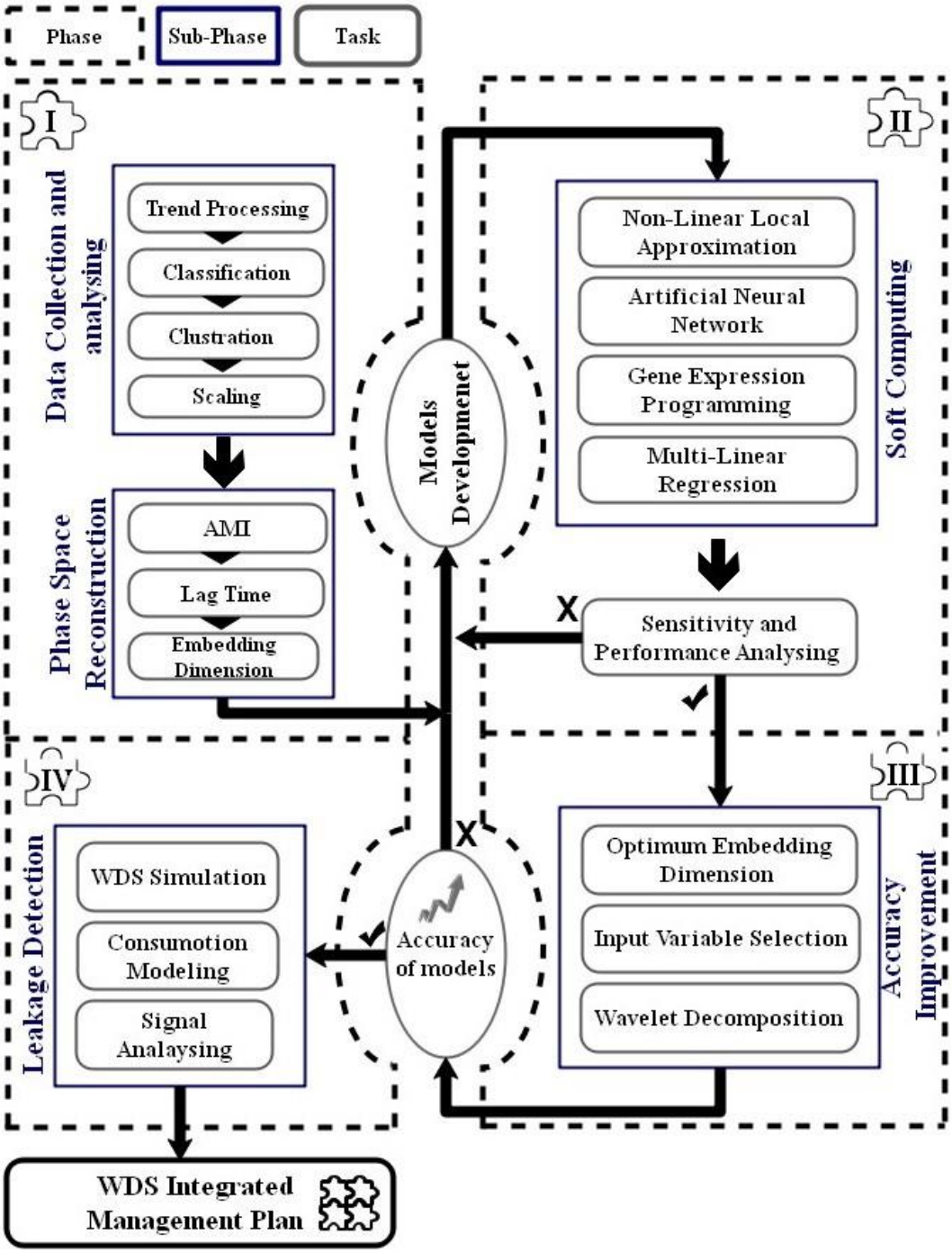


Figure 1.2 Brief scheme of the thesis structure.

Chapter 2: Nature of Explanatory Variables in Forecasting of Consumption[1]

2.1 Overview

Numerous variables have been used in water consumption modeling within the water distribution system (WDS). The nature of these variables is not similar. Therefore, the information about the dynamic of the nature variables that are used in modeling is beneficial. The main benefit of this information is to improve the performance of the models (e.g. accuracy). In forecasting models, factors such as accuracy, reliability, and flexibility are essential in the evaluation of the models' performance.

Regarding to the literature, the focus on improving the models' performance is mostly related to pre-processing techniques. Therefore, there is a limited number of studies in the investigation of the dynamic of consumption dataset in WDS modeling techniques. This chapter overviews investigating the existence of chaotic behavior in the case data. Followed by, the information about phase space reconstruction (PSR), false nearest neighbor and correlation dimension.

2.2 Background

Active variables in complex water consumption systems may benefit from chaos theory's techniques to improve forecasting models. Dimensional techniques based on chaos approach enhance the application of forecasting models to estimate future values in complex systems and improve the forecasting of nonlinearity within dynamic systems [9,71]. Descriptions of chaotic behavior have been used in various engineering applications. Chaotic systems are defined and characterized by significant changes in behavior resulting from small changes in initial conditions

[1] A version of this chapter has been accepted with minor revision for publication in the Journal of Water as a full paper: Yousefi, P., Curtice, G., Naser, G., Mohammadi, H. 2019. Nonlinear Dynamic Modeling of Urban Water Consumption - A Chaotic Approach, Journal of Water MDPI.

of the system [72]. Using the 'extent of complexity' of a chaotic system (defined primarily in the context of the variability of relevant data), chaotic systems are classified as low-, medium- or high-dimensional [73]. Also, the availability of noise in time series increases the complexity of the data and results in a high embedding dimension [74].

In water distribution systems, factors such as temperature, humidity, precipitation, the economic condition of consumers, population size, holidays, etc. have an impact on water consumption. By aggregating data at increasing temporal scales, the effects of scaling time series on deterministic chaos can be found [75,76] and any missing data can be generated [77,78]. This approach has been used in solving problems in various fields of study such as river discharge [79–81], sedimentation [82–85], climate [86], lake level variability [87], rainfall [88–91], traffic speed [92], finance [93], image processing [94] and ship motion prediction [95] and existents of chaos in hydrological variables [96]. However, there is a paucity of study on the application of chaos theory on water consumption modeling techniques. Oshima [9] developed a real-time forecasting model system for water consumption based on chaos theory. To support operations with labor saving, facility maintenance and efficient use of energy, he introduced "information integration type chaos theory-based demand forecasting." The report showed the application of chaos-based techniques in improving the accuracy of the results. Recent studies in water consumption forecasting primarily used various hydrological variables in modeling water consumption for different proposes. Therefore, this study investigated the impact of chaotic behavior of consumption data in models.

2.3 Methodology

2.3.1 Phase Space Reconstruction

This research employed the concept of phase space recounstruction to better understand the dynamic nature of a municipal water consumption dataset for the City of Kelowna. The dynamics

of a water consumption system are represented by data points along a trajectory, whereby each position in time represents a system state. The lag-embedding technique can be used on deterministic, dynamic systems such as the present water consumption dataset to reconstruct phase-space from time series. The fundamental dynamics of a system can be studied by reconstructing an *m*-dimensional phase-space of D_t that is defined by [97,98]:

$$D_t = \left\{d_t, d_{t-\tau}, d_{t-2\tau}, \dots, d_{t-(m-1)\tau}\right\},\ t = 1,2,3, \dots, N \tag{2.1}$$

where D_t is a vector of the consumption data of $\{d_t\}_{t=1,\dots,\ N}$, N is the number of recorded consumption data points, τ is the lag time, and m is the number of embedding dimension that in this chapter generally varies from 1 to 20 [82,84,87,99]. In the case when m is greater than the minimum embedding dimension, the trajectory of reconstructed vectors can display the true state of the chaotic system. Often, the lag time (τ) is arbitrary as the data are assumed to have infinite precision. The lag time should not be too small given the difference between various elements of the delay vectors; and, it should not be too large as this can result in low coordinate correlation [100]. If the dynamics of the system can be reduced to a set of deterministic laws, trajectories will converge towards a subset of the phase-space with a fractional dimension called the attractor [99]. The lag-embedding method is sensitive to both embedding parameters of τ and m. Average mutual information (AMI), is a well-known method for estimating the lag time [101]. To estimate the lag time, this research employed AMI:

$$I(d_t, d_{t+\tau}) = \sum_{d_t} \sum_{d_{t+\tau}} P(d_t, d_{t+\tau}) \log_2 \frac{P(d_t,\ d_{t+\tau})}{P(d_t)P(d_{t+\tau})} \tag{2.2}$$

where the sum is extended over the total number of samples in the time series, $P(d_t)$ and $P(d_{t+\tau})$ are the marginal probabilities for measurements, d_t and $d_{t+\tau}$ and $P(d_t, d_{t+\tau})$ is their joint probability.

The optimal τ minimizes the value of the function $I(d_t, d_{t+\tau})$ for $t = \tau$. AMI considers the first local minimum as the lag time [102]. In addition to AMI method, autocorrelation function (ACF) is considered as another well-known method in literature to estimate the lag time [101]. More details about ACF and different functions are presented in chapter 3. This chapter employed AMI in the process of investigating the availability of chaotic behavior in consumption time series. To compare the impact of both AMI and ACF in models' results, in chapter 3, both AMI and ACF are considered.

2.3.2 Contribution of Chaos in the Problem

The term "chaos" is defined as turmoil and disorder. Therefore, chaos theory studies systems that at first sight, seem to have random and stochastic behavior, while specific rules govern these systems. Such systems are susceptible to initial conditions, that apparent imperfect and arbitrary inputs can extensively affect them. Integrating chaos theory is not only to focus on the irregularities in outputs of another system or attributed random sources but also to analyze very determined but nonlinear and chaotic dynamics of the system itself. Then, the chaotic data can be categorized between regular and predictable signals, or irregular and random signals. Chaos theory is mainly based on the fact that an order is hidden in each irregularity. Chaotic processes are intrinsically definite. In addition, chaos theory is the study of unstable and non-periodic behavior in the nonlinear dynamic systems. The scientific method, which is provided by this theory for us, is the change of aspect in look at the events so as their structural order can be discovered. Of course, a new view of this reasoning behind the order also challenges many traditional controversies about order proof, etc., in the philosophy. As an example, the result of a coin toss at a time is random and indefinite, since it has local amplitude. However, the expected results of the phenomenon are permanent and predictable when it is repeated many times. Edward Lorenz arranged twelve

equations that are followed the probabilistic weather forecasting to simulate with computer. Once for securing the accuracy of the result, he re-entered the data into the computer and obtained a different result. He saw that the data was rounded when entered into the computer and instead of considering data with 6 decimal places; data was considered with 3 decimal points, i.e., data were not similar to previous data, and there were partial differences. For example, the difference in the response of two near changes in the input, their behavior is similar at first, and it is entirely different at the end [103]. From the aspect of water distribution systems, many variables are involved in the water consumption value, which may look random at first sight in different periods (peak hours, local party nights, holiday nights, etc.). The chaotic investigation can study the behavior of these variables. Therefore, having information about the nature of different variables, give insight into whether these variables are influential explanatory for models' output. Regarding the nature of the variables, potential variables play a beneficial role in models' performance to accurate simulations. Although the increase of the accuracy is necessary, it can strikingly decrease unpredicted costs in long term prediction for exact management in networks designing and proper strategies of urban development, which nowadays is crucial in the field of water engineering.

2.3.3 Correlation Dimension

The dimension of a system designates its complexity and indicates the number of required variables that specify a deterministic system. Kermani classified different dimensions in a system including topological, Hausdorf, box counting, point-wise, and correlation dimensions [80]. Additionally, the correlation dimension is used as an indicator of a deterministic or stochastic process [85,104]. The below function calculates the correlation integral value [105].

$$C_m(r) = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=1}^{N} H(r - \|X_i - X_j\|) \quad (2.3)$$

here, N is the number of data points, where H is the Heaviside step function ($H(u)$ =1 for $u > 0$, $H(u) = 0$ for $u \leq 0$ and $u = r - \|X_i - X_j\|$, X_i is the i^{th} state vector, r is the radius of a sphere with the content of X_i or X_j as the center. $C_m(r)$ is proportional to r for stochastic time series, whereas for chaotic time series it scales with r as:

$$C_m(r) \propto r^{c_e} \quad (2.4)$$

where c_e is the correlation exponent defined by approximating the slope of $\log C_m(r)$ versus $\log(r)$ in logarithmic scale. If the calculated c_e is unchanged by increasing the number of embedding dimensions, c_e can be considered as the correlation dimension of the attractor in the system. But, if c_e is not stable as a function of embedding dimensions, this system can be considered non-chaotic [75,106,107].

2.3.4 False Nearest Neighbor

False nears neighbor method obtain the optimum embedding dimension to reconstructing the phase space [108]. The neurorthid of the points embedded in increasing n-dimensional manifolds, the method eliminates false neighbors. I other word, the points are apparently laying close together because of separated projections in higher number of dimensions. If the ratio of false neighbors between dimensions of m+1 and m gets below 5% (a given threshold), the next higher number of embedding dimensions would be the optimum one ($m' > m+ 1$) [85].

2.4 Case Study

Water consumption modeling relies upon many different factors including population size, climatic, hydrologic, economic, and customers' consumption patterns. Forecasting future values

of consumption is one of the goals in this research to reach the objective. Consequently, consumption time series data are used to train the forecasting models. National offices have provided the collected data of the test case. The consumption pattern in different temporal resolution are studied. The water supply for the region comes from various resources, which include Lake Okanagan, Mission Creek, Mill Creek, Scotty Creek, Hydraulic Creek and numerous wells [109]. There is one primary distribution system that services 99% of the population via Poplar Point, Eldorado (seasonal intake uses only utilized during peak consumption time period) and Cedar Creek pump stations [109], in addition to, plus Swick road pump station which services approximately 300 residents. This study considered Poplar Point, Eldorado and Cedar Creek stations. Table 1.A shows the summary of water production and consumption rate since 2012 to 2016. The information in the appendix A is reported by Kelowna City Utility as "Annual Water and Filtration Exclusion Report" in July 2017 [110]. The SCADA software obtains information remotely from sensors installed at the intake locations. The software platform facilitates tracking of historical system performance, for auditing and future decision making, to optimize the system [109]. Table 2.1 shows the characteristics of the dataset in the test case.

Table 2.1 Characteristics of the water consumption values in test case.

Property	**Daily**	**2-Day**	**4-Day**	**7-Day**	**14-Day**	**Monthly**
Number of Data	2186	1092	552	312	156	72
Max. value (m^3)	114597.2	210740.3	410428.3	656173.6	1255211	2475026
Min. value (m^3)	14124	31477.3	69655.5	124112.9	252704.1	557066.8
Average (m^3)	43046.4	86102	170332.4	301357.3	602714.7	1291944
Standard deviation (m^3)	20074.5	39897	79304.3	136626.8	268733.2	552701.5
Coefficient of variation	0.46	0.46	0.46	0.45	0.44	0.42
Skew	0.73	0.71	0.72	0.66	0.63	0.54
Kurtosis	-0.38	-0.45	-0.51	-0.63	-0.79	-0.91

The data is synchronized in 1-, 2-, 4-, 7-, 14-days and monthly temporal resolution. Also, all dataset is divided into two groups as the training and the test period. The nature of data fed to define the models are studied in all temporal scales. This study used six temporal scales of the consumption, including daily, 2-,4-,7-,14- and 30 days since January 1st, 2010 to December 30th, 2016.

2.5 Results and Discussion

Figure 2.1 presents the time variation of water consumption dataset for a six-year period (from January 2011 to December 2016), and the average 24-hour consumption based on City of Kelowna Utility report. The data was parsed into hourly maximum, minimum, and mean values for a 24-hour period (Figure 2.1b). Figure 2.1b also presents a boxplot of the total mean, minimum, and maximum consumption value distributions for the six year period. Figure 2.1b indicates that the average consumption and minimum values have low frequency, demonstrating the highly deterministic behavior of the data. consumption estimation and forecasting of the average and minimum values are not as complex as compared to the peak consumption values. However, the maximum values exhibit a non-linear behavior which does not follow a specific pattern (i.e. appears to be random), unlike average and minimum values. Further, the maximum consumption values in a water distribution system are very important. It is because of estimation of peak consumption to supply customers' consumption, and optimize WDS pipeline to make WDS more reliable such as; managing pipeline failures, improving peak pressure, reducing leakage, etc.

Average mutual information (AMI) was used to identify the proper lag time in this study. Figure (4a) presents the first local minimum lag time of τ = 17 days for daily time series as a function of lag time. The first minimum values of AMI were at lag times of 17, 12, 10, 6, 3 and 2 for daily, 2-, 4-, 7-, 14- and 30-days (monthly) data series of the water consumption, respectively. Table 2

summarizes the results for all other timescales. By using the lag time of τ = 17 days, the phase-space was reconstructed, as presented in Figure (4b) for daily scale. The vertical axis shows the time series and the other two axes show the same time series delayed by 17-(τ) and 34-days (2τ). This figure shows reconstructions in three dimensions; the projection attractor on the plane $\{x_i, x_{i+\tau}, x_{i+2\tau}\}$ with the lag time of 17 days. When comparing the two PSR graphs in Figure (4b), the presence of attractors becomes clear for τ = 17 days (black line) comparing to τ = 1 (blue dots). Indeed, the phase space is more spread out for 1-day lag time and more concentrated on attractors for 17-days lag time. Based on the figure, PSR can make a consistent set of inputs for the model which makes better results in comparison with other sets having less consistency (e.g., peak consumption represented in figure 3b). The results for other resolutions are available in the appendix (I).

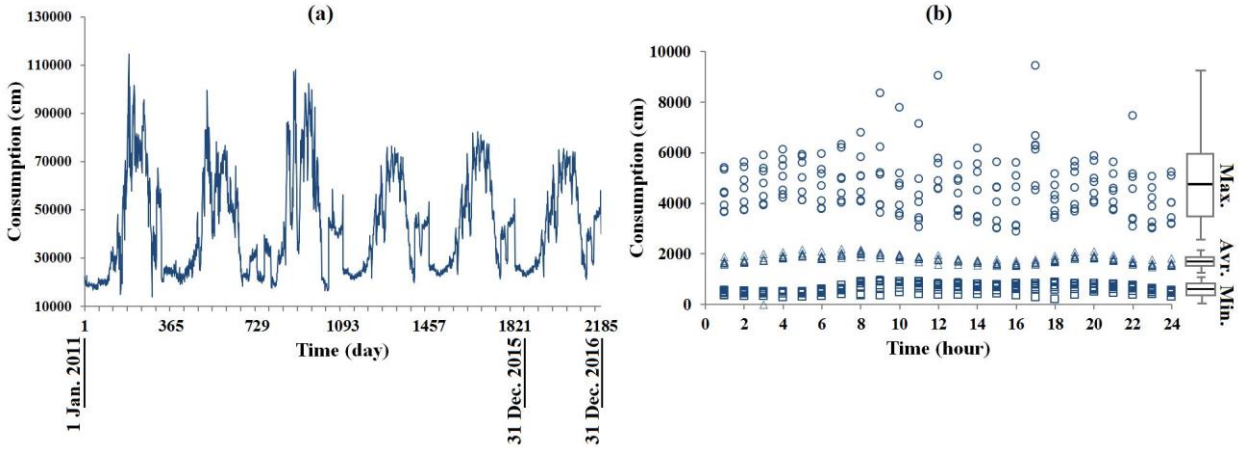


Figure 2.1 Time series plot of (a) daily water consumption; (b) 6-years consumption pattern within a 24-hour period.

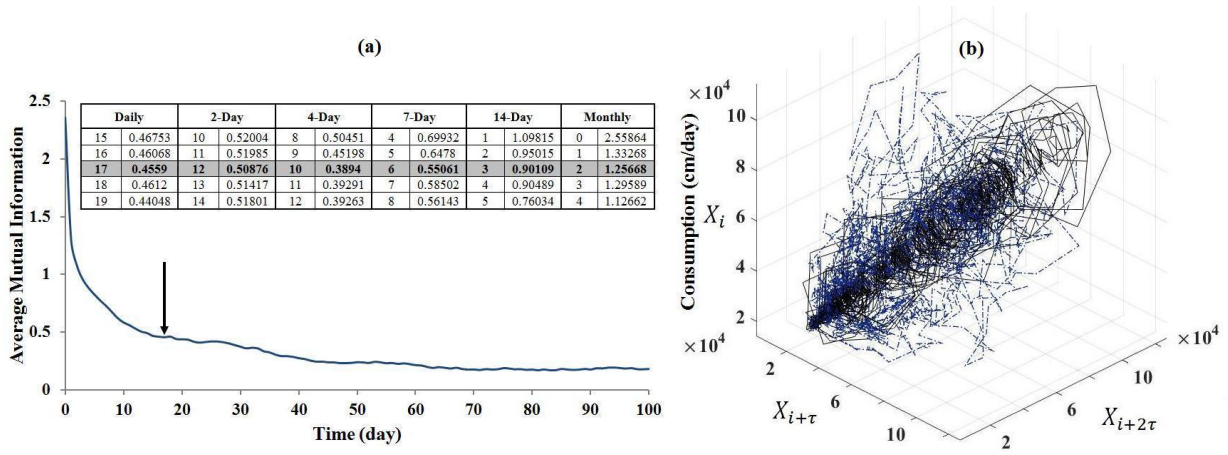


Figure 2.2 (a) Average mutual information (τ); (b) reconstructed phase space by (τ and 2τ -day lag time).

Figure 2.3 plots the results for correlation function versus log (r) for different embedding dimensions (m) varies in the range of 1 to 20 for two temporal scales of daily (a) and 2-days (b).

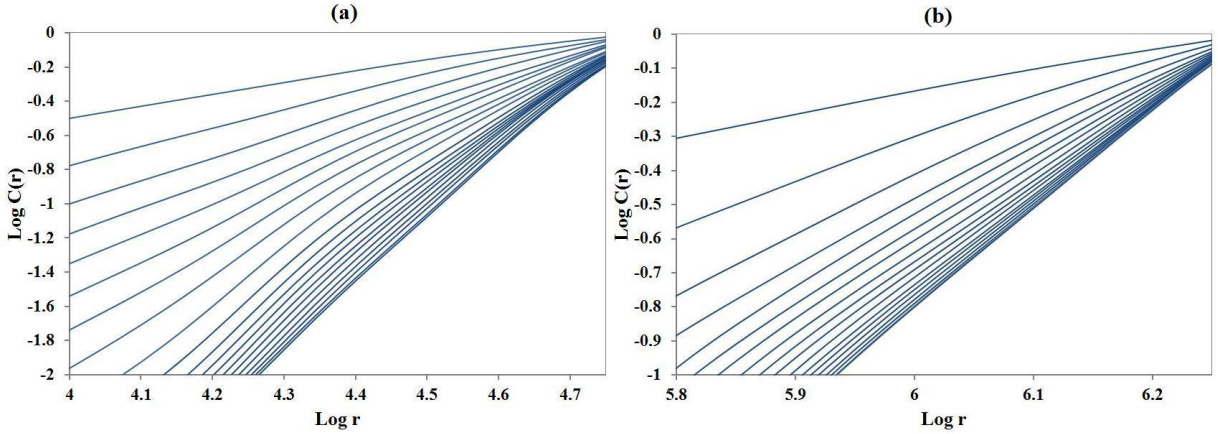


Figure 2.3 The relation between correlation function C(r) and r by different embedding dimensions, (a) daily; (b) 2-days.

Figure 2.4a reveals the value of false nearest neighbor (FNN) at the first local minimum of m = 17 for daily time series. m=17 is considered as the optimum embedding dimension (m_{opt}) for the daily temporal resolution. The first minimum values of FNN were at embedding dimension of 17, 17, 9, 6, and 4, for daily, 2-, 4-, 7-, 14- and 30-days (monthly) dataset, respectively.

To investigate the availability of chaotic behavior, the correlation exponents, C_e (m), were determined. The results are shown in Figure (2.4b). As the figure shows, the correlation exponent increases with the embedding dimension up to a certain value and then remains steady. The figure reveals that the slope of larger (m) would become constant for all temporal scales. The saturation of the correlation exponent in a specific embedding dimension, is an indication of the presence of chaos in the dataset. The correlation exponent was saturated approximately after the embedding dimension of 16 for all temporal scales.

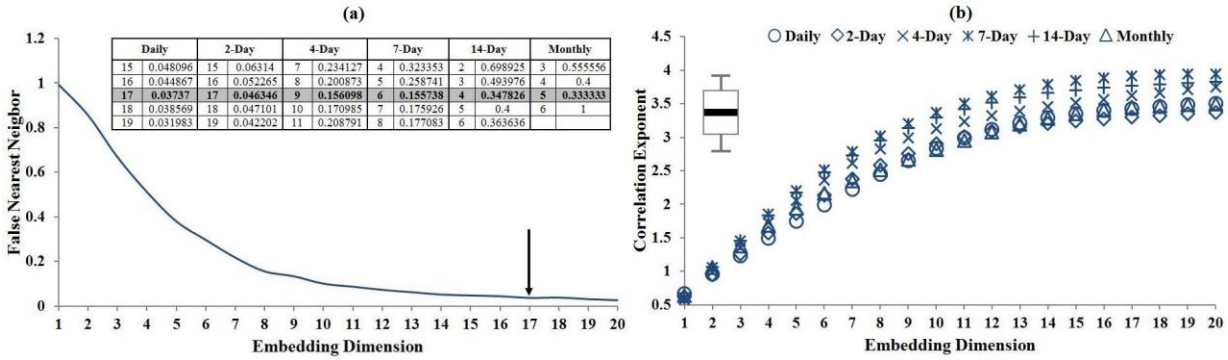


Figure 2.4 (a) False nearest neighbor values of embedding dimensions for daily temporal scale; (b) saturation of correlation dimension C_e (m) with embedding dimension m for different temporal scales.

The results of correlation exponent and FNN showed that the 17th embedding dimension is potentially the optimum embedding dimension of the system. The results of correlation exponent and FNN in the temporal scales for more than 4-days are not the same. Thus, it is evident that having a similar result by both methods clarified that the 17th embedding dimension is the optimum dimension in daily temporal scale. Also, the saturation and constant slope of correlation exponent for all temporal resolution revealed the high dimensional chaotic behavior of the total system. No study investigates the chaotic behavior of a total system by studying all temporal scales from high resolution to low-resolution temporal scale. Many studies that investigated the chaotic behavior of

a system concluded that high-resolution timescales have chaotic behavior, meanwhile the low-resolution timescale of the same time series does not.

Regarding the fractal law, it can be a debatable issue whether a chaotic system should have chaotic behavior in all temporal scales, though this study would bring this question into the area about the issue. This study would not conclude the issue because of the difference in the dynamic nature of the data set in this case study. Moreover, Regonda et al. suggested that to determine the chaotic behavior of a time series; it is better to investigate more than one temporal scale [75]. For the moment, investigation of the chaotic behavior in different time scales of the test case of this study is sufficient for following the objectives of this research. However, further researches are needed to make a reasonable conclusion about the issue, whether suggestions of Regonda et al. [75] and this study are encouraging to investigate the chaotic behavior of different time scales of a test data.

Back to the result of this study, the embedding dimension of 17 seems to be sufficient to explain the dynamic of the system (1- and 2-day consumption). The saturation values of the correlation exponent for all temporal resolutions are 3.5, 3.37, 3.74, 3.94, 3.83, and 3.49 for daily, 2-, 4-, 7-, 14- and monthly, respectively. Low correlation exponents obtained for all the temporal scales' series reveal that the dynamic behavior of the consumption values may be in low dimensional deterministic. Regarding the concept of PSR and correlation exponent, the nearest above integer value of correlation exponent indicates the number of dominant variables in the system. Therefore, it can be interpreted as the number of variables that are dominantly governing the temporal dynamics of water consumption are about 4 for all temporal scales. Table 2.2 shows the results of lag time and correlation exponent for all the temporal resolutions. The condition of $C_e < 2\log N$ (with N as the number of data) was satisfying for the chaotic temporal scales [111].

Table 2.2 Average mutual information and correlation exponent values in different temporal resolutions.

Time Scale	Average Mutual Information	Correlation Exponent	2 Log N > D_2
Daily	17	3.50	6.67
2-Day	12	3.37	6.07
4-Day	10	3.74	5.48
7-Day	6	3.94	4.98
14-Day	3	3.83	4.38
30-Day	2	3.49	3.71

2.6 Summary

Becoming steady value for the correlation exponent (no changes were shown in the slope of graph) in a certain embedding dimension (the dimension that the slop of graphs became steady) is an indication of chaotic behavior in all six temporal scales. The correlation exponent became steady after an embedding dimension of 17 for the temporal scales. The correlation exponent results showed the 17^{th} embedding dimension may be considered the optimum embedding dimension for this system; however, we have conducted additional investigation to better understand the optimum embedding dimension for the test case. Moreover, the saturation and constant slope of correlation exponent for all temporal scales revealed a highly chaotic behaviour exhibited by the whole system. Many studies that investigated the chaotic behaviour of the system concluded that high-resolution timescales demonstrate chaotic behaviour, while the low-resolution timescale of the same time series does not exhibit chaotic behaviour. Regarding the fractal law, it remains uncertain whether a chaotic system should have chaotic behaviour in all temporal scales; we speculate this is an important factor to consider under this application. Based on the presented results, this study suggests that to investigate the availability of chaos in any dataset, investigation for different temporal scales is needed to indicate if the system has chaotic

behaviour, as this conclusion supports the findings presented by other researchers. Considering the paucity of studies available for comparison, the present investigation of chaotic behaviour in different time scales using the test case provides enough data to evaluate the objectives of this research. However, additional study is required to make definitive conclusions regarding the findings of this study and more generally the chaotic behaviour of consumption data. Nevertheless, the study findings provide encouraging evidence to further investigate the chaotic behaviour of water consumption values over different time scales of a dataset.

Chapter 3: Soft-Computing Methods in Forecasting of Water Consumption[2]

3.1 Overview

Factors like urban area development, population growth, and industry expansion are also variables which are useful in supplying water resources. Therefore, it is necessary to understand the relationship between available water resources and the influential variables for assisting the governments in developing long-term plans for water-related problems (drinking water) in the period. Future consumption values are necessary for developing a practical and dynamic management plan for an efficient operation of water distribution system (WDS) to supply proper drinking water. The consumption forecasting is categorized as long-term (up to 25 years), mid-term (up to 2 years) and short-term (up to 2 days) based on essential factors like water supply plan, pipeline maintenance, and water distribution system optimization [5].

3.2 Background

Many variables are affecting the forecast of consumption values, but not as active variables in other hydrological forecasting problems (e.g., river discharge, sedimentation, rainfall, temperature, etc.). In context to literature, commonly used input variables to forecast are; temperature, humidity, precipitation, and recorded consumption series [13,26,27,112,113].

Different deterministic and probabilistic techniques are being listed in the literature to forecast drinking water consumption. Conventional methods, autoregressive integrated moving average (ARIMA), autoregressive integrated moving average with explanatory variable (ARIMAX),

[2] A version of this chapter has been published as papers: Yousefi, P., Naser, G., Mohammadi, H. 2018. Surface Water Quality Model: Impacts of Influential Variables. Journal of Water Resources Planning and Management 144 (5), 04018015 and, Yousefi, P., Shabani, S., Mohammadi, H., Naser, G. 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering 186, 544-550 and, Yousefi, P., Curtice, G., Naser, G., Mohammadi, H. 2019. Nonlinear Dynamic Modeling of Urban Water Consumption - A Chaotic Approach, accepted with minor revision to be published in Journal of Water MDPI.

artificial neural networks (ANNs), support vector machines (SVM), gene expression programming (GEP), neuro-fuzzy systems and hybrid techniques are commonly used to forecast drinking water consumption values [15,33,114–118].

Chaos approach is successfully employed to understand the dynamics of a non-linear system. The deterministic nature of a chaotic system is dependent on initial conditions and leads to entirely different behaviors for the next time slots. Non-linear local approximation method (NLA), is one of the techniques in this section to forecast water consumption values. NLA was applied in different hydrological studies, however, there is a paucity of studies on the application of chaos theory on water consumption forecasting methods. Oshima [9] developed a real-time forecasting system for water consumption based on chaos theory. The report showed the application of chaos-based techniques in improving the accuracy of the results. Yousefi et al. [13] studied the dynamic of explanatory variables by the chaos approach for urban consumption time series. They studied the different temporal resolution of the time series for a better understanding of the connections among the resolutions in a chaotic time series. Yousefi et al. [119] investigated the performance of NLA in water consumption forecasting. The results of their study showed that NLA has better performance compared to artificial neural networks and gene expression programming. Moreover, they concluded that phase space reconstruction (PSR) and pre-processing techniques (e.g., wavelet decomposition) improve the performance of models in forecasting a chaotic dataset and increase the forecasting horizon. Beside the advantages of NLA, the limitation of this method is employing univariate input variable.

Evolutionary techniques have become popular in modeling and optimizing fields. Genetic programming (GP) and gene expression programming (GEP) are included in the category of heuristic algorithms that are found in Darwin's evolution theory. Evolutionary techniques adjust

the population of a specific solution. In each stage, individuals are selected randomly from the current population. Then, the chosen individual plays the parent's role to be reproduced for another population for the next generation of solutions. This reproduction goes toward an optimal answer which has been defined as the goal values. The evolutionary techniques are used for optimization problems where standard methods are not suitable, such as discontinuity, nondifferentiable, stochastic, or highly nonlinear objective functions. Among evolutionary techniques, GP and GEP are used for modeling problems with the same abilities as genetic algorithm (GA). For example, informing data gaps and forecasting time series [22,120–123] used the GP technique to model suspended sediment load in the Tongue River (United States) and reported better performance from GP compared to sediment rating curves and multiple linear regressions (MLR). Since GEP provides a tree-structure scheme, it makes GEP more convenient to interpret the results in comparison with GP. Moreover, GEP presents mathematical equations that clarify the relationship between input and output variables by a factor of 100-10000 [124–126]. The superiority of GEP and the advantages of this technique interested researchers to develop more sophisticated models with hybrid methods, such as combining the extended Kalman filter [127], clustering the consumption values [128], Wavelet decomposition [113], and phase space reconstructed GEP (PSR-GEP) [13] in forecasting urban drinking water consumption. The results showed that GEP models are highly sensitive to wavelet decomposition when attempting to improve the performance of the models. The GEP models can be extracted based on the defined arithmetic operations. It is one the advantages of GEP.

Among the variety of examined methods, Artificial Neural Networks (ANNs), have been applied to the various period in the wide variety of hydrological issues. The main reason of ANNs frequent usage is its ability to overcome the relationship in determining the complexity of time series, even

with the limited data available to train the models. Therefore, most of the studies applicable in areas of water resource demands applies ANNs to forecast short, mid and long-term consumption values [27,129,130].

ANN offers several advantages and disadvantages. Advantages include; 1) ANN often requires less formal statistical training, 2) ANN implicitly detects complex nonlinear relationships among input and output variables, 3) ANN evaluates all possible interactions among effective variables, 4) ANN is computationally fast and requires fewer input variables in comparison with other data-driven methods, 5) ANN has powerful pattern classification and recognition capabilities [15], and 6) ANN is a self-learning technique suitable for predicting future states of a system [131]. Whereas disadvantages, ANN requires a large set of data for training and it is susceptible to overfitting. In contrast to multilayer perceptron (MLP) ANN, single layer perceptron (SLP) ANN often produces unreliable results due to the stochastic nature and the complexity of chemical and biological processes in water, and high variance and inherent non-linear relationships among influential variables [132]. This research employed the well-known MLP-ANN for the modeling of consumption values.

NLA, GEP, MLP-ANN and MLR are the methods that were used in this research. Each method is classified in different modeling techniques categories. To select the model with the highest performance, wide variety of techniques were compared. Therefore, the present study employed these four techniques for the modeling phase. MATLAB V2019 was used for running the models to extract the results.

3.3 Methodology

3.3.1 Non-linear Local Approximation

Non-linear local approximation (NLA) can forecast a system's future without development of an analytical model [103]. This research applied NLA to (I) test the chaotic nature of consumption data and (II) forecast the consumption in the case study. The results for the first section are reported in chapter 2. It was concluded that the consumption dataset for the city of Kelowna has chaotic behaviour for all temporal resolution with the optimum embedding dimension of 17 for the daily and 2-day resolution. This was done by reconstructing the phase space of the dataset. It is critical in chaos analysis to reconstruct the multi-dimensional phase space that provides a conceptual pattern for the time series. The presence of attractors in phase space indicates the possibility of chaos in the data set. A phase-space reconstruction in a dimension *m* facilitates an interpretation of the underlying dynamics in the form of an *m*-dimensional map, f_T, by:

$$X_{j+T} = f_T(X_j) \tag{3.1}$$

where X_j is the vector of dimension *m* at the current state of the system at time *j* and X_{j+T} is the vector of dimension *m* at the future state of the system at time $j+T$. NLA entails the subdivision of the f_T domain into many subsets. In other words, the dynamics of the system was described step-by-step locally in the phase-space [133]. In an *m*-dimensional space, estimating the change of trajectory with time would lead to forecasting. Considering the relation between two states $\vec{X}_t$ and $\vec{X}_{t+p}$, the behaviour at a future time (*p*) on the attractor was forecasted by the mapping $\vec{F}$ as [99]:

$$\vec{X}_{t+p} \cong \vec{F}(\vec{X}_t) \tag{3.2}$$

where the evolving dynamic of state $\vec{X}_t$ is influenced by nearby states. The future state $\vec{X}_{t+p}$ was determined by the first-order polynomial mapping $\vec{f}$ [134]:

$$\vec{X}_{t+p} \cong \vec{F}(\vec{X}_t) = \vec{a} + \vec{f}(\vec{X}_t, \vec{X}_{t-\tau}, \dots, \vec{X}_{t-(m-1)\tau}) \tag{3.3}$$

while the mapping $\vec{f}$ is linear, the forecasted value is nonlinear [135]. This is because every state on the trajectory belongs to a different subset defined by different expressions for $\vec{F}$.

3.3.2 Gene Expression Programming

Evolutionary soft computing techniques may be applied to solve engineering problems. Among these evolutionary techniques, genetic algorithm (GA), genetic programming (GP), and gene expression programming (GEP) are considered, inspired from Darwin's theory of evolution [136–139]. GA and GP work with a string of numbers and a specific length that is known as a "chromosome." GEP defines an equation that shows the relationship between input and output values. Moreover, the process of the algorithm's learning starts with the generation of chromosomes for a given random raw dataset that works with chromosomes and expression trees. Expression trees demonstrate the relationship among variables connecting with arithmetic operators. Chromosomes contain genes that provide a series of symbols of two parts, head and tail, which have functions, terminals, and only terminals, respectively. GEP follows the genome restructuring process by mutation, recombination, transposition, and gene duplication randomly. This random reputation will deliver the best model to be selected. This cycle will be continued by the reproduction of randomly generated chromosomes to reach the model with satisfying results based on defined evaluation criteria. In this research, 30 chromosomes with the head size of 8 and 3 genes are reproduced with the linking functions of $\{+, -, \times, x, x^2, \sqrt{x}, \log x\}$. All models are

selected with a fair assessment by considering stopping condition of 5000 generations. Figure 3.1 shows the scheme of the process with GEP method.

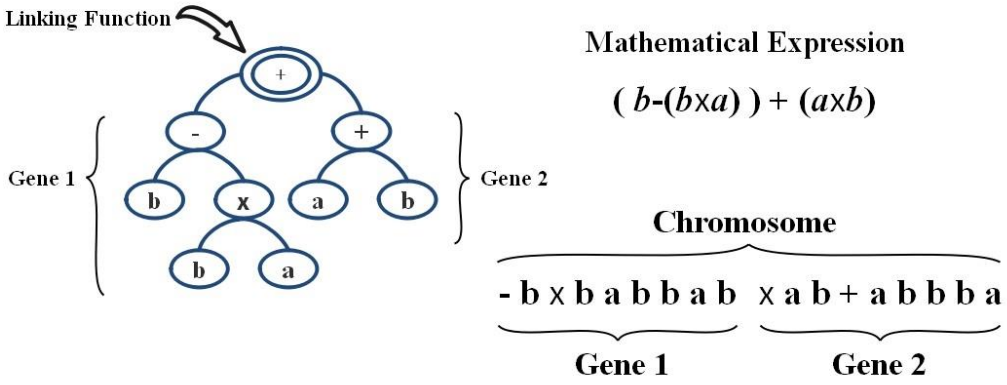


Figure 3.1 Expression tree and mathematical function of two gene chromosome.

