DATA PRE-PROCESSING TECHNIQUES FOR IMPROVING RIVER WATER LEVEL PREDICTION: A CASE STUDY OF THE DUNGUN RIVER, TERENGGANU, MALAYSIA

By

ERVIN TIU SHAN KHAI

A dissertation submitted to the Department of Civil Engineering,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Master of Engineering Science
April 2019

ABSTRACT

DATA PRE-PROCESSING TECHNIQUES FOR IMPROVING RIVER WATER LEVEL PREDICTION: A CASE STUDY OF THE DUNGUN RIVER, TERENGGANU, MALAYSIA

Ervin Tiu Shan Khai

An accurate river water level prediction model is vital for the development of flood mitigation plans in a river basin, and the accuracy of input data is important for ensuring good predictions. In this research study, a Support Vector Regression (SVR) model was applied to predict river water levels at the Dungun River, Terengganu, Malaysia. A major challenge of this study is to process the observed rainfall series which may tend to be incomplete and inconsistent. A better observed rainfall series data can lead to improved performance of the river water level prediction model. This study adopted the data pre-processing techniques to first improve the observed rainfall series. Three data pre-processing techniques namely; the Variational Mode Decomposition (VMD) method, the Boosting method and the Boosting-VMD method, were adopted to pre-process the observed rainfall series at the basin prior to applying them into the prediction model. Rainfall series and river water levels from November to February (Northeast Monsoon period) for the period of 1996-2016 were used. The three pre-processed rainfall series and the

original observed rainfall series were then separately and individually, used to predict river water levels using the SVR model. The predicted river water levels from the four modified SVR models namely; the Ori-SVR (using original observed rainfall) model, the VMD-SVR (using VMD method) model, the B-SVR (using Boosting method) model, and the B-VMD-SVR (using Boosting-VMD method) model, were assessed against the observed river water levels, using non-parametric statistics (Model prediction errors' range, standard deviation and confidence interval range) and parametric statistics (Bias, Root-Mean-Square Error, Mean Absolute Percentage Error, Nash-Sutcliffe Efficiency Coefficient and Mean Absolute Error) where appropriate, to assess the three data pre-processing techniques in enhancing the SVR model. Results indicated that the VMD, the Boosting and the Boosting-VMD methods afforded data improvements that boosted the performance of SVR models. Further statistical analyses showed that the SVR model cum the Boosting-VMD method that is, the (B-VMD-SVR) to be the most robust, with results of model prediction error's range (4.27; Min: -2.95, Max: 1.32), model prediction error's standard deviation (0.4200), model prediction error's confidence interval range ([-0.02, 0.01]; Standard error: 0.00868), Bias (0.00), Root-Mean-Square Error (0.42), Mean Absolute Percentage Error (4.36), Nash-Sutcliffe Efficiency Coefficient (0.96) and Mean Absolute Error (0.28).

ACKNOWLEDGEMENTS

I would like to thank everyone who had in one way or another, contributed to the successful completion of this project. First of all, I would like to express my sincere gratitude to my research supervisor, Ir. Dr. Huang Yuk Feng for his invaluable advice, guidance and his enormous patience throughout the development and progress of the research. His continuous support, patient, enthusiasm, motivation and the strong knowledge in the field of hydrology helped me to stay focussed throughout the duration of my master study. I would also like to record my appreciations to Dr. Ling Lloyd, for not only being my co-supervisor and my personal advisor in the statistical work but also for his useful and inspiring comments.

This research was funded by UTAR Research Fund and thus, I sincerely would also like to express my utmost appreciation to UTAR for the financial support without which my study would not have been possible.

Finally, I would also like to express my gratitude to my parents for their patience, encouragement and understanding and also to my friends who had helped and had given me encouragement to persevere in this project, apart from making my study at UTAR most memorable.

APPROVAL SHEET

This dissertation entitled "**DATA PRE-PROCESSING TECHNIQUES FOR IMPROVING RIVER WATER LEVEL PREDICTION: A CASE STUDY OF THE DUNGUN RIVER, TERENGGANU, MALAYSIA"** was prepared by ERVIN TIU SHAN KHAI and submitted as partial fulfillment of the requirements for the degree of Master of Engineering Science at Universiti Tunku Abdul Rahman.

Approved by:

_____
(Ir Dr Huang Yuk Feng)
Date:……..
Supervisor
Department of Civil Engineering
Lee Kong Chian Faculty of Engineering Science
Universiti Tunku Abdul Rahman

_____
(Dr Ling Lloyd)
Date:……..
Co-supervisor
Department of Civil Engineering
Lee Kong Chian Faculty of Engineering Science
Universiti Tunku Abdul Rahman

LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: _____

SUBMISSION OF DISSERTATION

It is hereby certified that ***Ervin Tiu Shan Khai*** (ID No: ***16UEM07563***) has completed this dissertation entitled "*DATA PRE-PROCESSING TECHNIQUES FOR IMPROVING RIVER WATER LEVEL PREDICTION: A CASE STUDY OF THE DUNGUN RIVER, TERENGGANU, MALAYSIA*" under the supervision of Ir Dr Huang Yuk Feng (Supervisor) from the Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, and Dr Ling Lloyd (Co-Supervisor) from the Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science.

I understand that University will upload softcopy of my dissertation in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

_____

(*Ervin Tiu Shan Khai)*

DECLARATION

I ERVIN TIU SHAN KHAI hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

_____

(ERVIN TIU SHAN KHAI)

Date _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

AD	Anderson-Darling test
ADMM	Alternate Direction Method of Multipliers
AI	Artificial Intelligence
AM-FM	Amplitude-Modulated-Frequency-Modulated
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Networks
AR	Autoregressive
ARIMA	Auto Regressive Integrated Moving Average
B-SVR	Support Vector Regression Model cum Boosting method
B-VMD-SVR	Support Vector Regression Model cum Boosting-Variational Mode Decomposition method
BCa	Bias Corrected and Accelerated
BR	Buishand Range test
CART	Classification and Regression Trees
CVM	Cramer-Vin Mises
CWA	Continuous Wavelets Analysis
DWA	Discrete Wavelets Analysis
DID	Department of Irrigation and Drainage
EMD	Empirical Mode Decomposition
ERM	Empirical Risk Minimization
H_o	Null Hypothesis
H_i	Alternative Hypothesis
IFAS	Integrated Flood Analysis System
IMF	Intrinsic Mode Function

KS	Kolmogorov-Smirnov test
LF	Lilliefors test
LRF	Linear Recurrent Formulae
LSSVM	Least Square Support Vector Machine
MA	Moving Average
MAE	Mean Absolute Error
MANN	Modular Artificial Neural Networks
MAPD	Mean Absolute Percentage Deviation
MAPE	Mean Absolute Percentage Error
MRE	Mean Relative Error
NADMA	National Disaster Management Agency
NLR	Nonlinear Regression
NLR-R	Nonlinear Regression with Regionalization
NSEC	Nash-Sutcliffe Efficiency Coefficient
OLS	Ordinary Least Squares
Ori-SVR	Support Vector Regression Model cum Original Observed Rainfall Series
PASW	Predictive Analytic Software
PCA	Principal Component Analysis
PeT	Pettitt Test
RBF	Radial Basis Kernel
RMSE	Root-Mean Square Error
RRNN	Rainfall-Runoff Neural Network
R	Coefficient of Correlation
SNHT	Standard Normal Homogeneity Test

SPSS	Statistical Package for the Social Science
SRM	Structural Risk Minimization
SSA	Singular Spectrum Analysis
SVM	Support Vector Machine
SVR	Support Vector Regression
SW	Shapiro-Wilk test
WSVR	Wavelet-Support Vector Regression
VMD	Variational Mode Decomposition
VMD-SVR	Support Vector Regression Model cum Variational Mode Decomposition method
VNR	Von Neumann Ratio test

CHAPTER 1

INTRODUCTION

1.1 Background

The Dungun River Basin is one of seven districts in the state of Terengganu, Malaysia. It is categorized as one of the more severe flood prone areas in Peninsular Malaysia. Some limited research had previously been carried out by researchers at the Dungun River Basin, due to the regular flooding events in the Dungun River Basin.

Hafiz et al. (2013) had reported on the flood forecasting and early warning system for the Dungun River Basin. An Integrated Flood Analysis System (IFAS) model has been implemented for the research study. It was found out that the Dungun River Basin suffers from annual regular severe flooding due to the prolong rainfall events caused by the Northeast monsoon at the East Coast Region of Peninsular Malaysia. Severe flooding will result in the loss of properties, crops and even deaths. People around the flood prone areas should be warned through the flood forecasting and early warning system so that they can evacuate immediately.

Lariyah (2014) had investigated the hydrological extreme flood event in Dungun River Basin region. In 20^{th} century, the hydrological extreme events have been associated with the climate changes and variabilities, causing

major flooding which resulted in severe economic losses and lives. Due to the geographical location of Malaysia, the Northeast Monsoons will bring abundant amounts of rainfall. The Department of Irrigation and Drainage (DID) Malaysia reiterated that major floods had occurred annually since 2001.

Wu et al. (2009) stated that suitable data pre-processing technique can help in improving the performance of data-driven models. Hybrid models which combine both data pre-processing techniques and Artificial Intelligence models have been introduced to estimate river water levels. During the late 1990s, the Empirical Mode Decomposition (EMD) which was introduced by Huang et al. (1998) was widely used in recursively decomposing signal into different modes of unknown signals. However, another entirely non-recursive model which is more robust to sampling and noise namely, the Variational Mode Decomposition (VMD) was introduced later to overcome the limitations of the EMD (Dragomiretskiy and Zosso, 2014).

Ensemble learning techniques have also employed widely in the hydrological modelling and estimation of hydrologic variables in recent years (Anctil and Lauzon, 2004; Snelder et al., 2009). Boosting is one of the ensemble methods that was used to improve the performance of weak predictors. A linear combination of the models can be created using Boosting method (Hancock et al., 2005). However, the ensemble methods have still yet to be explored in the field of hydrology field, especially in river water level prediction.

Since early 20th century, artificial intelligence was introduced into engineering and science problems, especially in the prediction and modelling of non-linear hydrological applications. In the last two decades, the Support Vector Machine (SVM) was introduced as a brand-new statistical learning method. It was a robust and effective method for both classification (Vapnik, 1995) and regression (Collobert and Bengio, 2001; Drucker el al., 1996) for a noisy set of data. The SVM was originally used for classification procedure purposes, and recently another version of the SVM called the Support Vector Regression (SVR) was introduced (Vapnik et al. 1997). The SVR has also used extensively in hydrology and other fields (Moradkhani et al. 2004; Yu et al. 2006; Lin et al. 2006; Wu et al. 2008; Lin et al. 2009; Chen et al. 2010; Yoon et al. 2011; Yousefi et al 2015).

1.2 Problem Statement

As late as 2016, the National Disaster Management Agency (NADMA), Malaysia have stated that continuous heavy rain had frequently caused severe inundating floods in the state of Terengganu, Malaysia forcing hundreds to evacuate. Some of the affected areas in the Terengganu state had recorded almost 250 mm of rainfall over a 24-hour period, which was considered a severe case. The worst affected areas in the state are the districts of Kuala Terengganu and Kemaman. Then, the districts of Hulu Terengganu, Marang and Dungun have also been pounded with heavy rains since 28 November 2016. In early 2017, the flooding became even serious in some other states in

Peninsular Malaysia. In Kelantan and Terengganu, approximately 25,000 people had evacuated from their homes. Areas of neighbouring southern Thailand was also affected severely, and the latest reports claimed that 96 people have died due to the flooding since early of the year (FloodList, 2017).

The river water level prediction model is used to calculate the conversion of rainfall series into river water level. The model must be trained and validated before it can be used. The observed rainfall data tend to be incomplete, inconsistent and lacking in certain behaviours or trends which more likely are a lead cause to many errors. Hence, data pre-processing techniques should be applied to improve the observed rainfall series data. This can then allow for improving the performance of river water level prediction model. And, more advance and latest modelling techniques such as the incorporation of artificial intelligence in the models are needed for more accurate river water level prediction.

1.3 Aim and Objectives

The aim of this study is to determine the best data pre-processing technique to be used in tandem with the flood level prediction model in order to improve the performance of the selected river water level prediction model for the Dungun River Basin, Terengganu state. The specific objectives of this study are stated as below:

i. To process observed rainfall series using data pre-processing techniques to be used as the input for the river water level prediction model,

ii. To train and validate the river water level prediction model using the processed rainfall series and the observed rainfall series, and

iii. To assess the performance of all the data pre-processing techniques using appropriate model performance assessment measures.

1.4 Significance of Study

It is quite important to conduct research to gain more accurate river water level prediction results for the water resources system in the Dungun River Basin. Through this study, the selected best improved river water level prediction model can be employed for the development of a flood mitigation emergency management plan and to ensure sustainable management of the Dungun River Basin. The data pre-processing techniques for analysing and denoising the observed rainfall series is used to improve the river water level prediction model. The results obtained from this study are beneficial to the hydrology field, especially in the field of flood forecasting and flood management. It can be used to predict river water level for flood mitigation purposes.

1.5 Scope of Work

The scope of work includes the application of data pre-processing techniques to improve the observed rainfall series for the three rainfall stations in the Dungun River Basin and the application of river water level prediction model to predict river water level based on the processed rainfall series. Based on the literature review, only three data pre-processing techniques will be chosen as the techniques for this research study. The data pre-processing techniques (VMD method, Boosting method, Boosting-VMD method) are employed to improve the performance of chosen river water level prediction model, which is the SVR model. In this research study, only the SVR model will be implemented for predicting river water level using the processed and observed rainfall series.

CHAPTER 2

LITERATURE REVIEW

2.1 Data Pre-Processing Techniques

2.1.1 Empirical Mode Decomposition (EMD)

The EMD method, which was introduced by Huang et al. (1998) is a multi-resolution signal pre-processing technique which is able to decompose time series into its intrinsic components adaptively. The incoming signal can be decomposed into a series using EMD, namely the Intrinsic Mode Function (IMF). The EMD can used to analyse the input signal without using any pre-determined basis functions as done for the Fourier and wavelet transforms (Huang et al., 1998; Lee and Ouarda, 2010). An IMF is a function that need to satisfy the two conditions stated below:

i. In the whole data set, the number of extrema (both maxima and minima) and the number of zero crossings must either equal or differ at most by one, and
ii. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The IMF are amplitude-modulated-frequency-modulated (AM-FM) signals whose equation is shown in Equation 2.1:

$$u_k(t) = A_k(t)\cos(\phi_k(t)) \quad (2.1)$$

where the phase $\phi_k(t)$ is a non-decreasing function, $\phi'_k(t) \geq 0$, the envelope is non-negative $A_k(t) \geq 0$, and both the envelope $A_k(t)$ and the instantaneous frequency $\omega_k(t) = \phi'_k(t)$ vary much slower than the phase $\phi_k(t)$.

Furthermore, the EMD is self-adaptive and data driven. Hence, this method is suitable to be used for both non-linear and non-stationary time series data analysis. Due to the attractive characteristics of EMD, it was involved extensively in several area signal processing, such as in mechanical engineering (Ricci and Pennacchi, 2011), signal de-noising (Li et al., 2011; Lahmiri and Boukadoum, 2014a; 2015a), speaker identification (Wu et al., 2011), biomedical image analysis (Ai et al., 2011; Lahmiri and Boukadoum, 2014b; 2015b), DNA sequence analysis (Zhang et al., 2012), and machinery fault diagnosis (Cheng et al., 2012). In addition, the EMD was used in the economic and financial data analysis as well, such as modelling and predicting crude oil price (Zhang et al., 2008; Zhang et al., 2009), stock market (Cheng et al., 2014), electricity price (An et al., 2013; Lisi and Nan, 2014), and foreign exchange rate (Lin et al., 2012; Premanode and Toumazou, 2013).

Apart from that, EMD has also applied widely in the fields of hydrological and water resources such as river water level prediction, which is the main focus of this research study. Some researches regarding EMD cum streamflow forecasting model had been done to yield the data for streamflow forecasting purpose (Napolitano et al., 2011, Huang et al., 2014 and Wang et

al., 2015). Zhu et al. (2016) had used the support vector machine model to predict the monthly streamflow. They had using the discrete wavelet analysis (DWA) and the EMD for time series decomposition and found that both methods help in enhancing the accuracy of streamflow forecasting.

2.1.2 Variational Mode Decomposition (VMD)

As stated by Aneesh et al. (2015), the VMD uses calculus of variation to decompose the signal into various modes of IMFs. Dragomiretskiy and Zosso (2014) was the first to introduce the multi-resolution VMD model as an alternative to the EMD method in order to overcome its limitations. As previous mentioned, the EMD was an adaptive technique by Huang et al. (1998) to decompose signal into different modes of unknown recursively. However, the limitations of the EMD while lacking in sensitivity to noise and sampling as well as algorithm ad-hoc nature lacking mathematical theory, also could only be partially solved by using more mathematical attempts to decomposition problem like synchro-squeezing, empirical wavelets or recursive variational decomposition (Dragomiretskiy and Zosso, 2014).

Hence, an entirely non-recursive VMD method was introduced as an alternative to EMD, which was able to extract the modes concurrently, properly balancing errors between them and separate tones of similar frequencies (Dragomiretskiy and Zosso, 2014). An optimization methodology called the Alternate Direction Method of Multipliers (ADMM) was proposed

in the VMD to search the central frequencies and IMFs centred on those frequencies concurrently. However, the EMD is sifting recursively and backward error correction is not allowed (Aneesh et al., 2015).

Apart from the statement by Aneesh et al. (2015), Dragomiretskiy and Zosso (2014) stated that VMD was able to show an optimized and positive results using ADMM approach. In their study, the preliminary results showed the attractive performance with respect to the previous EMD. The EMD was mostly limited when it is unable to cope with noise properly. As stated previously, the EMD is sifting recursively in most methods while not allowing for backward error correction. This will specifically inhibit the coping of the noise. Here, the VMD showed a more promising practical decomposition result as it is more robust to sampling and noise and it can address the presence of noise in the input signal specifically as compared to that of EMD (Dragomiretskiy and Zosso, 2014).

In summary, the tones of similar frequencies can be separated using VMD (Dragomiretskiy and Zosso, 2014). Dragomiretskiy and Zosso (2014) stated that VMD outperforms the EMD using simulated harmonic functions. Salim (2016) performed the VMD approach which integrated with the general regression neural network model to analyse and forecast the economic and financial series. He stated that VMD had better performance as compared to the EMD in terms of noisy signals analysis. The VMD can be processed faster and had faster convergence while dealing with large dataset. It is noteworthy that VMD outperforms EMD in some ways.

Besides that, the VMD was also found to contribute widely and possess outstanding performance in different field of studies. It had been applied to the fields of biomedical signal denoising (Lahmiri and Boukadoum, 2014c; 2015c) as well as the international stock markets analysis (Lahmiri, 2015d). Furthermore, Aditya et al. (2016) presented a novel method for denoising knee joint vibration (VAG) signals using VMD. They proved the VMD superiority in denoising raw VAG signals as compared to previous methods such as the wavelet-soft thresholding, the EMD-detrended fluctuation analysis and the EMD-filtering. However, the effectiveness of VMD against EMD in modelling especially river water level prediction modelling is still yet to be explored.

2.1.3 Discrete Wavelet Analysis (DWA)

Wavelet analysis is considered as an interesting tool for analysing time series variation. Through changing the scale and shifting factors of a mother wavelet (w), the incoming signal can be decomposed into multiple lower resolution levels. Both high-pass filter and low-pass filter were used to perform the wavelet analysis (Nalley et al., 2012).

Continuous wavelets analysis (CWA) and the discrete wavelets analysis (DWA) are the two main types of wavelet transform. CWA decomposes signal on all scales whereas the DWA decomposes on discrete number of scales. For CWA, it requires more computation time and data

during decomposition. However, as compared to the CWA, the DWA needs lesser computation time and it is easier to develop (Adamowski and Chan, 2011). Thus, for practical applications, the DWA is usually preferred and it is widely used in engineering applications. Percival (2008) stated that DWA is more suitable for rainfall data analysis as CWA was unable to generate the time series information. Chou (2007) stated that the transformation by DWA is more simplified due to the dyadic calculation of position and scale of a signal. Not to mention, the signal can even be decomposed into approximation and detail series using DWA (Misiti and Misiti, 1996).

As stated by Lee and Yamamoto (1994), the Haar, Meyer, Morlet and Daubechies wavelets are the four common mother wavelets. Maheswaran et al. (2013) also stated that there are several wavelets in the family of wavelets which are distinguished by their respective number of decompositions such as the Haar wavelet, Daubechies 2 wavelet (db2), Symlet (sym) or Spline. The appropriate settings for the filter bank are based on the selection of a wavelet family. According to Maheswaran and Khosa (2011b), the wavelets' type and order, as well as the levels of decomposition for the time series will have diverse features. However, the db wavelet is the most common DWA family which has been applied in hydro-meteorological studies because of its characteristics. Two types of evaluation criteria namely the mean relative error (MRE) and the relative error were used to assess the type of mother wavelet and the appropriate level of decompositions (de Artigas et al., 2006; Nalley et al., 2012).

2.1.3.1 Haar Wavelet

The Haar wavelet is a type of discrete wavelet function and sequence of rescaled square shape function. It possesses two functions which are the scaling function, ϕ (father wavelet) and wavelet, ψ (mother wavelet) that generated a family which used to break up or reconstruct a signal. The wavelet analysis which is based on the Haar scaling function is simple as its building blocks are translations and dilations. de Chazal et al. (2000) stated that the Haar wavelet (also called as Daubechies 1) was suitable for the classification task as it required the least computation. It is simple for applications as the process of Haar wavelet is sharp and compact (Kumar and Georgiou, 1997).

Maheswaran and Khosa (2012) had compared different kind of wavelets for hydrologic forecasting. In the model application, the researchers had dealt with three case studies with different time series. In the study, they stated that the Haar wavelet's feature which possesses relatively narrow span over was found to be applicable in signals that have sharp changes. However, it was categorized as a wavelet which have transient features as compared to other wavelets in wavelet family. It is only recommended for time steps that have short term features. It is expected to be implemented in daily streamflow time series.

Through this several criteria, some best wavelet family had been chosen. By using several quantitative standard statistical performance evaluation measures for performance and accuracy checking, Haar wavelet

possessed a better performance as compared to other wavelets. Therefore, Haar wavelet is recommended to be used for time series of daily streamflow (Maheswaran and Khosa, 2012).

2.1.3.2 Daubechies Wavelet (db)

Daubechies wavelet (db) was discovered by Ingrid Daubechies. This also includes the Haar wavelet as previous mentioned, which is written as db1, as a special case. The concepts of the Daubechies wavelet is similar to the Haar wavelet in terms of trend and fluctuation. However, the Daubechies wavelet has different properties in how scaling functions and wavelets are defined.

The high-pass and low-pass filters are the function sets that were operated by the Daubechies wavelet for time series decomposition purpose. The incoming signal are processed by using both high-pass and low-pass filters and eventually the signal are separated into two different scales which are details and approximation. For the approximation time series, the signal is decomposed into high-scale and low frequency components and vice versa for the details time series.

Maheswaran and Khosa (2012) had investigated the suitability of wavelets for different time series in their research's case studies to compare the different wavelets for hydrologic forecasting. In the model application, they had dealt with the seasonal and monthly streamflow time series analysis.

Throughout the case studies, they found out that db2 are recommended for time series having non-linear features and long-term memory such as seasonal and monthly time series. The underlying long-term features can be extracted more effectively using db2 wavelet as it possessed good frequency localization capabilities. Therefore, db2 wavelet is recommended to be used for time series of seasonal and monthly streamflow.

2.1.3.3 B-spline Wavelet

B-spline wavelet is well-known for its compactly supported characteristics and it can be formulated in an explicit form analytically. According to Chui (1992), the m^{th} order of B-spline function was presented in the recursive form while the 0^{th} order of B-spline corresponds to the well-known Haar wavelet, as shown in Equations 2.2 and 2.3.

$$N_m(x) = \frac{x}{m-1} N_{m-1}(x) + \frac{m-x}{m-1} N_{m-1}(x-1), m \geq 2 \qquad (2.2)$$

where

$$N_1(x) = \chi_{[0,1)}(x) = \begin{cases} 1 & if\ x \in [0,1) \\ 0 & otherwise \end{cases} \qquad (2.3)$$

Based on Wei and Billings (2006), the B-splines wavelets are unique among other various basis functions due to their three integrating remarkable properties which are stated below:

i. B-splines have compact support,

ii. B-splines can be analytically formulated, and

iii. B-splines are suitable for multi-resolution analysis.

2.1.3.4 Sym Wavelet

Sym wavelet, also known as symlet (in a compact form) is nearly symmetrical and it possesses similar properties to the general Daubechies wavelets, dbNs, as abovementioned. It was proposed in the Daubechies as modifications to the db family. At a given support width, the symlet possessed the highest number of vanishing moments. The associated scaling filters of symlet are close to linear-phase filters (Nibhanupudi, 2003). The symlet be able to both orthogonal and biorthogonal and it provides the compact support as well, as similar to B-spline wavelet as abovementioned. Several types of sym wavelet are clearly shown in Figure 2.1.

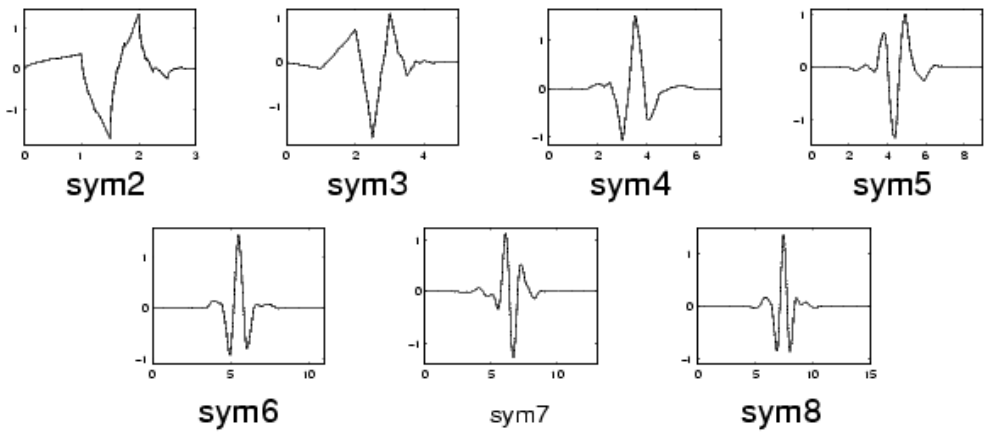


Figure 2.1: Types of Sym Wavelets (The MathWorks, 2005)

2.1.4 Moving Average (MA)

Moving Average (MA) can smoothen the input data by replacing one-to-one data point by using the average of the *k* neighbouring data points. The *k* neighbouring data points are the length of memory window. By getting the average points with the immediate neighbours' data points, those large irregular components at any point of time can exert a smaller effect which will in turn decrease the inaccuracy of data obtained (Newbold et al., 2003). The MA method is carried out by smoothing the results with same weight of data value. MA includes the centering, backward and forward types. Only backward mode was used as another two types may necessitate future observed values. The *k*-term unweighted moving average y_t^* is shown in Equation 2.4 when the backward moving mode is implemented (Lee et al., 2000).

$$y_t^* = \frac{(\sum_{i=0}^{k-1} y_{t-i})}{k} \tag{2.4}$$

where $t = k, \ldots, N$. The choice of the window length *k* is by a trial and error procedure with a minimization of the loss of the objective function.

de Vos and Rientjes (2005) proposed the MA method to be coupled with the Artificial Neural Networks (ANN). The MA method as a data pre-processing technique was found to be able to enhance the performance of ANN. As stated by the researchers, the lagged predictions of ANN results were caused by the previous ANN observed input data. The MA method was

used as a new relatively effective solution to obtain a new model inputs as compared to that of original data series.

Wu et al. (2009) had predicted the monthly streamflow using data pre-processing techniques cum data-driven models. Singular Spectrum Analysis (SSA) and Moving Average (MA) were used in the study. In the research study, they found out that both the data pre-processing techniques improved the model performance, where the MA method had better result. The correlation between the input and output components of models were adjusted and enhanced by using the MA method.

2.1.5 Principal Component Analysis (PCA)

Pearson (1901) has introduced the Principal Component Analysis (PCA) and this method has been developed independently by Hotelling (1933). Recently, this method has been used widely for data analysis. The dimensionality of a data set with many interrelated variables can be reduced using PCA and the variation present was retained in the data set as much as possible.

Hu et al. (2007) adopted the PCA method in Rainfall-Runoff Neural Network (RRNN) model. The generalization performance of RRNN model will decrease if there are too many input neurons in RRNN model. The principal components can be significantly extracted from the lagged input

hydro meteorological data with incorporation of the PCA method. Hence, the method improved the prediction accuracy of the RRNN model.

2.1.6 Singular Spectrum Analysis (SSA)

SSA can be used to decompose the incoming signal into the sum of its independent components, in which the independent components are identified as a trend, oscillatory behaviour or noise (Rocco, 2013). Based on Golyandina et al. (2001), the SSA method is categorized into two stages which are decomposition and reconstruction stages. In the research study, the researchers established the Linear Recurrent Formulae (LRF) model that coupled with SSA technique and found that the hybrid models outperform the original modes, showing results with better performance.

SSA was performed as an efficient and effective data pre-processing technique in avoiding the effect of discontinuous or intermittent signals. It was coupled with different types of streamflow forecasting models (Lisi et al., 1995; Sivapragasam et al., 2001; Baratta et al., 2003; Ba et al., 2017). Lisi et al. (1995) implemented SSA to extract the important and significant components in the study of the incoming signal and used the Artificial Neural Networks (ANN) for prediction. Throughout the process, the first "p" significant components were summed up to reconstruct the original time series. Besides that, Sivapragasam et al. (2001) used SSA coupled with SVM for hydrological analysis. The hybrid model was shown to be superior to the original SVM

model as once again showed that SSA technique helped in improve the performance of hydrological model.

Zhang et al. (2011) used SSA cum the Auto Regressive Integrated Moving Average (ARIMA) for hydrological forecasting. The annual streamflow data are collected from Biliuhe Reservoir and Dahuofang Reservoir, China for runoff analysis. Another hybrid model which consists of SSA and LRF model are used for comparison purpose. In the research study, the SSA-ARIMA model gained the popularity and had a better performance compared to that of the SSA-LRF model. , The SSA-ARIMA hybrid model was superior to that of ARIMA single model itself. The SSA technique has significantly improved the runoff forecasting results.

2.1.7 Boosting Method

The Boosting method is able to enhance the prediction accuracy of weak predictors such as decision (regression) trees and artificial neural networks. Schapire (1990) has introduced the first boosting algorithm. The prediction accuracy of original datasets can be optimized by replicating datasets through fitting an additive model of base functions using Boosting method (Hancock et al., 2005). Chou et al. (2011) has also stated that the Boosting method can inherit the benefits of regression trees while overcoming their errors in obtaining the best possible solution in any given application.

Deng et al. (2005) had incorporated two famous ensemble methods (bagging and boosting) into the SVM model for nonlinear time series prediction. They stated that these two ensemble methods had successfully improved the performance of the single SVM model. Halil and Onur (2013) had adopted the ensemble techniques (bagging and boosting) to enhance the monthly streamflow prediction accuracy of streamflow forecasting model. The monthly streamflow time series can be forecasted using tree-based ensemble models. However, there has been little research regarding ensemble methods conducted previously, especially in the prediction of river water level.

2.1.8 Summary

Seven types of data pre-processing techniques were used for analysing data and denoising purpose as abovementioned, which are EMD, VMD, DWA which including Haar wavelet, Daubechies wavelet B-spline wavelet and Sym wavelet, MA, PCA, SSA and Boosting method. Some model performance measures had been implemented to evaluate the performances of the data pre-processing techniques. The statistical method will then be decided after the observed rainfall series have been assessed using following normality tests.

For comparison purposes, some researches regarding the seven types of data pre-processing techniques had been carried out. Wu et al. (2010) had used several data pre-processing techniques namely MA, PCA and SSA together with the Modular Artificial Neural Networks (MANN) model to

predict the rainfall time series. In the research study, the results indicated that SSA showed more noticeable advantages as compared to that of MA and PCA and the enhancement of model performance using SSA was more superior. Hence, SSA outperforms the MA and PCA methods (Wu et al., 2010).

Aman (2015) had enhanced the streamflow using the Least Square Support Vector Machine (LSSVM) model coupled with SSA and DWA. In their research, both data pre-processing techniques could enhance the performance of LSSVM model as compared to that of original model. However, performance evaluation done showed that the DWA technique could performed better than the SSA technique. The DWA approach had determined the input variables of model output in a more accurate way.

Zhu et al. (2016) proposed a streamflow estimation by using SVM coupled with different methods of times series decomposition. They found out that both DWA and EMD could enhance the streamflow forecasting. However, DWA outperforms the EMD in terms of performance evaluation. Furthermore, another research which was done by Amar and Guennoun (2012) had indicated that wavelets transform performed better as compared to that of EMD. They had done research of measurement of the U.S core inflation using wavelet transformation and EMD. The wavelets transformation has performed well in the case of more complicated trend as compared to that of EMD.

Salim (2016) had analysed and forecasted the economic and financial time series using VMD. He stated that EMD has been successfully used in

forecasting purpose. However, VMD, an alternative to EMD was proposed to perform more effectively and efficiently as compared to EMD. The upgraded version of VMD possessed a better performance in terms of MAE, MAPE and RMSE and this indicated that VMD performed better than EMD in decomposition approach.

Halil and Onur (2013) had used the ensemble methods to enhance the monthly streamflow accuracy. In this research study, ensemble methods (Bagging and boosting methods) had showed a remarkable prediction in monthly streamflow forecasting. The Boosting method was chosen as one of the data pre-processing techniques for this research study.

In summary, the VMD, DWA and Boosting methods obtained the good performance and accuracy as compared to that of EMD, MA, PCA and SSA methods as abovementioned. Throughout the years, there were quite several research studies regarding DWA applied in hydrological time series analysis as abovementioned. However, the ensemble method (Boosting method) is quite new in this field. Hence, Boosting method can be a great exploration on analysing and denoising data for river water level prediction. Therefore, the VMD and the Boosting method will be implemented for analysing data and denoising purpose prior to predicting with the respected river water level prediction model. Both the data pre-processing techniques have their own characteristics and properties for the data analysis. The selected data pre-processing techniques will be evaluated using the respective model performance measures to evaluate their accuracy and performance.

2.2 River Water Level Prediction Models

Malaysia is prone to natural disasters that are only associated with extreme high or low rainfall such as floods and droughts. Here, the river water level prediction model plays an important role in predicting the future river water level in order to aid mitigating efforts against the calamities mentioned. Several models had been reviewed to obtain a better river water level prediction method. The river water level prediction models chosen for this study are the Support Vector Regression (SVR), the Artificial Neural Network (ANN) and the Adaptive Neuro Fuzzy Inference System (ANFIS).

2.2.1 Support Vector Machine (SVM)

The SVM was effective for both classifications (Vapnik, 1995) and regression (Collobert and Bengio 2001; Drucker et al. 1996); for a noisy set of data. Till then, Kisi and Cimen (2011) stated that the SVM had been in use in different fields (Smola, 1996; Vapnik et al., 1997; Gao et al., 2001; Yoon et al., 2004; McNamara et al., 2005; Awad et al., 2007, Kaheil et al., 2008; Matías 2004). Khan and Coulibaly (2006) used the SVM to forecast future water levels and they found that the SVM showed a better performance compared to two other streamflow forecasting models for 3 to 12 months-lead predictions by evaluating using the performance measures.

As mentioned above, the SVM was originally used for classification procedure purposes, and then, recently another version of the SVM called the SVR, has become another application form of the SVM (Vapnik et al., 1997). The regression problems with SVM can solved using the SVR as formulated in Equation 2.5.

$$\hat{y} = \mathrm{f(x)} = \omega_i \,.\, \phi_i(x) + b \tag{2.5}$$

where ω_i is a weight vector, b is a bias, and ϕ_i denotes a nonlinear transfer function that maps the input vectors into a high-dimensional feature space in which theoretically a simple linear regression able to match with the complex nonlinear regression of the input space. It is also crucial to determine the parameters used for model training and validation. For the SVR, the regularization parameter, C and insensitive loss function, ε are the two main input parameters once the kernel function and data in training set are fixed.

For reconstruction of the SVM, the structural risk minimization (SRM) induction principle (Vapnik, 1995) is implemented instead of the empirical risk minimization (ERM). The SRM induction principle can reduce both the empirical risk and the model complexity, simultaneously, in turn improving the generalization ability of the SVMs. This can aid in preventing the possibilities of under-fitting or over-fitting of model, as shown in Figure 3.2.

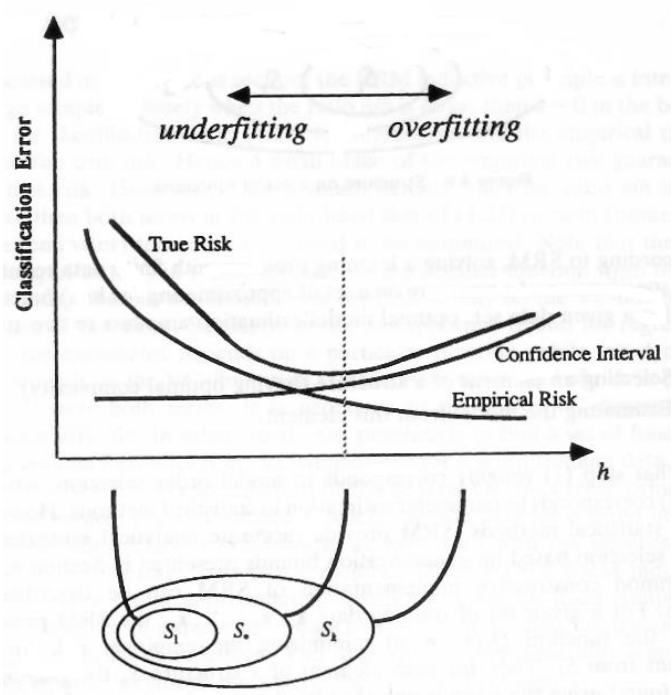


Figure 2.2: Structural Risk Minimization (SRM) Induction Principle (Cherkassky and Mulier, 1998)

Bray and Han (2004) used the SVR to produce an appropriate model structure for forecasting streamflow accurately. They stated that SVM model has been applied in many fields especially in classification tasks such as pattern recognition. Despite the SVM having gained a great success in the field of classification tasks, the SVM still produce an appropriate model structure for rainfall and runoff modelling. Yu et al. (2006) successfully predicted flood stages using SVR in Lan-Yang River, Taiwan. Employment of the SVR has been successfully used to solve the forecasting problems with hydrologic variables. Apart from that, Kisi and Cimen (2011) had used the Wavelet-Support Vector Regression (WSVR) conjunction model to forecast the monthly streamflow and showed that the DWA was able to enhance the performance of the SVR model.

Again, the SVR has been successfully applied in different problems and contributed widely to the hydrological field (Moradkhani et al., 2004; Yu et al., 2006; Lin et al., 2006; Wu et al., 2008; Lin et al., 2009; Chen et al., 2010; Yoon et al., 2011; Yousefi et al., 2015; Ghorbani et al., 2016; Zhu et al., 2016; Wang et al., 2013). Dibike et al. (2001) studied the use of the SVM for rainfall-runoff modelling and found that the SVM was able to provide quite an accurate prediction of streamflow. Liong and Sivapragasam (2002) used the SVR analysis for flood stage forecasting and they stated that the SVR is based on the principal of Structural Risk Minimization (SRM) as opposed to the principal of Empirical Risk Minimization (ERM). This statement was also supported by other researchers (Yu et al., 2006, Wu et al., 2012). The researchers had compared the SVR analysis with that of ANN and they stated the prediction accuracy of SVR analysis was much better than that from ANN and could overcome a lot of limitations of the ANN. Hence, the SVR appeared to be a very promising prediction tool in water level prediction.

However, the SVR model has its own limitations. Using the SVR model, the non-stationary of original hydrological data may differ over a range of scales (Cannas et al., 2006; Adamowski and Chan, 2011). Hu et al. (2013a) stated that the fluctuation and intrinsic complexity of time series data can be the problem to the assumption of results obtained. Another concern is that the design of the input vector is usually based on the experience, and in an arbitrary way, despite it being an important parameter for the time series prediction engine.

2.2.2 Artificial Neural Network (ANN)

The ANN possessed the capabilities to imitate the biological neural networks of human brain, by mapping the complex non-linear relationships and processes that are inherent among several influencing variables through a model structure (Haykin, 1999). In short, ANN is a form of a non-linear regression model that uses a set of weights to perform the input-output mapping, through several neurons in a hidden layer. Determination of the number of hidden layers, the number of neurons in the connection and the type of training algorithm can be crucial to the model performance. Thus, these need to be pre-determined precisely before running the whole ANN model. The ANN has been shown to have contributed significantly to the hydrological field (Tingsanchali and Gautam, 2000; Chen et al., 2006).

As stated by Abrahart et al. (2012), ANN has been used widely in time series predictions and it can be applied in noisy environments and dense interconnections of simple computational elements. Solomatine and Ostfeld (2008) stated that the ANN has been widely used in river basin management and other related problems. It has been used widely in several streamflow forecasting events (Tiwari and Chatterjee, 2009; Feng 2013; Awchi 2014). The characteristics of ANN, which is non-linear, can even solve complicated problems more accurately than linear techniques do (Cameron et al., 1999).

ANN can also adjust their behaviour according to the environment in order to produce consistent responses; the network is able to mimic the

patterns directly from their instances even though the data may be distorted and incomplete. It can even abstract the important characteristics that contain irrelevant data in inputs. ANN is able to act faster than other alternative methods due to its highly parallel characteristic which makes it able to execute operations simultaneously (Cameron et al., 1999).