3.3.3 Multi-Layer Perceptron Artificial Neural Network

This research employed multi-layer perceptron (MLP) artificial neural network (ANN), including input, hidden and output layers (Figure 3.2). The neurons and hidden layers are the most fundamental parts of ANN. In general, the number of neurons in the input and output layers vary with the number of the variables of the system under study. The numbers of neurons in the input and output layers are equal to the numbers of input and output variables, respectively.

In hidden layers, each neuron calculates the sum of the values (x_i) with weight (w_{ij}) and uses a transfer function ($\emptyset$) to determine the output signal (u_j). Figure 3.3 shows a neuron model with forecasting of consumption values as the output (O_j) defined by Equations (3.4) and (3.5):

$$u_j = \sum_{i=1}^{p} w_{ij} x_i \tag{3.4}$$

$$O_j = \emptyset(u_j - \theta_j) \tag{3.5}$$

where θ is a threshold limit [140–142].

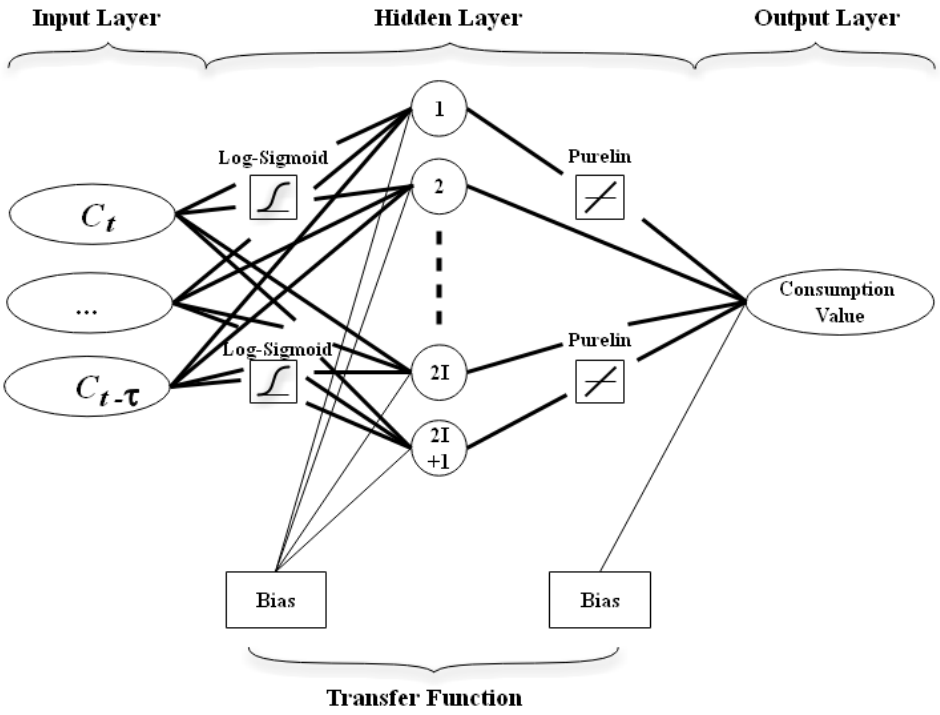


Figure 3.2 Simple configuration of multilayer perceptron neural network.

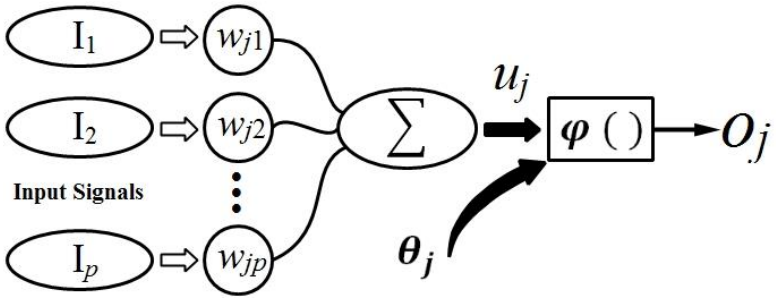


Figure 3.3 Nonlinear model of a neuron.

The transfer function is often monotonically increasing, continuous, differentiable, and bounded. Theoretically, the transfer function may take any form including sigmoid shape, piecewise function, step function, linear function (Purelin), and non-linear function. The logistic sigmoid and Purelin transfer functions are common in the literature [140–143]. This research studied a large number of input variables. Thus, the research applied no transfer function at the input layer in order to make the forecasting less computationally demanding. While the research applied all the above

transfer functions at the output and hidden layers, the logistic sigmoid and Purelin transfer functions provided the most accurate predictions. Back propagation learning technique was used to compare the predictions with measured data and to compute bias with (3.6) that was then fed back through the network [144].

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(R_i - F_i) \tag{3.6}$$

where N, R_i and F_i are the number of observations, recorded values and forecasted values, respectively. Note that error is any difference between the predicted values, and the measured (true) values. While error makes up all imperfections and faults in results, bias is the error that is systematic in nature [142,143]. Representing the mean of all individual errors, bias reveals how over- or under-estimated the predictions are [144].

3.3.4 Multiple Linear Regression

When multiple linear regression (MLR) corresponds to a linear combination of the components of multiple signals x (e.g. Recorded consumption value, delayed consumption, or combination of them) to a single output signal y (water consumption) by:

$$y = b + \sum_{i=0}^{N} a_i x_i \tag{3.7}$$

where x_i is the defined input (reconstructed phase space for consumption values) and a_i is regression coefficient determined by the least square method with the residual r defined by:

$$r = y - a_1 x_1 - a_2 x_2 - \cdots - b \tag{3.8}$$

3.3.5 Largest Lyapunov Exponent

The rate of convergence or divergence in different dimensions is measured using the Lyapunov exponent. A deterministic system contains at least one positive Lyapunov exponent [145]. This research employed the approach proposed by Rosenstein et al. [145] to extract the value of Lyapunov exponent for the test case. After reconstruction of phase-space, a point Y_{n0} was selected and all the points in the neighborhood of Y_n, with the closer distance of r to that point were found. r is the radius of a sphere with the content of Y_n as the center. The procedure was repeated for N points in the route to find a stretch factor S:

$$S = \frac{1}{N}\sum_{n_0=1}^{N} \ln\left[\frac{1}{\left|u_{Y_{n0}}\right|}\sum|Y_{n0} - Y_n|\right] \tag{3.9}$$

where $\left|u_{Y_{n0}}\right|$ is the number of neighbors to point Y_{n0}. The plot of S verses N consists of linear and nonlinear components. Literature provides two methods for calculating the largest Lyapunov exponent (λ_{max}). Shang et al. [146] determined λ_{max} as the slope of the linear part of the curve, while Rosenstein et al. [145] determined λ_{max} as the average of the slope of the first part and that of the second part [146,147]. This research applied the second approach to study different temporal scales. The forecasting horizon (Δt) is a time in which the consumption dataset sustains its dynamics in the most accurate forecasting; Δt was obtained as the inverse of the largest Lyapunov exponent:

$$\Delta t = \frac{1}{\lambda_{max}} \tag{3.10}$$

3.3.6 Evaluation of Model's Performance

This research measured the models' accuracy by coefficient of determination (CD), root mean squared error (RMSE), and mean absolute error (MAE) defined as:

$$CD = \left[\frac{\sum_{i=1}^{N_t}(R_i - \bar{R})(F_i - \bar{F})}{\sqrt{\sum_{i}^{N_t}(R_i - \bar{R})^2}\ \sqrt{\sum_{i}^{N_t}(F_i - \bar{F})^2}}\right]^2 \tag{3.11}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{N_t}(R_i - \bar{R})\,(F_i - \bar{F})}{\sqrt{\sum_{i}^{N_t}(R_i - \bar{R})^2}\ \sqrt{\sum_{i}^{N_t}(F_i - \bar{F})^2}} \tag{3.12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_t}(R_i - F_i)^2}{N_t}} \tag{3.13}$$

$$MAE = \frac{1}{N_t}\sum_{i=1}^{N_t}|R_i - F_i| \tag{3.14}$$

where N_t is the number of the recorded consumption values in the dataset, R and F are the recorded and forecasted values of water consumption, respectively. $\bar{R}$ and $\bar{F}$ are the mean of the recorded and forecasted consumption values, respectively. Note that the range of CD is between 0 and 1 with higher positive values indicating better agreement. The range of CC is -1 and 1, where larger positive value of CC and a lower value of RMSE and MAE indicates better agreement between the observed and forecasted values. It should be noticed that to evaluate the performance of the models, only three of the fitness criteria is considered. The results are based on three different publications, which in each of the different criteria is used to select the models.

3.4 Data Information and Test Case

3.4.1 Understandings

Unlike natural water resources like rainfall, the lower percentage of drinking water, which is changed to wastewater after use, goes back to the cycle. Water pressure in a pipeline, water quality, supply peak consumption time, pipeline maintenance, maintenance cost, specialist and educated human resources, pipeline failure management, etc. are the variables that should be under control at the same time. Also, to develop an integrated long-term plan, the availability of resources is crucial. Therefore, knowing about the value of consumption in a specific period is the first step for any management plan beyond urban drinking water supply and allocation systems. This chapter investigates the first step of every long-term plan development in urban drinking water, as discussed below. Water utility management needs drinking water long-term forecasted values in several terms. 1) water distribution network design; 2) supply and consumption management; 3) efficient application of distribution network; 4) pipeline pressure management; 5) network development; 6) optimizing the cost of water supply and network maintenance.

3.4.2 Study Area

This section selected water consumption of the City of Kelowna (BC, Canada) as the test case. The City of Kelowna water utility provides services for approximately 79000 residents in water district boundary [110]. Poplar Point, Eldorado, Cedar Creek, and Swick Road pump stations cover services for 99% of the population of the area [109]. However, few areas in the boundary are named as “Future City” which do not contain any population yet, but land development plan shows water servicing is considered in the area. Monitoring of water quality, the operation of the pumps, water level in reservoirs, and pipeline pressure are conducted by the use of Supervisory Control and Data Acquisition system (SCADA).

3.4.3 Review of Data Records

Hourly water consumption for the stations, as mentioned above, has been made available by the city utility of Kelowna. The data used a time period of six years (approximately 52,464 hourly consumption) starting on January 1st, 2011 to December 30th, 2016. Concerning the six years water consumption samples of daily scale (2186 points), the first five years (1882 points) are used for calibrating the models, and the last year (365 points – 2016) is considered as the test period. Also, cross correlation and K-fold methods were used for training and testing splitting periods. Since, the periodicity of the time series are quite steady, it did not affect the results of models in selection of the best modeling technique. Therefore, to make consistence condition for selection phase, the last years was considered as the test period for the modeling phases. Table 3.4 shows the characteristics of the dataset in the test case for the daily temporal resolution that has been applied to evaluate the performance of models in the methodology.

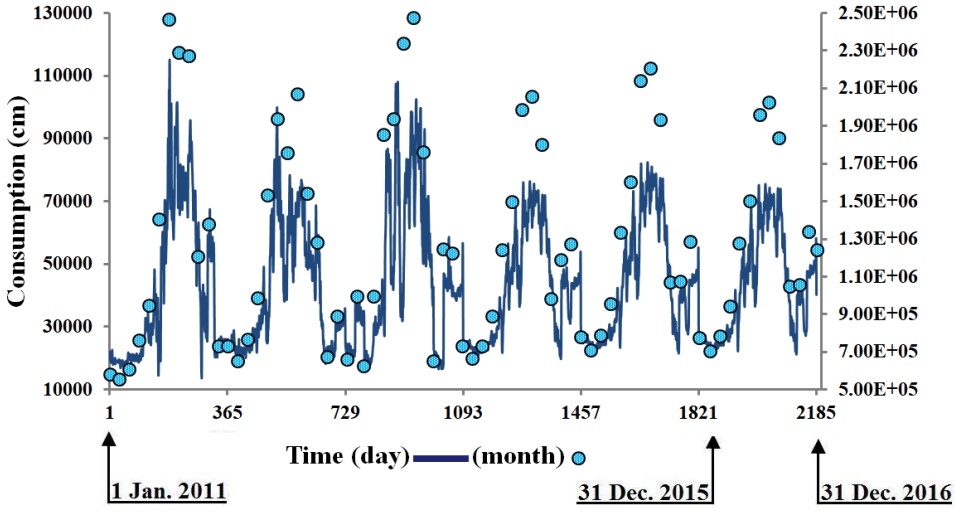


Figure 3.4 Time series plot of daily and monthly temporal scale of water consumption in City of Kelowna.

Table 3.1 Statistics of water consumption of Kelowna City in different temporal resolutions (*m^3).

Property	Number of Data	Max. Value*	Min. Value*	Average*	Standard deviation*	Coefficient of variation	Skew	Kurtosis
Data	2186	114597.2	14124	43046.4	20074.5	0.46	0.73	-0.38

3.5 Models Development and Results

***Non-linear Local Approximation* (*NLA*):** The present research studied whether the embedding dimension and lag time influence the accuracy of forecasted consumption values. First, forecasted values were evaluated for the embedding dimensions (ranging from 2 to 20) at lag times 1 and 17 days to predict 1-day-ahead. For m = 2 for a lag time of 1-day, two variables D_t and x_{t-1}, and for m = 2 for a lag time of 17 days two variables D_t and D_{t-17} were used as input variables to predict 1-day ahead (D_{t+1}). Also, for m = 3 for the lag time of 1 and 17 days, three variables D_t, D_{t-1}, D_{t-2} and D_t, D_{t-17}, D_{t-34} were used as input variables to forecast 1-day-ahead. Moreover, phase space was reconstructed for m>3. Table 3.2 presents a summary of the 1-day ahead forecasted values that reconstructed phase-space in dimensions ranging from 2 to 20 for τ = 1 and 17 days. The overall average for all embedding dimensions CC > 0.96, RMSE< 4200 (m3/day) (8% of daily average consumption) and MAE < 49 indicate reasonable forecasting for each series.

Moreover, the table reveals the best embedding dimension for the most accurate forecasted value in bold. Using CC, RMSE, and MAE, the optimum embedding dimension (m_{opt}) was found to be 18 and 19 for the lag time of 1 and 17 days, respectively. The results of NLA forecasting methods admit the same results that correlation exponent (or correlation dimension presented in chapter 2) and false nearest neighbor resulted in optimum embedding dimension (m=17 for CD and FNN for daily value). To evaluate the performance of selected m and τ, the results of the best models for m = 18, τ = 1 and m = 19, τ = 17 have been applied to forecast 2-, 4-, 7-, 14-, 30- and 60-days time step ahead (lead time) water consumption. Figure 3.5 shows the results of each model (NLA and PSR-NLA) in comparison with actual recorded values.

Table 3.2 Fitness values for NLA and PSR-NLA methods in different embedding dimension lag time and lead time.

NLA , τ =1 T=1				PSR-NLA , τ =17 T=1				τ =1 m=18			
m	**CC**	**RMSE***	**MAE**	***m***	**CC**	**RMSE***	**MAE**	**T**	**CC**	**RMSE***	**MAE**
1								**1**	0.9842	2855.6	43.07
2	0.9759	3532.1	47.85	**2**	0.9751	3581.7	49.09	**2**	0.9783	3351.6	48.02
3	0.9771	3424.8	47.34	**3**	0.9773	3425.3	48.01	**4**	0.9331	5883.5	63.03
4	0.9762	3495.5	47.74	**4**	0.9789	3302.9	47.30	**7**	0.8932	7445.8	72.49
5	0.9785	3332.0	47.08	**5**	0.9785	3331.7	47.10	**14**	0.7877	10555.0	87.76
6	0.9795	3248.6	46.13	**6**	0.9795	3257.8	46.79	**30**	0.6735	13307.6	100.44
7	0.9802	3187.3	45.49	**7**	0.9829	2967.1	45.80	**60**	0.2523	20189.1	129.71
8	0.9805	3176.4	45.65	**8**	0.9838	2887.9	45.22	τ =17 m=19			
9	0.9806	3164.1	45.65	**9**	0.9849	2792.8	43.95	**T**	**CC**	**RMSE***	**MAE**
10	0.9803	3193.6	45.09	**10**	0.9846	2828.2	44.52	**1**	0.9852	2772.8	43.83
11	0.9813	3098.7	44.56	**11**	0.9850	2792.2	43.95	**2**	0.9898	2295.7	39.59
12	0.9763	3495.1	48.11	**12**	0.9804	3189.4	46.69	**4**	0.9415	5504.2	61.47
13	0.9752	3578.0	48.11	**13**	0.9768	3457.4	48.32	**7**	0.9002	7211.2	71.46
14	0.9779	3378.9	47.14	**14**	0.9788	3303.6	47.65	**14**	0.8048	10147.5	86.61
15	0.9806	3169.7	46.06	**15**	0.9790	3290.2	47.23	**30**	0.6776	13265.1	95.31
16	0.9765	3491.0	47.24	**16**	0.9796	3250.2	46.70	**60**	0.4363	17784.9	118.53
17	0.9810	3139.6	45.21	**17**	0.9825	2995.5	45.95				
18	**0.9842**	**2855.6**	**43.07**	**18**	0.9838	2894.0	45.21				
19	0.9685	4088.5	44.69	**19**	**0.9852**	**2772.8**	**43.83**				
20	0.9661	4209.2	45.29	**20**	0.9846	2833.1	44.43				
Tot	**0.9775**	**3394.2**	**46.30**	**Tot**	**0.9807**	**3142.9**	**46.34**				
Best	**0.9842**	**2855.6**	**43.07**	**Best**	**0.9852**	**2772.8**	**43.83**				
EM	**18**	**18**	**18**	**EM**	**19**	**19**	**19**				

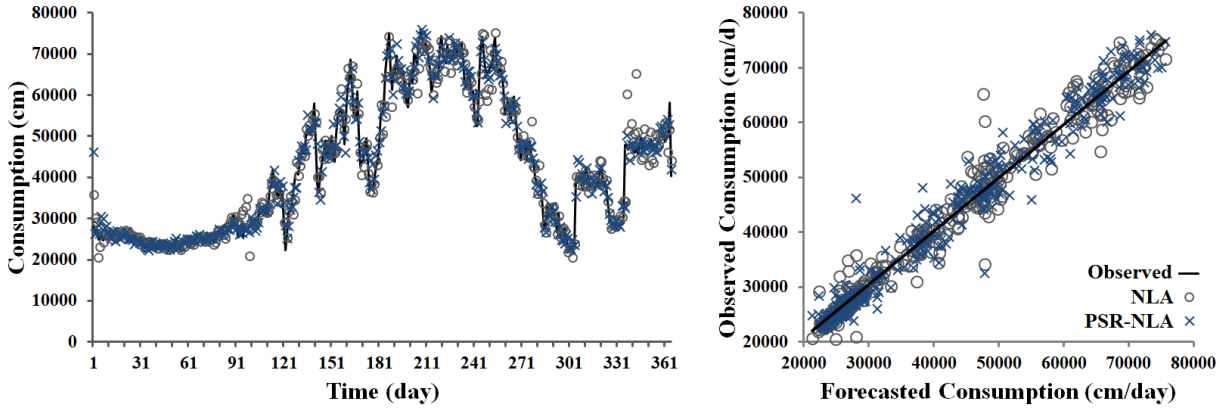


Figure 3.5 Forecasted values for water consumption by NLA and PSR-NLA in comparison with actual recorded values.

The research indicated that the results were more sensitive to optimum embedding dimension and lag time calculated by AMI than to the reconstructed phase space by 1-day lag time. However, the difference between the results of NLA and PSR-NLA were not considerable, the performance of PSR-NLA was better than NLA to forecast consumption values in different lead time. As it is shown in Figure 3.5, PSR-NLA values are more concentrated on the actual values. Moreover, m_{opt} = 19 provided more accurate results than m_{opt} = 18 in different lead time (Table 3.2).

***Gene Expression Programming* (*GEP*):**

One of the most important steps in developing an accurate model is the selection of the input variables. The combinations were selected in a way that they included daily consumption data with lag times of τ = 1 and 17 days. Different combinations of the time series of daily consumption were used to structure a policy for input dataset. Combinations of D_t, D_{t-1}, D_{t-2}, …, D_{t-20} variables were used as input data with D_{t+1} (1-day step ahead) as output of the GEP, and combinations of D_t, $D_{t-\tau}$, $D_{t-2\tau}$, …, $D_{t-20\tau}$ variables were used as input data with D_{t+1} as output of the PSR-GEP (forecasting of 1 day step ahead). The combinations of arithmetic functions of $\{+, -, \times, x, x^2, \sqrt{x}, \log x\}$ were used for GEP and PSR-GEP and PSR-MLR models, respectively. Ultimately, the best combination was selected using the criteria of CC, RMSE, and MAE. Table 3.3 reveals the 20 combinations of inputs and their performance with GEP and PSR-GEP models. On the table, the three criteria indicate the fourth combinations (m = 4) for the lag time of 1-day, and (m=8) for the lag time of 17 days as the best combination of input data (reconstructed phase space) and arithmetic functions for GEP and PSR-GEP. The study revealed the followings for both GEP and PSR-GEP, respectively:

$$D_{t+1} = D_{t-1} + \log[\log(\log(9.58D_{t-4})) \times (D_{t-2}D_{t-3})] + 8.253 \quad (3.15)$$

$$D_{t+1} = D_{t-\tau} - 0.04D_{t-4\tau} + 8.03\sqrt{D_{t-4\tau}} + 0.623 \quad (3.16)$$

To investigate the accuracy of the forecasted values by equations (3.15) and (3.16), consumption values with different lag times were calculated. The results were not found more sensitive to lag time calculated by AMI (17 days) than to the 1-day lag time. However, PSR made the performance of the model better than m=4 with 1-day lag time. Moreover, regarding the equations (3.15) and (3.16), equation (3.15) used all four arithmetic operators made by *m*=4 ($D_{t-1}, D_{t-2}, D_{t-3}, D_{t-4}$)

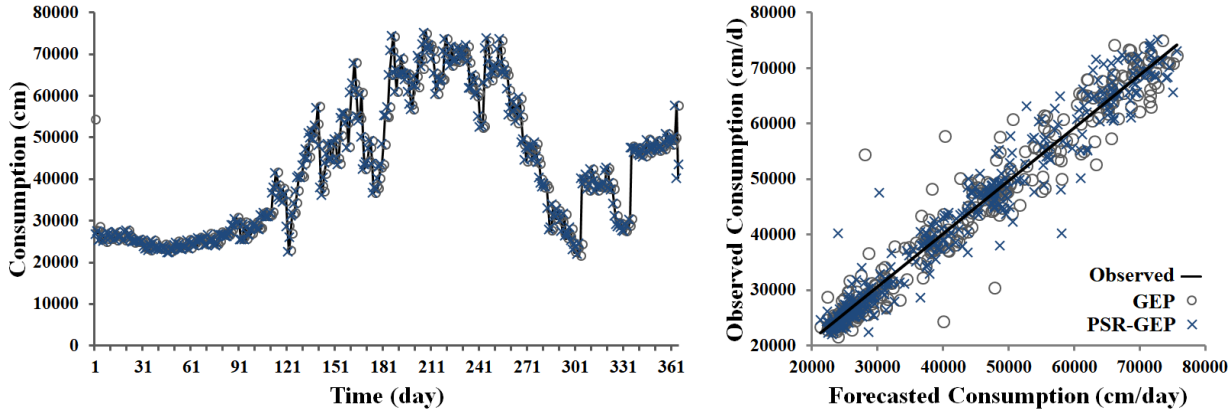


Figure 3.6 Forecasted values for water consumption by GEP and PSR-GEP in comparison with actual recorded values.

while equation (3.14) only used two of the arithmetic operators out of 8 operators (m=8; $D_{t-\tau}, D_{t-4\tau}$). Table 3.3 compares further details for GEP and PSR-GEP models for different lead time forecasted values.

Multi-Layer Perceptron Artificial Neural Network (ANN)

Artificial Neural network (ANN) is another approach to model the consumption values represented in section 3.3.4. ANN's structures have different hidden layer neurons (HLN) from 1 to 10 with 100 epochs for each model. Table 3.4 represents the result of ANN for both 1-day

Table 3.3 Fitness values for GEP and PSR-GEP methods in different embedding dimension lag time and lead time.

GEP , τ =1 T=1				PSR-GEP , τ =17 T=1				τ =1 m=4			
m	**CC**	**RMSE***	**MAE**	***m***	**CC**	**RMSE***	**MAE**	**T**	**CC**	**RMSE***	**MAE**
1								**1**	0.9764	3486.6	47.83
2	0.9757	3543.7	48.14	**2**	0.9789	3636.9	48.57	**2**	0.9494	5112.2	57.92
3	0.9761	3517.8	47.91	**3**	0.9788	3644.6	48.59	**4**	0.9130	6716.4	67.48
4	**0.9764**	**3486.6**	**47.83**	**4**	0.9789	3647.1	48.56	**7**	0.8652	8376.4	76.57
5	0.9760	3519.0	47.95	**5**	0.9789	3635.5	48.68	**14**	0.7810	10734.8	88.95
6	0.9760	3520.5	48.37	**6**	0.9788	3649.5	48.71	**30**	0.6548	13649.0	97.03
7	0.9760	3500.9	47.91	**7**	0.9788	3649.0	48.70	**60**	0.2345	20411.3	130.23
8	0.9760	3521.4	47.97	**8**	**0.9789**	**3631.6**	**48.62**	τ =17 m=8			
9	0.9760	3511.9	47.89	**9**	0.9788	3653.9	48.63	**T**	**CC**	**RMSE***	**MAE**
10	0.9760	3514.7	47.89	**10**	0.9788	3650.6	48.73	**1**	0.9789	3631.6	48.62
11	0.9760	3514.6	47.88	**11**	0.9788	3656.6	48.65	**2**	0.9553	5267.3	58.52
12	0.9760	3514.7	47.89	**12**	0.9787	3657.4	48.69	**4**	0.9227	6894.5	68.47
13	0.9760	3510.1	47.91	**13**	0.9789	3645.4	48.55	**7**	0.8713	8848.2	78.11
14	0.9760	3516.6	47.91	**14**	0.9787	3655.7	48.69	**14**	0.7782	11571.8	91.84
15	0.9760	3510.4	47.86	**15**	0.9788	3650.5	48.61	**30**	0.6334	14631.5	105.98
16	0.9760	3498.7	47.88	**16**	0.9789	3644.4	48.54	**60**	0.3864	18670.2	126.20
17	0.9759	3515.1	47.89	**17**	0.9789	3638.9	48.56				
18	0.9760	3509.7	47.85	**18**	0.9789	3646.6	48.57				
19	0.9759	3514.8	47.93	**19**	0.9787	3650.6	48.74				
20	0.9759	3514.2	47.90	**20**	0.9789	3642.5	48.51				
Tot	**0.9759**	**3519.0**	**47.96**	**Tot**	**0.9788**	**3646.5**	**48.62**				
Best	**0.9764**	**3486.6**	**47.83**	**Best**	**0.9789**	**3631.6**	**48.51**				
EM	**4**	**4**	**4**	**EM**	**8**	**8**	**20**				

delay and PSR values. The results in the table for each *m* are extracted from the result of various HLN and epochs. Figure 3.7 shows the example for selecting *m*=14 among (10×100=1000) for CC and RMSE. This calculation has been done for all *m* from 1 to 20 for both 1-day delay and the delay with the lag time (PSR). (1000×20×2=40000) number of calculations where the best 20 values have been selected (Table 3.4). For all the models, *m*=1 to 20 is considered because of the results in chapter 2. The embedding dimensions with the higher fitness value were among the range of m=1 to 20. Then, the table provides the results for the *m*=1 to 20.

Selection of ANN structures are represented in Table 3.4 for the test period. Statistical indices for the fitness values showed *m*=17 for 1-day delay and *m*=14 for PSR, with the values of (CD= 0.9802, RMSE=3183.7 and MAE=46.2) and (CD=0.9817, RMSE=3405.8 and MAE=48.4), respectively. Regarding the results, PSR-ANN mostly dominates in all embedding dimensions for the fitness accuracy indices. Figure 3.8 shows the comparison of the recorded and modeled values in the test period for both ANN and PSR-ANN in *m*=17 and 14, respectively.

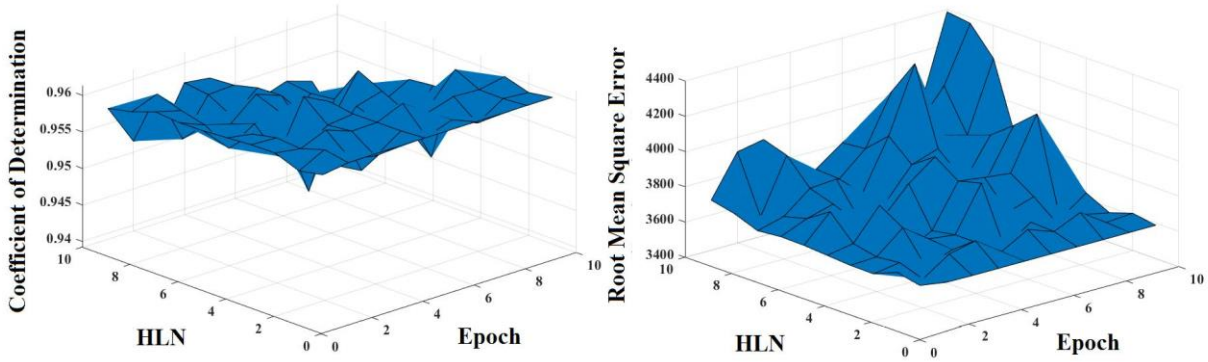


Figure 3.7 The results of ANN for τ =17 PSR by various HLN and Epochs.

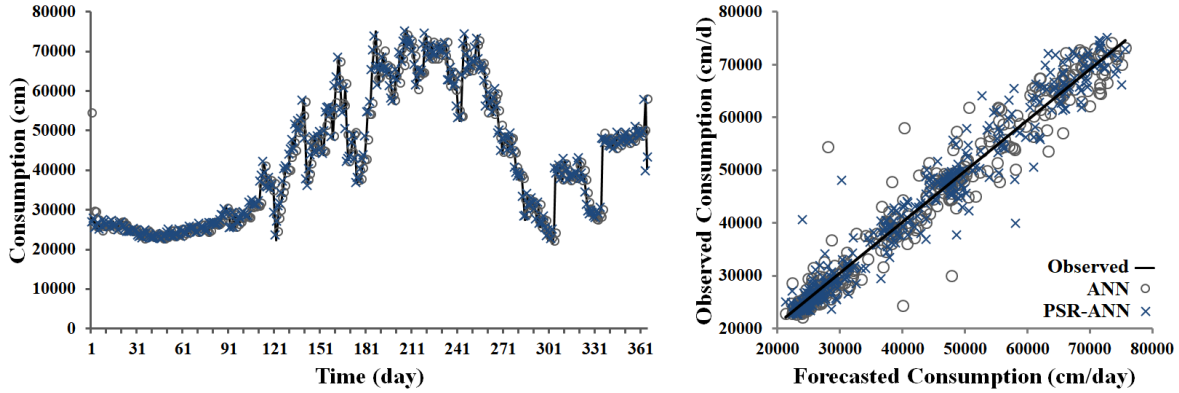


Figure 3.8 Forecasted values for water consumption by ANN and PSR-ANN in comparison with actual recorded values.

Table 3.4 Fitness values for ANN and PSR-ANN in different embedding dimensions (*m^3/day).

	ANN , τ =1 T=1					PSR-ANN , τ =17 T=1					τ =1 *m*=17		
m	**ST**	**CC**	**RMSE***	**MAE**	***m***	**ST**	**CC**	**RMSE***	**MAE**	**T**	**CC**	**RMSE***	**MAE**
1										**1**	**0.9802**	**3183.7**	**46.2**
2	2-2-1	0.9769	3450.2	47.4	**2**	2-4-1	0.9794	3600.6	48.2	**2**	0.9403	5811.38	62.97
3	3-3-1	0.9774	3400.7	46.9	**3**	3-6-1	0.9792	3613.0	48.4	**4**	0.9127	7041.91	69.96
4	4-1-1	0.9779	3373.5	46.9	**4**	4-4-1	0.9794	3600.5	48.1	**7**	0.8650	8776.83	78.39
5	5-4-1	0.9776	3387.1	47.3	**5**	5-1-1	0.9795	3588.1	48.4	**14**	0.7873	11121.08	89.86
6	6-3-1	0.9781	3340.4	47.3	**6**	6-8-1	0.9793	3610.8	49.1	**30**	0.6291	14502.09	104.25
7	7-5-1	0.9766	3457.4	48.0	**7**	7-4-1	0.9791	3619.2	48.1	**60**	0.2222	20509.33	130.35
8	8-1-1	0.9782	3348.6	46.9	**8**	8-6-1	0.9798	3587.4	49.2	τ =17 *m*=14			
9	9-4-1	0.9793	3291.8	46.9	**9**	9-4-1	0.9799	3548.7	48.2	**T**	**CC**	**RMSE***	**MAE**
10	10-6-1	0.9791	3272.7	46.1	**10**	10-3-1	0.9799	3553.7	48.0	**1**	0.9817	3405.80	48.40
11	11-6-1	0.9785	3320.9	46.9	**11**	11-8-1	0.9808	3538.0	48.7	**2**	0.9390	4692.09	55.09
12	12-5-1	0.9790	3289.7	47.3	**12**	12-9-1	0.9805	3511.8	48.3	**4**	0.8655	7007.03	68.68
13	13-1-1	0.9791	3263.3	47.1	**13**	13-7-1	0.9809	3482.8	48.3	**7**	0.7708	9186.85	79.04
14	14-8-1	0.9791	3276.8	46.5	**14**	**14-4-1**	**0.9817**	**3405.8**	**48.4**	**14**	0.6244	11910.5	92.82
15	15-5-1	0.9795	3238.5	46.8	**15**	15-9-1	0.9809	3486.4	48.8	**30**	0.4877	15014.8	106.15
16	16-4-1	0.9799	3215.0	46.3	**16**	16-5-1	0.9801	3538.9	48.3	**60**	0.3862	18677.2	126.50
17	**17-3-1**	**0.9802**	**3183.7**	**46.2**	**17**	17-3-1	0.9808	3498.0	48.3				
18	18-9-1	0.9793	3259.9	47.1	**18**	18-3-1	0.9806	3524.3	49.7				
19	19-7-1	0.9768	3472.1	47.9	**19**	19-6-1	0.9806	3489.7	47.5				
20	20-8-1	0.9783	3346.5	48.2	**20**	20-8-1	0.9803	3528.3	49.2				
Tot		**0.9783**	**3335.3**	**47.1**	**Tot**		**0.9801**	**3547.6**	**48.5**				
Best		**0.9802**	**3183.7**	**46.1**	**Best**		**0.9817**	**3405.8**	**47.5**				
EM		**M17**	**M17**	**M10**	**EM**		**M14**	**M14**	**M19**				

***Multiple Linear Regression* (*MLR*):**

Microsoft Office Excel 365 was used to implement MLR model. The train period was used to derive regression coefficient from getting the value of variables in the linear equation. The availability of trained equation helped in testifying the previous year data as the test period. In the first fold, the 1-day delay was considered for m=1 to 20, and second fold applied 83-day delay. Table 3.5 shows the results of both MLR and PSR-MLR in the test period. Figure 3.9 shows the comparison of the recorded and modeled values in the test period for both MLR and PSR-MLR in m=17 and 3, respectively. The results showed (D_t, $D_{t-\tau}$, $D_{t-2\tau}$) as the best input combination for the models (m=3 for PSR-MLR). Statistical indices for the fitness values showed m=1 for 1-day delay and m=4 for the reconstructed phase space with the value of (CC=0.9789, RMSE=3638.45 and MAE=48.56) and (CC=0.9809, RMSE=3336.0 and MAE=48.17), respectively. The difference between the two models is not considerable, however, in the case of a large value of water consumption in long-term this difference can come into account. Moreover, the suggested equation for the best result by MLR is PSR-MLR with m=3 given by:

$$D_{t+1} = -0.00854D_t - 0.0366D_{t+\tau} - 0.0128D_{t+2\tau} + 0.9427D_{t+3\tau} \quad (3.17)$$

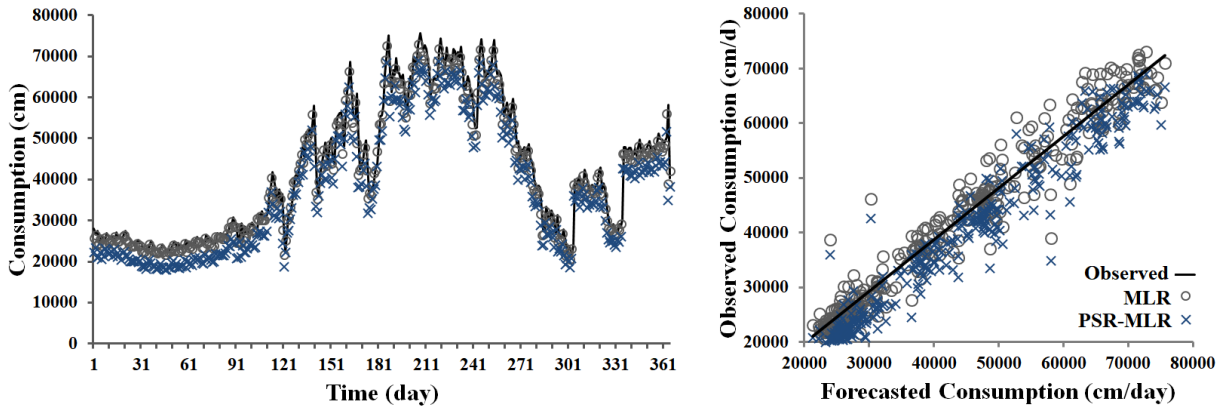


Figure 3.9 Forecasted values for water consumption by ANN and PSR-ANN in comparison with actual recorded values.

Table 3.5 Fitness values for MLR and PSR-MLR methods in different embedding dimensions.

MLR , τ =1 T=1				PSR-MLR , τ =17 T=1				τ =1 *m*=17			
m	**CC**	**RMSE***	**MAE**	***m***	**CC**	**RMSE***	**MAE**	**T**	**CC**	**RMSE***	**MAE**
1								**1**	0.9789	3638.9	48.56
2	0.7825	11658.6	92.37	**2**	0.9758	3763.4	50.48	**2**	0.9494	5139.4	58.65
3	0.9790	3762.3	49.89	**3**	**0.9809**	**3336.0**	**48.17**	**4**	0.9130	6701.6	68.26
4	0.9790	3792.3	50.19	**4**	0.9811	3443.9	49.41	**7**	0.8595	8511.9	77.68
5	0.9790	3814.1	50.42	**5**	0.9766	3905.5	52.04	**14**	0.7595	11204.4	91.66
6	0.9790	3958.0	52.11	**6**	0.9148	22846.5	145.73	**30**	0.6447	13707.5	101.02
7	0.9790	4133.6	54.24	**7**	0.9765	3568.2	48.60	**60**	0.2282	20211.2	129.01
8	0.9790	4187.1	54.85	**8**	0.9764	3811.2	51.08	τ =17 *m*=3			
9	0.9791	4432.7	57.65	**9**	0.9766	3568.2	48.71	**T**	**CC**	**RMSE***	**MAE**
10	0.9792	5024.0	63.46	**10**	0.9766	3680.4	49.81	**1**	0.9809	3336.0	48.17
11	0.9792	5576.0	68.14	**11**	0.9767	3610.6	49.13	**2**	0.9555	5336.5	59.61
12	0.9792	5921.4	70.92	**12**	0.9767	3603.9	49.04	**4**	0.9232	6889.3	69.17
13	0.9793	6276.8	73.59	**13**	0.9766	3601.4	49.13	**7**	0.8723	8841.9	78.12
14	0.9793	7267.0	80.61	**14**	0.9767	3584.5	48.92	**14**	0.7790	11549.8	91.70
15	0.9794	9128.5	92.30	**15**	0.9769	3560.6	48.79	**30**	0.6344	14504.2	104.84
16	0.9794	10115.3	97.81	**16**	0.9769	3550.4	48.73	**60**	0.3859	18351.2	125.12
17	**0.9789**	**3638.9**	**48.56**	**17**	0.9769	3550.5	48.73				
18	0.9794	10114.2	97.80	**18**	0.9769	3560.4	48.81				
19	0.9794	10115.3	97.81	**19**	0.9768	3561.6	48.87				
20	0.9795	9618.8	95.07	**20**	0.9769	3610.5	49.39				
Tot	**0.9595**	**6709.6**	**72.00**	**Tot**	**0.9738**	**4579.2**	**54.20**				
Best	**0.9795**	**3638.8**	**48.56**	**Best**	**0.9811**	**3336.0**	**48.17**				
EM	**20**	**17**	**17**	**EM**	**4**	**3**	**3**				

Figure 3.11 plots the stretching factor (*S*) versus the number of points (*N*=100). The figure shows an overall increase in *S* by increasing *N*. Furthermore, the figure reveals two components: the first part reveals a sudden increase in *S*, while the second part (after *N*=25) shows a more gradual increase in *S*. Figure 3.11 also indicates the best line (dashed line) fitted to the second part. Following Rosenstein et al. [145], the present study revealed λ_{max} of 0.014, 0.0102 and 0.0082 for *m* = 17, 18 and 19, respectively. Three different values of λ_{max} are considered because of the embedding dimensions range that were determined by CD, FNN, NLA and PSR-NLA, which were between 17 and 19. Note that *m*=18 and 19 were determined by NLA and PSR-NLA, respectively, which gave higher accuracy than the other embedding dimensions (see Table 3.2).