However, ANN also have some drawbacks for certain applications. There is no learnable function and insufficient size of data set in ANN, so it may fail to produce a promising solution. Besides that, a trial-and-error problem needs to be applied as the optimum network geometry of ANN is considered as problem dependent. ANN is unable to respond accurately if there were major changes as they are calibrated on historical data and an assumption has been made that the relationship learned will be applied in future (Cameron et al., 1999). Furthermore, ANN is not able to extrapolate beyond the training data range (Flood and Kartam, 1994; Minns and Hall, 1996; Tokar and Johnson, 1999). This is due to the characteristics of ANN model being unable to understand the underlying physical processes (Bowden et al., 2012).

2.2.3 Adaptive Neuro Fuzzy Inference System (ANFIS)

The characteristics of the decision making of ANFIS are based on a set of logical rules, so-called fuzzy logic and fuzzy set theory (Zadeh, 1965). In the past, fuzzy logic-based modelling approaches had been introduced into the

hydrological modelling and water resources field (See and Openshaw, 2000; Xiong et al., 2001; Shrestha et al., 1996; Fontane et al., 1997; Yu et al., 2000; Chang and Chen, 2001; Lohani et al., 2006, 2007a, 2007b; Badrzadeh et al., 2017). The ANFIS model was proposed by combining both the ANN and fuzzy logic, and this newly proposed model has been applied widely in many researches. For ANFIS model, the type of membership function and the determination of number of membership functions for each input are needed.

According to Jang (1992, 1993), ANFIS used the learning ability of ANN for the purpose to express the input-output relationship. The fuzzy rules are constructed by determining the input structure. The results of the system obtained are from the reasoning capability of fuzzy logic (Firat and Gungor, 2008). These methods are used to adapt themselves and learn to do better when it comes to a new environment (Jang, 1992, 1993).

Through the years, ANFIS had been widely deployed in the hydrological and water resources field. Chang and Chang (2001) reported that the ANFIS model helped to provide a higher efficiency of reservoir operation as compared to the classical models based on the rule curve. The researchers claimed that once information is sufficient to construct the fuzzy rules using ANFIS, it will perform more effective than the rule-curve-based model.

El-Shafie et al. (2007) had adopted ANFIS model to forecast the inflow of river on a monthly time series basis. The ANFIS model was able to divide the input space into fuzzy sub-spaces and maps the output using a set of

linear function. In their research study, the ANFIS model was used to compare with that of the ANN model and they found that the ANFIS is able to enhance forecasting accuracy especially in extreme inflow events.

Shu and Ouarda (2008) had applied the ANFIS model for estimating the regional flood frequency at ungauged sites. Therein, an ANFIS model was used to compare with that of ANN model as well as the Nonlinear Regression (NLR) approach and the Nonlinear Regression with Regionalization approach (NLR-R). The results showed that the ANFIS possessed a better generalization capability than those of the ANN, NLR and NLR-R approaches.

A neuro-fuzzy network-based inflow forecasting model has also been developed for Brazilian hydroelectric plants (Ballini et al., 2001). Ballini et al., (2001) used a class of neuro-fuzzy network that was applied to the problem of seasonal streamflow forecasting as well as the monthly streamflow forecasting. The ANFIS model was compared with the ANN model and autoregressive-moving average model and showed that the ANFIS model performed much better than the two models in streamflow forecasting.

2.2.4 Summary

There are three potential river water level prediction models that can be used for river water level prediction as previously mentioned, and they are: the SVR, the ANN and the ANFIS models. Some methods of scrutinizing model performance measures had been discussed to evaluate the effectiveness

of the data pre-processing techniques. The engaging statistical methods will be decided after the observed rainfall series have been assessed using established normality tests.

Some researches regarding the three specific river water level prediction models had been studied for comparison purpose. Lohani et al. (2012) had developed autoregressive (AR), ANN and ANFIS in hydrological time series modelling and forecasting. AR models, which is one of the types of stochastic models was not taken for comparison. By comparing the ANFIS and ANN models, the coefficient of correlation (R) and Nash-Sutcliffe Efficiency Coefficient (NSEC) values of ANFIS model were higher than that of ANN model as well as the Root-Mean-Square Error (RMSE) value was lower than that of ANN model. It was proven that ANFIS model does better than the ANN model in their study.

However, another research study carried out by He et al. (2014) compared the ANN, ANFIS and SVM (using SVR) for river flow forecasting. As compared to ANN and ANFIS, the SVM model was shown to have performed better through their evaluating using performance measures and had achieved a better accuracy in the forecasting of river flow.

In a nutshell, data pre-processing techniques which are specifically here for this current study, the VMD and the Boosting method are used for denoising and analysing observed rainfall series before the pre-processed data input are used for predicting using the water level prediction model, the SVR

model. The resulting river water level output data from the SVR model will be further evaluated using the respective model performance measures that have already been delineated. Daily, monthly, seasonal and annually time series analysis of observed rainfall series will be implemented.

2.3 Model Assessment

2.3.1 Homogeneity Test

For river water level prediction, the homogeneity of the observed rainfall series collected from different rainfall stations is extremely important. Several methods have been employed by different researchers to assess the homogeneity of observed rainfall series collected, and they are the standard normal homogeneity test, the Buishand range test, the Pettitt test and the Von Neumann ratio test (Wijingaard et al., 2003; Sahin and Cigizoglu, 2010; Kang and Yusof, 2012).

These four absolute tests were used to assess the homogeneity of European daily temperature and precipitation series (Wijingaard et al., 2003). It was found that relocation of stations as well as the way of observing and measuring techniques can affect the homogeneity of the series. The homogeneity of the rainfall series in Peninsular Malaysia had been investigated by Kang and Yusof (2012). In their research study, 33 stations over the Damansara, Johor and Kelantan areas were selected, with less than 10%

missing data. All four tests were put to the annual mean, maximum and median of each of the rainfall station, tested at 95% level of significance. With the classification of homogeneity of the rainfall series, the annual mean and annual maximum series of all stations were found to be 'useful'. However, 4 stations were considered as 'doubtful' and 1 station was considered as 'suspect' for the annual median series. It was observed that the percentage of missing data will not affect the homogeneity of the rainfall series.

2.3.2 Normality Test

Several methods were adopted to assess the normality of observed rainfall series (input data set). They are mainly based on the graphical and statistical methods. For graphical methods, it is less accurate and not useful for small sample size. Instead, statistical methods can provide a more accurate and precise results as actual probabilities are calculated.

Nornadiah and Yap (2011) compared the power of four formal tests of normality which are; the Shapiro-Wilk (SW) test, Kolmogorov-Smirnov (KS) test, Lilliefors (LF) test and the Anderson-Darling (AD) test. In their research study, the normality statistical tests were compared with the respective critical values to obtain the power of each test. From the results obtained, it was shown that the most powerful normality test was from the SW test, followed by the AD test, LF test and then, lastly KS test. The findings were similar to

the findings of Mendes and Pala (2003) and Keskin (2006). All four tests were not performed well for small sample size.

2.3.3 Model Performance Measures

2.3.3.1 Non-Parametric Statistics

The non-parametric inferential statistics of Bootstrapping, Bias Corrected and the Accelerated (BCa) procedure was used to assess the model prediction errors. The optimum model prediction error's range, standard deviation and confidence interval range were selected using numerical analysis algorithm (Fattorini, 1999; Mordecai, 2003; Jon, 2004; Jorge and Stephen, 2006; Ruszczyński, 2006) and to assess two null hypotheses. Thus, the SPSS software was used in this study. The BCa procedure with 2,000 samplings in non-parametric Bootstrapping technique (Efron and Tibshirani, 1994; Efron, 2010) was conducted at 95% and 99% of confidence interval to derive the key parameters of the said model performance measures.

2.3.3.2 Parametric Statistics

The procedures of parametric statistics rely on the assumptions about the shape of the distribution, which are assumed to be normal distribution, and about the parameters, which are assumed to be means and standard deviations.

Both, parametric and non-parametric statistics, are commonly applied in order to determine the trends (Hamed and Rao, 1998). Examples of parametric tests are linear regression, Ordinary Least Squares (OLS) and student t-test. According to Corder and Foreman (2009), parametric methods are more powerful in nature as they have more assumptions associated than those of non-parametric methods. However, time series is assumed to be normally distributed and independent for parametric statistics. The parametric methods are not often robust when the input data does not satisfy the assumption of parametric methods to be normally distributed, the methods can be greatly misleading (Kundzewicz and Robson, 2004).

CHAPTER 3

METHODOLOGY

3.1 Flowchart

The flowchart for the research study is illustrated in Figure 3.1. The river water level prediction began with the collection of observed rainfall series and river water level from the DID, Malaysia. The collected observed rainfall series were sorted using the homogeneity test to produce a more homogenous observed rainfall series. The normality test was conducted on the observed rainfall series (input) to check whether the input data are normally distributed or non-normally distributed. The selected data pre-processing techniques are the VMD method, the Boosting method and the combined Boosting-VMD method whereas the chosen river water level prediction model is the SVR model.

Before the SVR procedure, the observed rainfall series were processed using different data pre-processing techniques. Then, the four models' river water levels were predicted based on the processed rainfall series using the Ori-SVR (observed rainfall series), the VMD-SVR (VMD method) series, the B-SVR (Boosting method) series and the B-VMD-SVR (Booting-VMD method) series. Last but not least, all the predicted river water levels from the SVR were assessed with the observed river water levels using the proven performance measures. The resulting predicted river water levels from SVR

were assessed against the observed river water levels, using non-parametric statistics (Model prediction errors' range and standard deviation) and parametric statistics (Bias, Root-Mean-Square Error, Mean Absolute Percentage Error, Nash-Sutcliffe Efficiency Coefficient and Mean Absolute Error).

Besides that, the normality test was once again conducted on the residuals of predicted river water levels (output) to decide on another performance measure to be employed. If the residuals of predicted river water levels are normally distributed, the model prediction errors' mean confidence interval range will be used; but if the residuals of predicted river water levels are non-normally distributed, then the model prediction errors' median confidence interval range will be used.

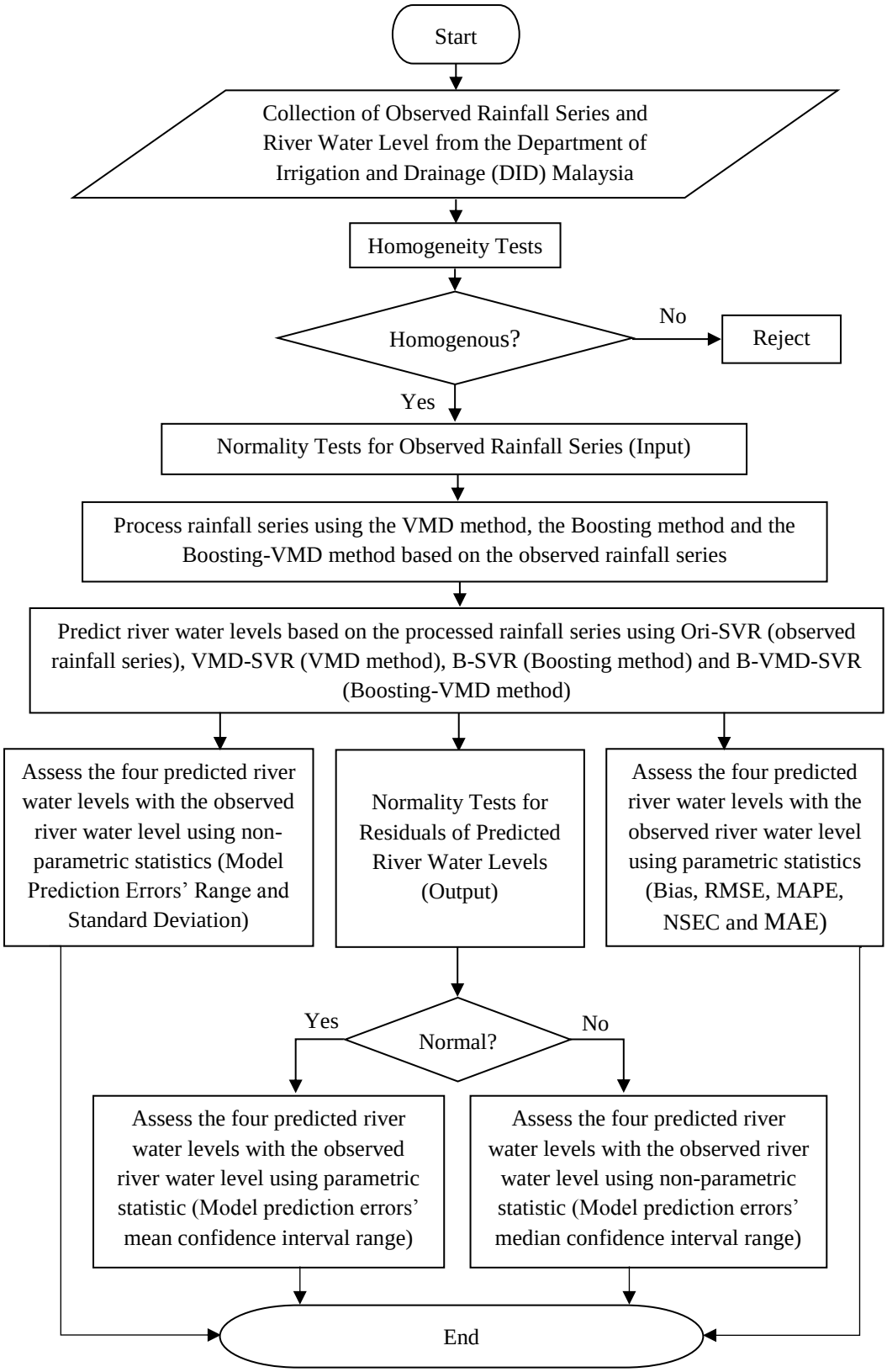


Figure 3.1: Overall Flowchart of the Research Procedures

3.2 Study Area

The Dungun River is the main river in the basin which is located between latitudes 4º36'10"N to 4º53'02"N and between longitudes 103º07'25"E to 103º25'50"E, and the Dungun district is one of the seven districts in the state of Terengganu, Malaysia. The area of the Dungun River Basin is approximately 1,800 km^2. Sungai Dungun flows for about 110 kilometres, flowing from Pasir Raja to Kuala Jengai, Jerangau, and finally discharging at Kuala Dungun. The river width ranges from 50 to 300 meters. The Dungun River receives runoffs from its main tributaries such as the Telemboh, Lok, Kelmin, Loh and Perlis rivers before discharging into South China Sea. The location of Dungun River Basin in Peninsular Malaysia and the three stations within the basin is shown in Figure 3.2.

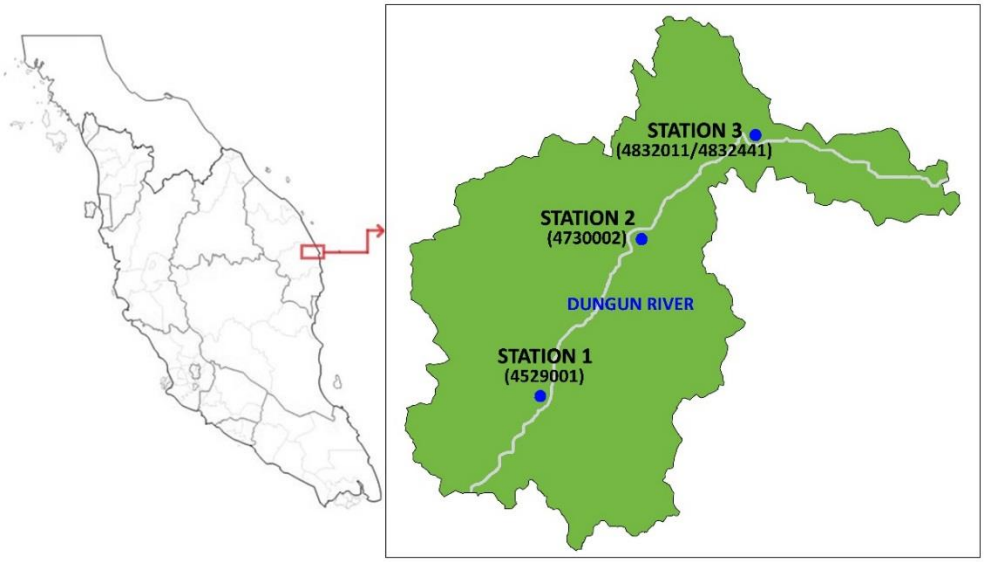


Figure 3.2: Location of the Dungun River Basin and the three rainfall stations

3.3 Data Acquisition

All the rainfall series and river water level were obtained from the DID, Malaysia. There are three rainfall stations within the catchment (Figure 3.2) each having a continuous good recording period. The hourly rainfall series at each station was accumulated to calculate the total daily, total monthly and total seasonal data series. A mean rainfall series from 1996 to 2016 (20 years) had been determined. Station 3 at downstream was selected as the only river water level station (4832441) within the basin. The details of the stations are clearly shown in Table 3.1.

Table 3.1: The name, station code and coordinates of the three stations

Station No.	Rainfall Station Code	Water Level Station Code	Station Name	Latitude	Longitude
1	4529001	-	Rumah Pam Paya Kempian at Pasir Raja	4°34'05"N	102°58'45"E
2	4730002	-	Kg. Surau at Kuala Jengai	4°44'05"N	103°05'15"E
3	4832011	4832441	Jambatan Jerangau at Terengganu	4°50'35"N	103°12'15"E

A mean rainfall data series within the catchment (3 rainfall stations) was computed using the Thiessen Polygon method. The area weightage for the respective 3 rainfall stations are shown in Figure 3.3 and Table 3.2.

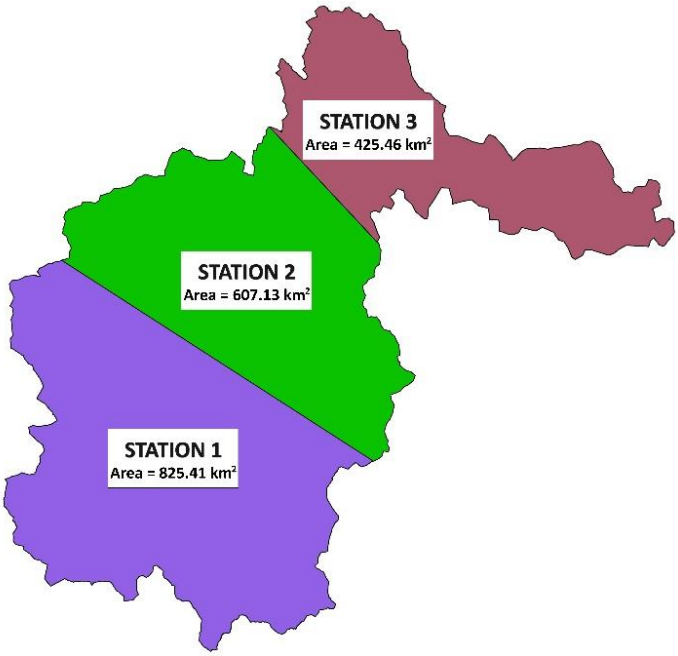


Figure 3.3: Map of area weighted of three rainfall stations using Thiessen Polygon method

Table 3.2: Area weighted of three rainfall stations using Thiessen Polygon method

Station No.	Rainfall Station Code	Station Name	Area, A_i (km^2)	Area Weighted (A_i/A_T)
1	4529001	Rumah Pam Paya Kempian at Pasir Raja	825.41	0.4442
2	4730002	Kg. Surau at Kuala Jengai	607.13	0.3268
3	4832011	Jambatan Jerangau at Terengganu	425.46	0.2290
		Total (A_T) =	1858.00	1.0000

3.4 Homogeneity Test

Observed rainfall series used for river water level prediction purpose are normally affected by weather and climate only. However, there are other factors such as location of rainfall station, environment of rainfall station;

different observing and practices might also affect the homogeneity of the recorded rainfall time series. Therefore, homogeneity tests are implemented for the observed rainfall series of every rainfall stations before further investigating, to obtain a more homogenous observed rainfall series. There are four homogeneity tests which are applied to the observed rainfall series, namely the Standard Normal Homogeneity Test (SNHT), the Buishand Range test (BR), the Pettitt test (PeT) and the Von Neumann Ratio test (VNR). All four absolute tests are applied to the observed rainfall time series concurrently. The critical values for the homogeneity tests are stated in Table 3.3

The criteria that was used in this research study is similar to the one proposed by Wijingaard et al. (2003) and which was used to determine the homogeneity of data series. The classification of homogeneity of observed rainfall series is stated as follow and all the four tests were tested at 95% significance level:

i. If one or none of the tests reject the null hypothesis: 'Useful',

ii. If any two tests reject the null hypothesis: 'Doubtful', and

iii. If three or all tests reject the null hypothesis: 'Suspect'.

Table 3.3: The 5% Critical Values for the homogeneity tests (Wijingaard et al., 2003)

n	T_0 (SNHT)	$R/\sqrt{n}$ (BR)	X_E (PeT)	N (VNR)
20	6.95	1.43	57	1.30
30	7.65	1.50	107	1.42
40	8.10	1.53	167	1.49
50	8.45	1.55	235	1.54

3.4.1 Standard Normal Homogeneity Test (SNHT)

The SNHT can detect breaks at the starting and ending of a series. The annual series are assumed to be normal. The statistic of the test can be computed as shown in Equations 3.1 and 3.2.

$$T_k = k\overline{z_1^2} + (n-k)\overline{z_2^2}\,, k = 1,2,\dots,n \tag{3.1}$$

where,

$$\overline{z_1} = \frac{1}{k}\sum_{i=1}^{k}\frac{(Y_i - \bar{Y})}{s} \text{ and } \overline{z_2} = \frac{1}{n-k}\sum_{i=k+1}^{n}\frac{(Y_i - \bar{Y})}{s} \tag{3.2}$$

Y_i = The annual series to be tested

i = The year from 1 to n

$\bar{Y}$ = The mean

s = The standard deviation

If there is a break in the series at year K, then T_k comes to a maximum near the year $k = K$. Thus, the test statistic is stated in Equation 3.3.

$$T_0 = \max_{1 \le k \le n} T_k \tag{3.3}$$

If T_0 is larger than the critical value, the null hypothesis is rejected. The critical values are contingent on the sample size n which stated in Table 3.3.

3.4.2 The Buishand Range Test (BR)

The BR test is more effective in detecting the break in the middle of a series. However, the Y_i values are assumed to be normally distributed. The adjusted partial sum is stated in Equation 3.4.

$$S_0^* = 0 \text{ and } S_k^* = \sum_{i=1}^{k} (Y_i - \bar{Y}), k = 1,2, \dots, n \tag{3.4}$$

If the series is homogeneous, the values of S_k^* will vary around 0. On the other hand, if there is a break in year *K*, the S_k^* achieves either a maximum or a minimum value. The 'rescaled adjusted range' R can be attained using the formula stated in Equation 3.5.

$$R = \frac{\left(\max_{0 \le k \le n} S_k^* - \min_{0 \le k \le n} S_k^*\right)}{s} \tag{3.5}$$

The R is used to calculate the significance of the mean deviation. The critical value of $\frac{R}{\sqrt{n}}$ stated in Table 3.3 is used to determine whether to accept or to reject the null hypothesis.

3.4.3 The Pettitt Test (PeT)

The PeT possesses similar capabilities to the BR test. It is more competent in detecting the break in the middle of a series. However, since the PeT is based on the rank, it does not involve any normality assumption of the data. Besides that, the non-parametric rank-based PeT is less sensitive to the outliers as compared to other tests. The statistic of the PeT can be assessed using formula stated in Equation 3.6.

$$X_k = 2\sum_{i=1}^{k} r_i - k(n+1)\,, k = 1,2,\dots,n \tag{3.6}$$

where r_i is the rank of Y_i.

X_k shows maximum or minimum if the break happens near the year k = E, as shown in Equation 3.7.

$$X_E = \max_{1\le k\le n} |X_k| \tag{3.7}$$

The critical values of X_E are stated in Table 3.3.

3.4.4 The Von Neumann Ratio Test (VNR)

The VMR test examines whether a series is randomly distributed instead of detecting the breaks in the series as other methods do. The test statistics of VMR test is shown in Equation 3.8.

$$N = \frac{\sum_{i=1}^{n-1}(Y_i - Y_{i+1})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3.8}$$

The projected value *N* is 2 when the series is homogeneous. If there is a break in the series, *N* will tend to be lower. When the series has instant variations in the mean, the value of *N* will increase above 2. The critical values of *N* are stated in Table 3.3.

3.5 Normality Test

For normal distribution of data, parametric statistic tests had been implemented to evaluate the observed rainfall series. If the observed rainfall series are non-normally distributed, then the non-parametric statistic tests are preferred. There are several tests that are used to assess the normality of observed rainfall series. For the normal distributed data, the observed rainfall series must pass all the stated normality tests.

In frequentist statistics statistical hypothesis testing, the observed rainfall series are tested against the null hypothesis that it is normally

distributed. The normality that calculated using statistical methods are shown as follow:

i. Significance Level > 95 % or $\alpha > 0.05$: Observed rainfall series are normal,

ii. Significance Level < 95 % or $\alpha < 0.05$: Observed rainfall series are not normal.

To pass the normality test, it is necessary to pass all the frequentist tests. If the results of any one of the frequentist tests are showing not normal, then the observed rainfall series is considered not normally distributed and non-parametric tests will then be implemented. The frequentist tests are the Shapiro-Wilk Test (SW), the Kolmogorov-Smirnov test (KS), the Jarque-Bera Test (JB) and the Anderson-Darling Test (AD).

3.5.1 Shapiro-Wilk Test (SW)

The SW test was proposed in 1965. It was originally restricted for a sample size of less than 50 as it works best for this sample size. For the SW test, the *W* statistics is calculated. The null hypothesis is that a random sample (x_1, x_2, x_3, … , x_n) is normally distributed. The small value of *W* is evidence that the data are not normally distributed. The *W* statistics is defined in Equation 3.9:

$$W = \frac{\sum_{i=1}^{n} a_i x_i{}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3.9)$$

where,

x_i = The i[th] order statistic

$\bar{x}$ = The sample mean

$a_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$

and $m = (m_1, \ldots, m_n)^T$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matric of those order statistics.

3.5.2 Kolmogorov-Smirnov Test (KS)

The KS test works best for a sample size of more than 50, totally opposite to that of the Shapiro-Wilk test above. The null hypothesis is that a random sample ($x_1, x_2, x_3, \ldots, x_n$) is normally distributed. Conover (1999) defined the test statistics, T proposed by Kolmogorov (1933) is shown in Equation 3.10.

$$T = sup_x |F^*(x) - F_n(x)| \quad (3.10)$$

where,

'sup' = Supremum, the greatest value

$F^*(x)$ = The hypothesized distribution function

$F_n(x)$ = The EDF estimated based on the random sample

In the KS test, the $F^*(x)$ is indicated the normal distribution with mean, μ and standard deviation, σ. The test interpretation of the KS test is as follow:

Null Hypothesis, H_o: $F(x) = F^*(x)$ for all x from $-\infty$ to ∞ (The data follow a specified distribution)

Alternative Hypothesis, H_a: $F(x) \neq F^*(x)$ for at least one value of x (The data do not follow the specified distribution)

If T exceed the 1-α quantile as given by the table of quantiles for the KS test statistics, then the H_o is rejected at the level of significance, α.

3.5.3 Jarque-Bera Test (JB)

The JB test is a goodness-of-fit test of whether the sample data have the skewness and kurtosis matching a normal distribution. It is usually used for large data sets, because some other normality tests are not reliable when the sample size is large. For example, the SW test is not reliable with sample size of more than 2,000 and it only works best for sample size of less than 50. This test was named after Carlos Jarque and Anil K. Bera. The statistics of the JB test is defined in Equations 3.11, 3.12 and 3.13.

$$JB = n(\frac{(k_3)^2}{6} + \frac{(k_4)^2}{24}) \quad (3.11)$$

$$k_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3} \quad (3.12)$$

$$k_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3 \quad (3.13)$$

where,

x is each observation

n is the sample size

s is the standard deviation

k_3 is skewness

k_4 is kurtosis

The JB statistics can be compared to the critical values of the Chi-Squared (χ^2) distribution to determine the critical value at an alpha level of 0.05.

3.5.4 Anderson-Darling Test (AD)

The AD test is further modified from the Cramer-Von Mises (CVM) test. Farrel and Stewart (2006) stated that the AD test gives more weight to the tails of the distribution. According to Arshad et al. (2003), the AD test is even more powerful than that of EDF tests.

The statistics for this test proposed by Anderson and Darling (1954) is defined in Equation 3.14.

$$W_n^2 = n\int_{-\infty}^{\infty}[F_n(x) - F^*(x)]^2\,\varphi\left(F^*(X)\right)dF^*(x) \qquad (3.14)$$

where,

φ is a nonnegative weight function that calculated through formula stated in Equation 3.15.

$$\varphi = [F^*(x)(1 - F^*(x))]^{-1} \qquad (3.15)$$

Arshad *et al.* (2003) had applied the formula as shown in Equation 3.16, to simplify the statistics computation.

$$W_n^2 = -n - \frac{1}{n}\sum(2i-1)\{logF^*(X_i) + \log(1 - F^*(X_{n+1-i})\} \qquad (3.16)$$

where,

$F^*(x_i)$ = The cumulative distribution function of the specified distribution

$x_i s$ = The ordered data

n = The sample size

3.6 Data Pre-Processing Techniques

3.6.1 Variational Mode Decomposition (VMD)

The VMD, proposed by Dragomiretskiy and Zosso (2014) decomposes the incoming signal into *k* discrete number of sub-signals (modes), u_k, where each mode possesses the limited bandwidth in the spectral domain. The modes, u_k have specific sparsity properties while reproducing the input. Therefore, each of the mode *k* is required to be compacted the most around a centre pulsation, ω_k determined along the decomposition process. The VMD method will find out the central frequencies and IMFs located at the centre of the frequencies concurrently using an optimization method namely the ADMM. The original formulation of the optimization problem is continuous in the time domain. The algorithm of the VMD which assess the bandwidth of a one-dimension signal is stated as follows:

i. For each mode, u_k, the associated analytic signal was computed by means of the Hilbert transform to obtain the unilateral frequency spectrum.
ii. For each mode, the mode's frequency spectrum was shifted to baseband by mixing with an exponential tuned to the respective estimated centre frequency.
iii. The bandwidth was estimated through the Gaussian smoothness of the demodulated signal.

The constrained variational problem which was given by Dragomiretskiy and Zosso (2014) is stated in Equations 3.17 and 3.18.

$$\min_{u_k\,,\,\omega_k}\left\{\sum_k \left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\} \tag{3.17}$$

subject to,

$$\sum_k u_k = f \tag{3.18}$$

where,

f = The signal

u = The mode

ω = The frequency

δ = The Dirac distribution

t = Time script

k = Number of modes

$*$ = Convolution

The mode u with high-order k represents the low frequency components.

3.6.2 Boosting Method

The Boosting method is an ensemble method which help to improve the accuracy of any given learning algorithm (Schapire 1990; Freund and Schapire 1996). The central idea of the boosting method is to create a

sequence of models so that each subsequent model concentrates on the training cases that were not well predicted by the previous one (Freund and Schapire 1996). It assists in boosting "weak" learning algorithms, which perform just slightly better than random guessing, into a relatively "strong" learning algorithm. The base learning algorithm generates a new weak prediction rule, and after many rounds of generation, the boosting algorithm combines all the so-called weak rules into a single prediction rule and this will result in higher accuracy than that of any one of the individual weak rules.

For the Boosting method, the 'fitensemble' function in the MATLAB was used to boost the rainfall series. As suggested by Friedman (1999), the Least Squares Boosting (LS-Boost) fits for the regression of time series. Hence, in MATLAB, the "LSBoost" with shrinkage were used to combine weak learners and generate a more accurate ensemble (Cordiner 2009). After all these, the boosted rainfall series were imported to the 'fitsvm' function in MATLAB for model training and validation using the SVR model.

3.6.3 Boosting-VMD Method

The Boosting-VMD method is a hybrid of the VMD and the Boosting methods. Processed rainfall series using the VMD method will be further processed and decomposed using the Boosting method. It will be a double decomposition method and the methodologies are similar to the steps stated a forehand.

3.7 River Water Level Prediction Model – Support Vector Regression (SVR)

The SVM which was proposed by Vapnik (1995), is known as a classification and was extended to include regression. Thus, the SVR is derived from the SVM and is to solve the regression problems with the SVM. The regression function of the SVM that relates the input vector x to the output $\hat{y}$ is formulated as stated in Equation 3.19.

$$\hat{y} = f(x) = \omega^T . \phi(x) + b \tag{3.19}$$

where,

ω^T = A weight vector

b = A bias

ϕ = A nonlinear transfer function

According to the SRM induction principle, ω_i and b are estimated by minimizing the structural risk function stated in Equations 3.20 and 3.21:

$$\mathrm{R} = \frac{1}{2}\omega^T\omega + C\sum_{i-1}^{N_d} L_\varepsilon(\hat{y}_i) \tag{3.20}$$

where C represents the regularization parameter that weighing between the model complexity and its empirical error and Vapnik's ε-insensitive loss function L_ε is defined as:

$$fL_{\varepsilon}(\hat{y}) = |y - \hat{y}|_{\varepsilon} = \begin{cases} 0 & for\ |y - \hat{y}| < \varepsilon \\ |y - \hat{y}| - \varepsilon & for\ |y - \hat{y}| \geq \varepsilon \end{cases} \quad (3.21)$$

Vapnik (1995) expressed the SVR problems through the optimization problem. The optimization problems are normally worked out in its dual form using Lagrange multipliers, where α and α' are dual Lagrange multipliers. The optimal Lagrange multipliers α^* are solved by the standard quadratic programming algorithm. The SVR function can then be rewritten as shown in Equations 3.22 and 3.23.

$$f(\mathrm{x}) = \sum_{i-1}^{N_d} \alpha_i^* K(x_i, x) + b \quad (3.22)$$

where the kernel function $K(x_i, x)$ is defined as:

$$K(x_i, x) = \emptyset(x_i)^T \emptyset(x) \quad (3.23)$$

In the case of nonlinear regression, an SVM uses Radial Basis Function, RBF kernel due to its ability in parameters tuning to cope with different pattern analysis (Kecman 2001). RBF is implemented for the kernel function as stated in Equation 3.24.

$$K(x_i, x) = exp\left(-\frac{1}{n_x}|x_i - x|^2\right) \quad (3.24)$$

where n_x is the number of components in input vector x.

Some of the solved Lagrange multipliers ($\alpha - \alpha'$) are zero and they must be excluded from the regression functions. After resolving the non-zero Lagrange multipliers and the corresponding input vectors of the training data, which are called as support vectors, into the regression function, the final regression function can then be rewritten as shown in Equation 3.25.

$$f(\mathrm{x}) = \sum_{k-1}^{N_{sv}} \alpha_k K(x_k, x) + b \tag{3.25}$$

where,

x_k = The *k*th support vector

N_{sv} = Number of the support vectors

In MATLAB, the 'fitrsvm' function was used for model training and validation. Rainfall series from November 1996 to February 2014 (90%) were for model training and November 2014 to February 2016 (10%) were used for validation thereafter.

3.8 Model Performance Measures

There are several model performance measures which can be used to evaluate the model performances. For this research, the models' performance was evaluated using both the non-parametric and parametric statistics.

3.8.1 Non-Parametric Statistics - Residual Modelling Statistics

Non-parametric, residual modelling statistics was adopted. Non-parametric inferential statistics of Bootstrapping, Bias Corrected and Accelerated (BCa) procedure were adopted to assess model prediction errors. The model prediction error's range, standard deviation and confidence interval range were used to assess the different predictive model's prediction error. Finally, the best model was determined based on the smallest range and standard deviation's error (in magnitude), as well as the narrowest prediction error's confidence interval range. A negative value indicated that the model tends to under predict while a positive value indicate that the model tends to over predict. All the residual modelling statistics were analysed and processed with the SPSS software.

3.8.2 Parametric Statistics

For parametric statistics, five standard performance measures were chosen, and they are the Bias Method, the Root-Mean-Square Error (RMSE), the Mean Absolute Percentage Error (MAPE), the Nash-Sutcliffe Efficiency Coefficient (NSEC) and the Mean Absolute Error (MAE). Finally, the best model was determined based on the lowest Bias (in magnitude), and the RMSE, MAPE, MAE and the highest NSEC.

3.8.2.1 Bias

The Bias is a measurement to check the differences between the predicted water levels and the observed water level, as shown in Equation 3.26. It can be used to check whether the model is under or over predicting. Therefore, positive values of small magnitudes are preferred (over predictions are preferable, on the safe side).

$$Bias = \sum_{i=1}^{n} \frac{Q_i' - Q_i}{n} \tag{3.26}$$

where,

n = The number of observations

Q_i = The observed water levels

Q_i' = The predicted water levels

3.8.2.2 Root-Mean-Square Error (RMSE)

The RMSE is the square root of the mean / average of the squared of all the errors. The RMSE is a very commonly used measurement and it produces an exceptionally general-purpose error metric for numerical predictions. However, the RMSE amplifies and produces a larger error as compared to the MAPE. This may result in misleading conclusions. The RMSE is defined in Equation 3.27.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - Q_i')^2} \quad (3.27)$$

where,

n = The number of observations

Q_i = The observed water levels

Q_i' = The predicted water levels

3.8.2.3 Mean Absolute Percentage Error (MAPE)

The MAPE which also known as the Mean Absolute Percentage Deviation (MAPD), is used to calculate the prediction precision of a forecasting method in statistics. The MAPE can enhance the error measurement as compared to that of the RMSE since it does not magnify large errors as does the RMSE. However, there are some drawbacks for the measurement of the MAPE. The MAPE will go to extreme when the actual observed value is small (approximately zero) as the denominator of the equation is the actual observed value, Q_i. The MAPE is defined in Equation 3.28.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Q_i - Q_i'}{Q_i}\right| x\ 100\ \% \quad (3.28)$$

where,

n = The number of observations

Q_i = The observed water levels

Q_i' = The predicted water levels

Lewis (1982) had acquired a scale of judgement for the accuracy of the forecasted model using the MAPE. The scale of judgement for the MAPE accordingly is as shown in Table 3.4.

Table 3.4: The Scale of Judgement for Mean Absolute Percentage Error (Lewis, 1982)

Mean Absolute Percentage Error (MAPE)	Judgement of Forecast Accuracy
Less than 10 %	Highly accurate
11 % to 20 %	Good Forecast
21 % to 50 %	Reasonable Forecast
51 % or more	Inaccurate Forecast

3.8.2.4 Mean Absolute Error (MAE)

The MAE measures the differences between the predicted water levels and the observed water levels as similar to the Bias, but absolute values of the difference will be taken for the MAE. The lower the value of the MAE, the better the performance of the model. The MAE is defined in Equation 3.29.

$$MAE = \sum_{i=1}^{n} \frac{|Q_i' - Q_i|}{n} \tag{3.291}$$

where,

n = The number of observations

Q_i = The observed water levels

Q_i' = The predicted water levels

3.8.2.5 Nash-Sutcliffe Efficiency Coefficient (NSEC)

The NSEC is used to assess the predictive power of hydrological models. It is a normalized statistic that determine the relative magnitude of the residual variance as compared to that of the measured data variance (Nash and Sutcliffe, 1970). The NSEC is defined in Equation 3.30.

$$NSEC = 1 - \frac{\sum_{i=1}^{n}(Q_i - Q_i')^2}{\sum_{i=1}^{n}(Q_i - \overline{Q_\iota})^2} \quad (3.30)$$

where,

n = The number of observations

Q_i = The observed water levels

Q_i' = The predicted water levels

$\overline{Q_\iota}$ = The mean for observed water levels

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Homogeneity Tests

The homogeneity tests were implemented on the observed rainfall series from the three rainfall stations to obtain a more homogenous observed rainfall series. If the significance level, $\alpha > 0.05$, it means that the data are homogenous. Based on the results from the homogeneity tests as shown in Table 4.1, all the significance level of four tests of station 1 and station 2 for the months of Jan, Feb, Nov and Dec are more than 0.05. With these results, the homogeneity of all months are categorized as useful. However, for station 3, only in one of the months that had failed the PeT test, which is at Feb ($\alpha = 0.045$). And despite it was shown that only in one test that had the homogeneity rejected, that is for the month of Feb at station 3, it was however accepted that the homogeneity of all months at station 3 were categorized as useful on an overall basis, if nothing else. The results of homogeneity tests for three rainfall stations are clearly shown in Table 4.1.

Table 4.1: Homogeneity Tests of Observed Rainfall Series for Station 1 (4529001), Station 2 (4730002) and Station 3 (4832011)

Station	Month	Homogeneity Test (Significance Level, α)				Homogeneity
		SNHT	BR	PeT	VNR	
Station 1 (4529001)	Jan	0.376	0.278	0.619	0.682	Useful
	Feb	0.419	0.788	0.156	0.668	Useful
	Nov	0.144	0.112	0.487	0.147	Useful
	Dec	0.142	0.161	0.484	0.120	Useful
Station 2 (4730002)	Jan	0.206	0.119	0.159	0.728	Useful
	Feb	0.494	0.741	0.974	0.740	Useful
	Nov	0.154	0.139	0.226	0.432	Useful
	Dec	0.215	0.127	0.484	0.382	Useful
Station 3 (4832011)	Jan	0.277	0.149	0.232	0.490	Useful
	Feb	0.518	0.780	**0.045**	0.717	Useful
	Nov	0.411	0.433	0.739	0.766	Useful
	Dec	0.130	0.083	0.327	0.140	Useful

4.2 Normality Test for Observed Rainfall Series (Input)

The normality tests such as the SW test, the KS test, the JB test and the AD test were first applied to the observed rainfall series (input). The test interpretations are stated as follow:

Null hypothesis, H_o: Normally distributed observed rainfall series ($\alpha > 0.05$)

Alternative hypothesis, H_a: Non-normally distributed observed rainfall series ($\alpha < 0.05$)

Based on the results from normality tests as shown in Table 4.2, all the significance level of four tests of stations 1, 2 and 3 are lesser than 0.05, thence the null hypothesis, H_o should be rejected and the alternative hypothesis, H_a accepted. Hence, the observed rainfall series were conclusively

non-normally distributed, and thus non-parametric statistics are to be used. The results of normality tests for observed rainfall series (input) are shown in Table 4.2.