***Largest Lyapunov Exponent and Forecasting Horizon*:**

The forecasting horizons ($\Delta t = 1/\lambda_{max}$) were 72, 98 and 122 days for m = 17, 18 and 19, respectively. Also, m=17 with forecasting horizon value of 72 was determined by CD. Figure 3.10b shows the frequency of the data is smooth from day 1 to 72. Even with the high data frequency between days 98 and 122, PSR-NLA increased the forecasting horizon from 98 days (NLA) to 122 days. Considering results of CD, NLA, PSR-NLA, and forecasting horizon, it is reasonable to select m=17 as the system's optimum dimension.

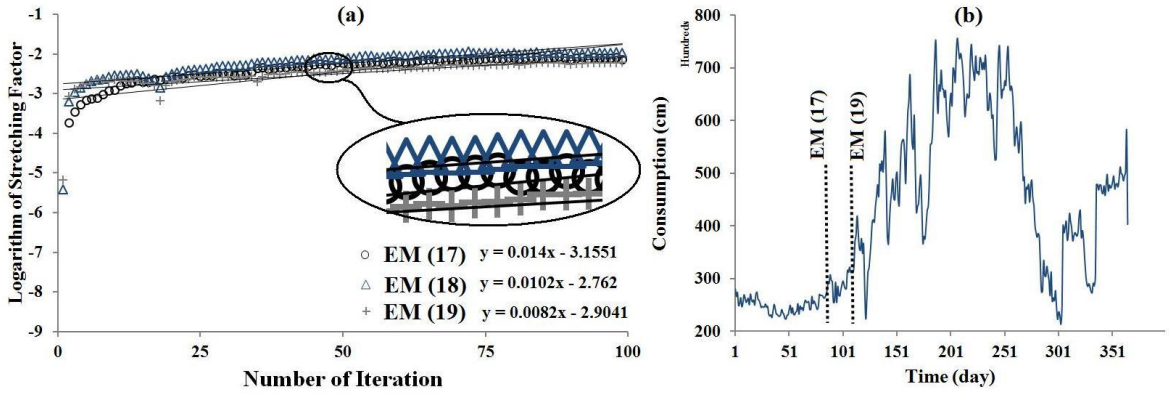


Figure 3.10 Estimation of largets Lyapunov exponent for daily consumption (a) for embedding dimension of 17, 18 and 19; (b) Observed values of water consumption in the test period.

3.6 Discussion

The results indicate that forecasting was more sensitive to optimum embedding dimension and lag time calculated by AMI (Chapter 2), than to the reconstructed phase space by 1-day lag time. It can be considered as the pre-processing method to improve the performance of the models. However, the difference between the results of NLA and PSR-NLA was not considerable; the performance of PSR-NLA was better than NLA to forecast consumption values in different lead time. Moreover, m_{opt} = 19 provided more accurate results than m_{opt} = 18 (Table 3.2). Figures 3.5 to 3.9 highlight the negligible error in the results of all models during the test period. Regarding

the average fitness results for all dimensions, the performance of GEP is better than PSR-GEP, while the application of PSR is shown in forecasting with equal dimension and different lead times. Considering the small difference among the forecasted values by different embedding dimensions in all models, it is recommended to test the optimum embedding dimension by various methods. However, their results are quite similar, the overview of approximation for all the techniques for multiple dimensions concluded that the optimum embedding dimension is between m=17 and m=19, demonstrating reliability of the methods. To compare the models, Table 3.6 shows the statistics of the forecasted and actual recorded data. The most accurate models which had the least percentage of error comparing to the actual recorded values were highlighted with a check mark. The results reveal the PSR technique had a positive impact on the accuracy of all models. As Figure 3.10 shows, the performance of NLA and PSR-NLA was approximately similar for daily and 2-days ahead forecasting, but PSR-NLA was more accurate than NLA in different lead time (1-, 2-months ahead). Nevertheless, the slope of the fitted line for the fitness functions shows that PSR-NLA can forecast more accurate values in comparison with NLA. The results of this application of PSR may also serve useful for application in long-term forecasting, which was previously identified in the literature as a topic requiring further study in drinking water consumption modeling.

Table 3.6 Statistics comparison of observed and forecasted consumption in test period by the selected models.

Property	Observed	NLA τ=1 m=18	PSR-NLA τ=17 m=19	GEP τ=1 m=4	PSR-GEP τ=17 m=8	MLR τ=1 m=17	PSR-MLR τ=17 m=3	ANN τ=1 m=6	PSR-ANN τ=17 m=3
Max. value	75620.26		✓						
Min. value	21313.72				✓				
Average	42500.82		✓						
Standard deviation	16117.34								✓
Coefficient of variation	0.38							✓	
Skew	0.43		✓						
Kurtosis	-1.13		✓						

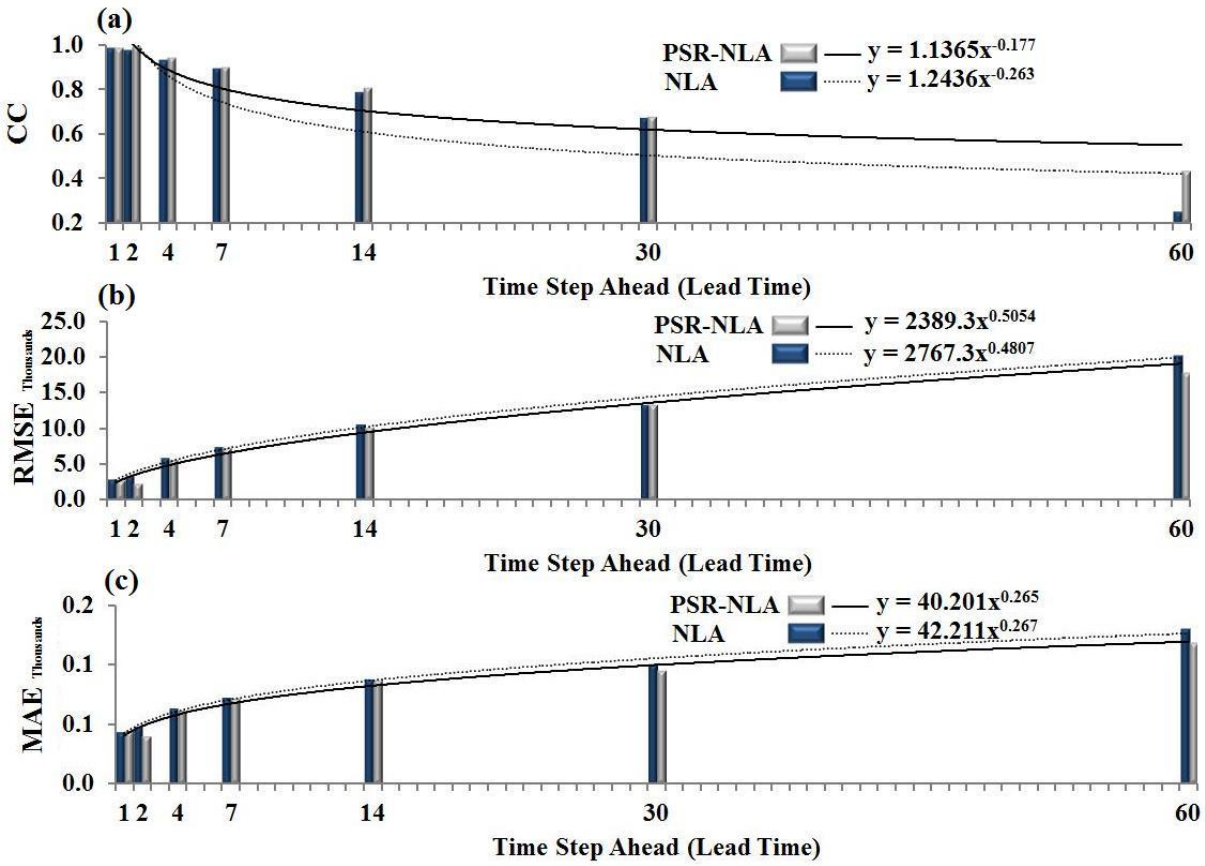


Figure 3.11 Performance of NLA and PSR-NLA in time ahead forecasting by the fitness functions of; (a) Correlation Coefficient; (b) Root Mean Square Error; (c) Mean Absolute Error.

3.7 Summary

An urban water consumption dataset from the City of Kelowna (British Columbia, Canada) was used as a test case to forecast future consumption values using varying lead times to identify models which may improve forecasting performance. NLA, MLP-ANN, GEP and MLR were the techniques for modelling the consumption values. Chaos theory techniques were compared with previously studied forecasting methods to assess the applicability of considering the chaotic behavior of urban consumption values to improve forecasting performance. Non-linear approximation, dynamic investigation, phase space reconstruction for input variables were considered to forecast various periodicity and lead time. Based on the selection criteria, NLA gave

the highest accuracy with 0.9852, 2772.8 and 43.83 for CD, RMSE and MAE, respectively. Also, PSR improved the accuracy of all models. Findings suggest that considering the chaotic behavior of consumption values (chapter 2) may improve forecasting performance. Further study of the chaotic behavior of consumption values may help inform forecasting methods to improve water distribution system management.

Chapter 4: Improvement of Soft-Computing Forecasting models' Accuracy[3]

4.1 Overview

The limited availability of water resources is becoming significant, especially in arid and semiarid regions, due to climatic change. Factors like urban area development, population growth, and industry expansion are also reasons responsible for water resources scarcity. Therefore, it is necessary to understand the relation of available water resources and the active variables for assisting the governments in developing long-term plans for water-related problems (drinking water) in the future. For a practical management plan and its operation to supply proper drinking water for the future, accurate estimation and consumption information is required. The accurate estimation of drinking water consumption can also reduce the water stress of an area [6]. Many variables are affecting the forecast of water consumption values, but not as influential variables in other hydrological forecasting problems (e.g., river discharge, sedimentation, rainfall, temperature, etc.). In context to literature, commonly used input variables to forecast are temperature, humidity, precipitation and recorded consumption series [148–150]. The present accepted knowledge for these factors is still limited and depends upon 1) accurate estimation and forecast water consumption and 2) determination of type and degree of nonlinearity among the effective variables [151]. While studies have advanced in the understanding of nonlinear characteristics and high complexity of water consumption factors, further research is still required. The deterministic

[3] A version of this chapter has been published as; Yousefi, P., Naser, G., Mohammadi, H. 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications, Intech; Yousefi, P., Shabani, S., Mohammadi, H., Naser, G. 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering 186, 544-550.; Yousefi, P., Naser, G., Mohammadi, H. 2018. Hybrid Wavelet and Local Approximation Method for Urban Water Demand Forecasting – Chaotic Approach. Published in WDSA / CCWI Joint Conference 2018.

approach is solely based on the input variables and their initial conditions, whereas, a probabilistic model relies on modeling uncertainties and randomness of the input variables. Given the significant challenges and complexity of probabilistic methods, and the fact that pre-processing methods can provide an accurate approximation to their probabilistic counterparts, this research focused on the application of pre-processing methods in improving the accuracy of the models.

4.2 Background

Midterm water consumption forecast helps the water management authorities to develop an integrated plan which balances supply and consumption in a given period. The applicability of this plan has direct relation with the accuracy of estimation for the future value of water consumption. In addition, water stress of an area can be reduced by accurate estimation of drinking water consumption [6,13,18,148,151]. Moreover, management can provide water sustainability based on their experience, as well as the accurate and reliable value of estimation of future consumption [28]. Literature enlists various deterministic and probabilistic techniques for forecasting urban drinking water consumption. In general, conventional methods were prevalent for a better understanding of determinants of water consumption which consider linear relationships between effective variables and water consumption, which is nonlinear. The mentioned studies are broadly categorized into two-fold: physical based and black box models. Without analyzing the physical processes, the second one applies artificial intelligence techniques (artificial neural networks, genetic programming, etc.), fuzzy-based (fuzzy logic, neuro-fuzzy, etc.), soft computing (support vector machine, etc.), and nonlinear deterministic (nonlinear local approximation, etc.) to identify the relationship between the input and output variables[4].

[4] See Chapter 3 for further details.

The dominant application of wavelet pre-processing is improving the accuracy of deterministic and probabilistic methods in forecasting problems [152]. Regarding the literature review reported by Nourani et al. [152], they concluded about the dominant application of wavelet-based models. Moreover, Labat [153] informed about the improving ability of wavelet in models' performance. Therefore, the application of wavelet brought researchers attention into areas such as denoising [154]; stream flow and water resources [155]; evaporation and climatic models [156]; groundwater level modeling [157]; water consumption forecasting [44,150]. In most of the mentioned studies combination of Wavelet-ANNs performed accurately over conventional models without hybrid wavelet models (e.g. ARIMA, MLR, ANN and etc.). This chapter applies non-linear local approximation method (NLA), artificial neural network (ANN), gene expression programming (GEP), and multiple linear regression (MLR) to determine their performance with/without pre-processing by phase space reconstruction (PSR) and wavelet decomposition in the case. To understand the dynamics of a nonlinear system, chaos theory could be successfully employed. The investigation about availability of chaotic behavior for the test dataset has been done in chapter 2. The results showed the chaotic behavior of the dataset for all temporal scales.

Input variable selection (IVS) is another popular pre-process technique that is highlighted in the literature as a fundamental step in creating an accurate and computationally efficient model [158,159]. Computational efficiency as well as complexity of a model strongly depends on the type and number of influential variables. Considering all influential variables in a water consumption model does not necessarily improve the model performance, however, a large set of input variables can cause many challenges in modeling. Selecting irrelevant variables or too many or too few variables as inputs can 1) increase the model's complexity and computational burden, 2) reduce computational efficiency of the model, and 3) decrease the model's accuracy [160–165].

The literature reveals a list of approaches for measuring the significance of variables and their relative importance [159,163,166]. The list includes correlation-based algorithm [162–164], connection weight approach [167], Garson's equation [168], partial derivatives [160], input perturbation [169], sensitivity analysis [170], forward stepwise addition, backward stepwise elimination, and improved stepwise selection [160], partial correlation input selection [171], and partial mutual information [166,172]. Olden *et al.* [167] compared several techniques and reported the connection weight approach as the best, followed by partial derivatives and input perturbation approaches. While researchers mainly applied IVS to reduce the complexities and to improve the accuracy of the models, they focused less on the most relevant and ecologically meaningful input variables [159]. Therefore, this chapter employed Garson's equation due to its simplicity in using the weights produced by MLP-ANN with no need for further information. It should be noted that the relevant literature provides applications of Garson's equation and ANN to find each variable's relative importance.

The objectives of the present study are four-fold: 1) evaluating the performance of NLA, ANN, GEP, and MLR methods by reconstructing phase space with different lag time, and 2) applying wavelet decomposition to the case data and investigating chaotic behavior of decomposed values; then, 3) forecasting the case data by the developed model with combination of PSR-Wavelet Decomposition and the selected high performance models.

4.3 Methodology

4.3.1 Nonlinear Local Approximation[5]

This chapter employs the correlation dimension since it is a general lower bound measure of the fractal dimension. The investigation of the time series has already been performed in chapter 2 and 3, and hence, is applied in the analysis of the objectives for the present chapter. This section applies NLA to forecast short-term period in the case study that is performed by reconstructing the phase space of the dataset. In chaos analysis, it is complicated to reconstruct the multi-dimensional phase space, which conceptually is responsible for designing the general pattern of the time series as the input. A phase-space reconstruction in m-dimension reflects an interpretation of the underlying dynamics in the form of an m-dimensional map. In other words, the dynamics of the system were explained locally step-by-step in the phase-space. The measure of change of trajectory with time in *m* dimensional place would result in forecasting.

4.3.2 Gene Expression Programming[6]

Initiating with the random generation of chromosomes, GEP is followed by different applications of genetic operators like replication, recombination, mutation, etc. The terminating condition for developing GEP depends upon the selection of maximum fitness. This section applied 30 chromosomes, eight head sizes, three genes, and arithmetic operators of $\{+,-, \times, x, x^2, \sqrt{x}\}$.

4.3.3 Artificial Neural Network[7]

ANN is based roughly on the neural layout of the human brain and is capable of non-linear modeling processes that can classify the patterns and recognize the capabilities. Regarding the

[5] Further details are available at section 3.3.1.
[6] Further details are available at section 3.3.2.
[7] Further details are available at section 3.3.3.

ability of multilayer perceptron (MLP)-ANN outperformance as a conventional ANN approaches [173,174], this section employed three-layer MLP-ANN (Input, Hidden and Output layers) and the different number of neurons. The number of neurons in the input layer varies from 1 to 10 (without decomposition) and 4 to 24 (with decomposition). Moreover, the neurons of the layers relate to the neurons in the next layer by weights. Also, to consider all optimal solutions with the highest probable accuracy, this study investigated the number of HLN from 1 to 20 in 1 to 200 epochs.

4.3.4 Multiple Linear Regression[8]

The common application of the linear regression (LR) modeling method is to model the goal variable (Y) with a single input variable (X). Multiple linear regression (MLR) describes the relationship between the goal value (Y) and the number of input variables (X_1, X_2, …, X_n). In other words, MLR finds a relationship with a linear combination of independent explanatory variables (e.g., reconstructed phase space with a different number of embedding dimensions for recorded values of water consumption) and response variable (time ahead values of water consumption).

4.3.5 Pre-Processing

4.3.5.1 Phase Space Reconstruction

The system dynamics (e.g., water consumption) can be described by a single point moving on a trajectory, where each point represents a state of the system under a given set of physical variables and their interactions. The lag-embedding method helps in the reconstruction of phase space from univariate or multivariate time series generated under the deterministic dynamic system [97]. The underlying dynamics could be understood by constructing m-dimensional space X_t defined by:

[8] Further details are available at section 3.3.4.

$$X_t = \{x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(m-1)\tau}\} \quad (4.1)$$

where X_t is a vector of the observed data of $\{x_t\}_{t=1,\ldots,N}$, N being the total number of observed data, τ is the lag time, and m is the embedding dimension. The lag-embedding method is sensitive to both embedding parameters of τ and m. The estimation of lag time can be calculated by the two popular methods, average mutual information (AMI) and autocorrelation function (ACF). To calculate the lag time in this research, ACF is used. In majority cases, wavelet transforms are employed for decomposition, de-noising, and compression of the time series [175]. Time series, based on the combination of low and high frequency represents improved features (e.g., cyclical trends) and chaotic element. Based on these frequencies, separated low and high frequencies entails the original pattern and behavior of the time series. Wavelet decomposition is used here to separate each level of frequencies in time series. The decomposition level shows the subseries. Following the study of [13], applying the same decomposition transforms and comparing the results of the present models with the previous results, i.e., second and fourth order Daubechies (db2, db4) with level 3, were applied for the pre-processing of input variables. db2 and db4 were the decomposition functions of the selected models with the highest accuracy. The software MATLAB 2019 and 2018 (https://www.mathworks.com) was employed for the analysis of the study.

4.3.5.2 Wavelet Decomposition

Commonly, wavelet transforms are used for decomposition, de-noising, and compression of the time series [175]. Time series have a combination of low and high frequency which represent improved features (e.g., cyclical trends) and chaotic element, respectively [176]. Considering these frequencies, separation of low and high frequency is helpful in studying the original pattern and

behavior of the time series. One of the mentioned methods is discrete wavelet transform (DWT) to separate per level of frequencies in time series. One of the common discretion ways proposed by Mallat [177] that this study is used the mentioned DTW method to separate the frequencies of the applied data. In which, the level of the decomposition shows the subseries. For example, for level 1 decomposition, the number of subseries is two. Therefore, the number of levels indicates the number of subseries plus one. Level 3 is considered as a suitable decomposition level in the present study regarding the number of data (2186 day) and following Nourani et al. [175] that offered:

$$L_n = int[log(N)] \tag{4.5}$$

where L_n is the number, the level of decomposition and N is the number of used data. Thus, the proper level in this study is considered as 3. However, increasing the level number does not necessarily improve the accuracy of the models. Therefore, the original data are discretized in a high-frequency subset (a_3) and three high frequencies as (d_1), (d_2) and (d_3), where the summation of all is equal to the value of original data. This research employed Haar, the second and fourth order Daubechies (db2, db4), and the second and fourth order Symlets (Sym2, sym4) wavelets to decompose daily water consumption time series into sub-series.

4.3.6 Evaluation of Models' Performance

This research measured the models' accuracy by coefficient of determination (CD), root mean squared error (RMSE) and mean absolute error (MAE) defined as:

$$CD = \left[\frac{\sum_{i=1}^{N_t}(R_i - \bar{R})(F_i - \bar{F})}{\sqrt{\sum_{i}^{N_t}(R_i - \bar{R})^2} \quad \sqrt{\sum_{i}^{N_t}(F_i - \bar{F})^2}} \right]^2 \tag{4.7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_t}(R_i - F_i)^2}{N_t}} \tag{4.8}$$

$$MAE = \frac{1}{N_t}\sum_{i=1}^{N_t}|R_i - F_i| \tag{4.9}$$

where N_t is the number of values, R and F are the recorded and forecasted values of consumption, respectively. $\bar{R}$ and $\bar{F}$ are the mean of the observed and forecasted consumption values, respectively. Note that the range of CD is between 0 and 1 with higher positive values indicating better agreement. A lower value of RMSE and MAE indicates better agreement between the observed and forecasted values.

4.4 Case Study and Dataset

Following previous sections, this chapter also selected water consumption of the City of Kelowna (BC, Canada) as the test case. Table 4.1 shows the characteristics of the dataset in the test case. Figure 3.4 and Table 3.1 shows the daily consumption values and statistics of the used data in training and test period.

4.5 Results and Discussion

Existence of chaotic behavior in the time series is shown in Figure 4.2. However, in this chapter the results are not entirely based on the proof of having chaotic behavior, as the figure only shows possible low-dimensional chaotic behavior for only one scale of temporal resolution. Although, the investigation beyond the availability of chaos in the test case is proved in chapter 2 based on

the results of all temporal resolution. Theoretically, several methods are well known for investigating the chaotic behavior, such as lag time calculation method (e.g., average mutual information (AMI), Autocorrelation function (ACF), correlation dimension, largest Lyapunov Exponent, etc.). This chapter investigates the chaotic behavior by applying ACF (chapter two's results were based on AMI) and correlation dimension. Having chaotic behavior allows the use of ACF to calculate the lag time of the time series. The value of lag time is considered as the first approach of ACF to 0 (Figure 4.21). The reason for considering two phase space techniques, AMI and ACF, is to investigate the performance of each methods in designing input combination for the forecasting models. Since, the results of both AMI and ACF shows the chaotic behavior in the

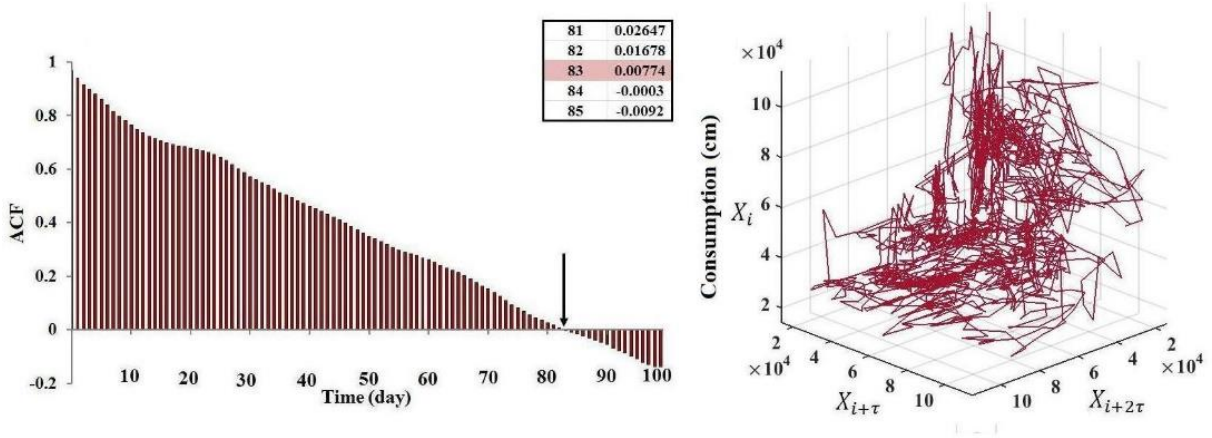


Figure 4.1 (a) Autocorrelation function (τ); (b) Reconstructed phase space by (τ and 2τ -day lag time).

data set, the results of fitness values for forecasting models can be the recognition criteria for selecting the appropriate methods for PSR. The results show 83-days as the lag time of the time series. Therefore, 83-days is used to design combinations of inputs as phase space for the time series. In this chapter, the difference between the 1st day and 83rd day is used as delay period for phase space reconstruction (PSR) varying embedding dimensions from 1 to 10 (m_1: D_t; m_2: D_t, $D_{t-\tau}$; m_{10}: D_t, …, $D_{t-10\tau}$). It should be noticed that several methods were introduced in literature to calculate the value of optimum embedding dimension, which may be more than 10 for the used

time series in this study. This study aims at showing the performance of embedding dimension and reconstructed phase space, where *m* is only considered 1 to 10 (further details are available in chapter two). Figure 4.1 shows the value of ACF for the consumption series and reconstructed phase space (τ=83). Figure 4.2a shows the relation between $C(r)$ and r and (4.2b) correlation exponent by varying *m*. Figure 4.2b shows that the value of correlation exponent increases by *m* and as *m*=17, the correlation exponent reaches a specific value (*Ce* =3.41). The value of *Ce* in chapter two reported as 3.50, which was based on AMI. However, for both of *Ce*, upper value of 4 is considered as the effective variable to define the system. Therefore, AMI and ACF both give reasonable results in the investigation of availability of chaos. This constant value of *Ce* at *m*=17 indicates the existence of the deterministic behavior of the time series.

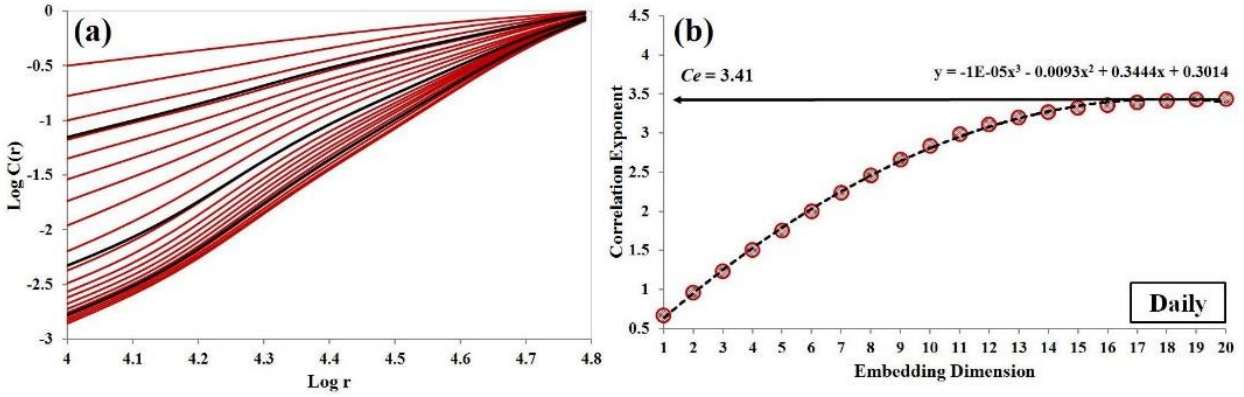


Figure 4.2 (a) The relation between correlation function C(r) and r by various m; (b) Saturation of correlation dimension Ce(m) with embedding dimensions.

4.5.1 Non-Linear Local Approximation

83-days and 1-day lag time (calculated value of lag time in the mentioned study) are used to reconstruct the phase space with various embedding dimension from *m*=1 to *m*=10 (m_1: D_t; m_2: D_t, $D_{t-\tau}$; m_{10}: D_t,…, $D_{t-10\tau}$ for τ=1 and 83). This chapter researches whether pre-processing influences the accuracy of forecasted consumption values. First, NLA is used to forecast the short-term daily consumption values by embedding dimensions ranging from 1 to 10 at lag times 1- and 83-days to

forecast short-term future values. As an illustration of input variables, D_t, D_{t-1} and D_t, D_{t-83} are used as input variables for embedding dimension (m=2) for two lag times τ=1 and τ=83, respectively, to forecast 1-day ahead (D_{t+1}) consumption value. Table 4.1 presents the summary of the forecasted values with/without PSR.

Table 4.1 Fitness values for NLA and PSR-NLA methods in different embedding dimensions.

NLA , τ =1				**PSR-NLA , τ =83**			
m	**CD**	**RMSE**(m^3/day)	**MAE**	***m***	**CD**	**RMSE**(m^3/day)	**MAE**
1	0.9735	3706.12	49.11	**1**	0.9735	3706.12	49.11
2	0.9759	3532.07	47.85	**2**	0.9741	3685.70	49.06
3	0.9771	3424.84	47.34	**3**	0.9745	3623.66	48.73
4	0.9762	3495.50	47.74	**4**	0.9773	3410.60	48.09
5	0.9785	3331.98	47.08	**5**	0.9778	3380.71	47.87
6	0.9795	3248.59	46.13	**6**	0.9785	3327.78	47.87
7	0.9802	3187.30	45.49	**7**	0.9794	3252.34	48.05
8	0.9805	3176.36	45.65	**8**	**0.9795**	**3242.26**	**47.38**
9	**0.9806**	**3164.13**	**45.65**	**9**	0.9778	3381.76	48.23
10	0.9803	3193.65	45.09	**10**	0.9770	3439.18	48.67

The average of fitness values for all dimensions is CD>0.97, RMSE<37100 (m^3/day) and MAE<50, which are the indication of reasonably calculated values. Moreover, the table shows the most accurate values in bold. The evaluation of the models for selected m and τ, showed CD=0.9806, RMSE=3164.13, MAE=45.65 and CD=0.9795, RMSE=3242.2, MAE=47.38 for τ=1, m=9 and τ=83, m=8, respectively. The results revealed that PSR pre-processing could not improve the performance of NLA in increasing the accuracy. However, PSR in MLR, ANN, and GEP models improved the accuracy of forecasted values, NLA and the mentioned methods are categorized into two different techniques. Although it appears to be due to the application of the methods, thorough exploring in needed to clarify which pre-processing methods are compatible with different techniques to improve the models' performance. Figure 4.3 compares the forecasted

values by NLA and PSR-NLA to the observed data that are bolded in Table 4.5. Also, the results revealed that the performances of the models are more sensitive to the number of dimensions comparing to with/without PSR. Moreover, the results of models' performance based on the designed input combination with AMI and ACF proved the priority of AMI in comparison with ACF. Previously, ACF and AMI were compared based on chaos investigation. Therefore, this research suggests using AMI for both chaos investigation and input combination design for forecasting models.

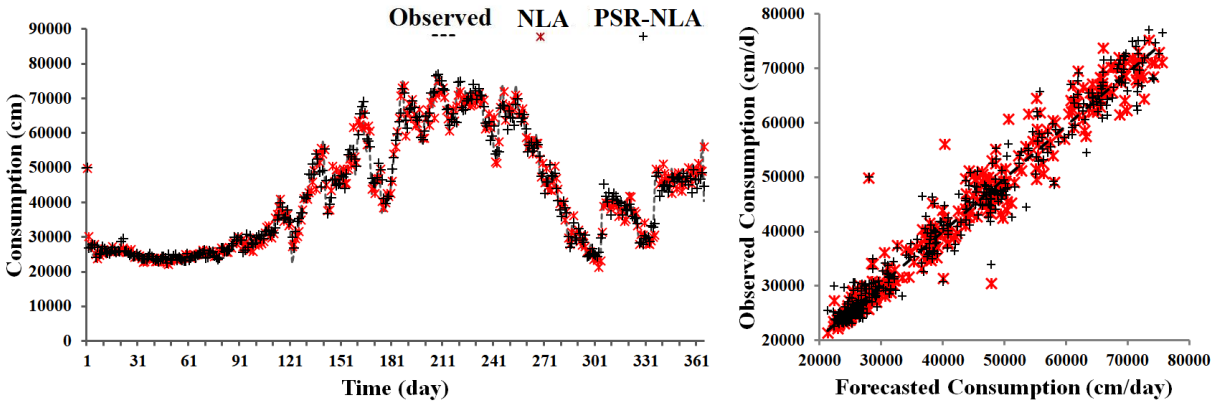


Figure 4.3 The performance of NLA and PSR-NLA in comparison with observed values.

4.5.2 Gene Expression Programming

GEP preliminarily investigates the relationship between input and output as discussed previously. Unlike the other models in this study, 1-day ahead is the output, and various combinations of input in terms of *m* are considered as input variables. The arithmetic operations used in this study are {+, - , ×, x, x_2, $\sqrt{x}$}, and GEP applies them to fit the best accuracy between input and output variables. Further details of GEP initial term values are in following of [22,44,85] to extract the GEP model for both 1-day delay and PSR. The results are shown in the table 4.2 for the test period. According to Table 4.2, there is not much difference among the different *m*. However, the difference in PSR-GEP results can be considered as a proof of sensitivity to the initial values of

specific time lags where the variations of the results for different *m* are more than 1-day delay. There is not a significant difference in the results of this study comparing to other alternative models, especially PSR-ANN is not an advantage of GEP. However, extracting the mathematical equation through GEP is one advantage of GEP comparing to other artificial models. As a result of the given model, the equation for *m*=3 (PSR-GEP) can calculate the consumption value for 1-day ahead by:

$$D_{t+1} = 0.0529\sqrt{D_{t+\tau} + D_{t+2\tau}} + D_t - 7.0838 \quad (4.11)$$

Table 4.2 Fitness values for GEP and PSR-GEP in different embedding dimensions.

GEP , τ =1				**PSR-GEP , τ =83**			
m	**CD**	**RMSE**(m^3/day)	**MAE**	***m***	**CD**	**RMSE**(m^3/day)	**MAE**
1	0.9494	3621.87	48.59	**1**	0.9565	3363.46	48.03
2	**0.9497**	**3609.82**	**48.37**	**2**	0.9565	3357.00	47.82
3	0.9494	3633.87	48.42	**3**	**0.9569**	**3343.36**	**47.50**
4	0.9494	3637.74	48.42	**4**	0.9566	3359.53	47.95
5	0.9494	3639.05	48.43	**5**	0.9562	3372.70	48.04
6	0.9494	3619.77	48.60	**6**	0.9566	3359.64	48.08
7	0.9495	3630.44	48.38	**7**	0.9564	3365.04	47.95
8	0.9494	3634.41	48.42	**8**	0.9567	3353.24	47.62
9	0.9494	3628.46	48.42	**9**	0.9562	3370.08	48.05
10	0.9494	3631.12	48.40	**10**	0.9565	3356.68	47.84

Although, variety of other arithmetic operations may have been applied here, focusing on the aim of study, only simple known operations were applied to extract the GEP equation. The results of PSR-GEP and alternative methods prove the advantage of PSR to improve the accuracy of the models. Statistical indices for the fitness values showed *m*=2 for 1-day delay and *m*=3 for the reconstructed phase space with the value of (CD=0.9497, RMSE=3609.82, and MAE=48.37) and (CD=0.9569, RMSE=3343.36, and MAE=47.50), respectively. Figure 4.4 shows the comparison

of observed and consumption values in the test period for both GEP and PSR-GEP in *m*=2 and 3, respectively.

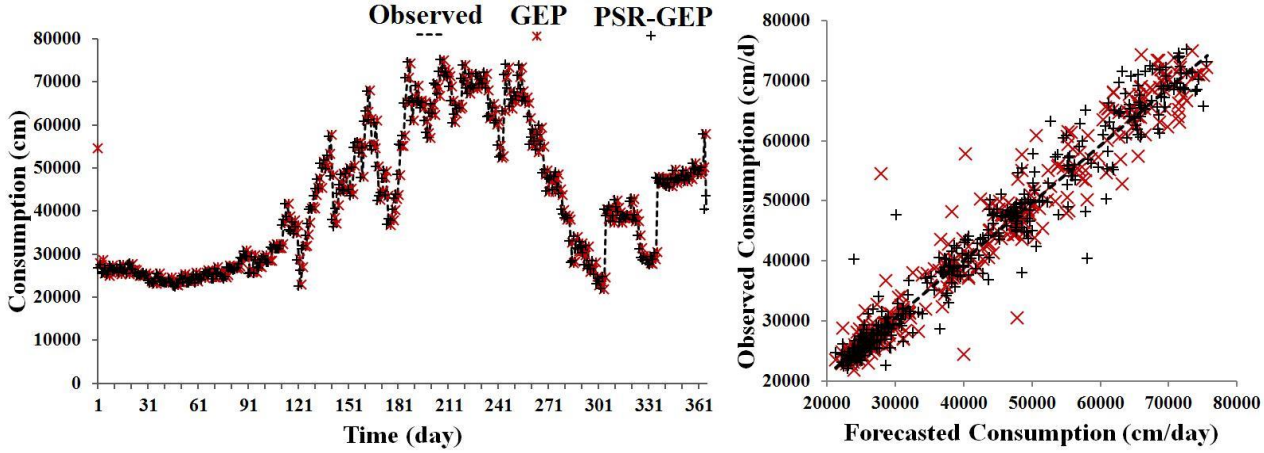


Figure 4.4 The performance of GEP and PSR-GEP in comparison with observed values.

4.5.3 Artificial Neural Networks

ANN is another approach to model the consumption values which were represented in section 4.3.3. ANN's structures have different hidden layer neurons (HLN) from 1 to 20 with 200 epochs for each model. Table 4.3 represents the result of ANN for both 1-day delay and PSR values.

Table 4.3 Fitness values for ANN and PSR-ANN in different embedding dimensions (*m3/day).