Table 4.2: Results of Normality Tests for Observed Rainfall Series (Input)

Station No.	Station Code	Normality Test (Significance Level, α)			
		Shapiro-Wilk Test	Kolmogorov-Smirnov Test	Jarque-Bera Test	Anderson-Darling Test
1	4529001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
2	4730002	< 0.0001	< 0.0001	< 0.0001	< 0.0001
3	4832011	< 0.0001	< 0.0001	< 0.0001	< 0.0001

4.3 Data Pre-Processing Techniques

The data pre-processing techniques (VMD method, Boosting method and Boosting-VMD method) were carried out on the daily, monthly and seasonal rainfall series of Northeast Monsoon (November to February) only. For the decomposition of the observed rainfall series using the different data pre-processing techniques, several input parameters for the certain techniques must be pre-determined precisely to produce the best processed rainfall series for predicting river water level using the SVR model in the later part. Then, in the later part of water level prediction, the regularization parameter, C and the insensitive loss function, ε will be the main input parameters for the model training and validation. The selection of C and ε value for model training and validation will be further discussed in the later. For now, the MAE in model

validation set was used as the determining factor to select the best processed rainfall data for each of the data pre-processing techniques.

4.3.1 Variational Mode Decomposition (VMD)

For the VMD method, the balancing parameter of the data fidelity constraint, *a* is the manipulated input parameter that need to be pre-determined to produce the most ideal case of processed rainfall using the VMD method. The optimal *a* value was chosen from the series of *a* = 10, 20, 30, 40, 50 by the trial-and-error approach, minimizing the MAE between the observed and predicted values. The MAE (validation set) of *a* = 10, 20, 30, 40, 50 are 1.2197, 1.1738, 1.2002, 1.1476, 1.1489, respectively. As shown in Figure 4.1, the lowest MAE occurred at *a* = 40. Therefore, the optimal *a* value is determined as 40 for the VMD method. The optimal balancing parameter of the data fidelity constraint, *a* of VMD method are stated in Table 4.3 and Figure 4.1.

Table 4.3: Optimal *a* of VMD method

Balancing Parameter of Data Fidelity Constraint, *a*	Regularization Parameter, *C*	Insensitive Loss Function, ε	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
10	140	0.1	0.9007	1.2197	0.3190
20	120	0.1	0.8944	1.1738	0.2794
30	3	0.1	0.9096	1.2002	0.2906
40	8	0.1	0.8925	1.1476	0.2551
50	900	0.1	0.8995	1.1489	0.2494

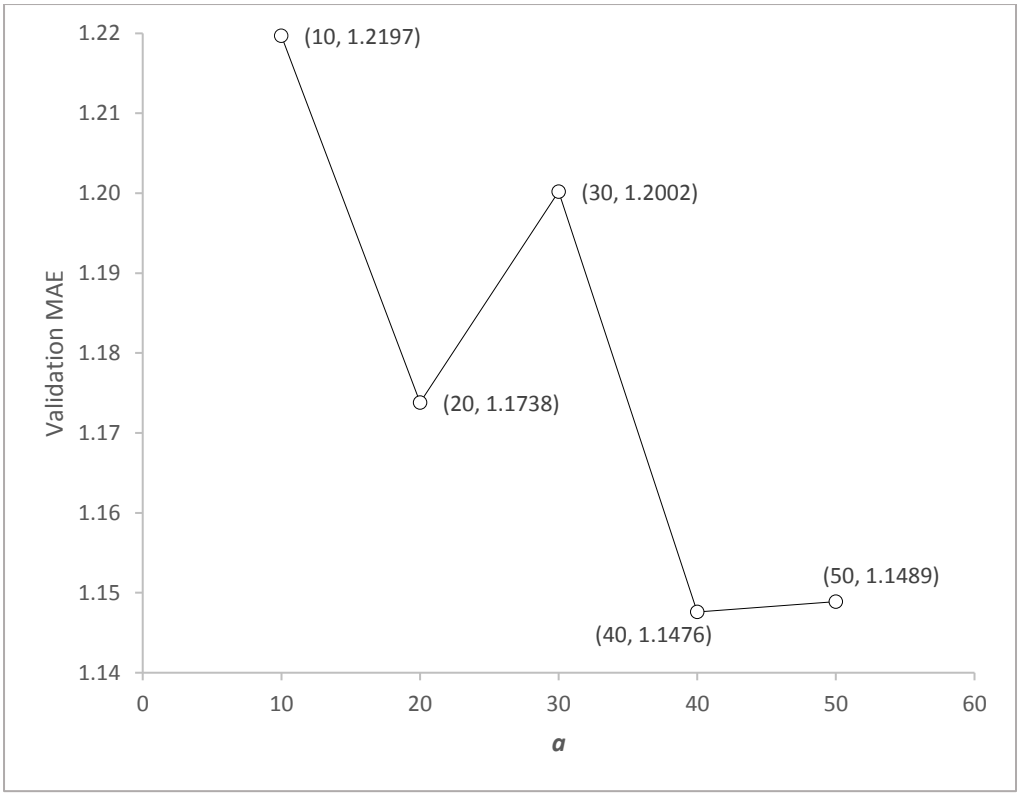


Figure 4.1: Validation MAE vs. *a* of VMD method

4.3.2 Boosting Method

For the Boosting method, the constant input parameters chosen are the LSBoost (which is used for the regression method) and the 'Tree' learner (which applicable to all methods except 'Subspace'). The optimal number of ensemble learning cycles, NLearn is the manipulated input parameter that need to be pre-determined to produce the most ideal case of processed rainfall using Boosting method. It was chosen from the series of NLearn = 1, 10, 100, 1000, 10000, 100000 by a trial-and-error approach, minimizing the MAE between the observed and predicted values. The MAE (validation set) of NLearn = 1, 10, 100, 1000, 10000, 100000 are 0.9391, 0.5305, 0.2987, 0.3704,

0.5111, 0.8686, respectively. As shown in Figure 4.2, the lowest MAE is 0.2850 which occurred at NLearn = 150. The total number of 2,405 rainfall data were used to enhance the accuracy of the learning algorithm. Only 150 number of "weak" learning algorithm needed to combine into a single prediction rule. Hence, the optimal NLearn is determined as 150 for the Boosting method. The optimal NLearn of Boosting method are stated in Table 4.4 and Figure 4.2.

Table 4.4: Optimal NLearn of Boosting Method

Number of Ensemble Learning Cycles, NLearn	Regularization Parameter, *C*	Insensitive Loss Function, ε	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
1	1000	0.1	0.7445	0.9391	0.1946
10	1000	0.8	0.4357	0.5305	0.0948
100	27	0.1	0.3360	0.2987	-0.0373
1,000	30	0.1	0.3603	0.3704	0.0101
10,000	1000	0.1	0.3980	0.5111	0.1131
100,000	150	0.1	0.7075	0.8686	0.1611

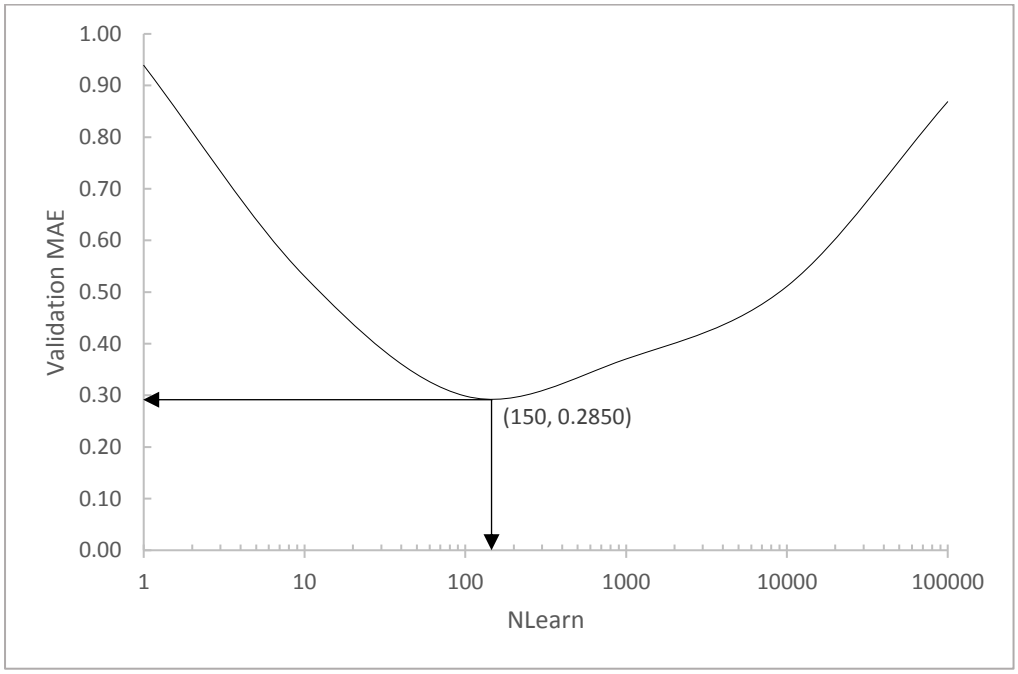


Figure 4.2: Validation MAE vs. NLearn of Boosting Method

4.3.3 Boosting-VMD Method

For the Boosting-VMD method, the input parameters chosen are similar to that of the Boosting method. The optimal number of ensemble learning cycles, NLearn need to be pre-determined to produce the most ideal case of processed rainfall using Boosting-VMD method. The MAE (validation set) of NLearn = 1, 10, 100, 1000, 10000, 100000 are 0.9419, 0.3032, 0.2515, 0.2841, 0.4298, 0.7911, respectively. As shown in Figure 4.3, the lowest MAE is 0.2501 which occurred at NLearn = 25. The total number of 2,405 rainfall data were used to enhance the accuracy of the learning algorithm. Only 25 number of "weak" learning algorithm needed to combine into a single prediction rule. Hence, the optimal number of ensemble learning cycles, NLearn is determined as 25. The optimal NLearn of Boosting-VMD method are stated in Table 4.5 and Figure 4.3.

Table 4.5: Optimal NLearn of Boosting-VMD Method

Number of Ensemble Learning Cycles, NLearn	Regularization Parameter, *C*	Insensitive Loss Function, ε	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
1	1000	0.1	0.7395	0.9419	0.2024
10	1000	0.8	0.3363	0.3032	-0.0331
100	90	0.1	0.2588	0.2515	-0.0073
1,000	1000	0.1	0.2541	0.2841	0.0300
10,000	1000	0.1	0.4088	0.4298	0.0210
100,000	25	0.1	0.6461	0.7911	0.1450

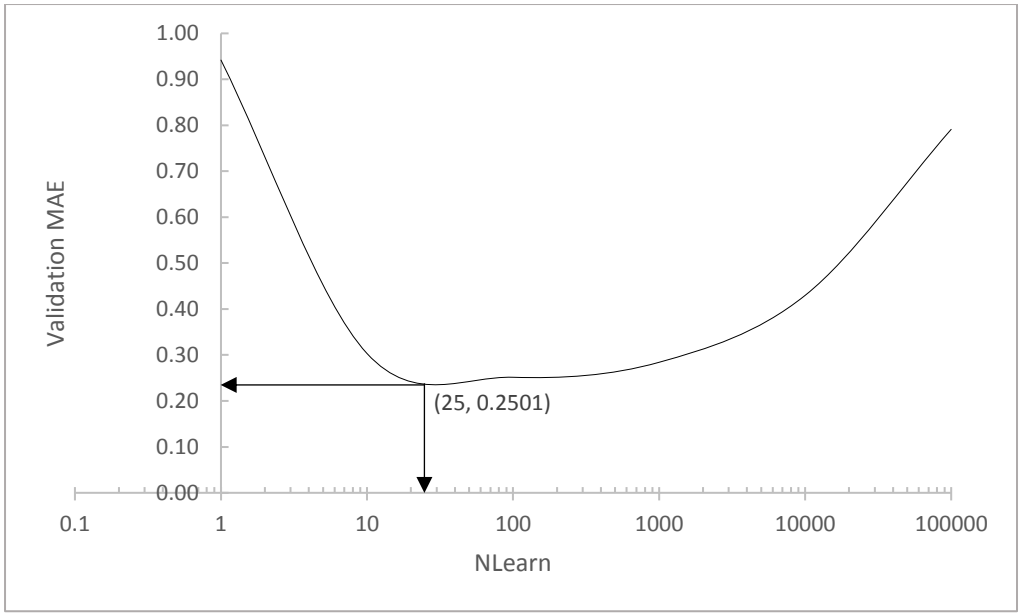


Figure 4.3: Validation MAE vs. NLearn of Boosting-VMD method

In conclusion, the three data pre-processing techniques had been successfully used to process the observed rainfall series, to be used as the input for the river water level prediction model (the SVR model). The processed rainfall series were the most ideal case for each of their own technique, based on the determining factor and pre-determined input parameters as abovementioned. The graphs of observed rainfall series and processed rainfall series using different data pre-processing techniques are clearly shown in Figure 4.4, Figure 4.5 and Figure 4.6.

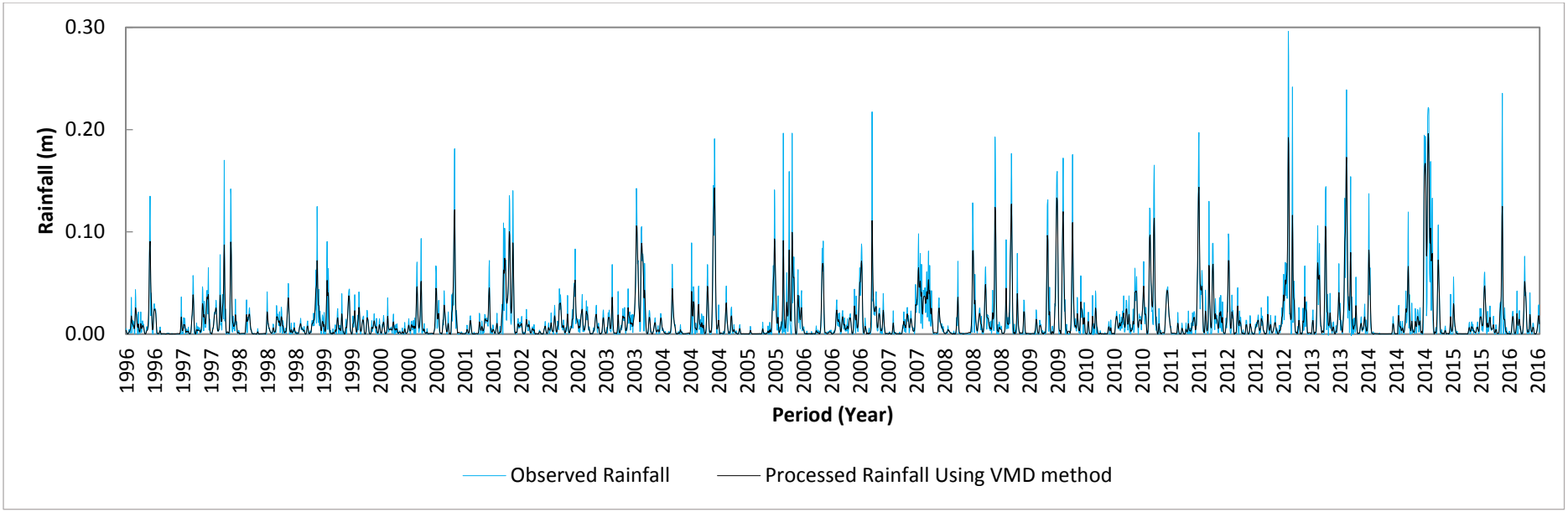


Figure 4.4: Graphs of observed and processed rainfall series using VMD method

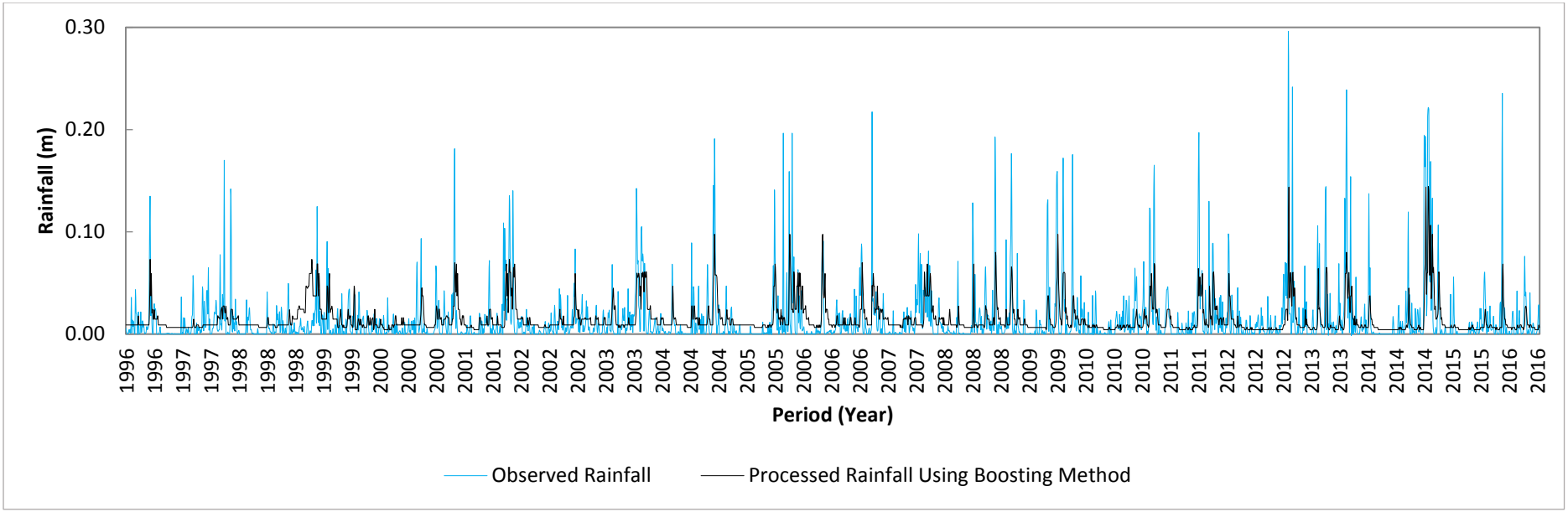


Figure 4.5: Graphs of observed and processed rainfall series using Boosting method

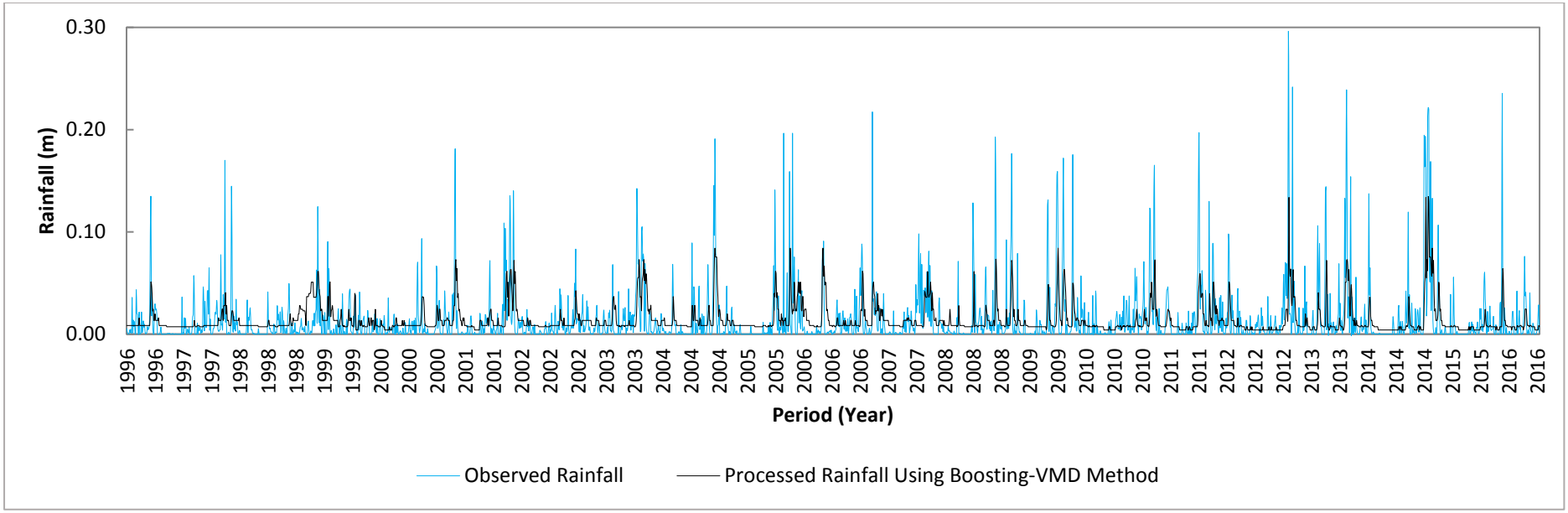


Figure 4.6: Graphs of observed and processed rainfall series using Boosting-VMD method

4.4 River Water Level Prediction Model – Support Vector Regression (SVR)

The SVR was used to predict river water levels based on, separately, the observed and the processed rainfall series. The data were split into a randomly selected ratio of 90:10 for model training and model validation, respectively. For the past records, there is no ideal ratio split for a training set and validation set and thus mostly, the split is arbitrary. For this research study, the data collected were from small data sets which possessed only 20 years of rainfall and river water level data. Hence, ratio of 90:10 split was preferable. Data from November 1996 to February 2014 (18-year) were allocated for model training and the remaining data from November 2014 to February 2016 (2-year) were to be used for model validation.