	ANN , τ =1						PSR-ANN , τ =83				
m	**Structure**	**Epoch**	**CD**	**RMSE***	**MAE**	***m***	**Structure**	**Epoch**	**CD**	**RMSE***	**MAE**
1	1-3-1	70	0.9505	3641.27	48.53	**1**	1-6-1	60	0.9565	3397.03	47.81
2	2-4-1	130	0.9508	3661.74	48.63	**2**	2-12-1	170	0.9564	3424.68	48.20
3	3-7-1	10	0.9514	3583.31	48.25	**3**	3-4-1	140	0.9569	3367.37	47.64
4	**4-11-1**	**40**	**0.9520**	**3564.67**	**47.76**	4	4-6-1	60	0.9570	3443.61	47.98
5	5-6-1	60	0.9512	3620.06	48.35	**5**	5-12-1	90	0.9567	3395.01	47.51
6	6-3-1	180	0.9514	3669.03	48.65	**6**	**6-9-1**	**90**	**0.9570**	**3357.65**	**47.31**
7	7-16-1	110	0.9510	3621.90	48.45	**7**	7-7-1	130	0.9564	3403.48	48.17
8	8-7-1	10	0.9509	3659.41	48.40	**8**	8-6-1	60	0.9565	3416.01	47.79
9	9-3-1	150	0.9509	3684.60	48.43	**9**	9-17-1	30	0.9564	3448.65	48.64
10	10-11-1	30	0.9515	3575.83	48.13	**10**	10-8-1	10	0.9566	3547.71	49.47

Selection of ANN structures are represented in Table 4.3 for the test period. Statistical indices for the fitness values showed *m*=4 for 1-day delay and *m*=6 for PSR, with the values of (CD= 0.9520, RMSE=3564.67 and MAE=47.76) and (CD=0.9570, RMSE=3357.65 and MAE=47.13), respectively. Regarding the results, PSR-ANN mostly dominates in all embedding dimensions for the fitness accuracy indices. Figure 4.5 shows the comparison of observed and consumption values in the test period for both ANN and PSR-ANN in *m*=6 and 3, respectively. The results showed (D_t, $D_{t+\tau}$, $D_{t+2\tau}$) as the best input combination for the models.

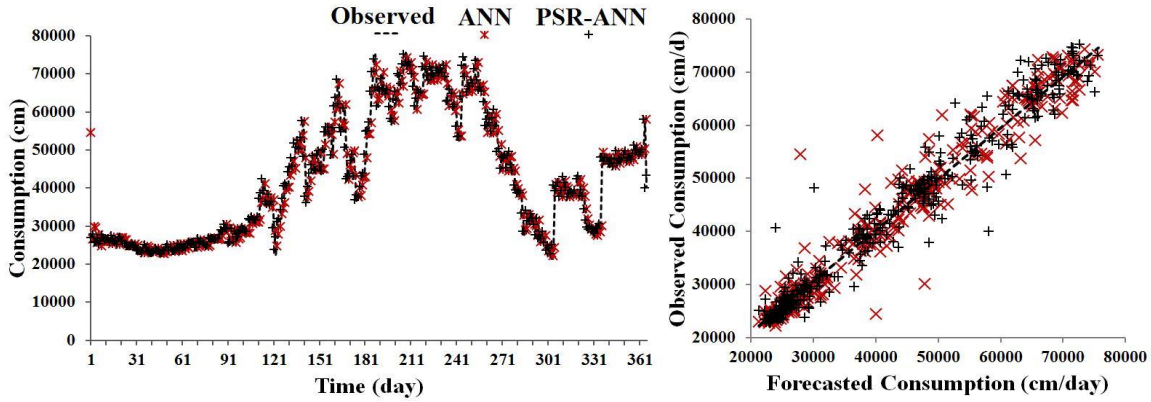


Figure 4.5 The performance of ANN and PSR-ANN in comparison with observed values.

4.5.4 Multiple Linear Regression

Excel 2010 was used to implement MLR model. The train period was used to derive regression coefficient from getting the value of variables in the linear equation. The availability of trained equation helped in testifying the previous years data as the test period. In the first fold, the 1-day delay was considered, for *m* 1 to 10, and second fold applied 83-day delay. Table 4.4 shows the results of both MLR and PSR-MLR in the test period. Statistical indices for the fitness values showed *m*=1 for 1-day delay and *m*=4 for the reconstructed phase space with the value of (CD=0.9565, RMSE=3642.89 and MAE=50.42) and (CD=0.9572, RMSE=3636.34 and MAE=51.04), respectively. However, the difference between the two models is not considerable,

in the large value of consumption in long-term this difference can come into account. Figure 4.6 shows the comparison of observed and modeled consumption values. Moreover, the suggested equation for the best result by MLR is given by:

$$D_{t+1} = -0.00854D_t - 0.0366D_{t+\tau} - 0.0128D_{t+2\tau} + 0.9427D_{t+3\tau} \quad (4.10)$$

Table 4.4 Fitness values for MLR and PSR-MLR methods in different embedding dimensions.

MLR , τ =1				**PSR-MLR , τ =83**			
m	**CD**	**RMSE**(m^3/day)	**MAE**	***m***	**CD**	**RMSE**(m^3/day)	**MAE**
1	**0.9565**	**3642.89**	**50.42**	**1**	0.9565	3642.89	50.42
2	0.9565	3804.14	52.14	**2**	0.9565	3804.14	52.14
3	0.9468	14106.70	112.82	**3**	0.9570	5319.51	66.90
4	0.9473	13174.97	108.82	**4**	**0.9572**	**3636.34**	**51.04**
5	0.9505	3724.99	49.81	**5**	0.9568	4167.55	56.45
6	0.9503	3746.33	50.09	**6**	0.9569	5907.90	71.65
7	0.9503	3747.49	50.10	**7**	0.9565	4370.03	58.86
8	0.9493	6058.34	70.88	**8**	0.9566	4581.10	60.89
9	0.9505	3736.33	50.02	**9**	0.9566	5023.16	64.71
10	0.9506	3738.35	50.07	**10**	0.9566	4327.34	58.48

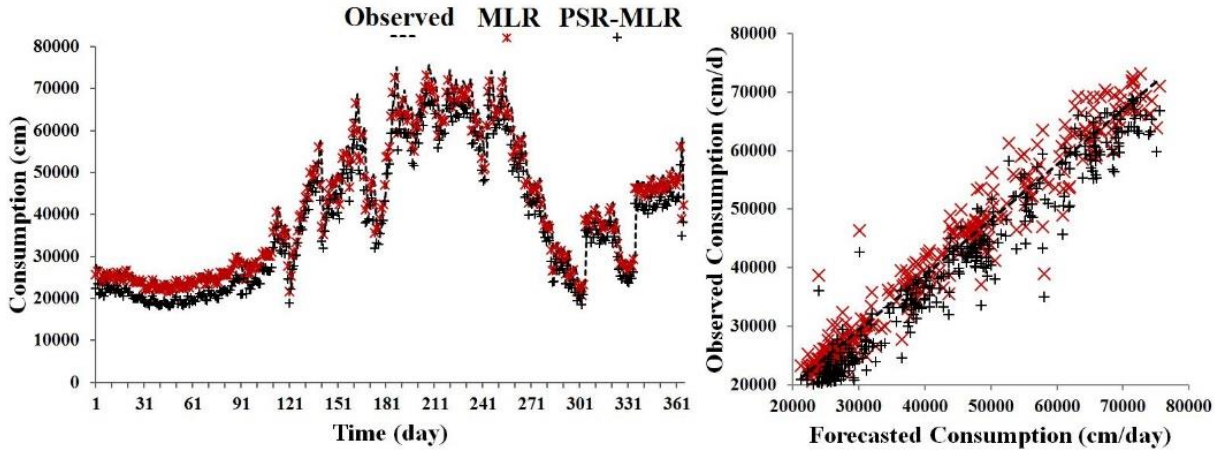


Figure 4.6 The performance of MLR and PSR-MLR in comparison with observed values.

4.5.5 Pre-Processing

The combination of models with wavelet decomposition is derived by adding the output of each wavelet to the input of the models. Figure 4.7 shows the example of the decomposed values for water consumption time series by *db*2 transform function. To discrete the consumption values, five wavelet transforms were applied (Section 4.3.6). As suggested by Nourani et al. [175], 3rd level decomposition is recommended for 2186 point data. To evaluate the availability of chaotic

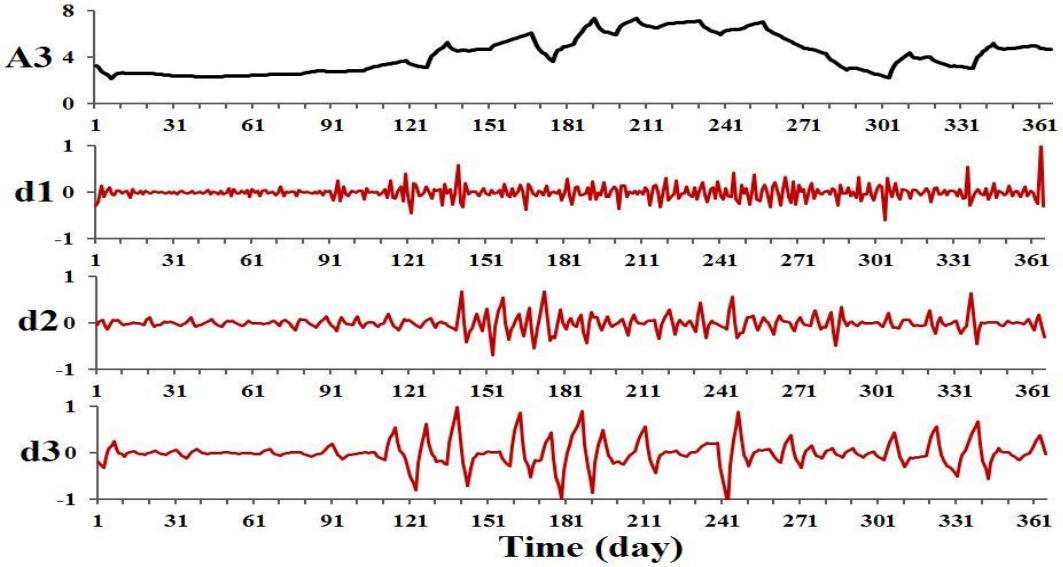


Figure 4.7 Three level DWT of daily water consumption time series of Kelowna City in 2016.

Table 4.5 Chaotic property identification of 3rd level decomposition of the daily water consumption.

TF	Wavelet	τ	*m*	*Ce*
db2	**A**	83	9	2.1
	D1	2	9	3.4
	D2	3	9	3.5
	D3	4	9	3.5
db4	**A**	83	8	2.2
	D1	2	8	3.3
	D2	3	8	3.4
	D3	4	8	3.5

behavior in decomposed values of the main test data, the third level of db2 and db4 are considered. Table 4.5 shows the results of the chaos investigation for approximation (A) and details (D) values

of case data. Figure 4.8 shows that the value of correlation exponent increases by *m*, but it reaches a steady value in *m*=9 for both db2 and db4. The value of the approximation (A) and detail 1 (D1) perform a chaotic behavior, but D2 and D3 of both db2 and db4 are not as steady as A and D1. Since each detail represents the different scale of the time series, the results reveal that to investigate the chaotic behavior of any time series, it is recommended to investigate all temporal

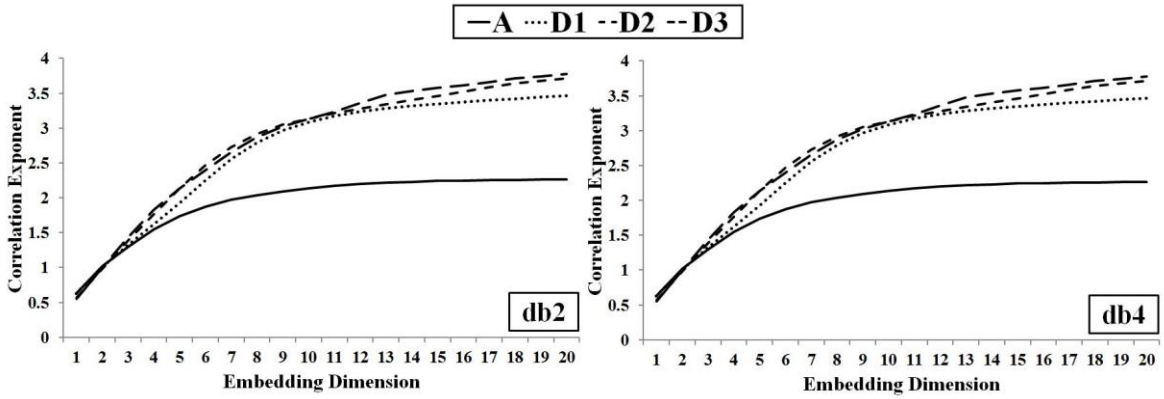


Figure 4.8 Saturation of correlation dimension $C_e(m)$ with embedding dimensions (a) db2 (b) db4.

scales to claim whether the series have chaotic behavior. While, all temporal resolutions were investigated in chapter 2, based on the results, it was concluded that the test data set has chaotic behavior. Regarding this fact, it seems AMI method is more appropriate than ACF in calculating lag time to be used in chaos investigation. The hybrid model with the combination of wavelet preprocessing and NLA model is derived by replacing the decomposed values as inputs. The reason for selecting the 3rd level of db2 and db4 in this section is because of these two functions' performance in the results of all models that resulted in the highest accuracy in the same case. Hence, db2 and db4 are applied in this study to compare the results of NLA and W-NLA models. Table 4.7 indicates the results of wavelet decomposition for the selected models in the previous section. As the table highlights, db4 and db2 are the transforms which resulted in the highest accuracy in W-MLR and W-PSR-MLR, with the value of (CD= 0.9697, RMSE=2804.44 and

MAE=42.11) and (CD=0.9745, RMSE=2699.83and MAE=43.61), respectively. After implying the decomposed inputs for MLR and PSR-MLR for result comparison improved the results in both models. Also, sym4 and db2 are the transforms which resulted in the highest accuracy in W-ANN and W-PSR-ANN, with the value of (CD= 0.9915, RMSE=1486.21 and MAE=30.06) and (CD=0.9756, RMSE=2517.24, and MAE=41.68), respectively. Also, calculations for W-ANN and W-PSR-ANN are done with HLN 1 to 20 and epochs 1 to 200, and the mentioned results in the table are selective of the highest among them. Unlike the results of MLR, W-ANN more accurately forecasted the inversion of the results of ANN and PSR-ANN than W-PSR-ANN." However, wavelet decomposition improved the results of W-ANN and W-PSR-ANN comparing to the alternative without decomposition. Moreover, db4 and db2 are the transforms which resulted in the highest accuracy in W-GEP and W-PSR- GEP, with the value of (CD= 0.9845, RMSE=2027.28 and MAE=36.62) and (CD=0.9753, RMSE=2532.21, and MAE=41.69), respectively. Following the results of ANN method, W-GEP more forecasted accurately than W-PSR-GEP. However, wavelet decomposition improved the results of W-GEP and W-PSR-GEP comparing to the alternative without decomposition. Also, the table shows the results of W-NLA for both with/without PSR. The optimum embedding dimensions of NLA and PSR-NLA are selected for wavelet pre-processing (m=9 and 8). As the table highlights, db2 is the function which resulted in highest fitness values for both W-NLA and W-PSR-NLA with the value of (CD= 0.9878, RMSE=3211.78 and MAE=45.61) and (CD=0.9924, RMSE=2759.18, and MAE=43.61), respectively. Unlike PSR, it seems wavelet pre-processing is compatible with the NLA method to improve the accuracy of the results. All PSR models resulted in the highest values which used the decomposed inputs by db2 transform. It is noticeable that PSR affects the inherent of the time series which the results of performance of all models are in common about improving the accuracy.

Considering this fact, PSR can be introduced as a pre-processing method like wavelet decomposition; however, complexity and accuracy of PSR cannot be compared with the higher result of wavelet decomposition. Figure 4.9 and 4.10 shows the comparison of all selected models with highest accuracy in forecast of short-term water consumption values. The figure shows that the performance of W-ANN and W-GEP is better than W-PSR-MLR, while

Table 4.6 Fitness values for decomposition of selection of models for the test period.

Models	Fitness	Transform Functions				
		haar	db2	db4	sym2	sym4
W-MLR	CD	0.9612	0.9477	**0.9697**	0.9677	0.9694
	RMSE(m^3/day)	3168.62	3681.17	**2804.44**	2893.60	2816.06
	MAE	44.48	49.04	**42.11**	43.54	42.24
W-PSR-MLR	CD	0.9670	**0.9745**	0.9719	0.9745	0.9712
	RMSE(m^3/day)	3008.34	**2699.83**	2811.58	2699.83	2845.69
	MAE	45.39	**43.61**	43.95	43.61	44.24
W-ANN	CD	0.9868	0.9816	0.9861	0.9856	**0.9915**
	RMSE(m^3/day)	1853.11	2189.15	2136.25	1948.28	**1486.21**
	MAE	33.78	36.91	39.50	33.86	**30.06**
W-PSR-ANN	CD	0.9685	**0.9756**	0.9723	0.9752	0.9715
	RMSE(m^3/day)	2867.87	**2517.24**	2677.44	2547.89	2724.56
	MAE	43.16	**41.68**	42.09	42.19	42.61
W-GEP	CD	0.9721	0.9766	**0.9845**	0.9297	0.9255
	RMSE(m^3/day)	2698.16	2492.46	**2027.28**	4311.60	4429.89
	MAE	41.21	39.05	**36.62**	54.23	55.25
W-PSR-GEP	CD	0.9667	**0.9753**	0.9721	0.9748	0.9704
	RMSE(m^3/day)	2937.76	**2532.21**	2689.20	2555.82	2770.80
	MAE	43.66	**41.69**	42.13	41.90	42.51
W-NLA	CD	-	0.9878	**0.9924**	-	-
	RMSE(m^3/day)	-	3211.78	**2759.18**	-	-
	MAE	-	45.61	**43.61**	-	-
W-PSR-NLA	CD	-	0.9840	**0.9877**	-	-
	RMSE(m^3/day)	-	3452.18	**2930.82**	-	-
	MAE	-	48.49	**45.59**	-	-

W-ANN's calculated values are more accurate than W-GEP in simulating peak points. This study eventually would be the extent of expectations in this study, but results are better than that for MLR, W-MLR, and GEP, W-GEP and ANN, W-ANN in forecasting the same case data. Also, the results of all four models presented revealed that the combination of both pre-processing methods (PSR and wavelet decomposition) at the same time would improve the accuracy, rather than

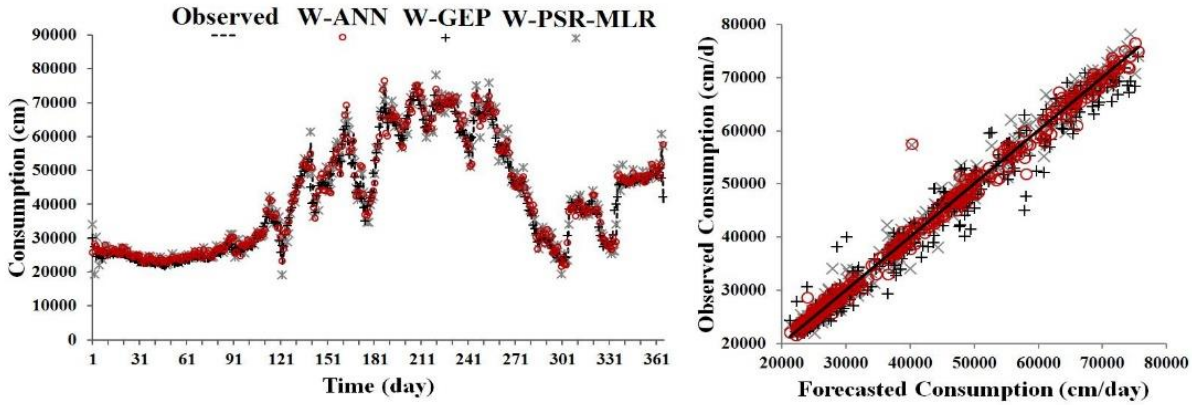


Figure 4.9 The performance of the ANN, GEP and MLR models in comparison with observed values

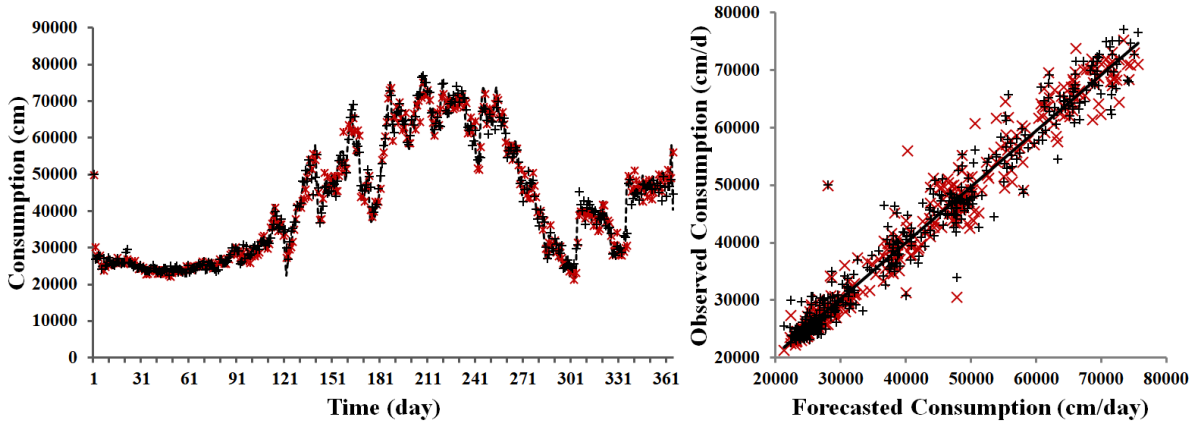


Figure 4.10 The performance of the NLA and Pre-Processed NLA in comparison with observed values.

applying one of them as is implied in Table 4.7. Figure 4.11 represents the performance of the models by residual values for the selected models with/without wavelet decomposition pre-processing. However, wavelet decomposition increased the values of evaluation criteria here, but the other models have advantages which makes them important. NLA simulated the lowest

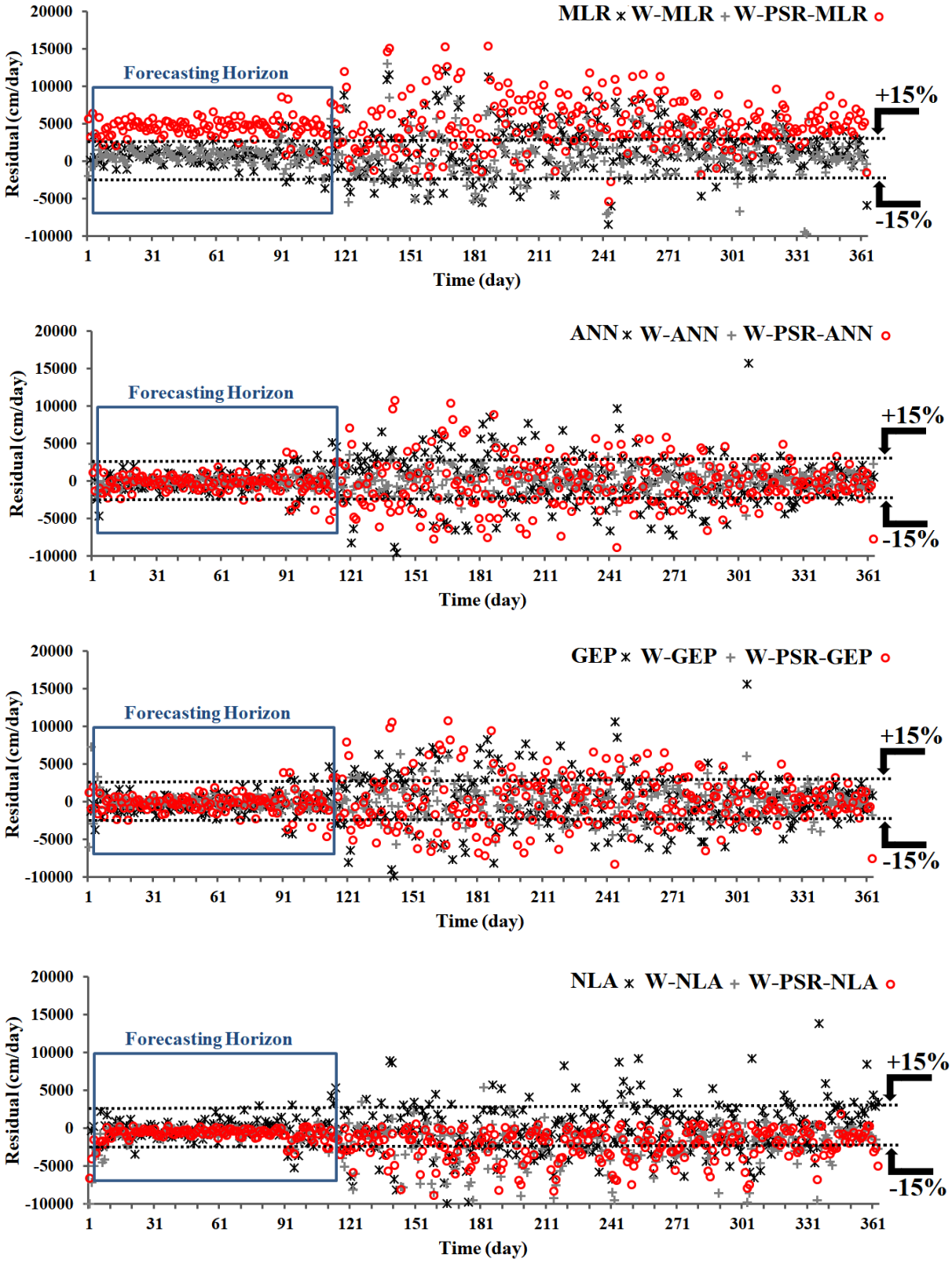


Figure 4.11 Residual values of the selected W-models.

consumption, whereas W-PSR-NLA simulated the peak consumption accurately. The peak values are decisive in managing water distribution network for satisfying allocation, and optimum pipeline design and the lowest values of consumption are essential in managing the optimum pressure within the network. Therefore, it can be concluded that the model with the highest accuracy is not an absolute optimum best model. To have an integrated dynamic WDS plan, knowledge of different models with their pros and cons in the different situation will improve the total efficiency of the application of any urban WDS. Therefore, it is recommended to evaluate models' performance in two separate parts as maximum values and minimum values along with evaluating criteria such as CD, RMSE, and MAE for the test period. The difference is not visible in Figures 4.9 and 4.10. Therefore, focusing on Figure 4.11, it shows the performance of models by residual values in the test period. In Figure 4.11, the residual values show the remarkable difference of performance of models. The results show the performance of NLA with both PSR and W-PSR pre-processing is the most accurate model among the other models based on the forecasting accuracy within the horizon.

4.6 Summary

Over the past decades, hydrologists have paid attention to data-driven modeling techniques. City governments and WDS operators are always looking for an accurate estimation of water consumption values, not only for the future but also focusing on probable failures like peak consumption and pressure values to manage the WDS pipelines. Therefore, the wide variety of modeling techniques such as artificial and evolutionary simulation methods are proposed by researchers. This chapter investigated the performance of four techniques (NLA, ANN, GEP, and MLR) in forecasting short-term water consumption of Kelowna City (BC, Canada). Six years daily dataset was employed for training and testing the models. The first five years were considered to

train the model and the last year as the test period. All three techniques performed considerably accurate, while the focus of this chapter was on improving the accuracy of the models for the same dataset. Firstly, the model was calibrated by different input combinations with 1-day lag time. Then, models were calibrated by the lag time of the data set (83-day) which was calculated by ACF method. WDT was combined with the models to capture multi-scale features of the signals by decomposing observed consumption values into sub-series. Five WDT functions (haar, Db2, db4, Sym2, and sym4) were employed to decompose the dataset. The results were then compared with the MLR, ANN, GEP and NLA when no pre-processing (PSR, WDT) was applied. The research results were more accurate than PSR. WDT have also improved the accuracy of models with PSR and without PSR. However, the impact of wavelet on the models with PSR was not as considerable as without PSR. The lowest error was reported by W-NLA among all alternative models in this chapter. Regarding the improvement of all models combining WDT and PSR, it is recommended to use the method in modeling and forecasting issues, especially about the dataset that the peak points are very critical in the case. The inherent behavior of dataset (deterministic or stochastic) can affect the performance of the pre-processing methods. Therefore, behavior of datasets should be investigated before deciding to combine any pre-process methods.

Chapter 5: Estimation of high-resolution water consumption time series[9]

5.1 Overview

To develop an integrated dynamic plan for water distribution systems (WDS), an approximate estimation for future values of water consumption is required. This plan is crucial for efficient customer satisfaction and the optimum allocation among industrial, commercial, and individual consumption. The estimation depends upon the increase of the rate of population and climatic changes. However, the information about influential variables (population growth rate, climate change, etc.) is limited. Hence, most of the research studies were done to evaluate water consumption for handling future probabilistic water disasters. Majority of well-known methods need to use high-resolution temporal scales of the data for reliable estimation of water consumption.

Nevertheless, only well-equipped supervisory control and data acquisition (SCADA) systems can record such high-resolution data (e.g., high-resolution temporal scale, proper time series of pressure values in pipelines, etc.). Therefore, development of a method to transfer low-resolution scales (e.g., annual and monthly values) to high-resolution (e.g., hourly or daily events) scales is necessary. This chapter reports a newly developed cascade method to transfer the resolution among the scales. Moreover, the results are compared with the previously introduced disaggregation technique. Finally, the role of resolution transfer in forecasting models is highlighted with the same

[9] A version of this chapter has been published as; Yousefi, P., Naser, G., Mohammadi, H. 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications, Intech. Yousefi, P., Naser, G., Mohammadi, H. 2018. Estimating High Resolution Temporal Scale of Water Demand Time Series – Disaggregation Approach (Case Study). HIC 2018. 13th International Conference on Hydro informatics 3, 2408-2416.

test case. Water consumption values of the City of Kelowna[10] is considered as the test case for this chapter.

5.2 Background

A reliable water distribution management plan of an urban area requires the understanding of the behavior of series and consumption patterns in different spatial and temporal resolutions [11,178]. Temporal high-resolution time series enables better estimation of peak consumption values that are very important in a short-term plan, alongside, developing a maintenance plan for pipeline systems in the long-term period to overcome probabilistic failures in the network [12,22,44]. These theories also have interpolation and extrapolation methods based on recorded values in SCADA systems. Those recorded values are time-dependent, therefore, making them unable to provide satisfying high temporal scales. This chapter aims to present a method that allows the transformation of the data from a low-resolution to high-resolution temporal scale. The concept of transformation from one temporal scale to another is a phenomenon and varies according to the particular case. Firstly, it is recommended to the researchers of this field to investigate the scaling behavior. Regarding literature about disaggregation techniques, the time series need to have scaling behavior. Sivakumar [179,180] considered the subject from the scaling point of view, and studied the possibility of scaling behavior in the suspended sediment load of rivers. Valencia [181] proposed disaggregation methods to transfer low-resolution time series to high-resolution temporal scales. They applied their proposed technique to disaggregate annual river discharge into seasonal values. Various downscaling and disaggregating methods were used to estimate high-resolution

[10] The test case in the articles are the City of Peachland (BC, Canada). To make this study consistence and compare the results off all techniques, the developed models are used to calculate the results for City of Kelowna (BC, Canada) as the test case.

values. Disaggregation including the value of rainfall in short-term temporal scales, down-scaling stream flow time series [182], evaluation of weather data [183–185] and river flow [186,187]. Random cascade method is one of the well-known disaggregation methods that has been used in the area of hydrology. Olsson [188] applied random cascade method to disaggregate rainfall values in Sweden. Burlando [189] evaluated two various stochastic disaggregation models based on random cascade method to gain observed values into a high-resolution scale. The mentioned methods were developed to apply in discontinuous data (e.g., rainfall).

This study aims to estimate the high-resolution temporal scales of urban drinking water consumption. A newly developed cascade method that is based on weights and probability statistics of the recorded data would be used to transfer continuous low-resolution, i.e., monthly values, into high-resolution daily values in Kelowna's public drinking water consumption (BC, Canada).

5.3 Methodology

5.3.1 Power Spectrum

The power spectrum is a standard tool in studying the properties scaling (fractal) time series [75,179,188]. If f is the frequency and β is an exponent, then the spectrum follows as power law equation:

$$E(f) = \propto f^{-\beta} \tag{5.1}$$

Equation (5.1) could indicate index of the existence or absence one-time scale specified in the scope exponent and therefore assuming behavior scaling (fractal) is possible. Power spectrum is also useful for studying the oscillations process. In the general stochastic process, power spectrum oscillates randomly for the constant value that indicates series varying frequency is greater than

other frequencies. For periodic or quasi-periodic series, peak values exist in special frequency only, and measured noise adds a continuous layer to the spectrum. Therefore, in the power spectrum, signals and noises are easily recognizable [179].

5.3.2 Random Cascade

To disaggregate the time series, the cascade model divides the value of the consumption (V) of each temporal scale into two scales (V_1, V_2). Each of the two scaled values resulted from multiplying the weight value (W_1, W_2; $W_1+W_2=1$) to (V) [190]**.** Figure 5.1 shows the brief scheme of the process in random cascade disaggregating.

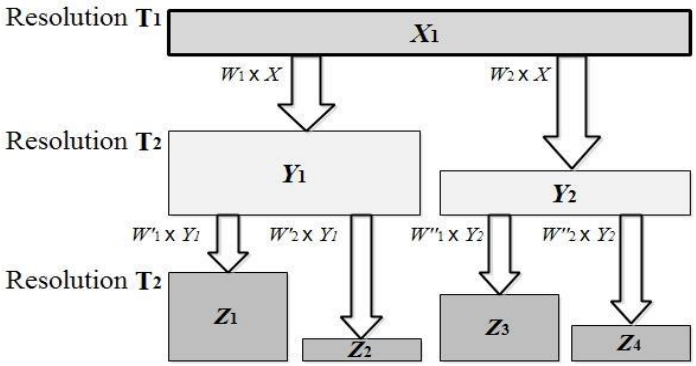


Figure 5.1 Schematic representation of cascade weight distribution

The idea of subdivision and boxes in cascade method has been taken from the proposed technique of [188]. This method disaggregates non-continuous rainfall data into short-term values. Since the consumption series are continuous values, this study suggests upgrading the technique to be applicable for continuous data (e.g., discharge, sedimentation, water consumption, etc.). Each box is subdivided into four; 1) have the value of demand be greater than both the previous and successive time; 2) lesser than the previous and greater than the successive time; 3) greater than the previous and lesser than the successive; 4) lesser than both previous and successive ones. The mean and median of the consumption values are considered as additional variables to divide different boxes. By combining all the information about the boxes, the probabilities of each box

P($x|x$) are estimated, and a proper weight (W_i) were given to each value to be divided into two different values.

5.3.3 Non-Linear Deterministic Method

To determine the degree of correlation of time series and the lag time, average mutual information (AMI) has been utilized (See Chapter 2). This function is a periodic phenomenon, which indicates stability of a relationship between repetitive values. It is assumed that $x_i, i = 1,2,3, \dots, n$ is a series of water consumption with a resolution of T_1 and aims to find the values of $(z_i)_k, k = 1,2,3, \dots, p$ in a high resolution, and also $p = T_1 \,/\, T_2$ and values of x_i according to $(z_i)_k, k = (W_i)_k = (W_i)_k \times x_i$ to $(z_i)_k$, $(W_i)_k$ is the distribution of transfer weights of X_i which is written to $(z_i)_k$ [75]. Figure 5.2 shows the scheme of the process.

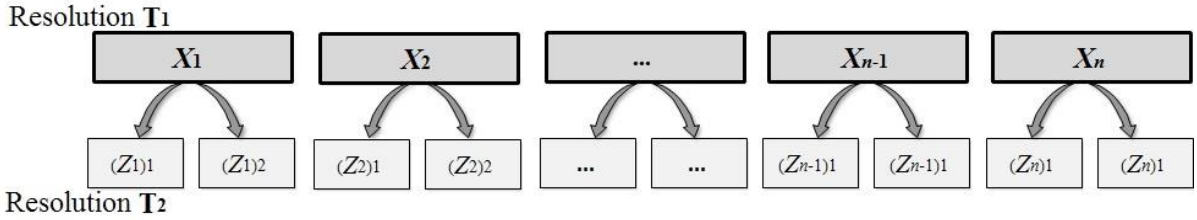


Figure 5.2 Schematic representation of distributions of weights between resolutions.

According to nonlinear behavior of demand data, the evolution of sub-mechanisms must be considered by using phase space reconstruction [97]. This reconstruction makes it possible to draw a relationship between the first and subsequent condition by a functional equation.

$$Y_{j+T} = f_T(Y_j) \tag{5.2}$$

By disaggregating the values of water consumption, the resolution increases from T_1 to T_2 and in this way the distributed weights of $(W_{n+1})_k$ can be determined.

5.3.4 Evaluation Criteria

This research measured the models' accuracy by correlation of coefficient (CC), root mean squared error (RMSE), and mean absolute error (MAE) defined as:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N_t}(R_i - \bar{R})\,(F_i - \bar{F})}{\sqrt{\sum_{i}^{N_t}(R_i - \bar{R})^2}\ \sqrt{\sum_{i}^{N_t}(F_i - \bar{F})^2}} \tag{5.3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_t}(R_i - F_i)^2}{N_t}} \tag{5.4}$$

$$MAE = \frac{1}{N_t}\sum_{i=1}^{N_t}|R_i - F_i| \tag{5.5}$$

where N_t is the number of the recorded consumption values in the dataset, R and F are the recorded and disaggregated/forecasted values of water consumption, respectively. $\bar{R}$ and $\bar{F}$ are the mean of the recorded and disaggregated/forecasted consumption values, respectively. The range of CC is -1 and 1, where a larger positive value of CC and a lower value of RMSE and MAE indicates better performance.

5.4 Case Study

The City of Kelowna is selected to evaluate the performance of the model based on their drinking water consumption (Table 5.1). Along with the consideration of scaling behavior of the used consumption data, embedded dimensions with phase space reconstruction, lag time, and disaggregation calculations have been done in five different temporal scales. Then the results of disaggregated values are compared with recorded data of the city. Six years of monthly data are considered for the test case. The test data in this chapter is the same as the previous chapters. As

seen in Table 5.1, the results of data characteristics are different from the previous chapters. Since the number of data that are used for disaggregation techniques should be equal to 2^n (n=1,2,3…), compared to previous chapters, the number of used data is different. Table 5.1 shows the results for the used number of data for Kelowna water consumption values in five temporal scales. Figure 5.3 show the time variations of the water consumption dataset for the selected number of data.

Table 5.1 Statistics of water consumption of Kelowna City in different temporal resolutions.

Property	Daily	2-Day	4-Day	8-Day	16-Day	32-Day
Number of Data	2048	1024	512	256	128	64
Max. value (m^3)	114597	210741	387912	748720	1370023	2617245
Min. value (m^3)	14124.00	31477	67642.00	153472	314633	655151
Average (m^3)	44466	88932	177865	355729	711458	1422916
Standard deviation (m^3)	19920	39518	77977	153522	295206	571961
Coefficient of variation	0.45	0.44	0.44	0.43	0.41	0.40
Skew	0.68	0.66	0.61	0.60	0.49	0.46
Kurtosis	-0.44	-0.54	-0.69	-0.74	-0.99	-1.04

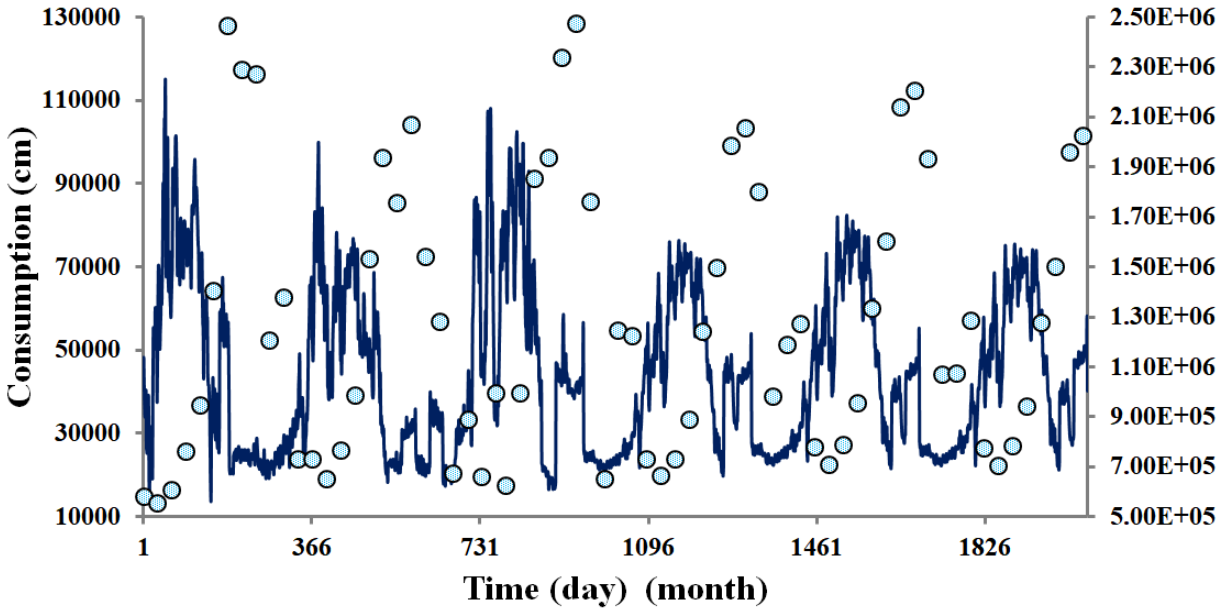


Figure 5.3 Time series plot of daily and monthly temporal scale of water consumption values

5.5 Results and Discussion

Figure 5.4 shows the result of power spectrum for daily scale of water consumption. The value of $\boldsymbol{\beta}$ can be calculated by the gradient of line, approximately 1.775; which is greater than 1, and represents scaling behavior of the time series [75].

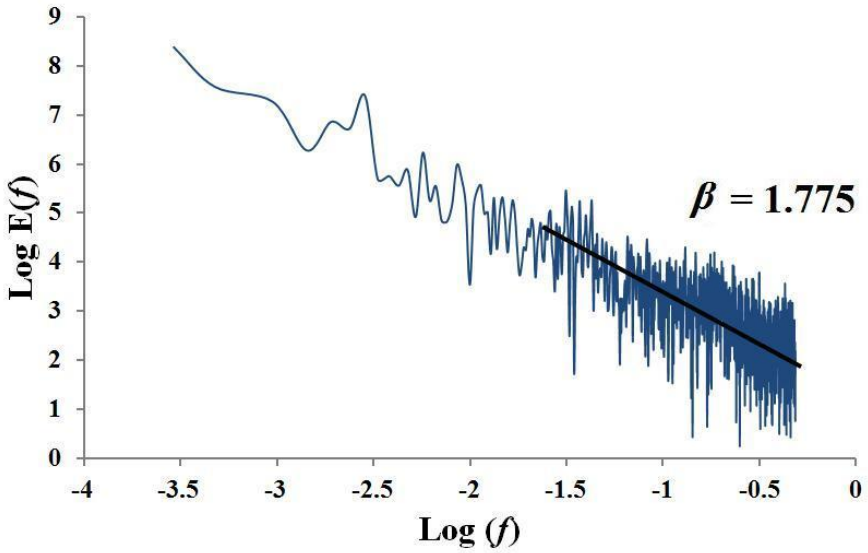
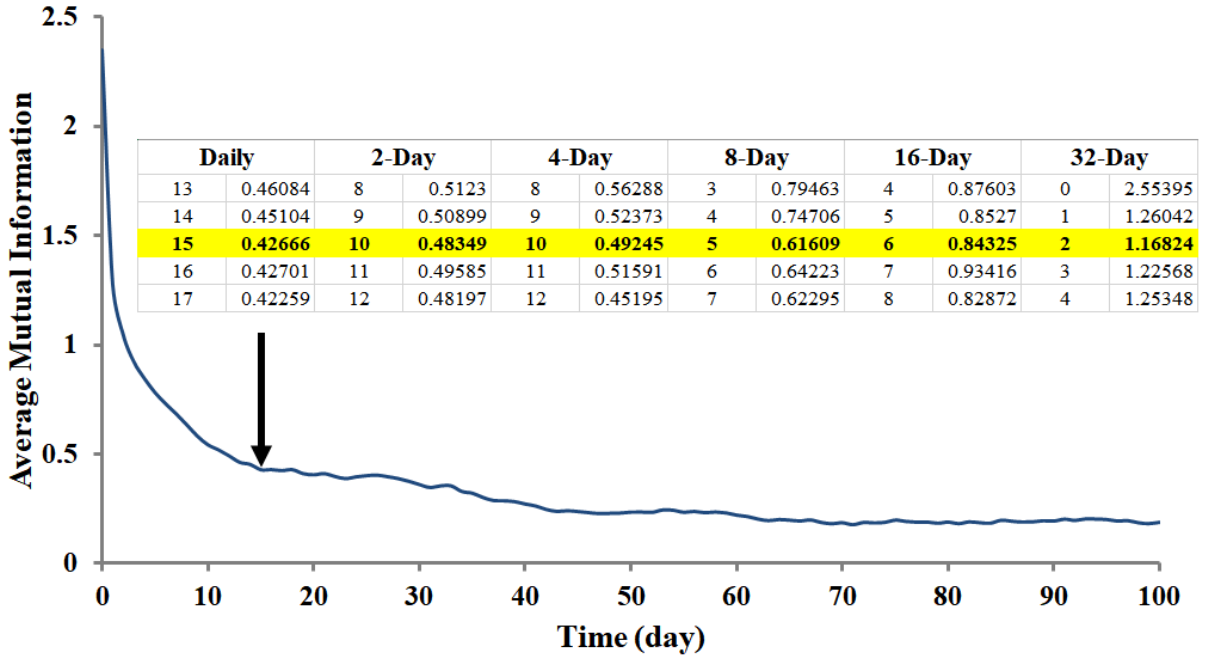


Figure 5.4 Scaling analysis of time series.

Therefore, the test case may potentially be used for transferring the temporal scales. This chapter aims to disaggregate the time series from the low-resolution scale into high resolution. Since the disaggregation techniques are based on phase space in relation with previous scales, availability of scaling behavior for the test case is necessary.

To compare the results of the continuous cascade method with another well-known method, the non-linear deterministic method is used to disaggregate the value of water consumption. Regarding the literature, the application of this method is well-known among researchers and was applied to various hydrology areas (discharge, sediment, climate, etc.). As mentioned previously, the number of used data in this chapter is changed. Therefore, investigation of the availability of chaos in the new set is needed. Firstly, to calculate the lag time for the series in each temporal scale, average

mutual information method (AMI) is applied. Figure 5.5 shows the first local minimum as the lag time for the scale. Figure 5.5 shows the results of time determining and the values of 15, 10, 10, 5, 6 and 2 are for the daily, 2-, 4-, 8-, 16-days and 32-days temporal scale consecutively.



Daily		2-Day		4-Day		8-Day		16-Day		32-Day	
13	0.46084	8	0.5123	8	0.56288	3	0.79463	4	0.87603	0	2.55395
14	0.45104	9	0.50899	9	0.52373	4	0.74706	5	0.8527	1	1.26042
15	**0.42666**	**10**	**0.48349**	**10**	**0.49245**	**5**	**0.61609**	**6**	**0.84325**	**2**	**1.16824**
16	0.42701	11	0.49585	11	0.51591	6	0.64223	7	0.93416	3	1.22568
17	0.42259	12	0.48197	12	0.45195	7	0.62295	8	0.82872	4	1.25348

Figure 5.5 Average mutual information (τ) for the daily values of consumption and the values of five temporal scales.

Figure 5.6a plots the results for correlation function versus log (r) for different embedding dimensions (m) varying in the range of 1 to 20 for the daily values of water consumption. To investigate the availability of chaotic behavior, the correlation exponents, C_e (m), were determined. The results are shown in Figure 5.6. Figure 5.6b shows that the correlation exponent increases with the embedding dimension up to a certain value and then remains steady. The figure reveals that the slope of larger (m) would become constant for all temporal scales. The saturation of the correlation exponent in a specific embedding dimension, is an indication of the presence of chaos in the dataset. Investigation of the chaotic behavior in different time scales for the test case of this

study is sufficient for following the objectives of this research. Furthermore, in chapter two, the results showed the chaotic behavior of the time series. By investigation of this time series, the embedding

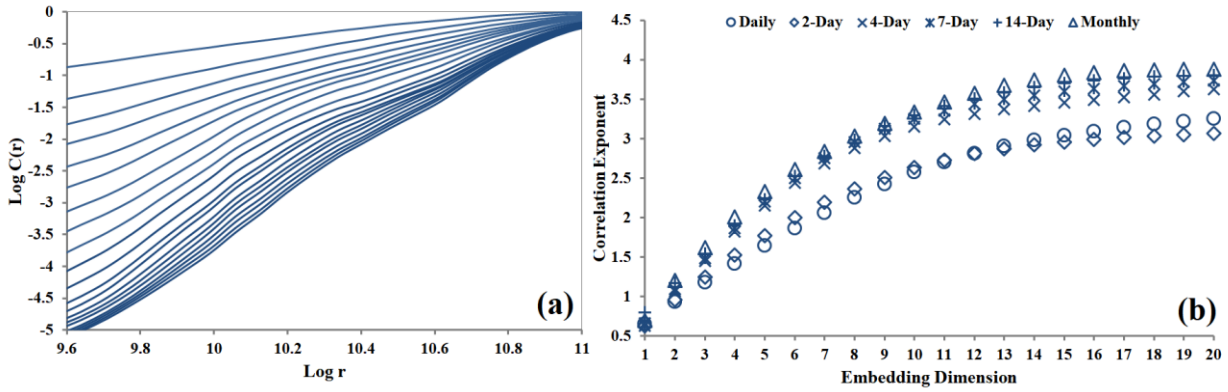


Figure 5.6 The relation between correlation function C(r) and r by different embedding dimensions for daily values; (b) saturation of correlation dimension C_e *(m)* with embedding dimension *m* for different temporal scales.

dimension of 17 seems to be sufficient to explain the dynamic of the system (the results are shown in chapter two). The saturation values of the correlation exponent for all temporal resolutions are 3.14, 3.01, 3.52, 3.66, 3.76, and 3.86 for daily, 2-, 4-, 8-, 16- and 32-days, respectively. Table 5.2 shows the results of lag time and correlation exponent for all the temporal resolutions. The condition of $C_e < 2\log N$ (with N as the number of data) was satisfied for all the chaotic temporal scales except 32-days resolution (C_e is larger than the condition) [111]. To the best knowledge of the authors, the lower number of samples for the 32-days resolution is the reason of having a lower chaotic behavior.

To transfer the resolutions from the lower scale to high resolution, the time series were divided into two groups. The first group is used for reconstruction of phase space and the second group, shows the performed fitness evaluation. Two disaggregation techniques are applied in transferring

Table 5.2 Average mutual information and correlation exponent values in different temporal resolutions.

Time Scale	AMI	Correlation Exponent ($C_{e)}$	2 Log $N > C_e$
Daily	15	3.14	6.62
2-Day	10	3.01	6.02
4-Day	10	3.52	5.41
8-Day	5	3.66	4.82
16-Day	6	3.76	4.21
32-Day	2	3.86	3.61

the resolutions. Based on the chaos approach, it is the non-linear deterministic approach [75]. The results of chaos and scaling behavior investigation proved the eligibility of the time series to be transferred by this method. Embedding dimension of m=1 to 20 shows the dynamic of reconstructed transportation; also, zero-degree approximation is utilized to find neighbors in the disaggregation process. The presented results are obtained from disaggregation of water consumption values (See Table 5.3). The results showed the acceptable performance of non-linear deterministic disaggregation method. Based on the selection criteria, the highest accuracy of the results is highlighted in Table 5.3 for the test period. 75% of the available data are used for training, and the remaining 25% is used for testifying the models. The embedding dimension of each resolution is different. The embedding dimension of 15 (m=15) is the selected dimension for the most accurate disaggregated among dimensions 1 to 20. Regarding the optimum embedding dimension of daily time series (see chapter 3), the results are reasonable. Figure 5.7 shows the calculated values in comparison with observed values for the daily temporal scale. The selected values are for the 15^{th} embedding dimension with CC= 0.954, RMSE= 5248 and MAE= 4260. Also, the number of samples for the test period for disaggregation of 2-days to 1-days is 512 (25% of 2048 samples).

Table 5.3 Accuracy of the disaggregated values in different temporal scales.

m	2 to 1			4 to 2			8 to 4			16 to 8			32 to 16		
	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE
1	0.948	5606	4379	0.949	10991	8852	0.941	23309	18714	0.930	49796	35463	0.894	122547	93522
2	0.947	5646	4524	0.950	10908	8336	0.939	23843	19641	0.936	47749	36990	0.906	112960	89986
3	0.942	5964	4664	0.959	9813	7719	0.938	24247	18948	0.938	46606	36380	0.907	113099	78802
4	0.948	5601	4498	0.946	11363	8908	0.943	23142	18493	0.938	46848	37150	0.905	114835	81567
5	0.941	6025	4810	0.942	11763	8818	0.942	23306	19071	0.937	47152	35793	0.949	83014	64150
6	0.947	5686	4431	0.942	11796	9264	0.929	25997	20538	0.947	43465	32850	0.928	98817	69262
7	0.949	5570	4520	0.937	12318	9657	0.940	23537	18129	0.940	45982	35222	0.945	85259	67723
8	0.948	5620	4423	0.936	12433	10149	0.942	23165	18786	0.940	45741	34265	0.946	85406	60441
9	0.943	5860	4568	0.948	11157	8785	0.932	25373	20501	0.941	45456	35936	**0.946**	**84829**	**55587**
10	0.945	5756	4621	0.941	11914	9184	0.938	24298	19899	0.947	42959	32496	0.933	94099	63758
11	0.946	5749	4543	0.947	11291	9012	0.931	25392	19538	0.938	47262	38678	0.920	104689	85177
12	0.945	5782	4575	**0.952**	**10672**	**8304**	0.927	26249	21165	0.942	45095	33464	0.935	94084	78779
13	0.939	6129	5012	0.948	11174	9077	0.927	26308	20781	0.937	47289	34186	0.911	109416	82365
14	0.937	6236	5117	0.951	10755	8370	**0.945**	**22576**	**17808**	0.921	53637	42791	0.924	100437	86124
15	**0.954**	**5248**	**4260**	0.951	10761	8128	0.929	25850	20102	0.930	49619	34933	0.927	97506	74771
16	0.953	5296	4268	0.950	10926	8430	0.931	25434	19727	0.917	54834	43224	0.935	92533	72886
17	0.953	5323	4244	0.942	11745	9025	0.942	23274	18767	0.927	50863	38814	0.923	103236	78302
18	0.951	5452	4290	0.944	11583	8884	0.935	24726	19327	**0.956**	**39343**	**32763**	0.934	93216	72018
19	0.948	5619	4529	0.938	12192	9416	0.926	26477	21918	0.931	49433	38821	0.936	94829	75587
20	0.943	5892	4783	0.937	12441	9790	0.927	26556	21272	0.956	38924	30803	0.933	94099	63758

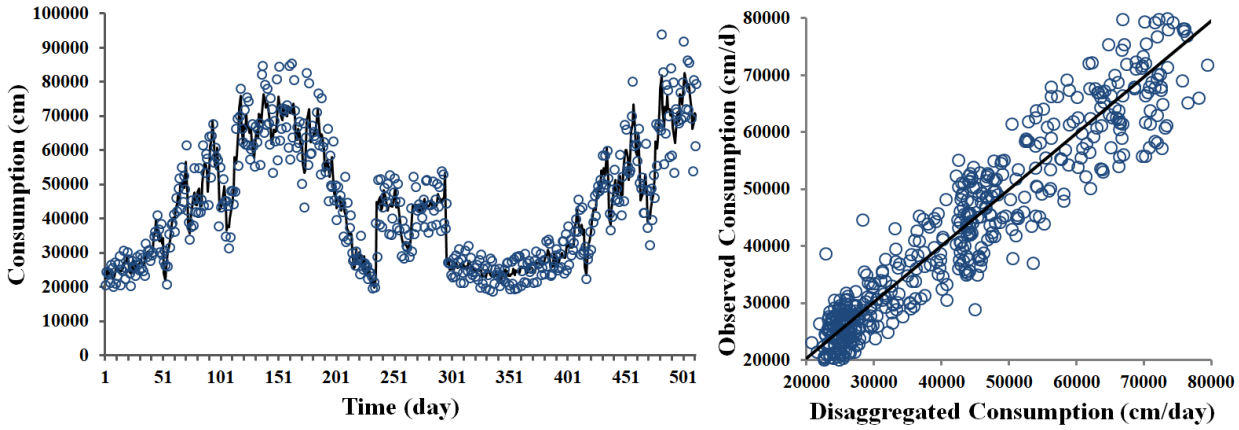


Figure 5.7 Observed in comparison with Disaggregated values of water consumption for daily temporal scale in the test period with continuous cascade.

However, the model used in the present study considered one variable (consumption values); applying the impact of related variables in water consumption such as weather information, holidays, peak hours, etc., seems to be more helpful for calibration of the model to estimate subs' weight more accurately. It has not yet been reported to consider the mentioned variables in the models if they need to have scaling behavior, in which this study suggests for future work in the field to upgrade the continuous disaggregation techniques to be applicable in different conditions. Considering the application of both methods that make the high-resolution temporal scale available from lower resolution, their accuracy is acceptable. Comparing the results of each technique, non-linear deterministic was more accurate than the continuous cascade technique. Figure 5.9 shows the residual of the selected embedding dimensions of non-linear deterministic and continuous cascade methods.

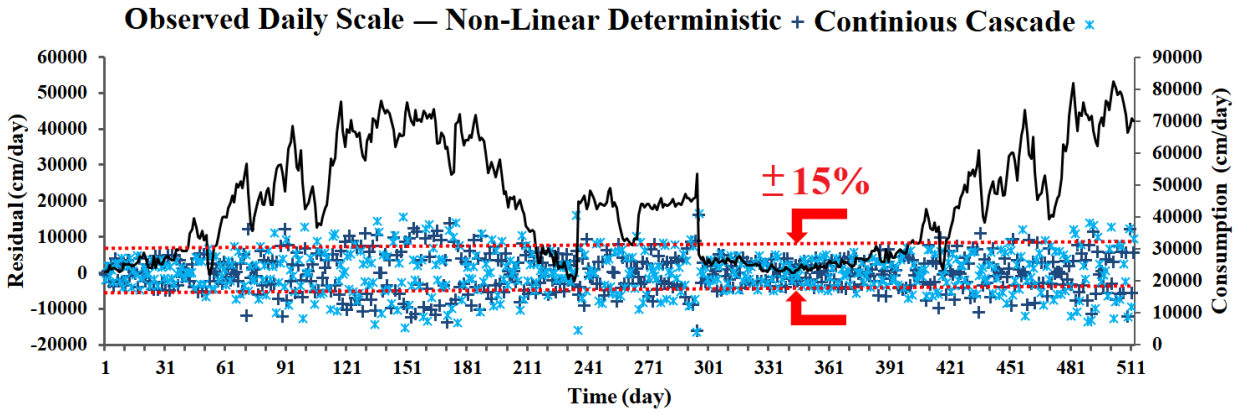


Figure 5.8 Residual values of the selected models for 2-Days to 1-Day in comparison with observed daily consumption value.

As it is seen in the figure, both disaggregation methods have weaknesses in simulation of high frequency consumption and the peak values. Neglecting this weakness, the ability of transferring the resolution from lower temporal scale to higher temporal scale bring the attention to the forecasting area. Regarding the literature and findings of this research, the accuracy of the forecasting models has laid upon the availability of higher resolution input values.

5.6 Summary

In this chapter, continuous random cascade and non-linear deterministic methods were considered in order to attain high-resolution temporal scale data for given water consumption of Kelowna. The power spectrum was used to investigate the scaling behavior of the time series. The value of the power spectrum was reported 1.775, which is more than 1 and represents the presence of scaling behavior of the time series. Downscaling was applied to convert monthly (32-Days) values into daily values for both methods. The domain of the given weight for random cascade method was in the range of 0.3 to 0.7 for each defined sub boxes, which resulted from the investigation of the relation for different temporal scales. It was suggested to use other active variables on consumption values (e.g., temperature, humidity, average income, population growth, etc.) in order

to calculate more accurate values of weight for the subs. The well-defined weight will improve the accuracy of continuous random cascade model in disaggregating. The value of CC, RMSE, and MAE for downscaling 2-days to daily scale was more accurate than the other temporal scales. Also, to disaggregate consumption values by non-linear deterministic method, various embedding dimensions (EDs) varying from 1 to 20 were considered for different temporal scales. The lag time for different temporal scales was calculated by average mutual information (AMI). The optimum ED was 15 for the daily resolution, that was similar to the results of chapter 2. The difference in the value of the optimum embedding dimension was because of the difference in the number of samples. By comparing the applications of the two techniques, it was suggested that non-linear deterministic approach is more effective than continuous random cascade method in simulating the peak consumption values in this stage of the research (without applying the role of influential variables). The focus of this chapter was to realize the application of disaggregation methods, which are mostly used for hydrology issues, and into urban management areas (i.e., drinking water consumption). Having the mentioned methods (or similar techniques) in the area will help the governors of the cities with low SCADA equipment to solve considerable problems. However, the value of fitness criteria of the used methods was not perfect; the total consumption in large time scale is precisely equal to the recorded values. The total values of consumption are considered in WDS to design the network and management planning. The disaggregation methods can be considered as an adequate replacement for any other non-linear evolutionary forecasting methods because the cumulative values of their results are less accurate than disaggregation methods.

Chapter 6: Anomaly (Leakage) detection within a water distribution system

6.1 Overview

Drinking water resources are becoming more scares and that directly affect humans' life. It is because of the direct relationship between population growth, climate change, industrialization and availability of fresh water. International water supply association (IWSA) estimate that 30% of fresh water is lost due to uncounted water including leakage, metering errors, pipeline burst and theft. This amount of loss is very considerable compared to the limited available drinking water resources. Therefore, it attracts municipalities' and other authorities' attention to control the loss. In addition to environmental and economic losses from leakage, it is risky for public health, as the water leaking from pipes can potentially take the contaminates into the resources. One of the methods to control these losses within a system is managing the distribution system to improve the reliability and availability of water supply and efficient operation of the system [191]. Implementing an audit system within a water distribution system (WDS) that is based on leakage control and detection program is a way to reduce the loss along with the aforementioned risks. This audit and detection system collect large amounts of data from the WDS. Interpreting this data is challenging due to influential variables on the water consumption, include seasonal variation, location of water supply, daily and weekly cycle, and peak demand cycles. Therefore, the reliable leakage management plan acts immediately, catching any anomalies on the allocation volume.

Leak detection systems work based on detecting changes in target variables (e.g. acoustic, pressure, etc.) that happen following leakage. Leakage management with the acoustic equipment are modern computer-based instruments that measure the difference of sound of vibration within two points of a pile line to determine the leakage. Though, this method helps to control the loss within a WDS, the large cost of maintenance, operation, online detection, equipment operation,

especially in a large WDS, make the method unsatisfactory. This section tries to suggest solutions to overcome the mentioned disadvantages.

6.2 Background

Current solutions and methodologies that are controlling and detecting the leak/burst in a WDS have various principals. Methods based on utilizing hardware and equipment's like noise detection, gas injection and acoustic sensing are used for these techniques of leak/burst detection [192]. However, these techniques are considered as expensive leak/burst detection methods because of professional operator resource and slow running operation system. Therefore, recent studies are focused on developing techniques that are not costly and time consuming.

Other methods that use fluid transient to examine the pipeline have been used, including transient analysis, frequency domain method and unsteady signal investigation frequently [193–197]. These techniques have lower cost compared to the equipment-based methods because they are based on numerical modelling methods. To the numerical models, they need pressure, or another dataset measured within the WDS that are sampled frequently. For example pressure abnormalities can be resulted from; 1) large water loss resulting from a burst, 2) power loss in a district which affects the performance of pumps in a network; 3) maintenance activities that results blocking a zone in a district; 4) operational activities such as reservoir, tank, pump, pipeline repair and change; 5) communication loss as a result of low power of signal which stop reading the data or sending the records to the cloud. These can affect the rate of water flow and velocity in pipelines, water quality in the network and the influential variables that are in relation to water consumption. To record these samples, large number of sensors should be implemented in the network. While, using numerical methods have pros about the costs, implementing the measuring equipment lead to high

expenditures. In total, these techniques have not succeeded in comparison with equipment-based methods [198].

Methods that use data maiming and artificial intelligence (AI) to process the operational factors (e.g. pressure, flow, quality, velocity etc.) are getting more useful compared to the previous equipment-based methods. Although data mining methods work based on the recorded data, the applicability and reliability of these techniques are still doubtful. Since, the anomaly detection element is on the soft-computing method, the devices for data acquisition are not necessarily fascinated and costly compared to the devices used for previous techniques. Artificial neural networks (ANN), fuzzy systems, Kalman Filter, and hierarchical rule-based approach are such techniques that are proposed for anomaly detection [199–203]. AI methods don't require high-performance devices, while also sampling less frequently compared those transient analysis methods [198]. Moreover, the required number of pressure or flow measurements are lower than acoustic recording devices, therefore, there is no need for large number of implemented devices within the network. These advantages highlight the popularity of AI methods in leakage detection. This chapter explore alternative soft-computing data processing method for anomaly detection on WDS. The method is selected based on the application of models stated in previous chapters. Also, the application of the models is investigated with/without pre-processing to probe into improvement of the reliability of detection models.

6.3 Methodology

6.3.1 Hydraulic Simulation

The detection system uses data collected by the WDS simulation by way of virtual leaks within the system. AI based techniques and statistical data analysis are used to evaluate the anomalies and detect virtual leakage(s) in the system. A section of the district is considered as the test case.

Figure 6.3 locates the selected zone of anomaly detection. As figure 6.3 reveals, two leaks are defined independently in the system that they happen with a time lag and then by combining available information from all available signals and techniques to detect the leak events. The simulated network is based on peak consumption to design the pipelines and the needs of tanks, pumps, and valves. EPANET V2 is used for simulating the peak consumption within the data. The designed details of the EPANET model for the WDS is provided by the City unitality.

The anomaly detection is simulated in seven phases: (1) perform the WDS without any defined leak within the target zone; (2) define virtual leaks in two different location (with the base of 5 LPS) and perform the WDS; (3) interpret the software output in combination with the actual values in the same resolution (6 hours); (4) analyzing the data and investigation of availability of chaos in the dataset and phase space reconstruction for the models; (5) applying the selected forecasting models and pre-processing analysis to define the forecasting horizon; (6) AI memory setting based on the forecasted models' output; and (7) anomaly detection from the simulated consumption values by using the data of the target reservoir and zone's main pipeline.

6.3.2 Models Implementation

Correlation dimension method is used to investigate the availably of chaos in the dataset. for phase space reconstruction, average mutual information (AMI) is employed to calculate the lag time (For further details see chapter 2). Non-linear approximation (NLA) and ANNs are the selected forecasting methods (For further details see chapter 3). To analyze the data and anomalies, wavelet decomposition is employed as the diagnosis and pre-processing approach (For further details see chapter 4). Figure 6.1 shows the two spots of artificial leakage simulation within the target district.

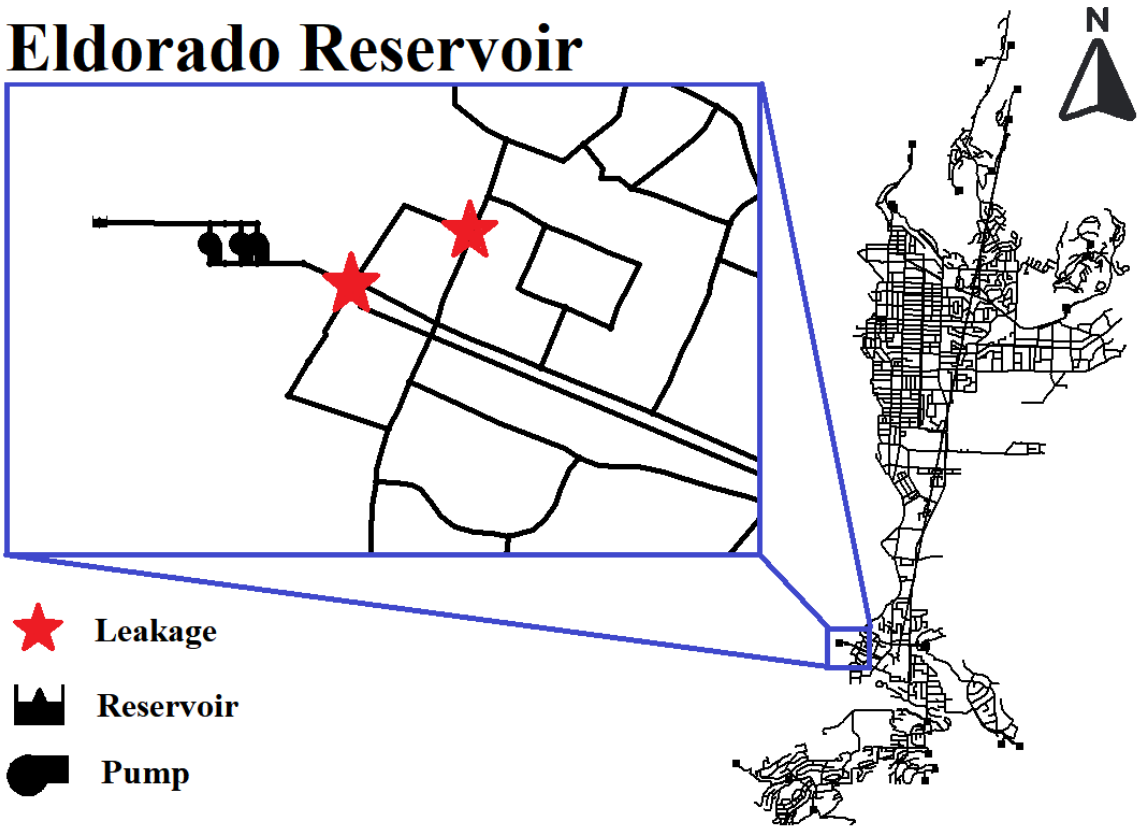


Figure 6.1 Location of the test zone and the artificial leakages within WDS of the City of Kelowna.

6.4 Case Study

Lake Okanagan, Mission Creek, Mill Creek, Scotty Creek, Hydraulic Creek, and numerous wells are supplying the City of Kelowna's water distribution system and provide service to approximately 65,000 residents plus commercial and business sectors [204]. Poplar Point, Eldorado, and Cedar Creek pump stations service 99% of the population. Additionally, the Swick road pump station services approximately 300 residents [204]. This study considered the hydraulic model that is simulated by EPANET V2. The hydraulic simulation of the district is provided by the City utility. The output data of the simulated WDS is synchronized in 6hour resolutions to cover four different demand times within 24 hours. The consumption pattern is defined based on actual data in this research. Figure 6.2 shows the district water consumption pattern which is based

on average consumption for the actual test case for hourly data since January 1, 2010 to December 30, 2016. All datasets are divided into two groups, as the training and the test period to calibrate the estimation models for data anomaly detection. Figure 6.3 presents the pipeline within the WDS district.

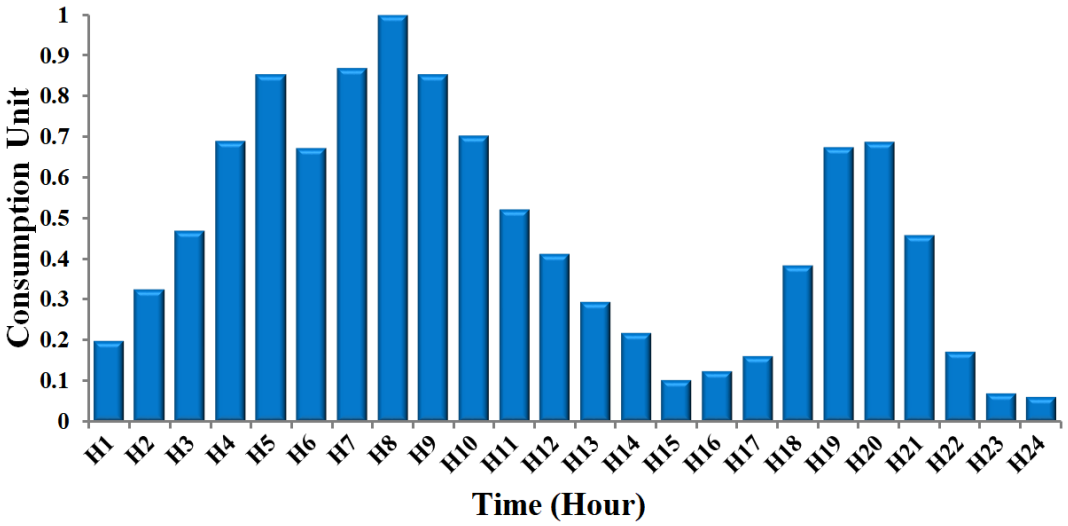


Figure 6.2 The City of Kelowna's water consumption pattern within 24 hours.

6.5 Results and Discussion

6.5.1 Review of Data Records

Hourly water consumption for the target zone, as mentioned above, have been selected from the hydraulic model of the WDS of the city of Kelowna. Three datasets have been simulated; (1) Consumption value supplied by Eldorado reservoir in the target district; (2) Consumption value of the reservoirs with one-leaked artificial node; (3) Consumption value of the reservoirs with two-leaked artificial node. Figure 6.4 shows the variation of 6-hour water consumption without any leakage within the district. Since the simulated WDS is based on the peak values to design the system, the time series is not based on actual consumption values. Therefore, the normalized unit

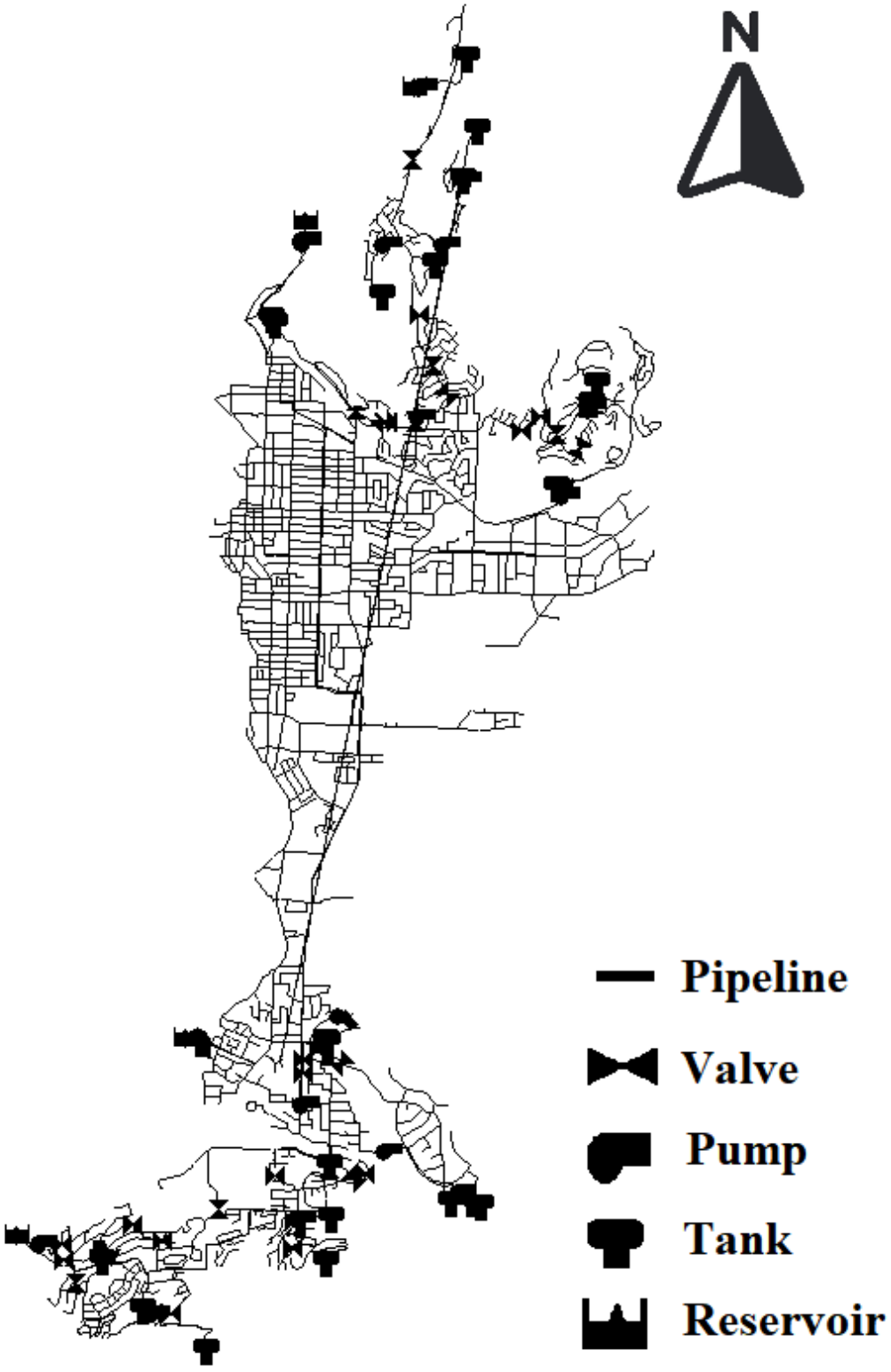


Figure 6.3 Water Distribution System of the City of Kelowna (EPANET Layer).

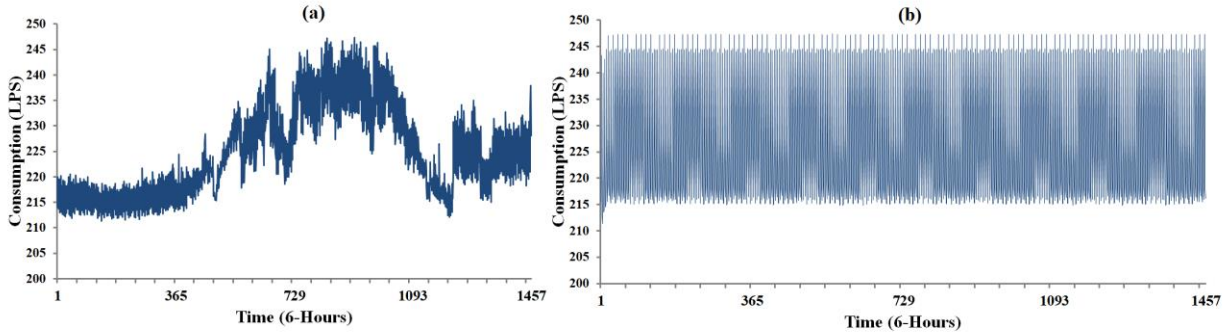


Figure 6.4 Time series plot of one-year (a) actual weighted consumption values; (b) simulated consumption values.

of the real data of the same time scale (6-hours) is used to transfer the hydraulic output into reliable actual data (Figure 6.4a, b). The data is divided into two parts for the training and testing period. The first 10 months (with the 6-hours resolution) is used for training the models and the last two months to evaluate the performance of trained models. The artificial leakages are employed in the second week of each month in the test period on the same day and same hour to analyze the performance of detection models and evaluate their accuracy (day 10^{th} at 12:00 pm). Table 6.1 shows the characteristics of the dataset in the test case for the three datasets.

Table 6.1 Characteristics of the water consumption values for three scenarios in target zone.

Property	Without Leak	1-Leak	2-Leak
Number of Data	1457	1457	1457
Max. value (L/S)	247.40	250.14	266.74
Min. value (L/S)	211.37	211.37	211.37
Average (L/S)	224.87	225.28	226.20
Standard deviation (L/S)	8.41	9.45	10.21
Coefficient of variation	0.04	0.04	0.05
Skew	0.57	0.74	0.74
Kurtosis	-0.57	-0.40	-0.31

6.5.2 Phase Space Reconstruction and Investigation of Chaotic[11]

The investigation beyond the availability of chaos in the test case is proved in chapter 2 based on the results of all temporal resolutions for the actual consumption values of the test case for the whole WDS districts. Although, the data used in the previous chapters was the total consumption of the day, this chapter uses the rate of consumption in a specific time represented as liter per second. This chapter investigates the chaotic behavior by applying average mutual information (chapter 2's results were based on AMI) and correlation dimension. The value of lag time is considered as the first local minimum (figure 6.5). Previously, both lag time calculation methods have been investigated, and the results showed the priority of AMI in comparison with autocorrelation function (AFC) in chaos investigation. Figure 6.6 presents the reconstructed phase space based on the lag time calculated by AMI.