During the model training and validation with the SVR, several input parameters must be pre-determined precisely as it is a vital phase for model training and validation. It also extremely affects the accuracy of the model. The regularization parameter, C and insensitive loss function, ε are the main input parameters for the model training and validation to predict water levels. The selection of C and ε as model inputs is discussed in the next section. Intermittently, the most appropriate kernel function found is a Radial Basis Kernel (RBF) function. The MAE in the validation set was used as the determining factor for choosing the optimal input parameters to be used. The kernel function and model input parameters are chosen by the trial-and-error approach, minimizing the MAE between the observed and predicted water levels.

4.4.1 Selection of C and ε

4.4.1.1 Model Ori-SVR (Observed Rainfall Series)

For the predicted river water levels based using Ori-SVR (the original observed rainfall series), the lowest MAE (validation set) was 1.2537 when C = 1.65 and ε = 0.1, as shown in Table 4.6 and Figure 4.7.

Table 4.6: MAE of Ori-SVR

Regularization Parameter, C	Insensitive Loss Function, ε	Mean Absolute Error, MAE		
		Training Set	Validation Set	Difference
0.1	0.1	0.9999	1.2911	0.2912
1	0.1	0.9567	1.2552	0.2985
10	0.1	0.9357	1.2622	0.3265
100	0.1	0.8989	1.3362	0.4373
1,000	0.1	0.9272	1.3185	0.3913

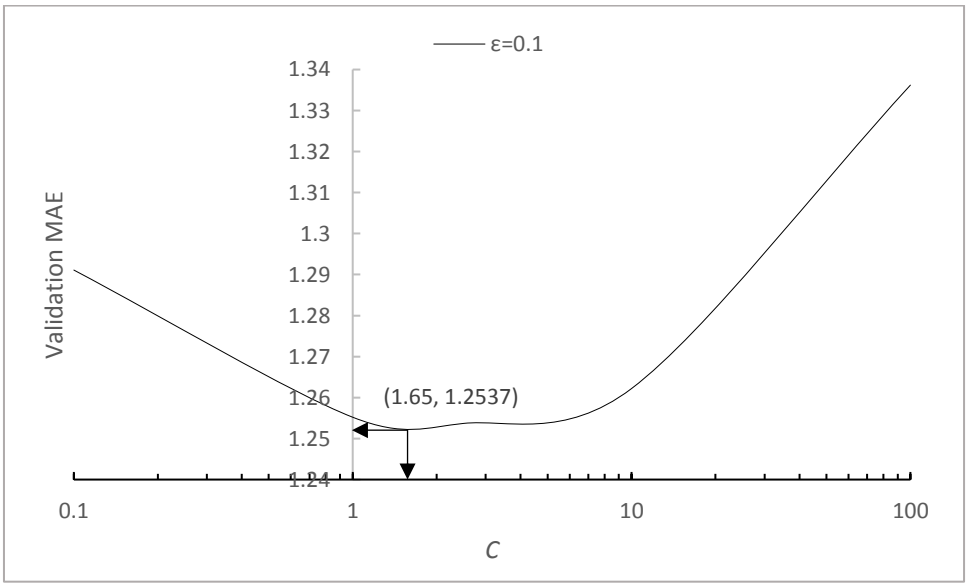


Figure 4.7: Validation MAE vs. *C* of Ori-SVR

4.4.1.2 Model VMD-SVR (VMD Method)

For the predicted river water levels based on the processed rainfall series using the VMD-SVR (VMD method), the lowest MAE (validation set) was 1.1476 when C = 8 and ε = 0.1, as shown in Table 4.7 and Figure 4.8.

Table 4.7: MAE of VMD-SVR

Regularization Parameter, *C*	Insensitive Loss Function, ε	Balancing Parameter, *a*	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
0.1	0.1	40	0.9608	1.2427	0.2819
1	0.1	40	0.9261	1.1704	0.2443
10	0.1	40	0.8909	1.1507	0.2598
100	0.1	40	0.8805	1.2006	0.3202
1,000	0.1	40	0.8695	1.2910	0.4215

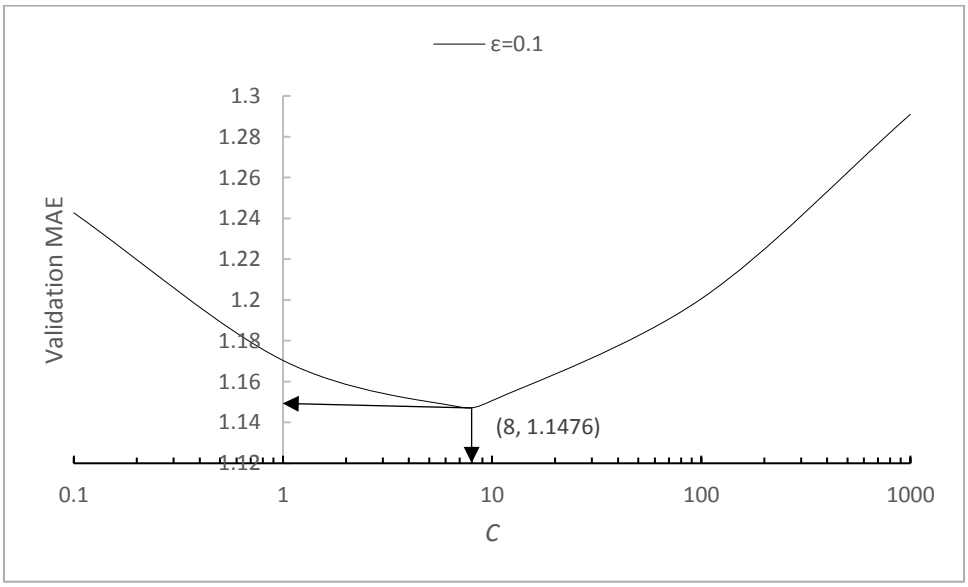


Figure 4.8: Validation MAE vs. *C* of VMD-SVR

4.4.1.3 Model B-SVR (Boosting Method)

For the predicted river water levels based on the processed rainfall series using the B-SVR (Boosting method), the lowest MAE (validation set) was 0.2850 when C = 27 and ε = 0.1, as shown in Table 4.8 and Figure 4.9.

Table 4.8: MAE of B-SVR

Regularization Parameter, *C*	Insensitive Loss Function, ε	Number of Ensemble Learning Cycles, NLearn	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
0.1	0.1	150	0.3845	0.5052	0.1207
1	0.1	150	0.3439	0.3525	0.0086
10	0.1	150	0.3368	0.3031	-0.0337
100	0.1	150	0.3355	0.3011	-0.0344
1,000	0.1	150	0.3308	0.3215	-0.0093

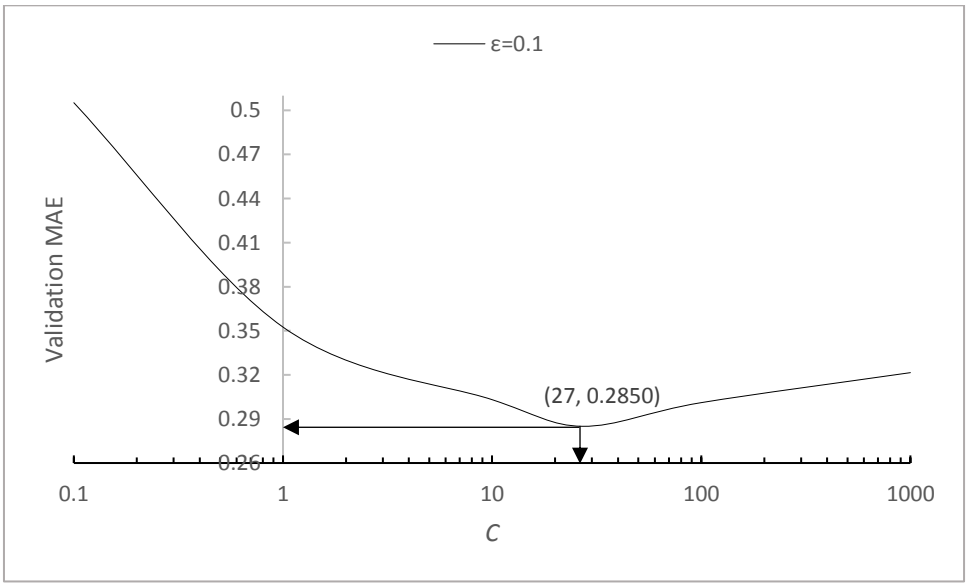


Figure 4.9: Validation MAE vs. *C* of B-SVR

4.4.1.4 Model B-VMD-SVR (Boosting-VMD Method)

For the predicted river water levels based on the processed rainfall series using the B-VMD-SVR (Boosting-VMD method), the lowest MAE (validation set) was 0.2501 when C = 90 and ε = 0.1, as shown in Table 4.9 and Figure 4.10.

Table 4.9: MAE of B-VMD-SVR

Regularization Parameter, C	Insensitive Loss Function, ε	Number of Ensemble Learning Cycles, NLearn	Mean Absolute Error, MAE		
			Training Set	Validation Set	Difference
0.1	0.1	25	0.3184	0.3870	0.0686
1	0.1	25	0.2913	0.3097	0.0184
10	0.1	25	0.2873	0.2791	-0.0082
100	0.1	25	0.2588	0.2515	-0.0073
1,000	0.1	25	0.2850	0.2795	-0.0055

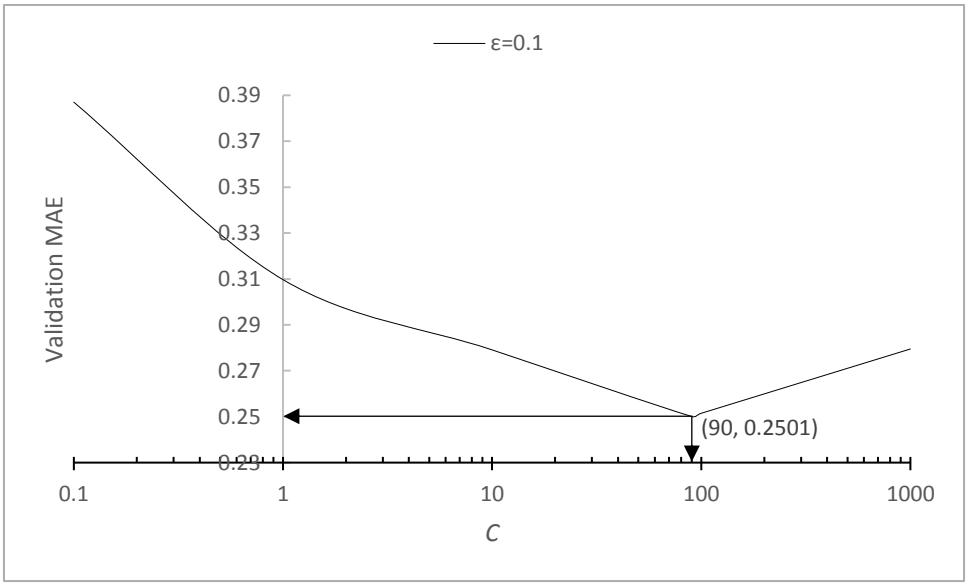


Figure 4.10: Validation MAE vs. C of B-VMD-SVR

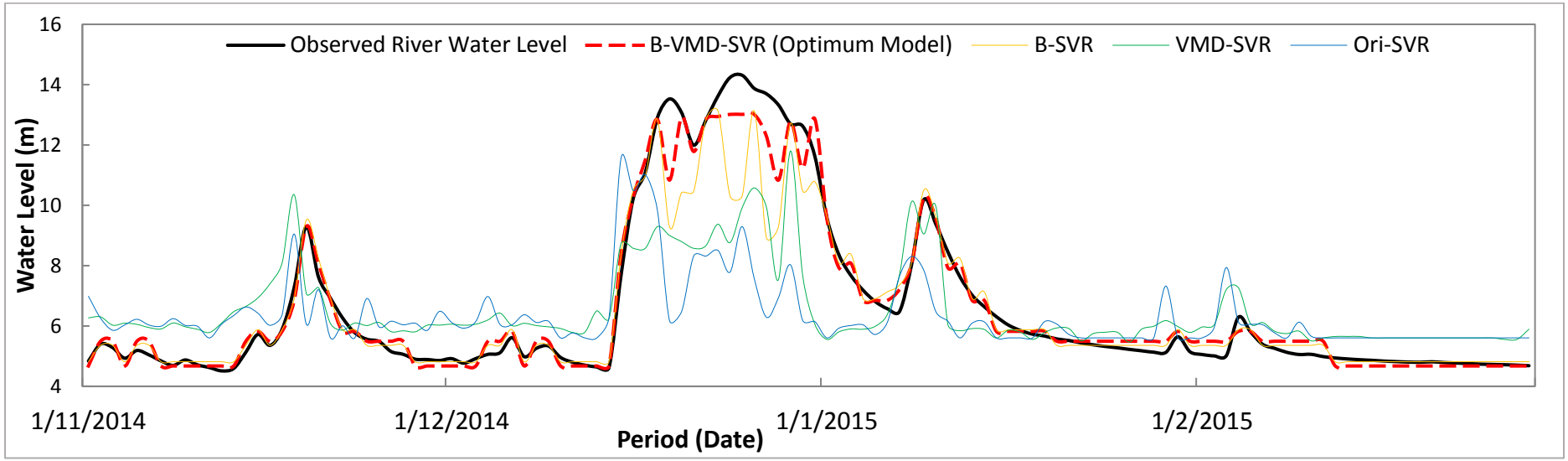
4.5 Model Development and Application

As mentioned above, during the study with SVR, the optimal input parameters for each of the models (Ori-SVR, VMD-SVR, B-SVR and B-VMD-SVR) were obtained. For Ori-SVR, C = 1.65 and ε = 0.1; for VMD-SVR, C = 8 and ε = 0.1; for B-SVR, C = 27 and ε = 0.1; and for B-VMD-SVR, C = 90 and ε = 0.1. From these optimal input parameters obtained, the optimal predicted river water levels for every model were determined.

The MAE (validation set) was used as the determining factor to select the best model for each of the models Ori-SVR, VMD-SVR, B-SVR and B-VMD-SVR, based on the optimal input parameters (C and ε). The lower the MAE value, the better the prediction of river water levels using river water level prediction model. During the study with SVR, according to the optimal input parameters for every model, the lowest MAE obtained were 1.2537, 1.1476, 0.2850, 0.2501 for the models Ori-SVR, VMD-SVR, B-SVR, B-VMD-SVR, respectively.

The hydrographs of observed river water levels and predicted river water levels using observed and processed rainfall series for validation sets (2 years of river water level series) are shown in Figure 4.11.

(a)


Observed River Water Level
B-VMD-SVR (Optimum Model)
B-SVR
VMD-SVR
Ori-SVR
Water Level (m)
16
14
12
10
8
6
4
1/11/2014
1/12/2014
1/1/2015
1/2/2015
Period (Date)

(b)

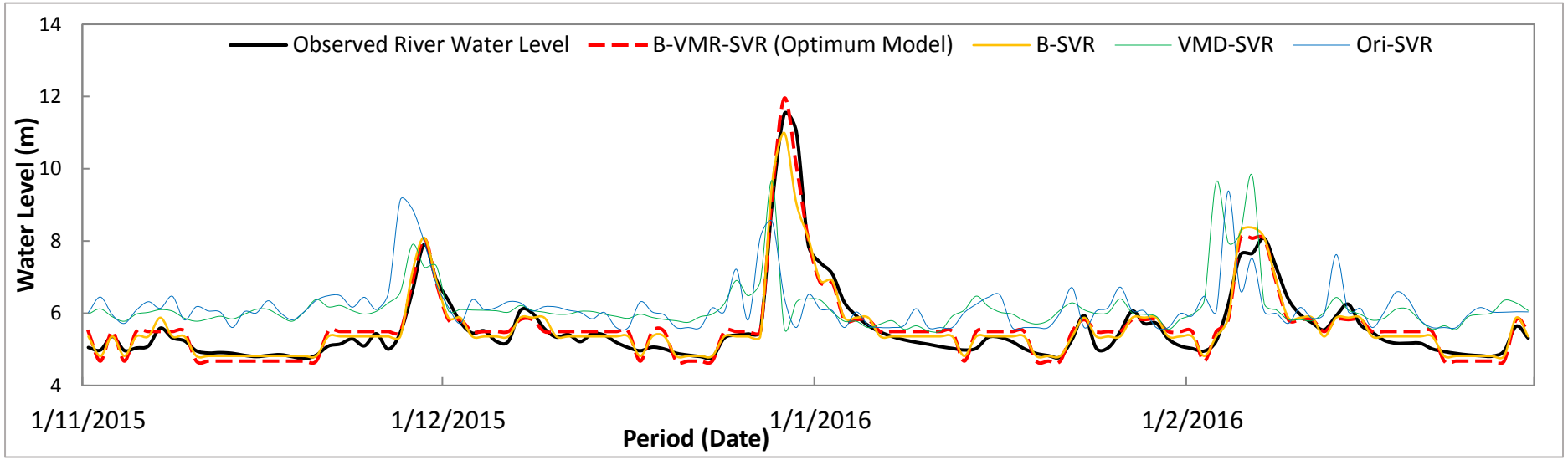


Figure 4.11: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for validation sets of 2014-2016 (a) year 2014-2015 (b) year 2015-2016

After the prediction of river water levels using the SVR, the differences between the predicted maximum river water levels for every model and the observed maximum river water levels; as well as differences between the predicted minimum river water levels for every model and the observed minimum river water levels, were determined.

The differences (in metre) between the observed maximum river water levels and the predicted maximum river water levels for the various models Ori-SVR, VMD-SVR, B-SVR, B-VMD-SVR are 6.76 m, 3.68 m, 4.46 m, 1.88 m, respectively (for training set) and 6.63 m, 3.78 m, 4.35 m, 1.75 m, respectively (for validation set).

Similarly, the differences (in metre) between the observed minimum river water levels and the predicted minimum river water levels for the various models Ori-SVR, VMD-SVR, B-SVR, B-VMD-SVR are 1.33 m, 1.24 m, 0.51 m, 0.42 m, respectively (for training set) and 1.16 m, 1.07 m, 0.34 m, 0.25 m, respectively (for validation set).

It can be observed that amongst all the models, the model B-VMD-SVR (Boosting-VMD method) gave the least differences for both the maximum and the minimum water level. The results are tabulated in Table 4.10 and Table 4.11.

Table 4.10: Differences between predicted maximum river water levels for every model and observed maximum river water levels

Period	Time (Year)	Model	Predicted Max Water Level (m)	Observed Max Water Level (m)	Differences	
					(m)	(%)
Training	18	Ori-SVR	7.68	14.44	6.76	46.81
		VMD-SVR	10.76	14.44	3.68	25.48
		B-SVR	9.98	14.44	4.46	30.89
		B-VMD-SVR	**12.56**	**14.44**	**1.88**	**13.02**
Validation	2	Ori-SVR	7.68	14.31	6.63	46.33
		VMD-SVR	10.53	14.31	3.78	26.42
		B-SVR	9.96	14.31	4.35	30.40
		B-VMD-SVR	**12.56**	**14.31**	**1.75**	**12.23**

Table 4.11: Differences between predicted minimum river water levels for every model and observed minimum river water levels

Period	Time (Year)	Model	Predicted Min Water Level (m)	Observed Min Water Level (m)	Differences	
					(m)	(%)
Training	18	Ori-SVR	5.67	4.34	1.33	30.65
		VMD-SVR	5.58	4.34	1.24	28.57
		B-SVR	4.85	4.34	0.51	11.75
		B-VMD-SVR	**4.76**	**4.34**	**0.42**	**9.68**
Validation	2	Ori-SVR	5.67	4.51	1.16	25.72
		VMD-SVR	5.58	4.51	1.07	23.73
		B-SVR	4.85	4.51	0.34	7.54
		B-VMD-SVR	**4.76**	**4.51**	**0.25**	**5.54**

4.6 Model Performance Measures

The predicted river water levels from both the processed and the observed rainfall series, were assessed against the observed river water levels, using certain assessment measures to evaluate the models' performance. Both non-parametric and parametric statistics were adopted.