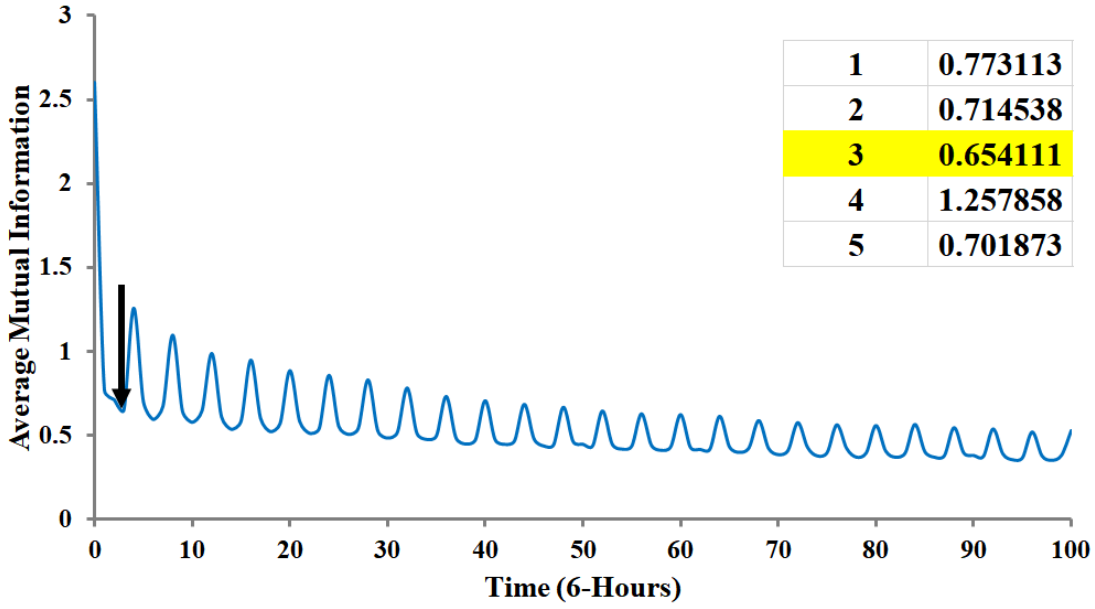


Figure 6.5 Average Mutual Information for the weighted consumption value.

[11] Further details of phase space reconstruction are available in sections 2.3.1 and 4.3.5.1.

The results show 3 scales (1-time scale = 6 hours) as the lag time of the time series. Therefore, 18-hours is used to design combinations of inputs as phase space for the time series. In this chapter, the difference between the 1st hour and 18th hour is used as a delay period for the phase space reconstruction (PSR) with varying embedding dimensions from 1 to 20 (m_1: D_t; m_2: D_t, $D_{t+\tau}$; m_{20}: D_t, …, $D_{t+20\tau}$).

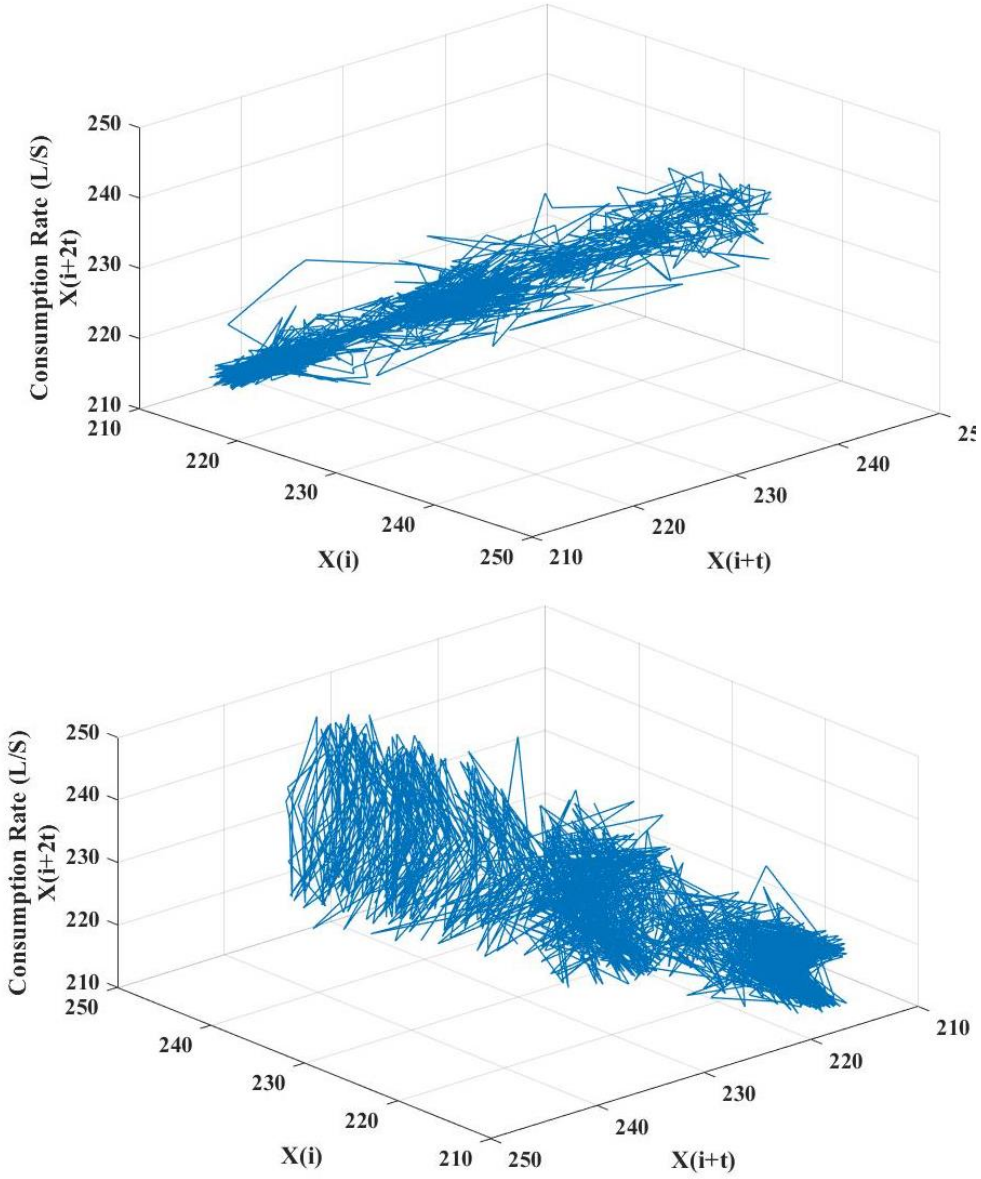


Figure 6.6 Reconstructed phase space by (τ and 2τ -day lag time), for the weighted consumption value.

Figure 6.7 shows the relation of $C(r)$ with r and correlation exponent by varying m. Regarding Figure 6.7, the value of correlation exponent increases by m, and as m=9, the correlation exponent reaches a specific value (*Ce* =2.83). The value of *Ce* in chapter 2 was reported as 3.50 and section three 3.41, which was based on AMI and ACF for the total actual consumption values of the district, respectively. However, for both *Ce*, the upper value of 4 is considered as the number of active variables to define the system, the lower value of the correlation exponent for the weighted simulated values is related to the small size of the district with the limited simulation joints. Therefore, the availably of chaos in the dataset gives opportunity to use the most accurate forecasting model in this research (Nonlinear local approximation).

6.5.3 Forecasting Model - Non-linear Local Approximation[12]

Forecasted values were evaluated for the embedding dimensions (ranging from 2 to 10) at lag times 1 and 3 unit (6 hours each unit) to forecast 6-hours-ahead (lead time T=1).

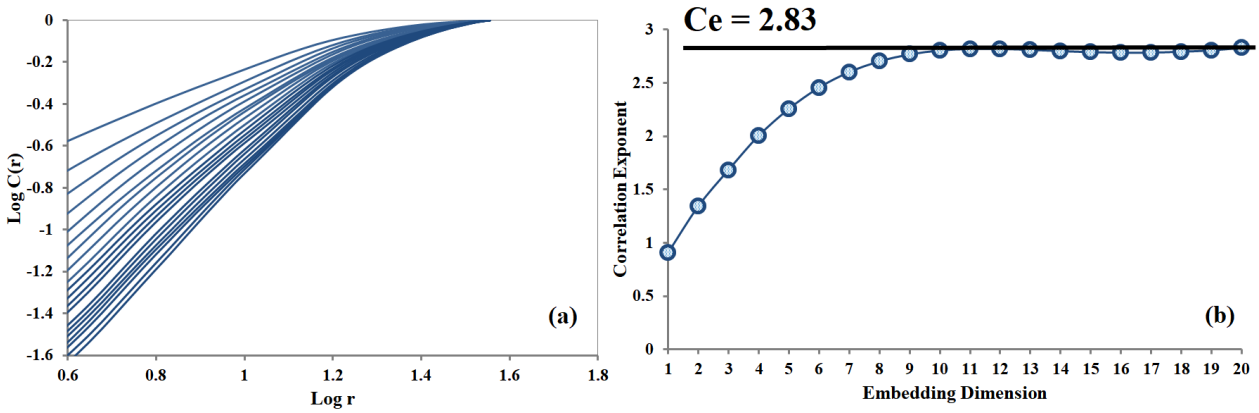


Figure 6.7 (a) The relation between correlation function C(r) and r by various m; (b) Saturation of correlation dimension Ce(m) with embedding dimensions.

[12] Further details of non-linear local approximation forecasting model are available in section 3.3.1.

For m = 2 for a lag time of 1 unit (6 hours), two variables D_t and D_{t-1}, and for m = 2 for a lag time of 3 units (18 hours) two variables D_t and D_{t-3} were used as input variables to forecast 1-day ahead (D_{t+1}). For m= 3 for the lag time of 1 and 3, three variables D_t, D_{t-1}, D_{t-2} and D_t, D_{t-3}, D_{t-6} were used as input variables to forecast 1unit ahead. Moreover, phase space was reconstructed for $m > 3$.

Table 6.2 Forecasted results of NLA and PSR-NLA for embedding dimensions in the test period.

NLA , τ =1 T=1				PSR-NLA , τ =3 T=1			
m	**CC**	**RMSE***	**MAE**	***m***	**CC**	**RMSE***	**MAE**
1							
2	-0.019	6.58	2.35	**2**	0.830	2.53	1.36
3	0.148	5.37	2.07	**3**	**0.830**	**2.53**	**1.34**
4	0.213	5.00	1.94	**4**	0.815	2.68	1.37
5	0.803	2.73	1.39	**5**	0.787	2.84	1.41
6	0.809	2.70	1.37	**6**	0.818	2.66	1.37
7	0.806	2.71	1.37	**7**	0.812	2.71	1.39
8	0.789	2.86	1.40	**8**	0.796	2.80	1.42
9	0.783	2.89	1.42	**9**	0.771	2.97	1.46
10	**0.817**	**2.67**	**1.36**	**10**	0.808	2.72	1.40
Tot	**0.594**	**3.63**	**1.61**	**Tot**	**0.807**	**2.72**	**1.39**
Best	**0.817**	**2.67**	**1.35**	**Best**	**0.830**	**2.53**	**1.34**
EM	**18**	**18**	**18**	**EM**	**19**	**19**	**19**

Table 6.2 presents a summary of the 6 hours ahead forecasted values that are reconstructed phase-space in dimensions ranging from 2 to 10 for τ = 1 and 3. The overall average of fitness values for all embedding dimensions are CC > 0.59, RMSE< 3.63 (l/s) and MAE < 1.61. Comparing to the results in chapter 3, the lower value of the fitness functions for the forecasting model is training the models with lower number of the data. As stated, previously, long-term training period will train the models to estimate the actual values accurately. Figure 6.8 shows the results of each model (NLA and PSR-NLA) in comparison to simulated weighted values.

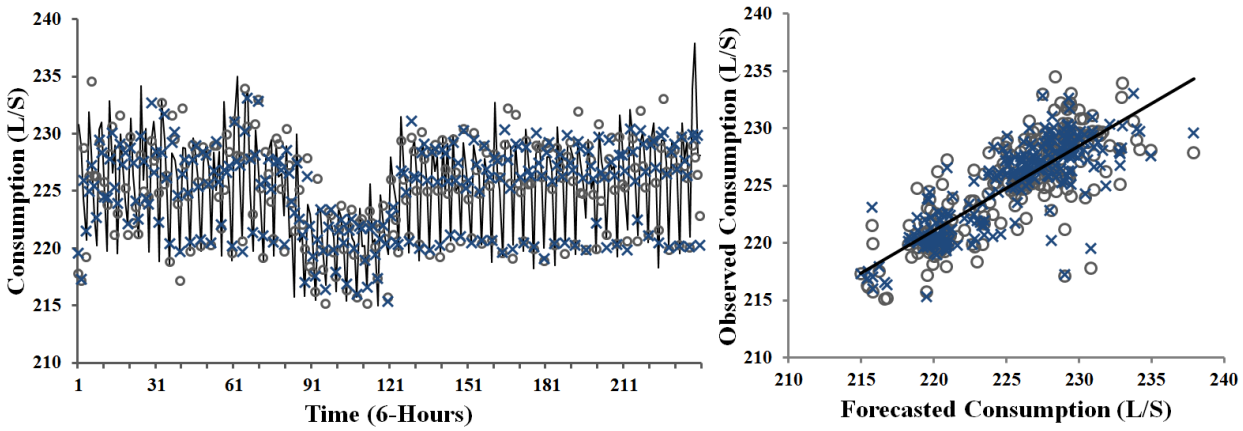


Figure 6.8 Forecasted values for water consumption by NLA and PSR-NLA in comparison with simulated weighted values.

The difference between the results of NLA and PSR-NLA were not considerable; the performance of PSR-NLA was better than NLA to forecast consumption values in lead time. As it is shown in Figure 6.8, PSR-NLA values are more concentrated on the actual values. Therefore, PSR-NLA is the potential forecasting model for processing this chapter. The aim of this section is employing an AI method to detect any probable anomalies. Phase space reconstruction is the only pre-processing method combined with the selected forecasting model.

6.5.4 Leakage Signal Analyzing[13]

The leakage events that are simulated on the same day and time of a week in the testing period are decomposed. Symlet 2 is used to decompose the signals in 5 levels. Symlet were the selected pre-processing method in improving the accuracy of all forecasting models. A higher number of levels are chosen to investigate further details on the changes within the signals. Figure 6.9 shows the signal in approximation value (A) and five details (5 Levels).

[13] Further details of wavelet decomposition are available in section 4.3.5.2.

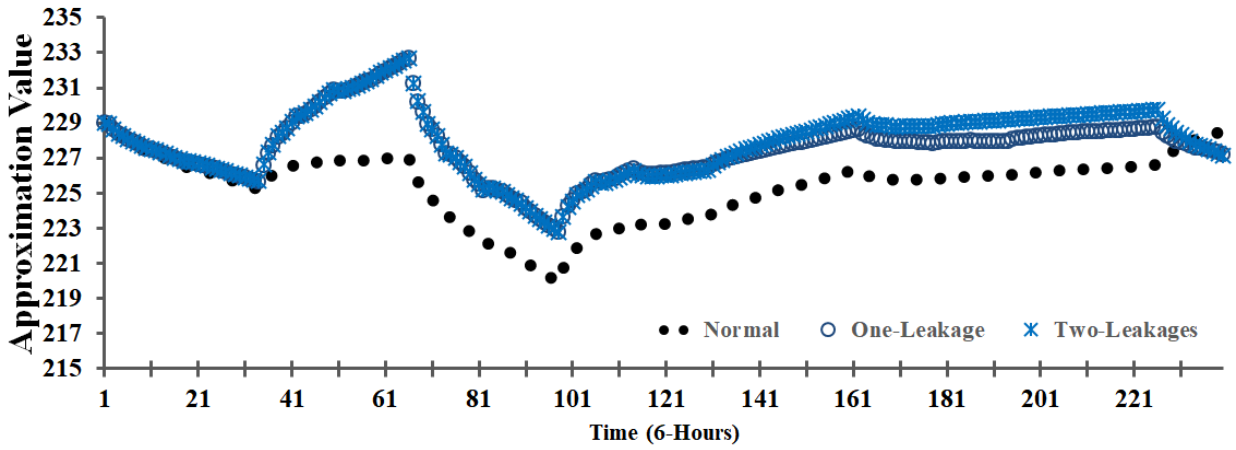


Figure 6.9 Decomposed values for approximation in 5 level with Symlet function for three scenarios.

To investigate the details within the time series of all scenarios, five levels of decomposition is employed. However, it was reported previously in this research that 3 levels of decomposition are sufficient (based on the number of data within the time series), a higher level of decomposition would give further details.

Figures 6.10-14 show the difference in the detailed decomposed time series for each scenario. All the graphs are for the testing period. The test period for the test case is the two last months. Scenario I is the application of WDS without any leakage in the target district, scenario II with one leakage and scenario III is the target district with two leakages. The locations of the leakages are highlighted in Figure 6.3.

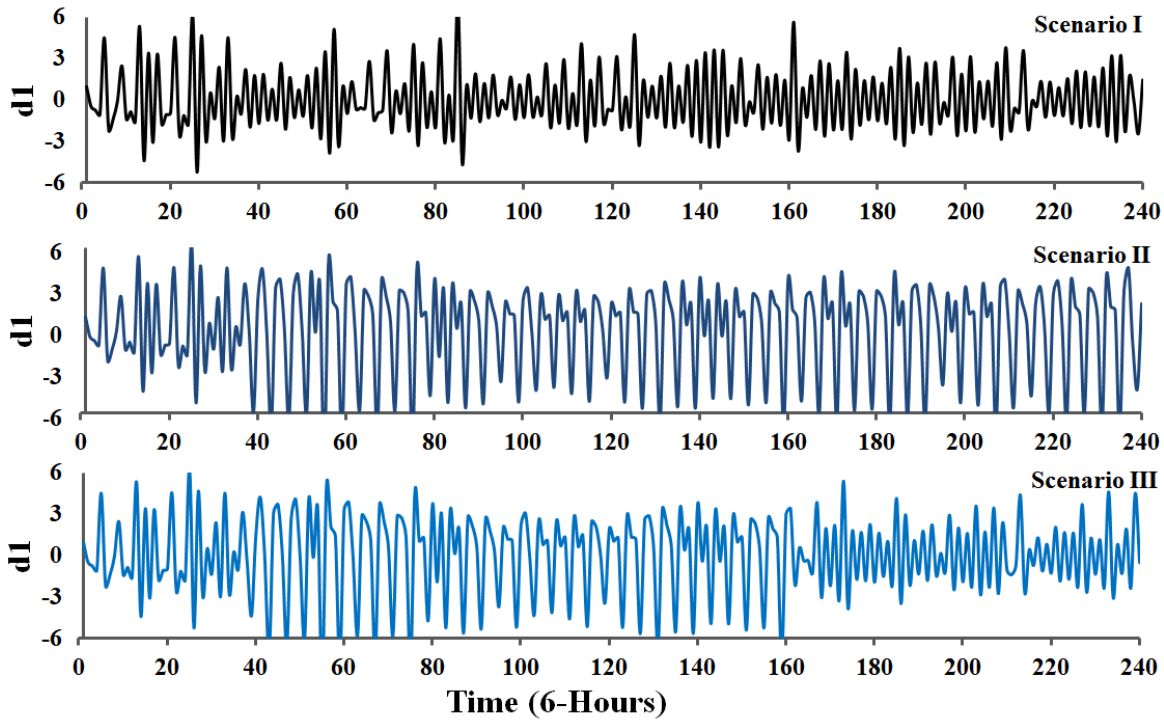


Figure 6.10 Decomposed values for detail (d1) in 1st level with Symlet function for three scenarios.

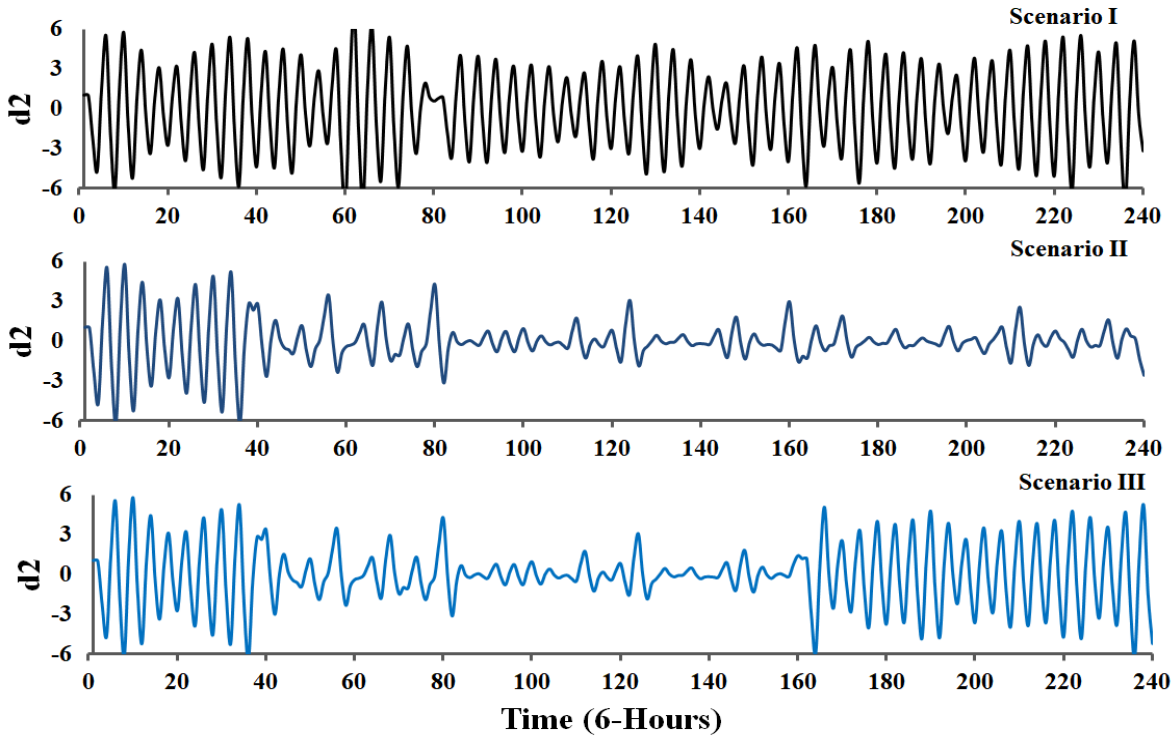


Figure 6.11 Decomposed values for detail (d2) in 2nd level with Symlet function for three scenarios.

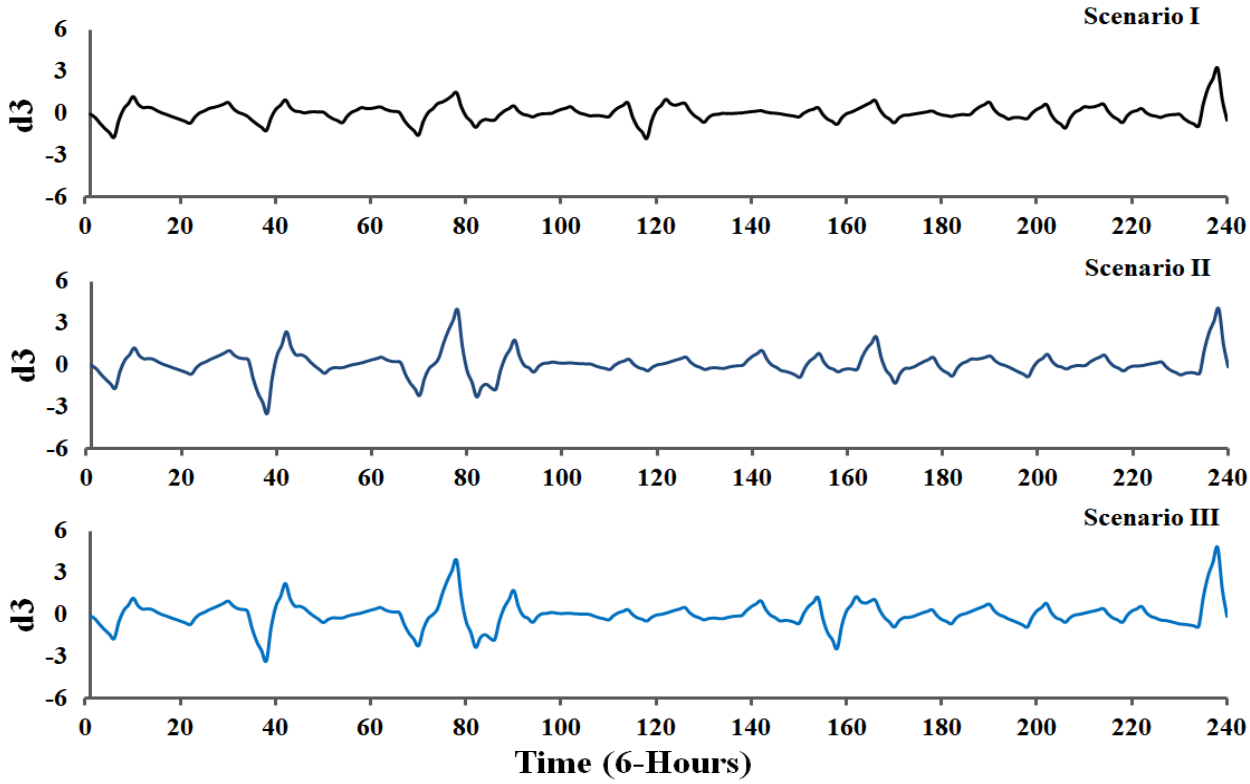


Figure 6.12 Decomposed values for detail (d3) in 3rd level with Symlet function for three scenarios.

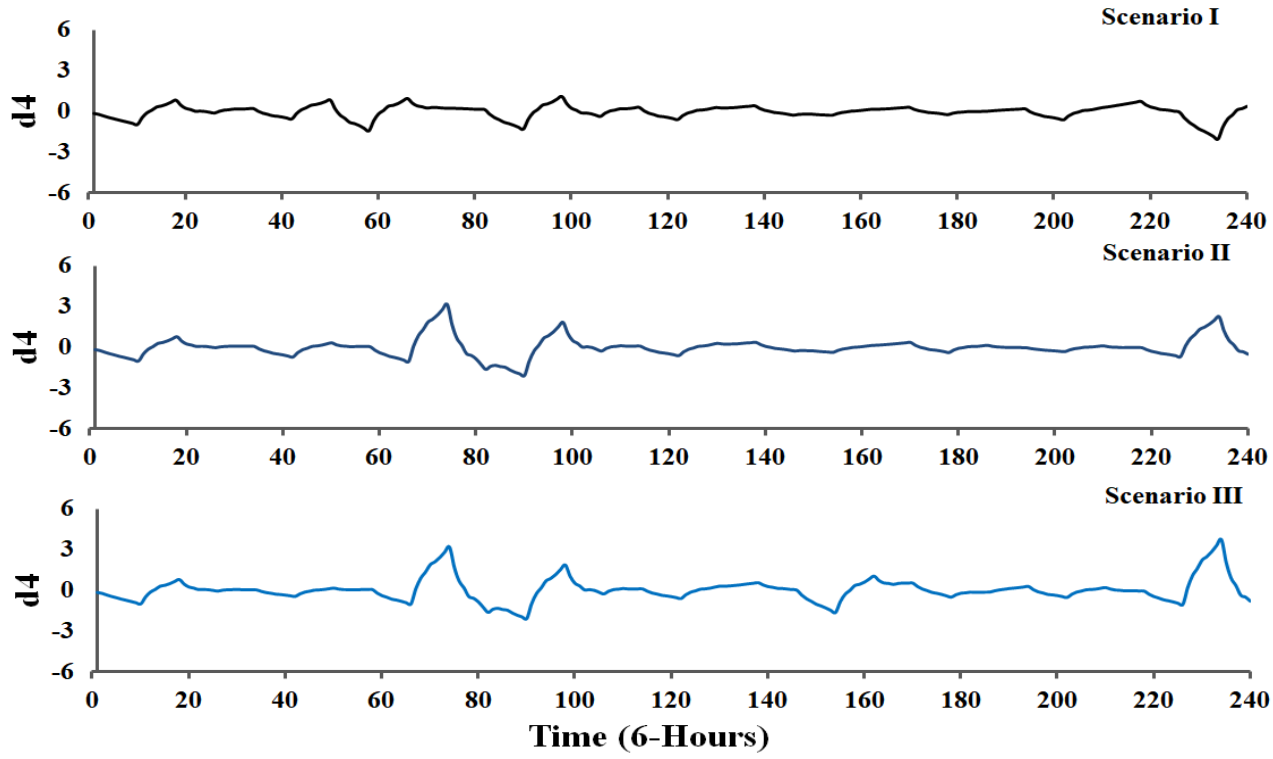


Figure 6.13 Decomposed values for detail (d4) in 4th level with Symlet function for three scenarios.

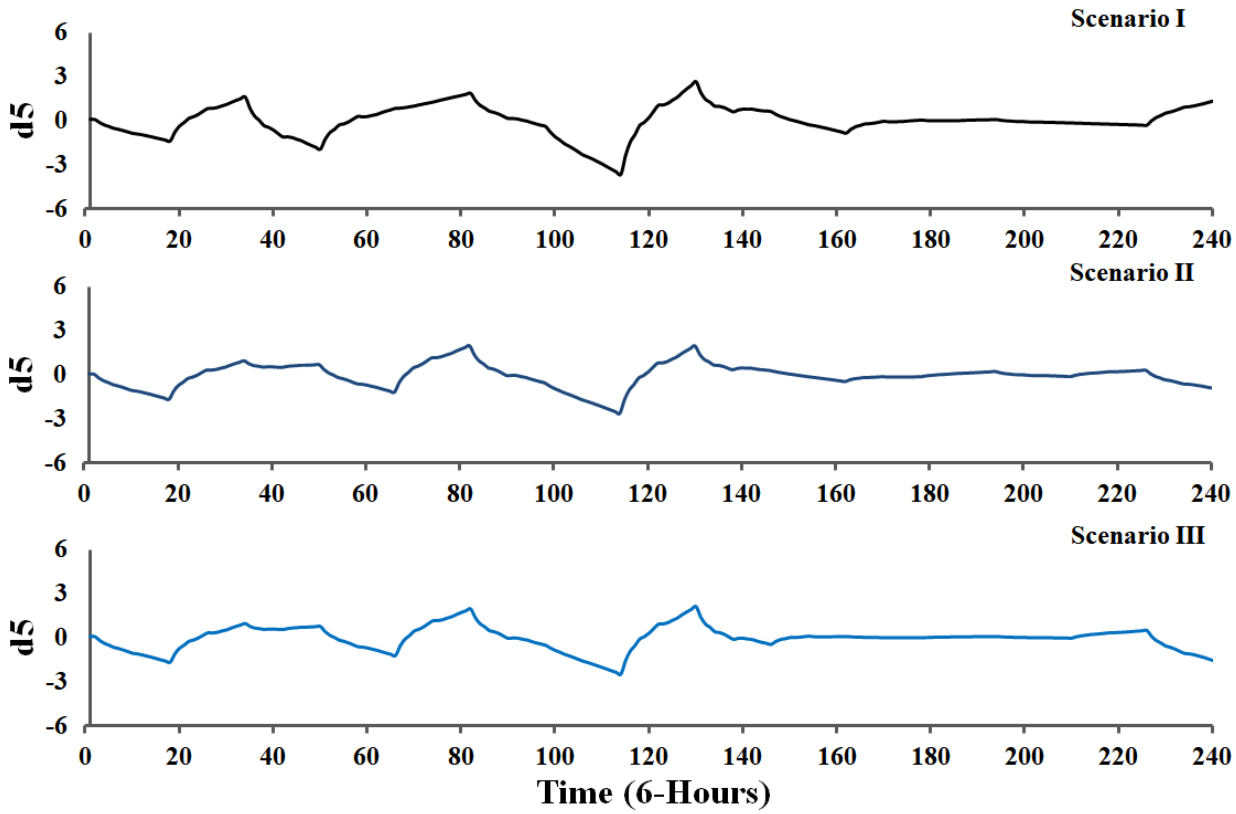


Figure 6.14 Decomposed values for detail (d5) in 5th level with Symlet function for three scenarios.

The AI tool combined with the SCADA system get updated by the new pulse of data. Therefore, the forecasting model forecasts the approximate values of the future (expected short-, mid-term values of consumption). The forecasted horizon is based on the previously recorded data. Meanwhile, if any changes happen within the system, the signal analyzing tool of SCADA (based on wavelet decomposition) detects the changes. Since the data has chaotic behavior, any small changes in initial condition will result in a significant difference after some time [103]. Figure 6.15-19 show the changes in signals of the forecasted values and the potential changes in each scenario. The results show the ability of the introduced technique in detecting the approximate time of the leakage.

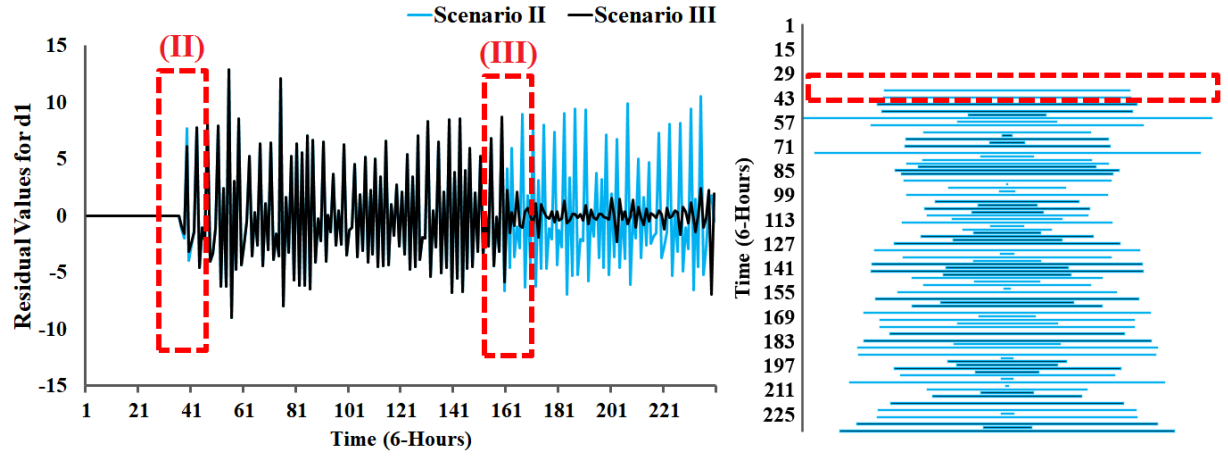


Figure 6.15 Residual of decomposed Values of forecasted consumption for d1, scenarios (II) and (III).

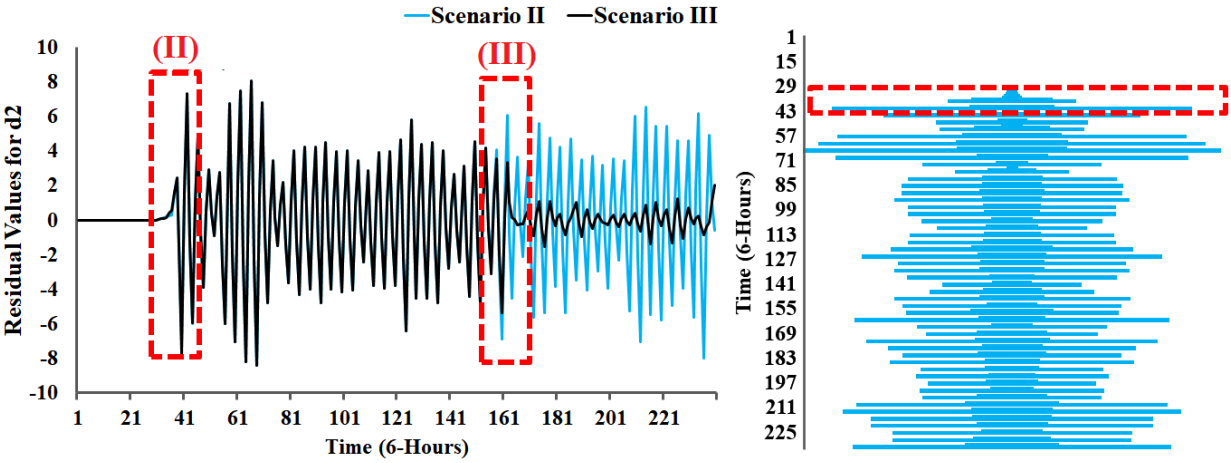


Figure 6.16 Residual of decomposed Values of forecasted consumption for d2, scenarios (II) and (III).

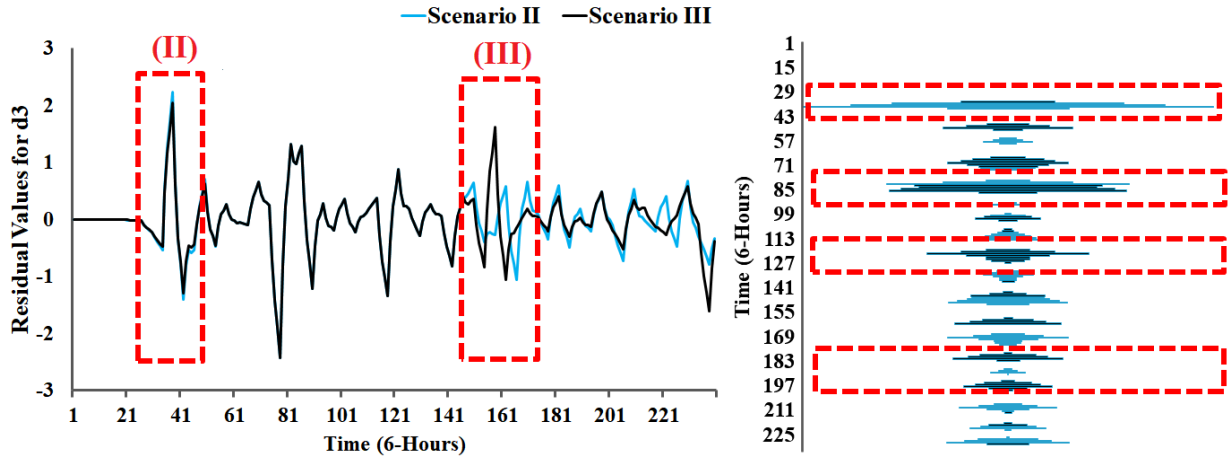


Figure 6.17 Residual of decomposed Values of forecasted consumption for d3, scenarios (II) and (III).

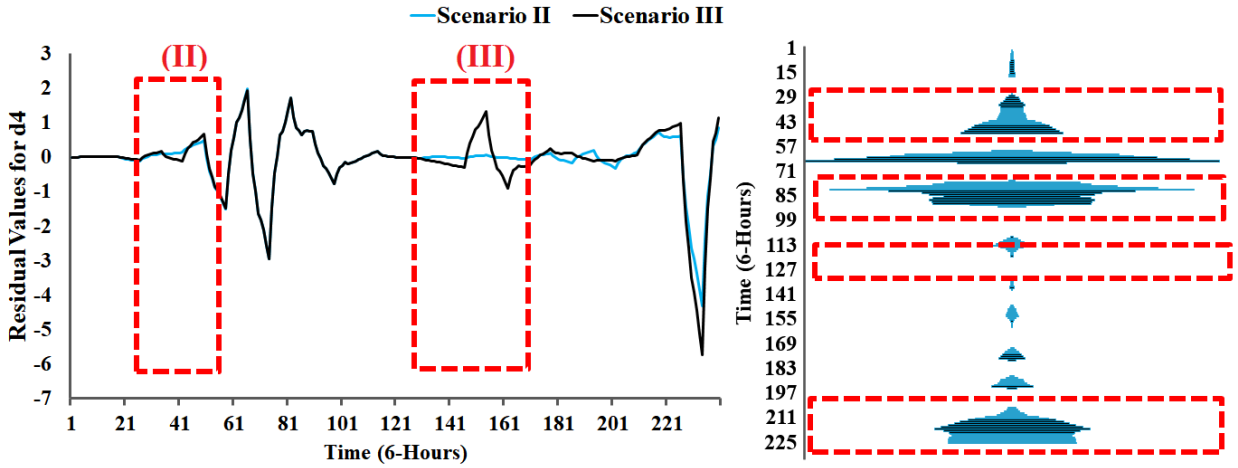


Figure 6.18 Residual of decomposed Values of forecasted consumption for d4, scenarios (II) and (III).

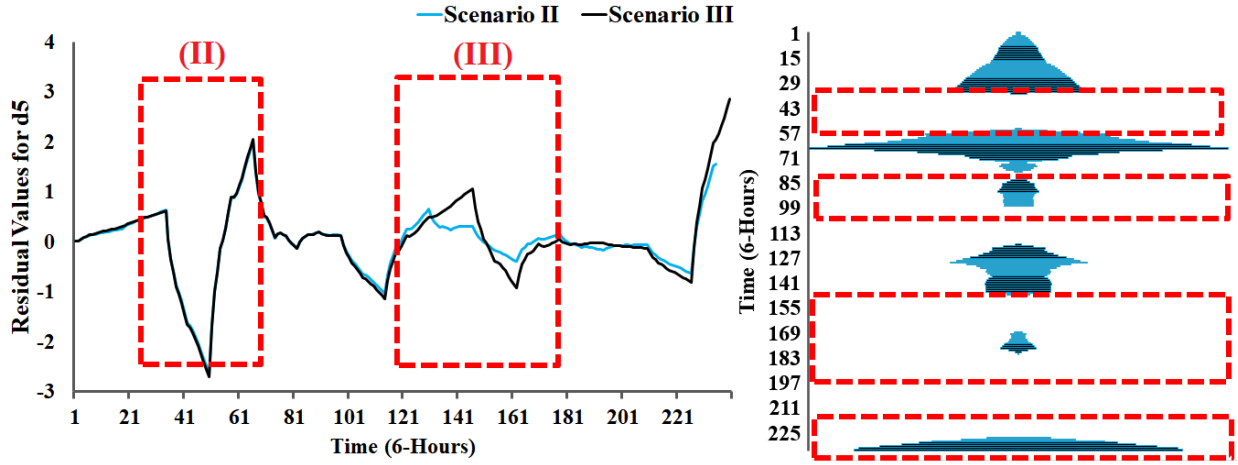


Figure 6.19 Residual of decomposed Values of forecasted consumption for d5, scenarios (II) and (III).

In the figures, the red boxes are the potential time range (6 hours each) for the probable anomalies within the decomposed signals. The results are based on the Symlet function in 5 levels. This research recommends employing different functions on different levels. The behavior of time series for different districts is not the same; therefore, to upgrade the leakage detection model, various functions and different levels would improve the performance of the detection method. Figure 6.20 shows all possible logical relations among the limited range of potential signal changes for different sets.

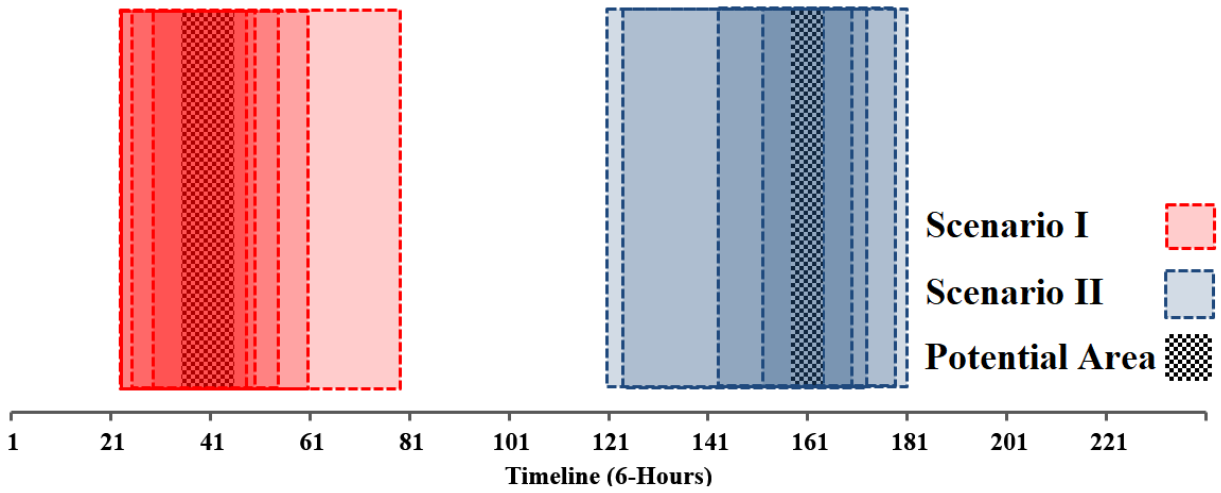


Figure 6.20 Logical relation for five levels of Symlet function for the residual values.

Figure 6.20 highlights the timeline with potential leakage. The results for the second scenario (blue color in Figure 6.20) are more accurate than the first scenario. To the best knowledge of the author, it is because of the length of training and testing period. From the first to the second scenario, there are more data employed to calibrate the models. In a chaotic time series, a higher number of data will improve the calibration of the models. This research estimated the approximate time for two simulated artificial scenarios. The estimation is 38-45 units and 158-163 units for first and second scenarios, respectively (each unit is 6 hours). The exact time for the leakage is on the 10^{th} day of each month in the test period at 12:00 pm. Based on the defined units, it is on the 41^{st} and 161^{st} units on the timeline. The results showed the developed AI in combination with SCADA devices gives the opportunity of detecting accurately the time that leakage appears within the target district.

6.6 Future Implementation

The presented methodology shows the ability to detect leakage in WDS districts from the aspect of the timeline. The results were simulated on a small target district within a large WDS for the test case. The test data were based on the reservoir that supplies the target district. The test case is

provided through different reservoirs and tanks that are presented in the section of the case study. To calibrate the presented models and techniques, online recording for the consumption values are necessary. The second step of Integrated WDS management is localizing the leakage. Having the approximate time for the event plus analyzing the signals pulsed from implemented devices in specific joints (location) within a WDS, provide enough information to detect the time and location of the leakage.

6.7 Summary

In this chapter, a supervised data-driven approach and signal tracking technique of hidden variables were employed for detection of anomalies in a WDS. The primary focus of this chapter was on the analysis of the consumption value of a reservoir in the test case. Many of the reported methods in the literature can recognize unusual patterns or anomalies in time series, the advantage of the presented technique is the detection of any changes in consumption pattern in a specific timeline. A target district was selected from the simulated hydraulic model of the City of Kelowna to evaluate the performance of the detection procedure. Two events were simulated within the district. In the first event, one artificial leakage, and in the second event, two leakages were defined. The events were defined at the same timeline of each month in the test period. NLA model employed to estimate the future expected values of consumption. Then, the value of consumption in the target district was analyzed by the wavelet decomposition approach. Symlet function was used to decompose the forecasted values and leakage events in 5 levels. The results showed that a combination of forecasting model and signal analysis technique recognized the pattern change within the system. Each level of decomposition gave the approximate range of timeline for the potential leakage. The combination of the outputs from analyzing each level detected the leakage events in the system accurately. t was suggested that by implementing recording devices within

the district, the presented technique not only identify the anomalies, but also localize them within the distribution network approximately.

Chapter 7: Conclusions and future work

7.1 Conclusions

To improve the practice in short-, mid- and long-term water consumption forecasting, besides leakage detection in a water distribution system, this study used an artificial intelligence approach in the forecasting and analytics of urban drinking water consumption. The outcomes of this research are listed below:

1. The results of the correlation exponent showed the existence of chaos in the water consumption dataset of Kelowna City, BC, Canada. Chapter two suggested that to investigate the availability of chaos in a dataset, an investigation for different temporal scales is needed to claim the existence of chaos. Besides, it was concluded that the dynamic nature of explanatory data provides valuable information about the selection of forecasting models. Average mutual information and correlation dimension methods are employed to calculate the lag time and existence of chaos for the test data, respectively.

2. The test case dataset was used to forecast future consumption values in various lead times in order to identify the model that had a better performance among all. Non-linear approximation (NLA), dynamic investigation, phase space reconstruction (PSR) for input variables were considered to forecast various periodicity, and lead time. Artificial neural network (ANN), gene expression programming (GEP), and multiple linear regression (MLR) models were calibrated based on the reconstructed phase space with average mutual information, and embedding dimensions varied from 1 to 20. Regarding the existence of chaos in the test case, the performance of NLA was the best among other conventional and soft-computing techniques. Moreover, the performance of NLA with PSR was reported to be better than the other models in forecasting far lead time.

3. The performance of four techniques (NLA, ANN, GEP, and MLR) was investigated in forecasting the short-term period of water consumption of Kelowna City (BC, Canada). All three techniques performed accurately without the combination of the pre-processing methods. WDT improved the accuracy of models with PSR and without PSR. However, the impact of wavelet on the models with PSR was not as considerable as without PSR. It was recommended to use the pre-processing techniques in modeling and forecasting issues, especially about the dataset that the peak points are very critical in the case. The inherent behavior of the dataset (deterministic or stochastic) can affect the performance of the pre-processing methods. Therefore, it was suggested that the behavior of datasets should be investigated before deciding to employ any pre-process methods.

4. By comparing the results of the models with/without pre-processing, based on their performance and the value of fitness functions, it was concluded that AMI gives a proper lag time for time series with chaotic behavior. The models that their embedding dimensions defined by the lag time, which was estimated with AMI, performed better than the other models with embedding dimension and ACF lag time.

5. The value of the power spectrum for the test case represented the presence of the scaling behavior within the test data. Fitness values for the models were reasonable; however, the models were not included as the models with the highest accuracy. The total consumption on a large time scale was precisely equal to the recorded values, which is an advantage for the disaggregation models in this research. Moreover, the disaggregation methods can be an adequate replacement for various non-linear techniques. This is because disaggregation techniques can estimate the target values on a large scale with the highest accuracy. It was reported that the non-linear deterministic approach was more accurate than the continuous random cascade method in simulating the peak values of water consumption.

6. A supervised data-driven approach and signal tracking technique of hidden variables were employed for the detection of anomalies at the test case. The presented method detected the artificial changes that were made in the pattern of consumption simulation. The results showed that the AI technique, which was the combination of forecasting and analyzing techniques, detected the approximate time of the leakage in the target district. Besides, it was suggested that by implementing information recording devices within the target district, the presented technique could detect the location of the leakage.

The consumption-based proposed method would result in an integrated dynamic management plan which can overcome possible unexpected failures in a WDS. Besides, it will allow the WDS operators to have failures information in real time. This information would lead to the automated management of the WDS as a smart management system. Pressure optimization, peak consumption supply, uncounted water management and optimizing pumps and tanks operation, are the expected advantages that the proposed technique offered.

7.2 Main Contributions

- Investigation of water consumption time series by the chaos approach to determine the explanatory variables of the system.
- Phase space reconstruction for the models based on the proper lag time and defining the combination of input variables based on the optimum embedding dimension.
- Application of PSR and Optimum embedding dimension in improving the performance of the models from the aspect of complexity and accuracy.
- Exploring the application of ACF and AMI in modeling problems and introducing their use for different modeling problems.

- Exploring the performance of conventional and soft-computing techniques combining with pre-processing methods to simulate the peak values accurately.
- Introduction to the application of an input variable selection technique in improving the complexities of the models by removing the non-active variables.
- Developing a new concept of a combination of the temporal scale transition with pre-processing methods to improve the models' accuracy.
- Defining a signal analyzing method in combination with AI modeling techniques to detect leakage time within a district.

7.3 Limitations

- The developed models do not have the same high accuracy in different case studies. This is because the time series of each test case has different explanatory variables. Therefore, the models should be calibrated based on the available explanatory variables of the test cases.
- Finding relevant data for input variable design is challenging. Besides, the relation of the variables to other variables and their effects on the target values are very important to calibrate the models. Therefore, having a variety of relevant data is not sufficient, but the study of their effects is necessary.
- The high accuracy of the calibrated models is the advantage of employing them in WDS management. But, complex calculations and the combination of different methods are time-consuming and costly.
- Implementing devices within a WDS to record accurate data is costly. Therefore, the cost of maintaining and implementing the tools versus savings from leakage and consumption control should be considered.

7.4 Future Work

New ideas can be developed by the findings of this research. The potential studies are as follow:

- All the developed models in this research were based on a univariate variable (the value of water consumption). It is suggested to employ active variables in the modeling of water consumption values. Temperature, humidity, holidays, population growth, and average income are all potential variables.
- This dissertation showed the ability of soft-computing techniques to detect leakage in WDS districts from the aspect of the timeline, without any device implemented within the target district. The simulation was done on a small district of the test case. Employing the introduced technique in different districts in a WDS can be an interesting topic for future research ideas.
- Detection of the location and time of the leakage at the same time needs implemented recording devices within the target district. The cost of implementation and maintenance are very considerable. Optimization methods locate the devices in only critical joints within a WDS. It will reduce the costs and the number of devices while providing useful information for better operational management. Further research can be done in investigating optimization methods to locate the devices.

Bibliography

[1] Billing, B., Jones, C., 2008. Forecasting Urban Water Demand. Second Edition, American Water Works Association.

[2] Ghalehkhondabi, I., Ardjmand, E., Young, W.A., Weckman, G.R., 2017. Water demand forecasting: review of soft computing methods. *Environmental Monitoring and Assessment* 189(7): 313, Doi: 10.1007/s10661-017-6030-3.

[3] Sastri, T., Valdes, J.B., 2008. Rainfall Intervention Analysis for On-Line Applications. *Journal of Water Resources Planning and Management* 115(4): 397–415, Doi: 10.1061/(asce)0733-9496(1989)115:4(397).

[4] Odan, F.K., Reis, L.F.R., 2012. Hybrid Water Demand Forecasting Model Associating Artificial Neural Network with Fourier Series. *Journal of Water Resources Planning and Management* 138(3): 245–56, Doi: 10.1061/(ASCE)WR.1943-5452.0000177.

[5] Iwanek, M., Kowalska, B., Hawryluk, E., Kondraciuk, K., 2016. Distance and time of water effluence on soil surface after failure of buried water pipe. Laboratory investigations and statistical analysis. *Eksploatacja i Niezawodnosc - Maintenance and Reliability* 18(2): 278–84, Doi: 10.17531/ein.2016.2.16.

[6] Ghiassi, M., Zimbra, D.K., Saidane, H., 2008. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. *Journal of Water Resources Planning and Management* 134(2): 138–46, Doi: 10.1061/(asce)0733-9496(2008)134:2(138).

[7] Jayawardena, A.W., Gurung, A.B., 2000. Noise reduction and prediction of hydrometeorological time series: Dynamical systems approach vs. stochastic approach. *Journal of Hydrology* 228(3–4): 242–64, Doi: 10.1016/S0022-1694(00)00142-6.

[8] Lisi, F., Villi, V., 2001. CHAOTIC FORECASTING OF DISCHARGE TIME SERIES: A

CASE STUDY. *Journal of the American Water Resources Association* 37(2): 271–9, Doi: 10.1111/j.1752-1688.2001.tb00967.x.

[9] Oshima, N., 2015. Information Integration Type Chaos Theory-Based Demand Forecasting for Predictive Control of Waterworks. *Water Purification Technologies* 164(2): 6–12.

[10] Jorgensen, B.S., Martin, J.F., Pearce, M., Willis, E., 2013. Some difficulties and inconsistencies when using habit strength and reasoned action variables in models of metered household water conservation. *Journal of Environmental Management* 115: 124–35, Doi: 10.1016/J.JENVMAN.2012.11.008.

[11] Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M., 2018. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environmental Modelling & Software* 102: 199–212, Doi: 10.1016/J.ENVSOFT.2017.11.022.

[12] Beal, C.D., Gurung, T.R., Stewart, R.A., 2016. Demand-side management for supply-side efficiency: Modeling tailored strategies for reducing peak residential water demand. *Sustainable Production and Consumption* 6: 1–11, Doi: 10.1016/J.SPC.2015.11.005.

[13] Yousefi, P., Naser, G., Mohammadi, H., 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications,.

[14] Cresswell, M., Naser, G., 2013. A water resources management strategy for small water districts — a case study of the South East Kelowna irrigation district. *Canadian Journal of Civil Engineering* 40(6): 499–507, Doi: 10.1139/cjce-2012-0398.

[15] Yousefi, P., Naser, G., Mohammadi, H., 2018. Surface Water Quality Model: Impacts of Influential Variables. *Journal of Water Resources Planning and Management* 144(5):

04018015, Doi: 10.1061/(ASCE)WR.1943-5452.0000900.

[16] Adamowski, J., Karapataki, C., 2010. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. *Journal of Hydrologic Engineering* 15(10): 729–43, Doi: 10.1061/(ASCE)HE.1943-5584.0000245.

[17] Lisi, F., Villi, V., 2001. Chaotic forecasting of discharge time series: A case study. *Journal of the American Water Resources Association* 37(2): 271–9, Doi: 10.1111/j.1752-1688.2001.tb00967.x.

[18] Kame'enui, A.E., 2003. Water demand forecasting in the Puget Sound Region: Short and long-term models: 1–97.

[19] Jain, A., Ormsbee, L.E., 2002. Short-term water demand forecast modeling techniques-CONVENTIONAL METHODS VERSUS AI. *Journal - American Water Works Association* 94(7): 64–72, Doi: 10.1002/j.1551-8833.2002.tb09507.x.

[20] Ghiassi, M., Zimbra, D.K., Saidane, H., 2008. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. *Journal of Water Resources Planning and Management* 134(2): 138–46, Doi: 10.1061/(ASCE)0733-9496(2008)134:2(138).

[21] Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology* 387(1–2): 141–50, Doi: 10.1016/J.JHYDROL.2010.04.005.

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

[23] Miaou, S.-P., 1990. A stepwise time series regression procedure for water demand model identification. *Water Resources Research* 26(9): 1887–97, Doi:

10.1029/WR026i009p01887.

[24] Jain, A., Kumar Varshney, A., Chandra Joshi, U., 2001. Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks. *Water Resources Management* 15(5): 299–321, Doi: 10.1023/A:1014415503476.

[25] Jain, A., Varshney, A.K., Joshi, U.C., 2001. Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks. *Water Resources Management* 15(5): 299–321, Doi: 10.1023/A:1014415503476.

[26] Bougadis, J., Adamowski, K., Diduch, R., 2005. Short-term municipal water demand forecasting. *Hydrological Processes* 19(1): 137–48, Doi: 10.1002/hyp.5763.

[27] Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research* 48(1), Doi: 10.1029/2010WR009945.

[28] Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J., 2000. Forecasting daily urban water demand: A case study of Melbourne. *Journal of Hydrology* 236(3–4): 153–64, Doi: 10.1016/S0022-1694(00)00287-0.

[29] Mukhopadhyay, A., Akber, A., Al-Awadi, E., 2001. Analysis of freshwater consumption patterns in the private residences of Kuwait. *Urban Water* 3(1–2): 53–62, Doi: 10.1016/S1462-0758(01)00016-4.

[30] Farah, E., Shahrour, I., 2017. Leakage Detection Using Smart Water System: Combination of Water Balance and Automated Minimum Night Flow. *Water Resources Management* 31(15): 4821–33, Doi: 10.1007/s11269-017-1780-9.

[31] Brekke, L., Larsen, M.D., Ausburn, M., Takaichi, L., 2002. Suburban Water Demand Modeling Using Stepwise Regression. *Journal - American Water Works Association* 94(10): 65–75, Doi: 10.1002/j.1551-8833.2002.tb09558.x.

[32] Polebitski, A.S., Palmer, R.N., 2009. Seasonal Residential Water Demand Forecasting for Census Tracts. *Journal of Water Resources Planning and Management* 136(1): 27–36, Doi: 10.1061/(asce)wr.1943-5452.0000003.

[33] Lee, S.-J., Wentz, E.A., Gober, P., 2010. Space–time forecasting using soft geostatistics: a case study in forecasting municipal water demand for Phoenix, Arizona. *Stochastic Environmental Research and Risk Assessment* 24(2): 283–95, Doi: 10.1007/s00477-009-0317-z.

[34] Ghiassi, M., Zimbra, D.K., Saidane, H., 2008. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. *Journal of Water Resources Planning and Management* 134(2): 138–46, Doi: 10.1061/(ASCE)0733-9496(2008)134:2(138).

[35] Jain, A., Ormsbee, L.E., 2002. Short-term water demand forecast modeling techniques-CONVENTIONAL METHODS VERSUS AI. *Journal - American Water Works Association* 94(7): 64–72, Doi: 10.1002/j.1551-8833.2002.tb09507.x.

[36] Cutore, P., Campisano, A., Kapelan, Z., Modica, C., Savic, D., 2008. Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. *Urban Water Journal* 5(2): 125–32, Doi: 10.1080/15730620701754434.

[37] Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. *Journal of Water Resources Planning and Management* 134(2): 119–28, Doi: 10.1061/(ASCE)0733-9496(2008)134:2(119).

[38] Firat, M., Yurdusev, M.A., Turan, M.E., 2009. Evaluation of Artificial Neural Network

Techniques for Municipal Water Consumption Modeling. *Water Resources Management* 23(4): 617–32, Doi: 10.1007/s11269-008-9291-3.

[39] Xu, Y., Zhang, J., Long, Z., Chen, Y., Xu, Y., Zhang, J., et al., 2018. A Novel Dual-Scale Deep Belief Network Method for Daily Urban Water Demand Forecasting. *Energies* 11(5): 1068, Doi: 10.3390/en11051068.

[40] Msiza, I.S., Nelwamondo, F. V., Marwala, T., 2007. Artificial neural networks and support vector machines for water demand time series forecasting. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, IEEE p. 638–43.

[41] Msiza, I.S., Nelwamondo, F. V., Marwala, T., 2008. Water demand prediction using artificial neural networks and support vector regression. *Journal of Computers* 3(11): 1–8, Doi: 10.4304/jcp.3.11.1-8.

[42] Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology* 387(1–2): 141–50, Doi: 10.1016/J.JHYDROL.2010.04.005.

[43] Shabani, S., Yousefi, P., Naser, G., 2017. Support Vector Machines in Urban Water Demand Forecasting Using Phase Space Reconstruction. *Procedia Engineering* 186: 537–43, Doi: 10.1016/J.PROENG.2017.03.267.

[44] Yousefi, P., Shabani, S., Mohammadi, H., Naser, G., 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. Procedia Engineering, vol. 186. p. 544–50.

[45] Shabani, S., Candelieri, A., Archetti, F., 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): 142, Doi: 10.3390/w10020142.

[46] Ambrosio, J.K., Brentan, B.M., Herrera, M., Luvizotto, E., Ribeiro, L., Izquierdo, J., 2019. Committee Machines for Hourly Water Demand Forecasting in Water Supply Systems. *Mathematical Problems in Engineering* 2019: 1–11, Doi: 10.1155/2019/9765468.

[47] Yousefi, P., Naser, G., Mohammadi, H., 2018. Application of Wavelet Decomposition and Phase Space Reconstruction in Urban Water Consumption Forecasting: Chaotic Approach (Case Study). Wavelet Theory and Its Applications, InTech.

[48] Yousefi, P., Naser, G., Mohammadi, H., 2018. Hybrid Wavelet and Local Approximation Method for Urban Water Demand Forecasting – Chaotic Approach. *WDSA / CCWI Joint Conference Proceedings* 1.

[49] Azadeh, A., Neshat, N., Hamidipour, H., 2011. Hybrid Fuzzy Regression–Artificial Neural Network for Improvement of Short-Term Water Consumption Estimation and Forecasting in Uncertain and Complex Environments: Case of a Large Metropolitan City. *Journal of Water Resources Planning and Management* 138(1): 71–5, Doi: 10.1061/(asce)wr.1943-5452.0000152.

[50] Papageorgiou, E.I., Poczeta, K., Laspidou, C., 2015. Application of Fuzzy Cognitive Maps to water demand prediction. 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE p. 1–8.

[51] Ahmadi, S., Alizadeh, S., Forouzideh, N., Yeh, C.H., Martin, R., Papageorgiou, E., 2014. ICLA imperialist competitive learning algorithm for fuzzy cognitive map: Application to water demand forecasting. IEEE International Conference on Fuzzy Systems, IEEE p. 1041–8.

[52] Navarrete-López, C., Herrera, M., Brentan, B., Luvizotto, E., Izquierdo, J., Navarrete-López, C., et al., 2019. Enhanced Water Demand Analysis via Symbolic Approximation

within an Epidemiology-Based Forecasting Framework. *Water* 11(2): 246, Doi: 10.3390/w11020246.

[53] Yousefi, P., Naser, G., Mohammadi, H., 2018. Estimating High Resolution Temporal Scale of Water Demand Time Series – Disaggregation Approach (Case Study). EPiC Series in Engineering, vol. 3. p. 2408–2398.

[54] Kozłowski, E., Kowalska, B., Kowalski, D., Mazurkiewicz, D., 2018. Water demand forecasting by trend and harmonic analysis. *Archives of Civil and Mechanical Engineering* 18(1): 140–8, Doi: 10.1016/J.ACME.2017.05.006.

[55] Campisi-Pinto, S., Adamowski, J., Oron, G., 2012. Forecasting Urban Water Demand Via Wavelet-Denoising and Neural Network Models. Case Study: City of Syracuse, Italy. *Water Resources Management* 26(12): 3539–58, Doi: 10.1007/s11269-012-0089-y.

[56] Hutton, C.J., Kapelan, Z., 2015. A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. *Environmental Modelling & Software* 66: 87–97, Doi: 10.1016/J.ENVSOFT.2014.12.021.

[57] Horielova, K.A., Zadachyn, V.M., 2016. Planning of City Water Supply System Modernization. *Ivan Kozhedub Kharkiv National Air Force* 4: 143–8.

[58] Tiwari, M., Adamowski, J., Adamowski, K., 2016. Water demand forecasting using extreme learning machines. *Journal of Water and Land Development* 28(1): 37–52, Doi: 10.1515/jwld-2016-0004.

[59] Altunkaynak, A., Nigussie, T.A., 2017. Monthly Water Consumption Prediction Using Season Algorithm and Wavelet Transform–Based Models. *Journal of Water Resources Planning and Management* 143(6): 04017011, Doi: 10.1061/(ASCE)WR.1943-5452.0000761.

[60] Pacchin, E., Alvisi, S., Franchini, M., 2017. A Short-Term Water Demand Forecasting Model Using a Moving Window on Previously Observed Data. *Water* 9(3): 172, Doi: 10.3390/w9030172.

[61] Brentan, B.M., Luvizotto Jr., E., Herrera, M., Izquierdo, J., Pérez-García, R., 2017. Hybrid regression model for near real-time urban water demand forecasting. *Journal of Computational and Applied Mathematics* 309: 532–41, Doi: 10.1016/J.CAM.2016.02.009.

[62] Mouatadid, S., Adamowski, J., 2017. Using extreme learning machines for short-term urban water demand forecasting. *Urban Water Journal* 14(6): 630–8, Doi: 10.1080/1573062X.2016.1236133.

[63] Candelieri, A., 2017. Clustering and support vector regression for water demand forecasting and anomaly detection. *Water (Switzerland)* 9(3): 224, Doi: 10.3390/w9030224.

[64] Xu, Y., Zhang, J., Long, Z., Lv, M., 2018. Daily Urban Water Demand Forecasting Based on Chaotic Theory and Continuous Deep Belief Neural Network. *Neural Processing Letters*, Doi: 10.1007/s11063-018-9914-5.

[65] Duerr, I., Merrill, H.R., Wang, C., Bai, R., Boyer, M., Dukes, M.D., et al., 2018. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. *Environmental Modelling & Software* 102: 29–38, Doi: 10.1016/J.ENVSOFT.2018.01.002.

[66] González Perea, R., Camacho Poyato, E., Montesinos, P., Rodríguez Díaz, J.A., 2019. Optimisation of water demand forecasting by artificial intelligence with short data sets. *Biosystems Engineering* 177: 59–66, Doi: 10.1016/J.BIOSYSTEMSENG.2018.03.011.

[67] Ghalehkhondabi, I., Ardjmand, E., Young, W.A., Weckman, G.R., 2017. Water demand forecasting: review of soft computing methods. *Environmental Monitoring and Assessment*

189(7): 313, Doi: 10.1007/s10661-017-6030-3.

[68] Gutzler, D.S., Nims, J.S., 2006. Interannual Variability of Water Demand and Summer Climate in Albuquerque, New Mexico. *Journal of Applied Meteorology* 44(12): 1777–87, Doi: 10.1175/jam2298.1.

[69] Donkor, E.A., Mazzuchi, T.A., Soyer, R., Alan Roberson, J., 2012. Urban Water Demand Forecasting: Review of Methods and Models. *Journal of Water Resources Planning and Management* 140(2): 146–59, Doi: 10.1061/(asce)wr.1943-5452.0000314.

[70] Pacchin, E., Alvisi, S., Franchini, M., Pacchin, E., Alvisi, S., Franchini, M., 2017. A Short-Term Water Demand Forecasting Model Using a Moving Window on Previously Observed Data. *Water* 9(3): 172, Doi: 10.3390/w9030172.

[71] Casdagli, M., 2018. Chaos and Deterministic Versus Stochastic Non-Linear Modelling. *Journal of the Royal Statistical Society: Series B (Methodological)* 54(2): 303–28, Doi: 10.1111/j.2517-6161.1992.tb01884.x.

[72] Lorenz, E.N., 2004. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences* 26(4): 636–46, Doi: 10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2.

[73] Sivakumar, B., Jayawardena, A.W., Li, W.K., 2007. Hydrologic complexity and classification: a simple data reconstruction approach. *Hydrological Processes* 21(20): 2713–28, Doi: 10.1002/hyp.6362.

[74] Ng, W.W., Panu, U.S., Lennox, W.C., 2007. Chaos based Analytical techniques for daily extreme hydrological observations. *Journal of Hydrology* 342(1–2): 17–41, Doi: 10.1016/J.JHYDROL.2007.04.023.

[75] Regonda, S.K., Sivakumar, B., Jain, A., 2004. Temporal scaling in river flow: can it be

chaotic? *Hydrological Sciences Journal* 49(3): 373–85.

[76] Salas, J.D., Kim, H.S., Eykholt, R., Burlando, P., Green, T.R., 2010. Aggregation and sampling in deterministic chaos: implications for chaos identification in hydrological processes. *Nonlinear Processes in Geophysics* 12(4): 557–67, Doi: 10.5194/npg-12-557-2005.

[77] Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2002. Noise reduction in chaotic hydrologic time series: Facts and doubts. *Journal of Hydrology* 256(3–4): 147–65, Doi: 10.1016/S0022-1694(01)00534-0.

[78] Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2002. Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology* 255(1–4): 123–33, Doi: 10.1016/S0022-1694(01)00513-3.

[79] Sivakumar, B., Wallender, W.W., 2005. Predictability of river flow and suspended sediment transport in the Mississippi River basin: a non-linear deterministic approach. *Earth Surface Processes and Landforms* 30(6): 665–77, Doi: 10.1002/esp.1167.

[80] Zounemat-Kermani, M., 2016. Investigating Chaos and Nonlinear Forecasting in Short Term and Mid-term River Discharge. *Water Resources Management* 30(5): 1851–65, Doi: 10.1007/s11269-016-1258-1.

[81] Ghorbani, M.A., Khatibi, R., Danandeh Mehr, A., Asadi, H., 2018. Chaos-based multigene genetic programming: A new hybrid strategy for river flow forecasting. *Journal of Hydrology* 562(10): 455–67, Doi: 10.1016/j.jhydrol.2018.04.054.

[82] Sivakumar, B., 2002. A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers. *Journal of Hydrology* 258(1–4): 149–62, Doi: 10.1016/S0022-1694(01)00573-X.

[83] Sivakumar, B., Chen, J., 2007. Suspended sediment load transport in the Mississippi River basin at St. Louis: temporal scaling and nonlinear determinism. *Earth Surface Processes and Landforms* 32(2): 269–80, Doi: 10.1002/esp.1392.

[84] SIVAKUMAR, B., JAYAWARDENA, A.W., 2010. An investigation of the presence of low-dimensional chaotic behaviour in the sediment transport phenomenon. *Hydrological Sciences Journal* 47(3): 405–16, Doi: 10.1080/02626660209492943.

[85] Ghorbani, M., Khatibi, R., Asadi, H., Yousefi, P., 2012. Inter-Comparison of an Evolutionary Programming Model of Suspended Sediment Time-Series with Other Local Models. Genetic Programming - New Approaches and Successful Applications, InTech.

[86] Petkov, B.H., Vitale, V., Mazzola, M., Lanconelli, C., Lupi, A., 2015. Chaotic behaviour of the short-term variations in ozone column observed in Arctic. *Communications in Nonlinear Science and Numerical Simulation* 26(1–3): 238–49, Doi: 10.1016/J.CNSNS.2015.02.020.

[87] Khatibi, R., Ghorbani, M.A., Aalami, M.T., Kocak, K., Makarynskyy, O., Makarynska, D., et al., 2011. Dynamics of hourly sea level at Hillarys Boat Harbour, Western Australia: a chaos theory perspective. *Ocean Dynamics* 61(11): 1797–807, Doi: 10.1007/s10236-011-0466-8.

[88] Rodriguez-Iturbe, I., Febres De Power, B., Sharifi, M.B., Georgakakos, K.P., 1989. Chaos in rainfall. *Water Resources Research* 25(7): 1667–75, Doi: 10.1029/WR025i007p01667.

[89] Jayawardena, A.W., Lai, F., 1994. Analysis and prediction of chaos in rainfall and stream flow time series. *Journal of Hydrology* 153(1–4): 23–52, Doi: 10.1016/0022-1694(94)90185-6.

[90] Sivakumar, B., Berndtsson, R., Olsson, J., Jinno, K., Kawamura, A., 2010. Dynamics of monthly rainfall-runoff process at the Gota basin: A search for chaos. *Hydrology and Earth*

System Sciences 4(3): 407–17, Doi: 10.5194/hess-4-407-2000.

[91] Maskey, M.L., Puente, C.E., Sivakumar, B., 2019. Temporal downscaling rainfall and streamflow records through a deterministic fractal geometric approach. *Journal of Hydrology* 568: 447–61, Doi: 10.1016/j.jhydrol.2018.09.014.

[92] Wang, J., Shi, Q., 2013. Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet Analysis-Support Vector Machine theory. *Transportation Research Part C: Emerging Technologies* 27: 219–32, Doi: 10.1016/J.TRC.2012.08.004.

[93] Ravi, V., Pradeepkumar, D., Deb, K., 2017. Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. *Swarm and Evolutionary Computation* 36: 136–49, Doi: 10.1016/J.SWEVO.2017.05.003.

[94] Abdechiri, M., Faez, K., Amindavar, H., Bilotta, E., 2017. The chaotic dynamics of high-dimensional systems. *Nonlinear Dynamics* 87(4): 2597–610, Doi: 10.1007/s11071-016-3213-3.

[95] Li, M.W., Geng, J., Han, D.F., Zheng, T.J., 2016. Ship motion prediction using dynamic seasonal RvSVR with phase space reconstruction and the chaos adaptive efficient FOA. *Neurocomputing* 174: 661–80, Doi: 10.1016/j.neucom.2015.09.089.

[96] Sivakumar, B., 2000. Chaos theory in hydrology: important issues and interpretations. *Journal of Hydrology* 227(1–4): 1–20, Doi: 10.1016/S0022-1694(99)00186-9.

[97] Takens, F., 1981. Detecting strange attractors in turbulence. Springer, Berlin, Heidelberg p. 366–81.

[98] SIVAKUMAR, B., BERNDTSSON, R., OLSSON, J., JINNO, K., 2001. Evidence of chaos in the rainfall-runoff process. *Hydrological Sciences Journal* 46(1): 131–45, Doi: 10.1080/02626660109492805.

[99] Khatibi, R., Sivakumar, B., Ghorbani, M.A., Kisi, O., Koçak, K., Farsadi Zadeh, D., 2012. Investigating chaos in river stage and discharge time series. *Journal of Hydrology* 414–415: 108–17, Doi: 10.1016/J.JHYDROL.2011.10.026.

[100] Meng, Q., Peng, Y., 2007. A new local linear prediction model for chaotic time series. *Physics Letters, Section A: General, Atomic and Solid State Physics* 370(5–6): 465–70, Doi: 10.1016/j.physleta.2007.06.010.

[101] Holzfuss, J., Mayer-Kress, G., 2011. An Approach to Error-Estimation in the Application of Dimension Algorithms. Springer, Berlin, Heidelberg p. 114–22.

[102] Hegger, R., Kantz, H., Schreiber, T., 1999. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos* 9(2): 413–35, Doi: 10.1063/1.166424.

[103] Lorenz, E.N., Lorenz, E.N., 1969. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences* 26(4): 636–46, Doi: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

[104] Zounemat-Kermani, M., Kisi, O., 2015. Time series analysis on marine wind-wave characteristics using chaos theory. *Ocean Engineering* 100: 46–53, Doi: 10.1016/J.OCEANENG.2015.03.013.

[105] Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena* 9(1–2): 189–208, Doi: 10.1016/0167-2789(83)90298-1.

[106] Islam, M.., Sivakumar, B., 2002. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. *Advances in Water Resources* 25(2): 179–90, Doi: 10.1016/S0309-1708(01)00053-7.

[107] Tongal, H., Berndtsson, R., 2017. Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. *Stochastic Environmental*

Research and Risk Assessment 31(3): 661–82, Doi: 10.1007/s00477-016-1236-4.

[108] Kennel, M.B., Brown, R., Abarbanel, H.D.I., 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A* 45(6): 3403–11, Doi: 10.1103/PhysRevA.45.3403.

[109] Strategic Value Solution, 2017. Kelowna Integrated Water Suply Plan. Kelowna.

[110] City of Kelowna, 2017. City of Kelowna Annual Water and Filtration Exclusion Report (December 2016): 85.

[111] Ruelle, D., 1990. The Claude Bernard Lecture, 1989. Deterministic Chaos: The Science and the Fiction. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 427(1873): 241–8, Doi: 10.1098/rspa.1990.0010.

[112] Jain, A., Kumar Varshney, A., Chandra Joshi, U., 2001. Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks. *Water Resources Management* 15(5): 299–321, Doi: 10.1023/A:1014415503476.

[113] Yousefi, P., Shabani, S., Mohammadi, H., Naser, G., 2017. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. *Procedia Engineering* 186: 544–50, Doi: 10.1016/J.PROENG.2017.03.268.

[114] Yousefi, P., Naser, G., Mohammadi, H., 2018. Hybrid Wavelet and Local Approximation Method for Urban Water Demand Forecasting – Chaotic Approach. *WDSA / CCWI Joint Conference Proceedings* 1.

[115] Adebiyi, A.A., Adewumi, A.O., Ayo, C.K., 2014. Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. *Journal of Applied Mathematics* 2014: 1–7, Doi: 10.1155/2014/614342.

[116] Young, C.-C., Liu, W.-C., Hsieh, W.-L., 2015. Predicting the Water Level Fluctuation in

an Alpine Lake Using Physically Based, Artificial Neural Network, and Time Series Forecasting Models. *Mathematical Problems in Engineering* 2015: 1–11, Doi: 10.1155/2015/708204.

[117] Adamowski, J., Karapataki, C., 2010. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. *Journal of Hydrologic Engineering* 15(10): 729–43, Doi: 10.1061/(ASCE)HE.1943-5584.0000245.

[118] Shabani, S., Yousefi, P., Naser, G., 2017. Support Vector Machines in Urban Water Demand Forecasting Using Phase Space Reconstruction. Procedia Engineering, vol. 186. Elsevier p. 537–43.

[119] Yousefi, P., Curtice, G., Naser, G.R., Mohammadi, H., 2019. Nonlinear Dynamic Modeling of Urban Water Consumption - A Chaotic Approach. *Water* In Press.

[120] Kalra, R., Deo, M.C., 2007. Genetic programming for retrieving missing information in wave records along the west coast of India. *Applied Ocean Research* 29(3): 99–111, Doi: 10.1016/J.APOR.2007.11.002.

[121] Ustoorikar, K., Deo, M.C., 2008. Filling up gaps in wave data with genetic programming. *Marine Structures* 21(2–3): 177–95, Doi: 10.1016/J.MARSTRUC.2007.12.001.

[122] Gaur, S., Deo, M.C., 2008. Real-time wave forecasting using genetic programming. *Ocean Engineering* 35(11–12): 1166–72, Doi: 10.1016/J.OCEANENG.2008.04.007.

[123] Aytek, A., Kişi, Ö., 2008. A genetic programming approach to suspended sediment modelling. *Journal of Hydrology* 351(3–4): 288–98, Doi: 10.1016/j.jhydrol.2007.12.005.