4.6.1 Non-Parametric Statistics – Residual Modelling Statistics

4.6.1.1 Model Prediction Error's Range and Standard Deviation

For non-parametric statistics, the residual modelling statistics were used to evaluate the models' performance (Ori-SVR, VMD-SVR, B-SVR and B-VMD-SVR) with the significance level of 99% ($\alpha = 0.01$). The statistics of residual modelling statistics (Model prediction errors' range and standard deviation) are shown in Table 4.12.

Table 4.12: Model Prediction Errors' Range and Standard Deviation

Model	Statistics			
	Model Prediction Errors' Range	**Ranking**	**Model Prediction Errors' Standard Deviation**	**Ranking**
Ori-SVR	12.05 (Min: -8.17, Max: 3.88)	3rd	1.4842	4th
VMD-SVR	13.62 (Min: -7.74, Max: 5.88)	4th	1.3523	3rd
B-SVR	6.66 (Min: -4.92, Max: 1.74)	2nd	0.5747	2nd
B-VMD-SVR	**4.27 (Min: -2.95, Max: 1.32)**	**1st**	**0.4200**	**1st**

Rankings were assigned to each model (ranking 1st for the best model and ranking 4th for the worst model). Based on the ranking of model prediction errors' range and standard deviation as shown in Table 4.12, amongst all the models, the predicted river water levels with the processed rainfall series using the B-VMD-SVR (Boosting-VMD method) achieved the best ranking (with smallest model prediction error's range and standard deviation). Low model

prediction error's range and standard deviation indicated that the fluctuation of error is small and model stability is high, respectively. Thus, there is higher capability for model to achieve higher accuracy.

4.6.1.2 Normality Test for Residuals of Predicted River Water Levels (Output)

The normality tests such as the SW test, the KS test, and the JB test were applied to the residuals of the predicted river water levels (output). For the JB test, the test interpretations are based on the critical values of the Chi-Squared (χ^2) distribution. However, for the SW test and the KS test, the test interpretations are stated as follow:

Null hypothesis, H_o: Normally distributed residuals of predicted river water levels ($\alpha > 0.05$)

Alternative hypothesis, H_a: Non-normally distributed residuals of predicted river water levels ($\alpha < 0.05$)

The number of observations, *n* is 2,405. Thus, SW test is not applicable here as it is only applicable for *n* which less than 2,000. However, the test results of SW were still shown in Table 4.13 as it was auto-generated from the SPSS software. Based on the results from normality tests as shown in Table 4.13, the significance level of the KS test for all the models are more than 0.05, following which, the null hypothesis, H_o should be accepted and reject the alternative hypothesis, H_a.

For the JB test, the degree of freedom (n – number of parameter) is 2,402. The critical χ^2 value is 9,007. Based on the results of the normality tests as shown in Table 4.13, the calculated JB statistics of Ori-SVR and VMD-SVR are less than 9,007 whereas the JB statistics of B-SVR and B-VMD-SVR are more than 9,007. Hence, the Ori-SVR and the VMD-SVR were conclusively normally distributed (use model prediction errors' mean confidence interval range as next model performance measure), whereas the B-SVR and the B-VMD-SVR were conclusively non-normally distributed (use model prediction errors' median confidence interval range as next model performance measure). The model prediction errors' confidence interval range was selected to evaluate the models' performance with the significance level of 99% ($\alpha = 0.01$). The results of the normality tests for residuals of predicted river water levels (output) are shown in Table 4.13.

Table 4.13: Normality Tests for Residuals of Predicted River Water Levels (Output)

Model	Statistics		
	Shapiro-Wilk Test	**Kolmogorov-Smirnov Test**	**Jarque-Bera Test**
Ori-SVR	0.846	0.167	418
VMD-SVR	0.898	0.129	691
B-SVR	0.739	0.171	5,599,525
B-VMD-SVR	0.892	0.102	199,849,857

4.6.1.3 Model Prediction Error's Confidence Interval Range

As abovementioned, the residuals of predicted river water levels (output) of Ori-SVR and VMD-SVR were normally distributed whereas the

residuals of predicted river water levels (output) of B-SVR and B-VMD-SVR were non-normally distributed. Therefore, the model prediction errors' mean confidence interval range was chosen for the Ori-SVR and the VMD-SVR, whereas the model prediction errors' median confidence interval range was chosen for the B-SVR and the B-VMD-SVR, as the model performance measure to evaluate the models' performance with the significance level of 99% ($\alpha = 0.01$). The statistics of residual modelling statistics (Model prediction errors' confidence interval range) is shown in Table 4.14.

Table 4.14: Model Prediction Errors' Confidence Interval Range

Model	Statistics			
	Model Prediction Errors		Standard Error	Ranking
	Mean Confidence Interval Range	Median Confidence Interval Range		
Ori-SVR	[-0.37, -0.21]	-	0.03027	4th
VMD-SVR	[-0.26, -0.11]	-	0.02757	3rd
B-SVR	-	[-0.01, 0.01]	0.00870	2nd
B-VMD-SVR	**-**	**[-0.02, 0.01]**	**0.00868**	**1st**

As shown in Table 4.14, the rankings of residual modelling statistics (Model prediction errors' confidence interval range) for residuals of predicted river water levels from both processed and observed rainfall series were ranked from the narrowest to the broadest model prediction errors' confidence interval range, as well as from the lowest to the highest standard error. The rankings from 1st to 4th were assigned to the B-VMD-SVR, B-SVR, VMD-SVR and Ori-SVR, respectively. On average, the model prediction error's confidence interval range of the B-VMD-SVR (Boosting-VMD method) was

nearer to zero value and achieved the best ranking (with narrowest model prediction error's mean confidence interval range); amongst all the models.

4.6.2 Parametric Statistics

For parametric statistics, the goodness-of-fit of every statistical test which was used to evaluate the models' performance (Ori-SVR, VMD-SVR, B-SVR and B-VMD-SVR) are shown in Table 4.15, for both the training and validation sets of predicted river water levels from both the processed and the observed rainfall series.

Table 4.15: Parametric Statistics of all Models

Period	Time (Year)	Model	Goodness-of-fit				
			Bias	RMSE	MAPE	NSEC	MAE
Training	18	Ori-SVR	-0.34	1.47	12.85	0.22	0.93
		VMD-SVR	-0.23	1.34	12.70	0.35	0.89
		B-SVR	-0.07	0.56	4.55	0.89	0.34
		B-VMD-SVR	**-0.07**	**0.43**	**4.11**	**0.93**	**0.29**
Validation	2	Ori-SVR	0.18	1.85	18.89	0.25	1.25
		VMD-SVR	0.28	1.56	17.81	0.46	1.15
		B-SVR	-0.07	0.71	3.89	0.89	0.30
		B-VMD-SVR	**0.00**	**0.42**	**4.36**	**0.96**	**0.28**

As shown in the model performance measures above, it was observed that the VMD-SVR, B-SVR and B-VMD-SVR obtained better results than the Ori-SVR. All processed rainfall series (using different data pre-processing techniques) resulted in improved performance of the river water level prediction model. The VMD-SVR obtained the third best results. Based on the parametric statistics, the VMD-SVR gave better results for both training (Bias

= - 0.23, RMSE = 1.34, MAPE = 12.70, NSEC = 0.35, MAE = 0.89) and validation (Bias = 0.28, RMSE = 1.56, MAPE = 17.81, NSEC = 0.46, MAE = 1.15) sets as compared to the Ori-SVR for both training (Bias = - 0.34, RMSE = 1.47, MAPE = 12.85, NSEC = 0.22, MAE = 0.93) and validation (Bias = 0.18, RMSE = 1.85, MAPE = 18.89, NSEC = 0.25, MAE = 1.25) sets. With the validation set, although the Ori-SVR obtained the lowest bias as compared to that of VMD-SVR, the VMD-SVR still outperformed Ori-SVR in terms of another four (RMSE, MAPE, NSEC and MAE) out of five tests. Besides that, based on the non-parametric statistics (Residual modelling statistics), the VMD-SVR was ranked as 4^{th}, 3^{rd} and 3^{rd} for model prediction error's range, standard deviation and confidence interval range, respectively; whereas the Ori-SVR was ranked as 3^{rd}, 4^{th} and 4^{th} for model prediction error's range, standard deviation and confidence interval range, respectively. Although the Ori-SVR and the VMD-SVR possessed the equal rankings for model prediction errors' range and standard deviation, the VMD-SVR (3^{rd} ranking) still was better than the Ori-SVR (4^{th} ranking) in terms of model prediction error's confidence interval range. The VMD method had extracted the modes of rainfall series concurrently and properly balancing the errors between them. The modes were extracted continuously and decomposed non-recursively. During the sifting process of IMF components that were used to eliminate the noises and smoothen the uneven amplitudes of the IMFs curve (observed rainfall series), the IMFs will be extracted concurrently, and this can cope with the noises more properly. When no more IMFs can be extracted, the final IMFs curve will be the decomposed / processed rainfall series. After the

decomposition process, several noises had been eliminated, and smoother rainfall series was easier learnt in the SVR model to predict river water levels.

The second best result came from the B-SVR. Based on the parametric statistics, B-SVR obtained better results for both training (Bias = -0.07, RMSE = 0.56, MAPE = 4.55, NSEC = 0.89, MAE = 0.34) and validation (Bias = -0.07, RMSE = 0.71, MAPE = 3.89, NSEC = 0.89, MAE = 0.30) sets as compared to the VMD-SVR. The B-SVR bested the VMD-SVR in terms of all five (Bias, RMSE, MAPE, NSEC and MAE) tests. At the same time, based on the non-parametric statistics (Residual modelling statistics), B-SVR was ranked 2nd for all model prediction error's range, standard deviation and confidence interval range; whereas VMD-SVR was ranked respectively 4th, 3rd and 3rd for model prediction error's range, standard deviation and confidence interval range. Obviously, the B-SVR possessed better rankings in terms of model prediction error's range, standard deviation and confidence interval range as compared to that of VMD-SVR. For model prediction error's confidence interval range, B-SVR obtained a narrower range and lower standard error. Hence, B-SVR was better than the VMD-SVR in term of residual modelling statistics. In the Boosting method, a new rainfall learning set is formed through making bootstrap replicas using multiple versions of rainfall series. The predicted rainfall series that was processed using the Boosting method is a mean over different versions of predictors (Grunwald et al. 2009). According to Zhang et al. (2008), the Boosting method combines the outcomes from various predictors and permutated the training data and to combine the approximation generated from the base learners. The Boosting

method assists in boosting “weak” learning algorithms, which performs just slightly better than random guessing, into a relatively “stronger” learning algorithm. After many rounds of generation, the boosting algorithm combines all the weak rules into a single prediction rule and this will be the new learning sets of rainfall series. As compared to the VMD method, the Boosting method assists in boosting “weak” learning algorithm more specifically and directly rather than just extracting the mode concurrently and smoothen the uneven amplitudes of rainfall series curve. This will be more effective in processing the so called “weak” rainfall series into a “stronger” rainfall series, which act as the learning algorithm for the later part of river water level prediction.

Last but not least, the best results originated from the B-VMD-SVR. Based on the parametric statistics, the B-VMD-SVR obtained better results for both training (Bias = -0.07, RMSE = 0.43, MAPE = 4.11, NSEC = 0.93, MAE = 0.29) and validation (Bias = 0.00, RMSE = 0.42, MAPE = 4.36, NSEC = 0.96, MAE = 0.28) sets when compared to the B-SVR. In training set, although B-VMD-SVR obtained the same bias value with B-SVR, the B-VMD-SVR still outperformed the B-SVR in terms of another four (RMSE, MAPE, NSEC and MAE) out of five tests. In validation set, although B-VMD-SVR obtained higher MAPE as compared to that of B-SVR, the B-VMD-SVR still outperformed B-SVR in terms of another four (Bias, RMSE, NSEC and MAE) out of five tests. At the same time, based on the non-parametric statistics (Residual modelling statistics), the B-VMD-SVR was ranked as 1st for all model prediction error’s range, standard deviation and confidence interval range; whereas the B-SVR was ranked as 2nd for all model

prediction error's range, standard deviation and confidence interval range, respectively. For model prediction errors's confidence interval range, the B-VMD-SVR obtained a narrower range and lower standard error when compared to those of the B-SVR. Hence, B-VMD-SVR outclassed the B-SVR in term of residual modelling statistics. The Boosting-VMD method was a hybrid of the VMD and the Boosting method. The generated IMFs curve from the VMD were further decomposed and boosted using the Boosting method. As shown in Table 4.4 and Table 4.5, Boosting method and Boosting-VMD method obtained the lowest MAE at NLearn = 150 and NLearn = 25, respectively. Only 25 number of "weak" learning algorithm was needed to combine into a single prediction rule for Boosting-VMD method, whereas 150 number of "weak" learning algorithm was needed to combine into a single prediction rule for Boosting method. Besides that, the Boosting-VMD method also obtained lower MAE (validation set) which is 0.2765 than that of Boosting method at 0.2987. These showed that a lesser error was obtained while boosting the "weak" learning algorithm of rainfall series which decomposed using VMD method earlier before Boosting method. Combining both the VMD and the Boosting method, the smoothen rainfall series curve will be easier for the Boosting method to boost the "weak" learning algorithm. This can help in processing the rainfall series more properly for further river water level prediction using the SVR.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Four data-pre-processing-technique-modified-SVR river water level prediction models namely; the Ori-SVR (using observed rainfall), the VMD-SVR (using VMD method), the B-SVR (using Boosting method) and the B-VMD-SVR (using Boosting-VMD method) to predict the river water levels, at the Dungun River, Terengganu were developed in this study. All the predicted river water levels were assessed against the observed river water levels, using non-parametric statistics (Model prediction errors' range, standard deviation and confidence interval range) and parametric statistics (Bias, RMSE, MAPE, NSEC and MAE) where appropriate, to assess the three data pre-processing techniques in enhancing the SVR model.

It was concluded that the accuracy of the SVR models were creditably improved with the using of the processed rainfall series that had gone through the VMD method (VMD-SVR), the Boosting method (B-SVR) or the Boosting-VMD method (B-VMD-SVR) data pre-processing techniques, when compared to those with the use of only the non-processed original observed rainfall series. Three methods were shown to be robust in processing the rainfall series a priori to inputting into the SVR model for river water level

prediction. For the VMD method, it was able to extract the modes of observed rainfall series concurrently and properly balancing the errors between them. Hence, the noise can be decomposed and eliminated, concurrently and non-recursively. The Boosting method was also investigated for use to enhance the performance of weak learners of the observed rainfall series. From the results shown, the optimal number of ensemble learning cycles, NLearn of Boosting method and Boosting-VMD method are 150 and 25, respectively. The number of ensemble learning cycles denoting the number of trained weak learners, the training difficulties were getting lesser when the observed rainfall series were first processed using VMD method in eliminating the errors beforehand. The double decomposition of observed rainfall series using both the VMD method followed with the Boosting method show better results as the optimal number of ensemble learning cycles was lesser as compared to the single Boosting method.

Each of the models Ori-SVR, VMD-SVR, B-SVR and B-VMD-SVR were well trained and validated with the optimal C and ε as input parameters. From these optimal input parameters obtained, the optimal predicted river water levels for every model were determined. The model B-VMD-SVR (in training and validation sets) closely mimic the pattern of observed river water levels, showing the best result of predicted river water levels; amongst all the models.

The predicted river water levels were assessed against the observed river water levels, using non-parametric statistics and parametric statistics to

assess the performance of three data pre-processing techniques. From the overall results shown, for both non-parametric statistics and parametric statistics, the accuracy of data pre-processing techniques was increased from the VMD method to the Boosting method and lastly, the Boosting-VMD method. All in all, the Boosting-VMD method (B-VMD-SVR) was the best data pre-processing technique for improving data series resulting in the best river water levels prediction with the SVR model for the Dungun River, Terengganu, with acceptable, reasonable and commendable results of model prediction error's range (4.27; Min: -2.95, Max: 1.32), model prediction error's standard deviation (0.4200), model prediction error's confidence interval range ([-0.02, 0.01]; Standard error: 0.00868), Bias (0.00), Root-Mean-Square Error (0.42), Mean Absolute Percentage Error (4.36), Nash-Sutcliffe Efficiency Coefficient (0.96) and Mean Absolute Error (0.28).

5.2 Recommendations

This study focused only on river water level prediction at the Dungun River, Terengganu, a tropical climate region. Other rivers with different physical characteristics and under different climate environment ought to be carried out in future using similar approaches to further examine the performance of the proposed model.

There were only three data pre-processing techniques adopted to process observed rainfall series in this study. Further studies on few more

other data pre-processing techniques are encouraged to adopt before applying in other study areas to help in increasing the scalability of the model used.

The SVR model was used as the base river water level prediction model in this study. Further studies on the VMD method, the Boosting method and the Boosting-VMD method with other machine learning methods to examine their performances can also be performed to see if similar results can be obtained as those of with the SVR models.

The period used for the river water level prediction in this study was 20 years (from 1996 to 2016). A longer period of rainfall series should be used in the future so that more training data set are applicable to increase the accuracy of river water level prediction model.

REFERENCES

Abrahart et al., 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, 36 (4), pp. 480-513.

Adamowski, J. and Chan, H.F., 2011. A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology*, 407, pp. 28-40.

Aditya, S., Chinmay, D. and Vivek, P., 2016. Denoising Knee Joint Vibration Signals Using Variational Mode Decomposition. *Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing*, 433.

Ai, L., Wang, J., and Yao, R., 2011. Classification of parkinsonian and essential tremor using empirical mode decomposition and support vector machine. *Digital Signal Processing*, 21, pp. 543-550.

Aman, M.K., 2015. Improving Forecasting Accuracy of Streamflow Time Series Using Least Squares Support Vector Machine Coupled with Data-Preprocessing Techniques. *Water Resources Management*.

Amar, A. and Guennoun, Z.E.A., 2012. Contribution of Wavelet Transformation and Empirical Mode Decomposition to Measurement of U.S Core Inflation. *Applied Mathematical Sciences*, 6 (135), pp. 6739-6752.

An, N., Zhao, W., Wang, J., Shang, D., and Zhao, E., 2013. Using multi-output feed-forward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. *Energy*, 49, pp. 279-288.

Anctil, F. and Lauzon, N., 2004. Generalisation for neural networks through data sampling and training procedures with applications to streamflow predictions. *Hydrology Earth System Science*, 8 (5), pp. 940-958.

Aneesh, C. et al., 2015. Performance comparison of Variational Mode Decomposition over Empirical Wavelet Transform for the classification of power quality disturbances using Support Vector Machine. *Procedia Computer Science*, 46, pp. 372-380.

de Artigas, M.Z., Elias, A.G. and De Campra, P.F., 2006. Discrete wavelet analysis to assess long-term trends in geomagnetic activity. *Physics and Chemistry of the Earth, Parts A/B/C*, 31, pp. 77-80.

Awad, M., Jiang, X. and Motai, Y., 2007. Incremental support vector machine framework for visual sensor networks. *Journal of Advance Signal Process*.

Awchi, T.A., 2014. River discharges forecasting in northern Iraq using different ANN techniques. *Water Resources Management*, 28(3), pp. 801-814.

Ba et al., 2017. Improving ANN model performance in runoff forecasting by adding soil moisture input and using data preprocessing techniques. *Hydrology Research*.

Badrzadeh, H., Sarukkalige, R. and Jayawardena, A.W., 2017. Intermittent stream flow forecasting and modelling with hybrid wavelet neuro-fuzzy model. *Hydrology Research*.

Ballini, R., Soares, S. and Andrade, M.G., 2001. Multi-step-ahead streamflow forecasting by a neurofuzzy network model. *Proceedings of the Annual Conference of the North Americal Fuzzy Information Processing Society*, 2, pp. 992-997.

Baratta, D., Cicioni, G., Masulli, F. and Studer, L., 2003. Application of an ensemble technique based on singular spectrum analysis to daily rainfall forecasting. *Neural Networks*, 16, pp. 375-387.

Beneki, C., Eeckels, B. and Leon, C., 2012. Signal extraction and forecasting of the UK tourism income time series: A Singular Spectrum Analysis approach. *Journal of Forecasting*, 31 (5), pp. 391-400.

Bowden, G.J., Maier, H.R. and Dandy, G.C., 2012. Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability. *Water Resources Research*, 48.

Bray, M. and Han, D., 2004. Identification of support vector machines for runoff modeling. *Journal of Hydroinformatics*, 6 (4), pp. 265-280.

Cameron, M.Z., Donald, H.B. and Slobodan, P.S., 1999. Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, 214, pp. 32-48.

Cannas, B., Fanni, A., See, L. and Sias, G., 2006. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. *Physics and Chemistry of the Earth*, 31 (18), pp. 1164-1171.

Chang, L.C. and Chang, F.J., 2001. Intelligent control for modeling of real time reservoir operation. *Hydrological Processes*, 15, pp. 1621-1634.