[124] Ghorbani, M.A., Kisi, O., Aalinezhad, M., 2010. A probe into the chaotic nature of daily streamflow time series by correlation dimension and largest Lyapunov methods. *Applied*

Mathematical Modelling 34(12): 4050–7, Doi: 10.1016/j.apm.2010.03.036.

[125] Ferreira, C., 2002. Gene Expression Programming in Problem Solving. Soft Computing and Industry, London: Springer London p. 635–53.

[126] Ferreira, C., 2003. Function Finding and the Creation of Numerical Constants in Gene Expression Programming. Advances in Soft Computing, London: Springer London p. 257–65.

[127] Nasseri, M., Moeini, A., Tabesh, M., 2011. Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming. *Expert Systems with Applications* 38(6): 7387–95, Doi: 10.1016/j.eswa.2010.12.087.

[128] Shabani, S., Candelieri, A., Archetti, F., Naser, G., 2018. Gene expression programming coupled with unsupervised learning: A two-stage learning process in multi-scale, short-termwater demand forecasts. *Water (Switzerland)* 10(2): 142, Doi: 10.3390/w10020142.

[129] Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. *Journal of Water Resources Planning and Management* 134(2): 119–28, Doi: 10.1061/(ASCE)0733-9496(2008)134:2(119).

[130] Jain, A., Ormsbee, L.E., 2002. Short-term water demand forecast modeling techniques-CONVENTIONAL METHODS VERSUS AI. *Journal - American Water Works Association* 94(7): 64–72, Doi: 10.1002/j.1551-8833.2002.tb09507.x.

[131] Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* 120(2–3): 65–73, Doi: 10.1016/S0304-3800(99)00092-7.

[132] Najah, A.A., El-Shafie, A., Karim, O.A., Jaafar, O., 2012. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. *Neural Computing and*

Applications 21(5): 833–41, Doi: 10.1007/s00521-010-0486-1.

[133] Farmer, J.D., Sidorowich, J.J., 1987. Predicting chaotic time series. *Physical Review Letters* 59(8): 845–8, Doi: 10.1103/PhysRevLett.59.845.

[134] Itoh, K.-I., 1995. A method for predicting chaotic time-series with outliers. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* 78(5): 44–53, Doi: 10.1002/ecjc.4430780505.

[135] Porporato, A., Ridolfi, L., 1997. Nonlinear analysis of river flow time sequences. *Water Resources Research* 33(6): 1353–67, Doi: 10.1029/96WR03535.

[136] Goldberg, D.E., Holland, J.H., 1988. Genetic Algorithms and Machine Learning. *Machine Learning* 3(2): 95–9, Doi: 10.1023/A:1022602019183.

[137] Holland, J.H., 1998. Genetic algorithms and the optimal allocation of trials. Evolutionary Computation: The Fossil Record, vol. 2. Society for Industrial and Applied Mathematics p. 443–60.

[138] Ferreira, C., 2013. Function Finding and the Creation of Numerical Constants in Gene Expression Programming. Advances in Soft Computing, London: Springer London p. 257–65.

[139] Ferreira, C., 2011. Gene Expression Programming in Problem Solving. Soft Computing and Industry, London: Springer London p. 635–53.

[140] Haykin, S.S., Simon, 1999. Neural networks : a comprehensive foundation. Prentice Hall.

[141] Melesse, A.M., Hanley, R.S., 2005. Artificial neural network application for multi-ecosystem carbon flux simulation. *Ecological Modelling* 189(3–4): 305–14, Doi: 10.1016/J.ECOLMODEL.2005.03.014.

[142] Ghorbani, M.A., Khatibi, R., Hosseini, B., Bilgili, M., 2013. Relative importance of

parameters affecting wind speed prediction using artificial neural networks. *Theoretical and Applied Climatology* 114(1–2): 107–14, Doi: 10.1007/s00704-012-0821-9.

[143] Najah, A., El-Shafie, A., Karim, O.A., El-Shafie, A.H., 2013. Application of artificial neural networks for water quality prediction. *Neural Computing and Applications* 22(S1): 187–201, Doi: 10.1007/s00521-012-0940-3.

[144] Basant, N., Gupta, S., Malik, A., Singh, K.P., 2010. Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water — A case study. *Chemometrics and Intelligent Laboratory Systems* 104(2): 172–80, Doi: 10.1016/J.CHEMOLAB.2010.08.005.

[145] Rosenstein, M.T., Collins, J.J., De Luca, C.J., 1993. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena* 65(1–2): 117–34, Doi: 10.1016/0167-2789(93)90009-P.

[146] Shang, P., Li, X., Kamae, S., 2005. Chaotic analysis of traffic time series. *Chaos, Solitons & Fractals* 25(1): 121–8, Doi: 10.1016/J.CHAOS.2004.09.104.

[147] Khatibi, R., Sivakumar, B., Ghorbani, M.A., Kisi, O., Koçak, K., Farsadi Zadeh, D., 2012. Investigating chaos in river stage and discharge time series. *Journal of Hydrology* 414–415: 108–17, Doi: 10.1016/j.jhydrol.2011.10.026.

[148] Jain, A., Kumar Varshney, A., Chandra Joshi, U., 2001. Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks. *Water Resources Management* 15(5): 299–321, Doi: 10.1023/A:1014415503476.

[149] Bougadis, J., Adamowski, K., Diduch, R., 2005. Short-term municipal water demand forecasting. *Hydrological Processes* 19(1): 137–48, Doi: 10.1002/hyp.5763.

[150] Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012.

Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research* 48(1), Doi: 10.1029/2010WR009945.

[151] Wang, W., Vrijling, J.K., Van Gelder, P.H.A.J.M., Ma, J., 2006. Testing for nonlinearity of streamflow processes at different timescales. *Journal of Hydrology* 322(1–4): 247–68, Doi: 10.1016/J.JHYDROL.2005.02.045.

[152] Nourani, V., Hosseini Baghanam, A., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. *Journal of Hydrology* 514: 358–77, Doi: 10.1016/J.JHYDROL.2014.03.057.

[153] Labat, D., 2005. Recent advances in wavelet analyses: Part 1. A review of concepts. *Journal of Hydrology* 314(1–4): 275–88, Doi: 10.1016/J.JHYDROL.2005.04.003.

[154] Chou, C.-M., 2014. Application of Set Pair Analysis-Based Similarity Forecast Model and Wavelet Denoising for Runoff Forecasting. *Water* 6(4): 912–28, Doi: 10.3390/w6040912.

[155] Labat, D., 2008. Wavelet analysis of the annual discharge records of the world's largest rivers. *Advances in Water Resources* 31(1): 109–17, Doi: 10.1016/J.ADVWATRES.2007.07.004.

[156] Partal, T., Cigizoglu, H.K., 2008. Estimation and forecasting of daily suspended sediment data using wavelet–neural networks. *Journal of Hydrology* 358(3–4): 317–31, Doi: 10.1016/J.JHYDROL.2008.06.013.

[157] Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. *Journal of Water Resources Planning and Management* 134(2): 119–28, Doi: 10.1061/(asce)0733-9496(2008)134:2(119).

[158] Tirelli, T., Pessani, D., 2011. Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example. *Ecological Informatics* 6(5): 309–15, Doi: 10.1016/J.ECOINF.2010.11.001.

[159] Fornarelli, R., Galelli, S., Castelletti, A., Antenucci, J.P., Marti, C.L., 2013. An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers. *Water Resources Research* 49(6): 3626–41, Doi: 10.1002/wrcr.20268.

[160] Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling* 160(3): 249–64, Doi: 10.1016/S0304-3800(02)00257-0.

[161] Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology* 301(1–4): 75–92, Doi: 10.1016/J.JHYDROL.2004.06.021.

[162] Maier, H.R., Morgan, N., Chow, C.W.., 2004. Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling & Software* 19(5): 485–94, Doi: 10.1016/S1364-8152(03)00163-4.

[163] Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software* 62: 33–51, Doi: 10.1016/J.ENVSOFT.2014.08.015.

[164] Wu, W., Dandy, G.C., Maier, H.R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environmental Modelling & Software* 54: 108–27, Doi:

10.1016/J.ENVSOFT.2013.12.016.

[165] Wu, Y., Zhang, D.Z., 2007. Demand fluctuation and chaotic behaviour by interaction between customers and suppliers. *International Journal of Production Economics* 107(1): 250–9, Doi: 10.1016/J.IJPE.2006.09.004.

[166] Li, X., Zecchin, A.C., Maier, H.R., 2015. Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation. *Environmental Modelling & Software* 71: 78–96, Doi: 10.1016/J.ENVSOFT.2015.05.013.

[167] Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178(3–4): 389–97, Doi: 10.1016/J.ECOLMODEL.2004.03.013.

[168] Garson, G.D., 1991. Interpreting neural-network connection weights. *AI Exper* 6(4): 46–51.

[169] Scardi, M., Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling* 120(2–3): 213–23, Doi: 10.1016/S0304-3800(99)00103-9.

[170] Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* 90(1): 39–52, Doi: 10.1016/0304-3800(95)00142-5.

[171] May, R.J., Dandy, G.C., Maier, H.R., Nixon, J.B., 2008. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. *Environmental Modelling & Software* 23(10–11): 1289–99, Doi: 10.1016/J.ENVSOFT.2008.03.008.

[172] Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data

driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology* 367(3–4): 165–76, Doi: 10.1016/J.JHYDROL.2008.10.019.

[173] Nourani, V., Entezari, E., Yousefi, P., 2013. ANN-RBF Hybrid Model for Spatiotemporal Estimation of Monthly Precipitation Case Study: Ardabil Plain. *International Journal of Applied Metaheuristic Computing* 4(2).

[174] Najah, A., El-Shafie, A., Karim, O.A., El-Shafie, A.H., 2013. Application of artificial neural networks for water quality prediction. *Neural Computing and Applications* 22(S1): 187–201, Doi: 10.1007/s00521-012-0940-3.

[175] Nourani, V., Alami, M.T., Aminfar, M.H., 2009. A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. *Engineering Applications of Artificial Intelligence* 22(3): 466–72, Doi: 10.1016/J.ENGAPPAI.2008.09.003.

[176] Zhou, T., Wang, F., Yang, Z., 2017. Comparative Analysis of ANN and SVM Models Combined with Wavelet Preprocess for Groundwater Depth Prediction. *Water* 9(10): 781, Doi: 10.3390/w9100781.

[177] Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7): 674–93, Doi: 10.1109/34.192463.

[178] Jorgensen, B.S., Martin, J.F., Pearce, M., Willis, E., 2013. Some difficulties and inconsistencies when using habit strength and reasoned action variables in models of metered household water conservation. *Journal of Environmental Management* 115: 124–35, Doi: 10.1016/J.JENVMAN.2012.11.008.

[179] Sivakumar, B., Wallender, W.W., 2004. Deriving high-resolution sediment load data using

a nonlinear deterministic approach. *Water Resources Research* 40(5), Doi: 10.1029/2004WR003152.

[180] Sivakumar, B., 2006. Suspended sediment load estimation and the problem of inadequate data sampling: a fractal view. *Earth Surface Processes and Landforms* 31(4): 414–27, Doi: 10.1002/esp.1273.

[181] R., D.V., Schakke, J.C., 1973. Disaggregation processes in stochastic hydrology. *Water Resources Research* 9(3): 580–5, Doi: 10.1029/WR009i003p00580.

[182] Rebora, N., Silvestro, F., Rudari, R., Herold, C., Ferraris, L., 2016. Downscaling stream flow time series from monthly to daily scales using an auto-regressive stochastic algorithm: StreamFARM. *Journal of Hydrology* 537: 297–310, Doi: 10.1016/J.JHYDROL.2016.03.015.

[183] Debele, B., Srinivasan, R., Yves Parlange, J., 2007. Accuracy evaluation of weather data generation and disaggregation methods at finer timescales. *Advances in Water Resources* 30(5): 1286–300, Doi: 10.1016/J.ADVWATRES.2006.11.009.

[184] Rajagopalan, B., Lall, U., 1999. A *k* -nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research* 35(10): 3089–101, Doi: 10.1029/1999WR900028.

[185] Yates, D., Gangopadhyay, S., Rajagopalan, B., Strzepek, K., 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research* 39(7), Doi: 10.1029/2002WR001769.

[186] Nowak, K., Prairie, J., Rajagopalan, B., Lall, U., 2010. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. *Water Resources Research* 46(8), Doi: 10.1029/2009WR008530.

[187] Prairie, J., Rajagopalan, B., Lall, U., Fulp, T., 2007. A stochastic nonparametric technique for space-time disaggregation of streamflows. *Water Resources Research* 43(3), Doi: 10.1029/2005WR004721.

[188] Olsson, J., 1998. Evaluation of a scaling cascade model for temporal rain- fall disaggregation.

[189] Burlando, P., 2005. Preservation of rainfall properties in stochastic disaggregation by a simple random cascade model. *Atmospheric Research* 77(1–4): 137–51, Doi: 10.1016/J.ATMOSRES.2004.10.024.

[190] Gaume, E., Mouhous, N., Andrieu, H., 2007. Rainfall stochastic disaggregation models: Calibration and validation of a multiplicative cascade model. *Advances in Water Resources* 30(5): 1301–19, Doi: 10.1016/J.ADVWATRES.2006.11.007.

[191] Wang, B., Liu, L., Huang, G.H., Li, W., Xie, Y.L., 2016. Forecast-based analysis for regional water supply and demand relationship by hybrid Markov chain models: a case study of Urumqi, China. *Journal of Hydroinformatics* 18(5): 905–18, Doi: 10.2166/hydro.2016.202.

[192] Mergelas and Henrich, G., B., 2005. Leak locating method for precommissioned transmission pipelines. *North American Case Studies in Leakage 2005* (50 mm): 1–7.

[193] Liggett, J.A., Chen, L., 1994. Inverse Transient Analysis in Pipe Networks. *Journal of Hydraulic Engineering* 120(8): 934–55, Doi: 10.1061/(ASCE)0733-9429(1994)120:8(934).

[194] Mpesha, W., Gassman, S.L., Chaudhry, M.H., 2001. Leak Detection in Pipes by Frequency Response Method. *Journal of Hydraulic Engineering* 127(2): 134–47, Doi: 10.1061/(ASCE)0733-9429(2001)127:2(134).

[195] Brunone, B., 1999. Transient Test-Based Technique for Leak Detection in Outfall Pipes. *Journal of Water Resources Planning and Management* 125(5): 302–6, Doi: 10.1061/(ASCE)0733-9496(1999)125:5(302).

[196] Kim, S.H., 2005. Extensive Development of Leak Detection Algorithm by Impulse Response Method. *Journal of Hydraulic Engineering* 131(3): 201–8, Doi: 10.1061/(ASCE)0733-9429(2005)131:3(201).

[197] Wang, X.-J., Lambert, M.F., Simpson, A.R., Liggett, J.A., Vtkovský, J.P., 2002. Leak Detection in Pipelines using the Damping of Fluid Transients. *Journal of Hydraulic Engineering* 128(7): 697–711, Doi: 10.1061/(ASCE)0733-9429(2002)128:7(697).

[198] Romano, M., Kapelan, Z., Savić, D.A., 2011. Burst Detection and Location in Water Distribution Systems. World Environmental and Water Resources Congress 2011, Reston, VA: American Society of Civil Engineers p. 1–10.

[199] Izquierdo, J., López, P.A., Martínez, F.J., Pérez, R., 2007. Fault detection in water supply systems using hybrid (theory and data-driven) modelling. *Mathematical and Computer Modelling* 46(3–4): 341–50, Doi: 10.1016/J.MCM.2006.11.013.

[200] Mounce, S.R., Boxall, J.B., Machell, J., 2010. Development and Verification of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows. *Journal of Water Resources Planning and Management* 136(3): 309–18, Doi: 10.1061/(ASCE)WR.1943-5452.0000030.

[201] Ye, G., Fenner, R.A., 2011. Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems. *Journal of Pipeline Systems Engineering and Practice* 2(1): 14–22, Doi: 10.1061/(ASCE)PS.1949-1204.0000070.

[202] Kühnert, C., Bernard, T., Arango, I.M., Nitsche, R., 2014. Water Quality Supervision of

Distribution Networks Based on Machine Learning Algorithms and Operator Feedback. *Procedia Engineering* 89: 189–96, Doi: 10.1016/J.PROENG.2014.11.176.

[203] Aminravan, F., Sadiq, R., Hoorfar, M., Rodriguez, M.J., Najjaran, H., 2015. Multi-level information fusion for spatiotemporal monitoring in water distribution networks. *Expert Systems with Applications* 42(7): 3813–31, Doi: 10.1016/J.ESWA.2014.11.014.

[204] Solutions, S.V., 2017. Kelowna Integrated Water Supply Plan. Kelowna: Strategic Value Solutions.

[205] Garson, D., 1991. InteInterpreting neural-network connection weightsrpreting neural-network connection weights. *AI Experts* 6(4): 46–51.

[206] Elmolla, E.S., Chaudhuri, M., Eltoukhy, M.M., 2010. The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. *Journal of Hazardous Materials* 179(1–3): 127–34, Doi: 10.1016/J.JHAZMAT.2010.02.068.

Appendices

Appendix A

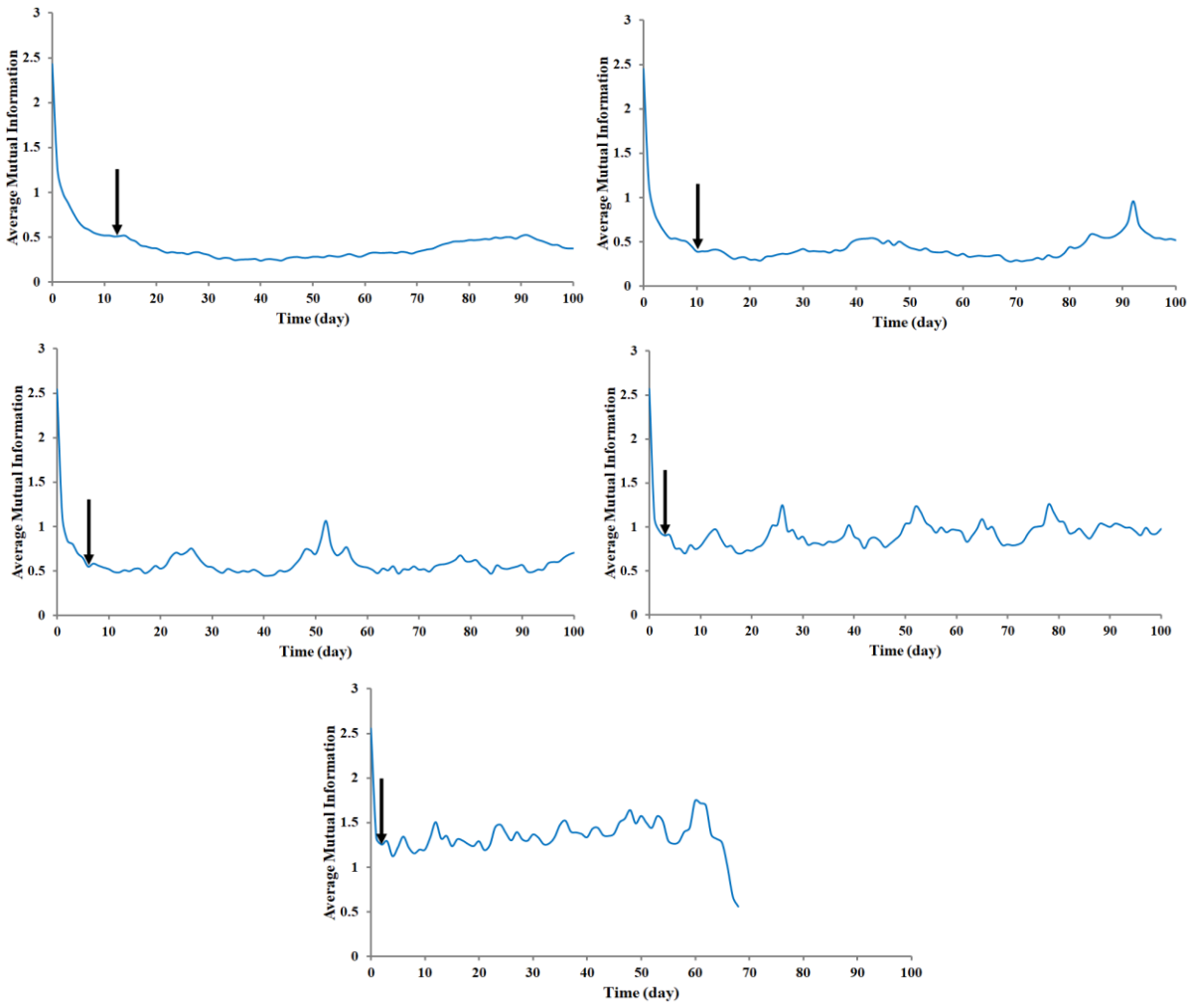
Table A.1 is published in “City of Kelowna Annual Water and Filtration Exclusion Report” that was published in July 2017 [110].

Table A.1. Summary of water production and consumption trends.

	2012	2013	2014	2015	2016	Unit
Production						
Total Pumped	14,983.2	16,962	15,243.2	16,083.6	15,512.8	1000 m3
Population within COK Water Utility Boundary						
2.7 per Single, 1.9 per multifamily	67,311	68,523	68,077	69,679	78,694	people
Climate						
Precipitation (April-September)	181	288	214	153	148	mm
Number of Services (active and inactive)						
Residential	14,087	14,408	14,425	14,084	14,365	services
Multifamily	396	409	387	342	351	services
Strata	658	660	651	648	657	services
Commercial	1,104	1,120	1,074	1,361	1,367	services
Parks	191	187	194	167	167	services
Total	16,436	16,784	16,980	16,910	16,907	services
Consumption (Metered)						
Residential	5,649.8	5,369.3	5,481.4	6,052.0	5,647.6	1000 m3
Multifamily	2,164.4	2,251.1	2,168.8	2,542.7	2,443.0	1000 m3
Strata	683.4	656.0	673.4	735.5	717.0	1000 m3
Commercial	3,058.8	3,250.9	3,304.9	3,666.0	3,867.8	1000 m3
Parks	479.5	402.2	495.6	585.3	498.0	1000 m3
Total	12,035.9	11,929.5	12,124.0	13,599.5	13,173.4	1000 m3
Indicator						
Other Use	2,446.9	2,728.0	829.9	2,421.5	2,201.1	1000 m3
Max Day Demand	82.3	89.4	87.7	87.9	77.8	1000 m3
Average Day Demand	41.1	41.0	41.7	44.1	42.4	1000 m3
Utility wide Peak Day Demand	1,223	1,304	1,392	1,261	989	L/capita/day
Utility wide Average Day Demand	611	598	662	633	539	L/capita/day
Single Family Dwelling Demand	392	371	420	436	403	L/capita/day
Monthly Residential Consumption	36	34	35	36	33	m3/month
Other Use	2,446.9	2,728.0	829.9	2,421.5	2,201.1	1000 m3
Peaking Factor (April-September)	1.5	1.6	1.6	1.5	1.4	
Peaking Factor (October-March)	1.9	1.3	1.5	1.7	1.7	

Appendix B

Figures B.1 plots the results for average mutual information (AMI) for different temporal scales. As it was reported in chapter 2, the first local minimum is considered as a function of lag time. The first minimum values of AMI were at the values of 17, 12, 10, 6, 3 and 2 for daily, 2-, 4-, 7-, 14- and 30-days (monthly) data series of the water consumption, respectively. Figure B.1 presents the AMI values for the temporal scales of 2-, 4-, 7-, 14- and 30-days resolutions.



A.5. Average mutual information (τ) for 2-, 4-, 7-, 14- and 30-days

Figure B.1 and B.2 plots the results for correlation function versus log (*r*) for different embedding dimensions (*m*) varies in the range of 1 to 20 for 4-, 7-, 14- and 30 days temporal scales.

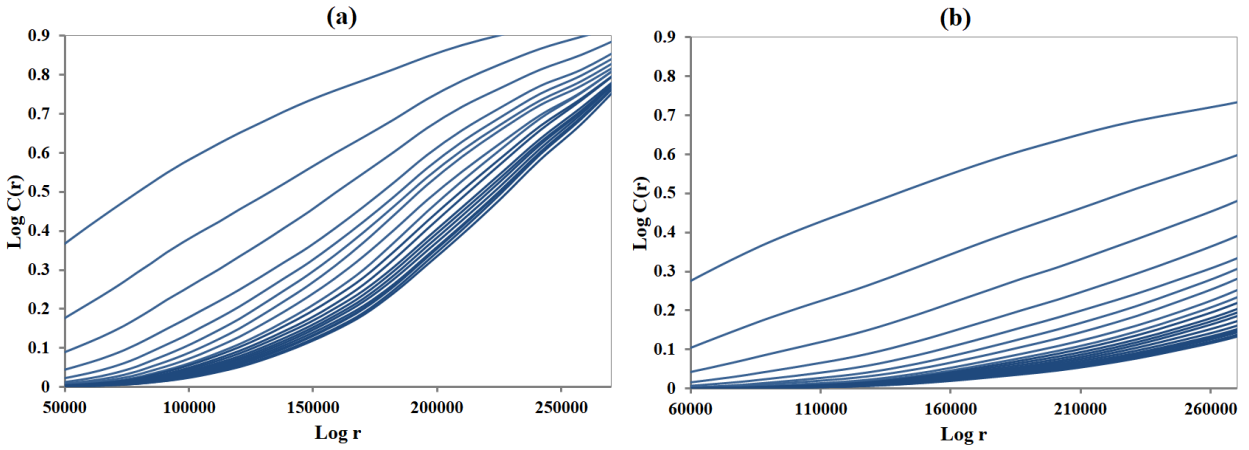


Figure B.1 The relation between correlation function C(r) and r by different embedding dimensions, (a) 4 days; (b) 7-days.

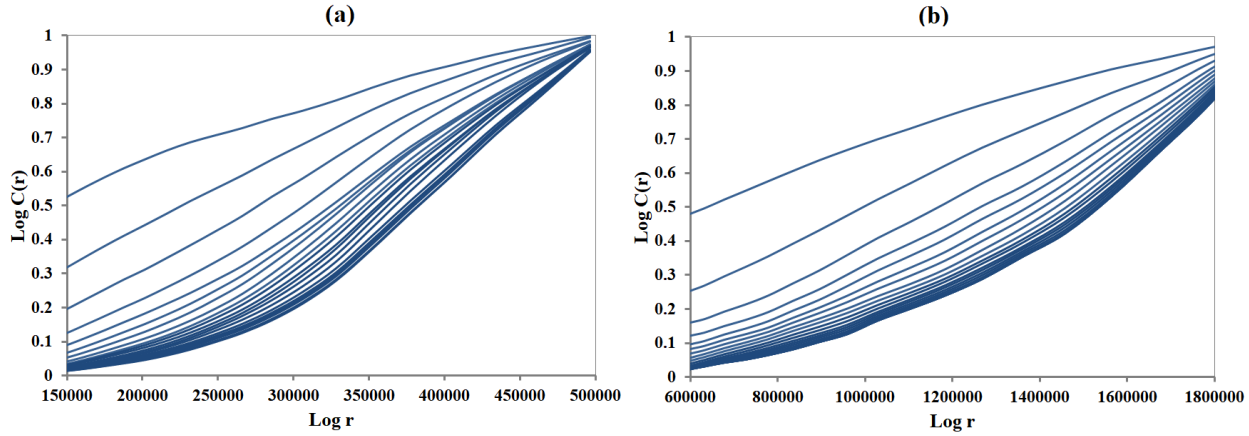


Figure B.2 The relation between correlation function C(r) and r by different embedding dimensions, (a) 14-days; (b) 30-days.

Appendix C [14]

C.1 Methodology

This research found the relative importance of each input variable by Garson's equation [205]:

$$\mathrm{Im}\, p_j = \frac{\sum_{m=1}^{N_h}\left[\left(\frac{\left|w_{j_m}^{i_h}\right|}{\sum_{k=1}^{N_i}\left|w_{k_m}^{i_h}\right|}\right)\times\left|w_{m_n}^{h_o}\right|\right]}{\sum_{k=1}^{N_i}\left[\sum_{m=1}^{N_h}\left(\frac{\left|w_{k_m}^{i_h}\right|}{\sum_{k=1}^{N_i}\left|w_{k_m}^{i_h}\right|}\right)\times\left|w_{m_n}^{h_o}\right|\right]} \quad \text{(C.1)}$$

where Imp_j is the relative importance of the j^{th} input variable, w is a connection weight, and N_i and N_h are the number of neurons at input layers and hidden layers, respectively, extracted from the developed ANN network. While the superscripts i, h and o refer to input, hidden and output layers, the subscripts k, m and n indicate input, hidden neurons and output neurons, respectively [206]. As equation C.1 indicates, the weights were normalized based on their absolute values. Thus, the signs of the weights did not influence the results. This method is considered as a technique for input variable selection (IVS) to improve the accuracy of the models while decreasing the complexity by considering a lower number of input variables. IVS methods employ the active variables that are the most influential on the target variables. The benefits of IVS is already shown in reducing the input variables for 40% in all [15] in water quality studies. The aim of this section is the investigation of the application of average mutual information (AMI) and autocorrelation function (ACF) in reconstructing the phase space for the time series with chaotic behavior.

---

[14] A version of this section has been published as papers: Yousefi, P., Naser, G., Mohammadi, H. 2018. Surface Water Quality Model: Impacts of Influential Variables. Journal of Water Resources Planning and Management 144 (5), 04018015

C.2 Implementation

AMI and ACF are the two methods that are considered to reconstructed phase space for the case data in chapter 3 and chapter 4, respectively. However, the application of AMI reported better than ACF based on the model's accuracy, and there is still a gap in interpreting the application of each method in reconstructing various phase spaces.

Garson equation, as the selected IVS method, is used to interpret the application of AMI and ACF separately in forecasting models' performance. The value of AMI and ACF for the daily temporal scale was reported as τ =17 and τ =83, respectively. Artificial neural network (ANN) is employed as the forecasting model. Embedding dimension of 10 (m=10) is the selected dimension for the phase space. The best structure of ANN is selected among 500 models with various hidden layer neurons and epochs. Table C.1 shows the results of the forecasting model for m=10.

Table C.1. Results of PSR-ANN with two ACF and AMI methods.

PSR	Structure	Epoch	CD	RMSE*	MAE
AMI	10-5-1	50	0.9800	3547.3	48.1
ACF	10-3-1	10	0.9784	3357.6	47.9

The results of Table C.1 shows that the performance of AMI, from the sight of accuracy, was better than ACF. But, in simulating the highest and lowest values, the errors in ACF was lower than AMI. Therefore, the fitness values of the forecasting model are not enough to conclude about the priority of AMI and ACF to each other in PSR. Tables C.2 and C.3 show the value of correlation and relative importance (RI) for input values and forecasted values for both PSR methods. Also, the highest values of correlation and RI for each input variable are selected to define a new combination of input variables for the same forecasting model. The results of this comparison reveal the effect of PSR with AMI and ACF on the models' accuracy, separately.

C.2. Correlation of input variables and forecasted values for PSR with AMI and ACF.

Par	V10	V9	V8	V7	V6	V5	V4	V3	V2	V1
Target (AMI)	-0.545	-0.422	-0.266	-0.132	0.008	**0.203**	**0.421**	**0.612**	**0.770**	**0.979**
Rank						**5**	**4**	**3**	**2**	**1**
Target (ACF)	**0.824**	**0.096**	-0.565	-0.408	**0.451**	**0.460**	-0.289	-0.539	-0.027	**0.978**
Rank	**2**	**5**			**4**	**3**				**1**

C.3. Relative importance of input variables and forecasted values for PSR with AMI and ACF.

Par	V10	V9	V8	V7	V6	V5	V4	V3	V2	V1
RI (AMI) %	4.65	2.13	**16.13**	9.35	**9.45**	7.47	1.19	**11.46**	**15.12**	**23.05**
Rank			**2**		**5**			**4**	**3**	**1**
RI (ACF) %	**4.83**	0.63	0.43	**1.38**	**1.52**	**2.69**	0.84	1.24	0.50	**85.94**
Rank	**2**			**5**	**4**	**3**				**1**

As seen in the figure, the distribution of correlation values and RI for AMI seems more reasonable based on their rank comparing to ACF. Figure C.1 shows the distribution of RI for AMI and ACF.

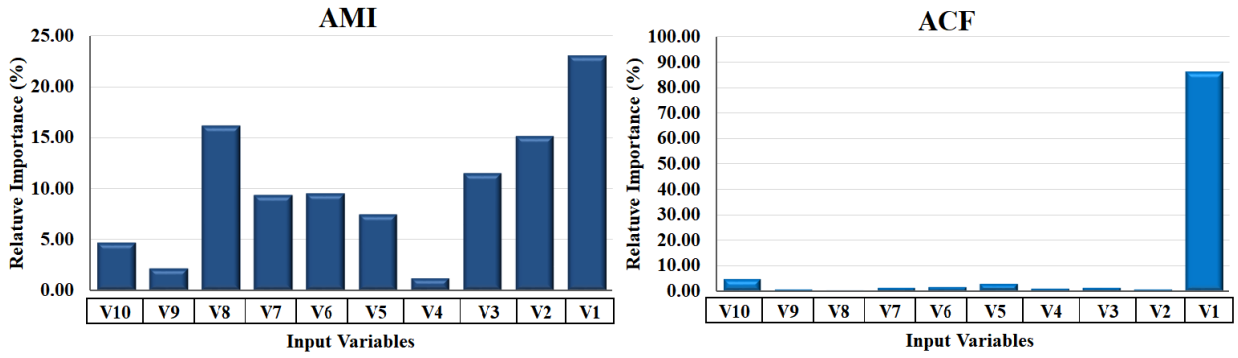


Figure C.1. Distribution of RI for input variables with AMI and ACF.

To investigate the performance of models in combination with the IVS method, the selected input variables (first five variables with higher RI) are used to define a new input combination. Table C.4 shows the results for the RI-PSR-ANN method with AMI and ACF. Table C.4 shows how the models, with the lower number of input variables and hidden layer neurons and epochs, gave the same results. Regarding Table C.1 and C.4, it is concluded that IVS methods, in combination with the models, can improve the performance of models.

Table C.4. Results of RI-PSR-ANN with two ACF and AMI methods.

PSR	Structure	Epoch	CD	RMSE*	MAE
AMI	5-3-1	10	0.9799	3556.8	47.7
ACF	5-2-1	50	0.9785	3348.1	48.3

Based on the outcomes in Appendix C, it was concluded that AMI is the appropriate method to calculate the lag time to phase space reconstruction in comparison with ACF. Also, the combination of IVS and forecasting models is a proper pre-processing approach to improve the performance of the models. For further investigation as the future work, this section suggests investigating other phase space reconstruction method. Moreover, the application of hybrid pre-processing techniques such as Wavelet-IVS and PSR-Wavelet-IVS are recommended to improve the performance of forecasting models and forecasting horizon in mid- and long-term periods.

Appendix D

To improve the accuracy of forecasting models, phase space reconstruction (PSR), wavelet decomposition and input variable selection (IVS) methods reported as the potential techniques. Also, it was reported in literature high resolution data improve the performance of the models in simulating the critical points (e.g. highest and lowest consumption values in water distribution networks). Also, it was reported in chapter one, majority of the studies are in forecasting short- and mid-term period in lower temporal scales such as monthly and weekly. Previously, application of PSR, wavelet decomposition, IVS and hybridizing the pre-processing methods in combination with forecasting models presented the better performance of the forecasting models in daily temporal scale. Along with the same results, this section presents the application of disaggregation models in forecasting low resolution consumption data in improving the accuracy of the models.

In the first step, monthly values (aggregated from the daily values) are used for forecasting monthly consumption values with artificial neural networks. Then, the disaggregated values (monthly to daily) for the same temporal scales are used as the input variables of the same forecasting models (same structure for ANN). The comparison of the results for the two input variables reveals the application of disaggregation methods in improving the performance of forecasting models.

The value of β was calculated by the gradient of line, approximately 1.775 (chapter 5); which is greater than 1, and represents scaling behavior of the time series [75]. Therefore, it was concluded that the test case dataset has scaling behavior. For the lag time, the values of 15, 10, 10, 5, 6 and 2 were reported (chapter 5) for the daily, 2-, 4-, 8-, 16-days and 32-days temporal scale consecutively. Therefore, $\tau = 2$, is the lag time to reconstruct the phase space for the monthly scale.

Embedding dimension of m=1 to 20 showed the dynamic of reconstructed transportation; also, zero-degree approximation is utilized to find neighbors in the disaggregation process. Following the results in chapter 6, 75% of the available data are used for training, and the remaining 25% is used for testifying the forecasting models. The embedding dimension of 15 (m=15) was the selected dimension for the most accurate disaggregated among dimensions 1 to 20. Table D.1 shows the input variables define for ANN Model.

Table D.1. Input variables definition for Daily and Monthly temporal scales.

Variables	Disaggregated (Daily)	Monthly
Input	C_{t-17}, $C_{t-(2\times17)}$, …, $C_{t-(15\times17)}$	C_{t-2}, $C_{t-(2\times2)}$, …, $C_{t-(15\times2)}$
Output	$C_{t\ (Daily)}$	$C_{t\ (Monthly)}$

The output for the daily resolution is aggregated to the monthly. The monthly scales for both calibrated models are compared. Artificial neural network (ANN) is employed as the forecasting model. Embedding dimension of 15 (m=15) is the selected dimension for the phase space (Selected dimension based on the value of fitness function (see Table 5.3). The best structure of ANN is selected among 2000 models with various hidden layer neurons (1 to 20) and epochs (1 to 100). Table D.2 shows the results of the forecasting model for m=15.

Table D.2. Results of monthly forecasting, daily forecasting and aggregated values from daily to monthly temporal scale.

Scale	Structure	Epoch	CD	RMSE*	MAE
Monthly	15-17-1	70	0.9532	98464	270
Daily	15-18-1	100	0.9123	5078	64
Daily to Monthly (Aggregated)	NA	NA	0.9770	79360	247

The results in Table D.1 are for the selected models with the most accurate fitness function values. The aggregated values have resulted from the transition of daily forecasting values to monthly

values. The results showed that employing disaggregation techniques that were presented in chapter 5 can improve the performance of the models by increasing accuracy. Therefore, besides the other pre-processing methods in the literature and chapter 5 of this research, a combination of disaggregation methods are potential methods to improve the accuracy of the forecasting models. Figure D.1 presents the comparison of observed values of water consumption with monthly forecasted and aggregated values.

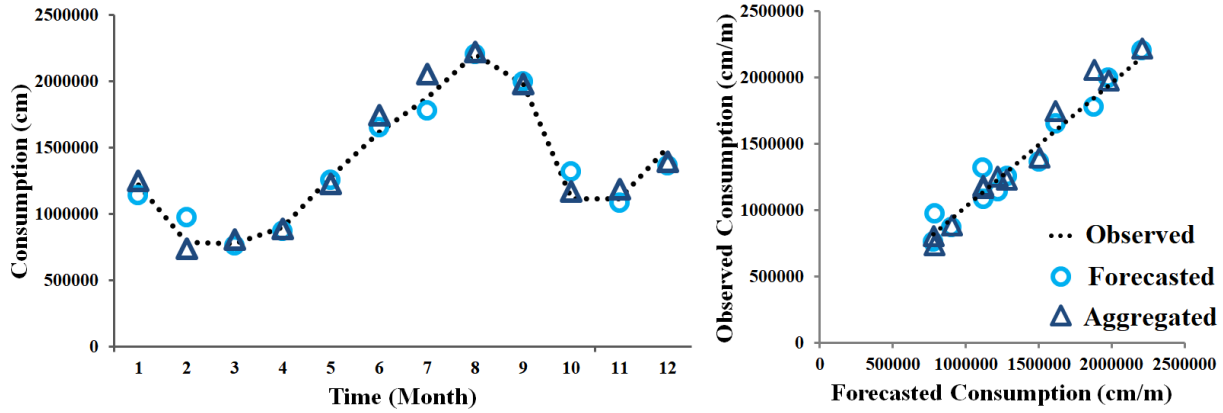


Figure D.1. Observed in comparison with monthly and aggregated values of water consumption.