Chang, F.J. and Chen, Y.C., 2001. A counterpropagation fuzzy-neural network modelling approach to real time streamflow prediction. *Journal of Hydrology*, 245, pp. 153-164.

de Chazal, P., Celler, B. and Reilly, R., 2000. Using wavelet coefficients for the classification of the electrocardiogram. *Proceedings of the 22nd Annual International Conference.*

Chen, S.H., Lin, Y.H., Chang, L.C. and Chang, F.J., 2006. The strategy of building a flood forecast model by neuro-fuzzy network. *Hydrological Processes*, 20, pp. 1525-1540.

Chen, S.T., Yu, P.S. and Tang, Y.H., 2010. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology*, 385, pp. 13-22.

Cheng, C.H., and Wei, L.Y., 2014. A novel time-series model based on empirical mode decomposition for forecasting TAIEX. *Economic Modeling*, 36, pp. 136-141.

Cheng, J., Yang, Y., and Yang, Y., 2012. A rotating machinery fault diagnosis method based on local mean decomposition. *Digital Signal Processing*, 22, pp. 356-366.

Chou, C.M., 2007. Efficient nonlinear modeling of rainfall-runoff process using wavelet compression. *Journal of Hydrology*, 332 (3–4), pp. 442-455.

Chou, J.S., Chiu, C.K., Farfoura, M. and Al-Taharwa, I., 2011. Optimizing the prediction accuracy of concrete compressive strength based on a comparison of datamining techniques. *Journal of Computing in Civil Engineering*, 25 (3), pp. 242-263.

Chui, C.K., 1992. An Introduction to Wavelets. Academic Press.

Collobert, R. and Bengio, S., 2001. Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1, pp. 143-160.

Conover, W.J., 1999. Practical Nonparametric Statistics. Third Edition, John Wiley & Sons, Inc. New York, pp. 428-433 (6.1).

Corder, G.W. and Foreman, D.I., 2009. Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. Wiley.

Cordiner, A., 2009. *Adaboost Toolbox: A Matlab Toolbox for Adaptive Boosting. Advanced Multimedia Research Lab Oratory Information Communications Technology Research Institute*. Australia: University of Wollongong.

Deng, Y.F., Jin, X. and Zhong, Y.X., 2005. Ensemble SVR for prediction of time series. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, pp. 3528-3534.

Department of Irrigation and Drainage (DID) Malaysia, 2001. National Register of River Basins. Kuala Lumpur: Department of Irrigation and Drainage Malaysia.

Department of Irrigation and Drainage (DID) Malaysia, 2016. National Register of River Basins. Kuala Lumpur: Department of Irrigation and Drainage Malaysia.

Dibike et al., 2001. Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering*, 15, pp 208-216.

Dragomiretskiy, K. and Zosso, D., 2014. Variational mode decomposition. *IEEE Transactions on Signal Processing*, 62, pp. 531-544.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. and Vapnik, V., 1996. Support vector regression machines. *Advances in Neural Information Processing Systems*, pp. 155-161.

Efron, B., 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Institute of Mathematical Statistics Monographs/Cambridge University Press. ISBN 9780521192491.

Efron, B. and Tibshirani, R., 1994. *An Introduction to the Bootstrap*. Chapman & hall/CRC. ISBN 978-0-412-04231-7.

El-Shafie, A., Taha, M.R. and Noureldin, A., 2007. A neuro-fuzzy model for inflow forecasting of the Nile River at Aswan high Dam. *Water Resources Management*, 21, pp. 533-556.

Farrel, P.J. and Stewart, K.R., 2006. Comprehensive Study of Tests for Normality And Symmetry: Extending The Spiegelhalter Test. *Journal of Statistical Computation and Simulation*, 76 (9), pp. 803-816.

Fattorini, H.O., 1999. *The Effect of Inflection Angle, Soil Proximity and Location on Runoff*. PhD Thesis, Pennsylvania State University, USA.

Feng, L.H., 2013. Flood forecasting at the Dadu River in China based on ANN. *Expert Systems with Applications*, 30(5), pp. 398-402.

Firat, M. and Gungor, M., 2008. Hydrological time-series modelling using an adaptive neuro-fuzzy inference system. *Hydrological Processes*, 22, pp. 2122-2132.

Flood, I. and Kartam, N., 1994. Neural networks in civil engineering I: Principles and understanding. *Journal of Computing in Civil Engineering*, 8(2), pp. 131-148.

FloodList, 2017. FloodList. Floodlist.com/tag/malaysia.

Fontane, D.G., Gates, T.K. and Moncada, E., 1997. Planning reservoir operations with imprecise objectives. *Journal of Water Resources Planning and Management*, 123 (3), pp. 154-168.

Friedman, J.H., 1999. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), pp. 1189-1232.

Gao, J.B., Gunn, S.R., Harris, C.J. and Brown, M., 2001. A probabilistic framework for SVM regression and error bar estimation. *Machine Learning*, 46, pp. 71-89.

Ghorbani, M.A., Khatibi, R., Goel, A., FazeliFard, M.H. and Azani, A., 2016. Modeling river discharge time series using support vector machine and artificial neural networks. *Environmental Earth Sciences*, 75 (685).

Golyandina, N., Nekrutkin, V. and Zhigljavsky, A.A., 2001. Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC, Berlin.

Grunwald, S., Daroub, S.H., Lang, T.A., Diaz, O.A., 2009. Tree-based modeling of complex interactions of phosphorus loadings and environmental factors. *Science of the Total Environment*, 407 (12), pp. 3772-3783.

Hafiz et al., 2013. Flood forecasting and early warning system for Dungun River Basin. *Earth and Environmental Science*, 16.

Hafiz, I. et al., 2013. Application of Integrated Flood Analysis System (IFAS) for Dungun River Basin. *Earth and Environmental Science*, 16.

Halil, I.E. and Onur, K., 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology*, 477, pp. 119-128.

Hancock, T., Put, R., Coomans, D., Vanderheyden, Y. and Everingham, Y., 2005. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemometrics and Intelligent Laboratory Systems*, 76 (2), pp. 185-196.

Haykin, S., 1999. Neural Network-a Comprehensive Foundation. Prentice-Hall, Englewood Cliffs, NJ.

Hu, J., Wang, J. and Zeng, G., 2013a. A hybrid forecasting approach applied to wind speed time series. *Renewable Energy*, 60, pp. 185-194.

Hu, T.S., Wu, F.Y. and Zhang, X., 2007. Rainfall–runoff modeling using principal component analysis and neural network. *Nordic Hydrology*, 38 (3), pp. 235-248.

He, Z., Wen, X., Liu, H. and Du, J., 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, 509, pp. 379-386.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, pp. 417-441.

Huang, N.E. et al., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 454, pp. 903-995.

Huang, S.Z., Chang, J.X., Huang, Q. and Chen, Y.T., 2014. Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology*, 511, pp. 764-775.

Jang, J.S.R., 1993. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23, pp. 665-685.

Jon, L., 2004. *A First Course in Combinatorial Optimization*. IBM T J Watson Research Center, New York: Cambridge University Press.

Jorge, N. and Stephen J.W., 2006. Numerical Optimization. 2nd ed. United States of America: Springer.

Kaheil, Y.H., Rosero, E., Gill, M.K., Mc Kee, M. and Basatidas, L.A., 2008. Downscaling and forecasting of evapotranspiration using a synthetic model of wavelets and support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 46 (9), pp. 2692-2707.

Kang, H.M. and Yusof, F., 2012. Homogeneity tests on daily rainfall series in Peninsular Malaysia. *International Journal of Comtemporary Mathematical Sciences*, 7, pp. 9-22.

Kecman, V., 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models. *Computing and Processing*, 1-14, pp. 608.

Keskin, S., 2006. Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation Based Small Samples. *Journal of Applied Science Research*, 2 (5), pp. 296-300.

Khan, M.S. and Coulibaly, P., 2006. Application of support vector machine in lake water level prediction. *Journal of Hydrology*, 11 (3), pp. 199-205.

Kisi, O. and Cimen, M., 2011. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology*, 399, pp. 132-140.

Kumar, P. and Foufoula-Georgiou., 1997. Wavelet analysis for geophysical applications. *Reviews of Geophysics*, 35 (4), pp. 385-412.

Kundzewicz, Z.W. and Robson, A.J., 2004. Change detection in hydrological records - A review of the methodology. *Hydrological Sciences Journal*, 49, pp. 7-19.

Lahmiri, S., 2014a. A comparative study of ECG signal denoising by wavelet thresholding in empirical and variational mode decomposition domains. *IET Healthcare Technology Letters Journal – The IET*, 1, pp. 104-109.

Lahmiri, S., 2015d. Long memory in international financial markets trends and short movements during 2008 financial crisis based on variational mode decomposition and detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 437, pp. 130-138.

Lahmiri, S., and Boukadoum, M., 2014b. Automated detection of circinate exudates in retina digital images using empirical mode decomposition and the entropy and uniformity of intrinsic mode functions. *Biomedizinische Technik/Biomedical Engineering*, 59, pp. 357-366.

Lahmiri, S., and Boukadoum, M., 2014c. Biomedical image denoising using variational mode decomposition. *IEEE Transaction on Biomedical Circuits and Systems*, pp. 340-343.

Lahmiri, S., and Boukadoum, M., 2015a. A weighted bio-signal denoising approach using empirical mode decomposition. *Biomedical Engineering Letters*, 5, pp. 131-139.

Lahmiri, S., and Boukadoum, M., 2015b. Pathology grading in retina digital images using student-adjusted empirical mode decomposition and power law statistics. *IEEE Latin American Symposium on Circuits and Systems*, pp. 1-4.

Lahmiri, S., and Boukadoum, M., 2015c. Physiological signal denoising with variational mode decomposition and weighted reconstruction after DWT thresholding. *IEEE International Symposium on Circuits & Systems*, pp. 806-809.

Lariyah, M.S., 2014. Hydrological Extreme Flood Event in Dungun River Basin Region. Conference: 13th International Conference on Urban Drainage, Sarawak, Malaysia.

Lee, C.F., Lee, J.C. and Lee, A.C., 2000. Statistics for Business and Financial Economics (2nd Version). World Scientific, Singapore.

Lee, D.T.L. and Yamamoto, A., 1994. Wavelet analysis: Theory and applications. *Hewlett-Packard Journal*, 45, pp. 44.

Lee, T. and Ouarda, T.B.M.J., 2010. Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. *Journal of Geophysical Research: Atmospheres*.

Lewis, C.D., 1982. Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting. Butterworth Scientific.

Li, C., Wang, X., Tao, Z., Wang, Q., and Du, S., 2011. Extraction of time varying information from noisy signals: An approach based on the empirical mode decomposition. *Mechanical Systems and Signal Processing*, 25, pp. 812-820.

Lin, C.S., Chiu, S.H., and Lin, T.Y., 2012. Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Economic Modeling*, 29, pp. 2583-2590.

Lin, G.F., Chen, G.R., Huang, P.Y. and Chou, Y.C., 2009. Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. *Journal of Hydrology*, 372, pp. 17-29.

Lin, J.Y., Cheng, C.T. and Chau, K.W., 2006. Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, 51, pp. 599-612.

Liong, S. and Sivapragasam, C., 2002. Flood stage forecasting with support vector machines. *Journal of the American Water Resources Association*, 38 (1), pp. 173-186.

Lisi, F., and Nan, F., 2014. Component estimation for electricity prices: Procedures and comparisons. *Energy Economics*, 44, pp. 143-159.

Lisi, F., Nicolis, O. and Sandri, M., 1995. Combining singular-spectrum analysis and neural networks for time series forecasting. *Neural Processing Letters*, 2 (4), pp. 6-10.

Lohani, A.K., Goel, N.K. and Bhatia, K.K.S., 2006. Takagi–Sugeno fuzzy inference system for modeling stage-discharge relationship. *Journal of Hydrology*, 331, pp. 146-160.

Lohani, A.K., Goel, N.K. and Bhatia, K.K.S., 2007a. Deriving stage–discharge–sediment concentration relationships using fuzzy logic. *Hydrological Sciences Journal*, 52 (4), pp. 793-807.

Lohani, A.K., Goel, N.K. and Bhatia, K.K.S., 2007b. Reply to comments provided by Z. Sen on ‘‘Takagi–Sugeno fuzzy system for modeling stage-discharge relationship’’ by A.K. Lohani, N.K. Goel and K.K.S. Bhatia. *Journal of Hydrology*, 337 (1–2), pp. 244-247.

Lohani, A.K., Rakesh, K. and R.D., Singh, 2012. Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. *Journal of Hydrology*, 442-443, pp. 23-35.

Maheswaran, R., Jan, A. and Rakesh, K., 2013. Multiscale streamflow forecasting using a new Bayesian Model Average based ensemble multi-wavelet Volterra nonlinear method. *Journal of Hydrology*, 507, pp. 186-200.

Maheswaran, R. and Khosa, R., 2011b. Comparative study of different wavelets for hydrologic forecasting. *Computers & Geosciences*, 46, pp. 284-295.

Maheswaran, R. and Khosa, R., 2012. Comparative study of different wavelets for hydrologic forecasting. *Computers & Geosciences*, 46, pp. 284-295.

McNamara, J.D., Scalea, F.L. and Fateh, M., 2005. Automatic defect classification in longrange ultrasonic rail inspection using a support vector machine-based 'smart system'. *Hydrological Sciences Journal*, 46 (6), pp. 331-337.

Mendes, M. and Pala, A., 2003. Type I Error Rate and Power of Three Normality Tests. *Pakistan Journal of Information and Technology*, 2 (2), pp. 135-139.

Miao, L., Jun, X. and Dejuan, M., 2012. Long-term trend analysis of seasonal precipitation for Beijing, China. *Journal of Resources and Ecology*, 3, pp. 64-72.

Minns, A.W. and M.J. Hall, 1996. Artificial neural networks as rainfallrunoff models. *Hydrological Sciences Journal*, 41 (3), pp. 399-417.

Misiti, M. and Misiti, Y., 1996. Rainfall trends in arid and semiarid regions of Iran. *Journal of Arid Environments*, 70, pp. 344-355.

Moradkhani, H., Hsu, K.-L., Gupta, H.V. and Sorooshian, S., 2004. Improved streamflow forecasting using self-organizing radial basis function artificial neural networks. *Journal of Hydrology*, 295, pp. 246-262.

Mordecai, A., 2003. *Nonlinear Programming: Analysis and Methods*. Mineola, New York: Dover Publications, Inc.

Moriasi et al., 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *American Society of Agricultural and Biological Engineers*, 50 (3), pp. 885-900.

Nalley, D., Adamowski, J. and Khalil, B., 2012. Using discrete wavelet transforms to analyze trends in streamflow and precipitation in Quebec and Ontario (1954–2008). *Journal of Hydrology*, 475, pp. 204-228.

Napolitano, G., Serinaldi, F. and See, L., 2011. Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: An empirical examination. *Journal of Hydrology*, 406, pp. 199-214.

Newbold, P., Carlson, W.L. and Thorne, B.M., 2003. Statistics for Business and Economics (Fifth Version). Prentice Hall, Upper Saddle River, NJ.

Nibhanupudi, S., 2003. Signal Denoising Using Wavelets. PhD Thesis, University of Cincinnati.

Nornadiah, M. R. and Yap, B. W., 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics*, 2 (1), pp. 21-33.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, pp. 559-572.

Percival, D.B., 2008. Analysis of geophysical time series using discrete wavelet transforms: An overview. *Springer, Berlin*, pp 61-79.

Premanode, B., and Toumazou, C., 2013. Improving prediction of exchange rates using Differential EMD. *Expert Systems with Applications*, 40, pp. 377-384.

Ricci, R.R., and Pennacchi, P., 2011. Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions. *Mechanical Systems and Signal Processing*, 25, pp. 821-838.

Rocco, C.M.S., 2013. Singular spectrum analysis and forecasting of failure time series. *Reliability Engineering & System Safety*, 114, pp. 126-136.

Ruszczyński, A., 2006. *Nonlinear Optimization*. Princeton University Press. ISBN 9781400841059.

Sahin, S. and Cigizoglu, H.K., 2010. Homogeneity analysis of Turkish meteorological data set. *Hydrological Processes*, 24 (8), pp. 981-992.

Salim, L., 2016. A vatiational mode decomposition approach for analysis and forecasting of economic and financial time series. *Expert Systems with Applications*, 55, pp. 268-273.

Schapire, E.R., 1990. The strength of weak learnability. *Machine Learning*, 5, pp. 197-227.

See, L. and Openshaw, S., 2000. A hybrid multi-model approach to river level forecasting. *Hydrological Sciences Journal*, 45 (4), pp. 523-536.

Shrestha, B.P., Duckstein, L. and Stakhiv, E.Z., 1996. Fuzzy rule based modeling of reservoir operation. *Journal of Water Resources Planning and Management*, 122 (4), pp. 262-269.

Shu, C. and Ouarda, T.B.M.J., 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology*, 349, pp. 31-43.

Sivapragasam, C., Liong, S.Y. and Pasha, M.F.K., 2001. Rainfall and runoff forecasting with SSA–SVM approach. *Journal of Hydroinformatics*, 3 (7), pp. 141-152.

Smola, A.J., 1996. Regression Estimation with Support Vector Learning Machines. M.Sc. Thesis, Technische Universität München, Germany.

Snelder, T.H. et al., 2009. Predictive mapping of the natural flow regimes of France. *Journal of Hydrology*, 373, pp. 57-67.

Solomatine, D.P. and A., Ostfeld, 2008. Data-driven modelling: Some past experiences and new approaches, *Journal of Hydroinformatics*, 10 (1), pp. 3-22.

Tingsanchali, T. and Gautam, M.R., 2000. Application of tank, NAM, ARMA and neural network to flood forecasting. *Hydrological Processes*, 14, pp. 2473-2487.

Tiwari, M.K. and Chatterjee, C., 2009. Daily discharge forecasting using WANNs coupled with nonlinear bias correction techniques. *IAHS-AISH Publication*, 331, pp. 98-108.

Tokar, A.S. and Johnson, P.A., 1999. Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 4 (3), pp. 232-239.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V., Golwich, S. and Smola, A.J., 1997. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems 9*, pp. 281-287.

de Vos, N.J. and Rientjes, T.H.M., 2005. Constraints of artificial neural networks for rainfall–runoff modeling: Trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences*, 9, pp. 111-126.

Wang, W., Chau, K., Qiu, L. and Chen, Y., 2015. Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *Environmental Research*, 139, pp. 46-54.

Wang, W-c., Xu, D-m., K-w, C. and Chen, S., 2013. Improved annual rainfall-runoff forecasting using PSO – SVM model based on EEMD. *Journal of Hydroinformatics*, 15 (4), pp. 1377-1390.

Wei, H. and Billings, S., 2006. Long term prediction of non-linear time series using multiresolution wavelet models. *International Journal of Control*, 79, pp. 569-580.

Wijingaard, J.B., Klein, T.A.M.G. and Können, G.P., 2003. Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*, 23, pp. 679-692.

Wu, C.L., Chau, K.W. and Fan, C., 2010. Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *Journal of Hydrology*, 389, pp. 146-167.

Wu, C.L., Chau, K.W. and Li, Y.S., 2008. River stage prediction based on a distributed support vector regression. *Journal of Hydrology*, 358, pp. 96-111.

Wu, C.L., Chau, K.W. and Li, Y.S., 2009. Predicting monthly streamflow using data-driven models coupled with data pre-processing techniques. *Water Resources Research*, 45.

Wu, J.D., and Tsai, Y.J., 2011. Speaker identification system using empirical mode decomposition and an artificial neural network. *Expert Systems with Applications*, 38, pp. 6112-6117.

Wu, M.C., Lin, G.F. and Lin, H.Y., 2012. Improving the forecasts of extreme streamflow by support vector regression with the data extracted by self-organizing map. *Hydrological Processes*.

Xiong, L., Shamseldin, A. and O'Connor, K., 2001. A non-linear combination of the forecasts of rainfall–runoff models by the first order Takagi–Sugeno fuzzy system. *Journal of* Hydrology, 245, pp. 196-217.

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O. and Lee, K.K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology*, 396, pp. 128-138.

Yoon, Y., Song, J., Hong, S.H. and Kim, J.Q., 2004. Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Hydrological Sciences Journal*, 41 (4), pp. 529-534.

Yu, P.S., Chen, C.J. and Chen, S.J., 2000. Application of grey and fuzzy methods for rainfall forecasting. *Journal of Hydrologic* Engineering, 5 (4), pp. 339-345.

Yu, P.S., Chen, S.T. and Chang, I.F., 2006. Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, 328, pp. 704-716.

Zadeh, L.A., 1965. Fuzzy sets. *American Journal of Infection Control*, 8, pp. 338-353.

Zhang, C.X., Zhang, J.S. and Wang, G.W., 2008. An empirical study of using rotation forest to improve regressors. *Applied Mathematics and Computation*, 195 (2), pp. 618-629.

Zhang, Q., Wang, B., He, B., Peng, Y. and Ren, M., 2011. Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. *Water Resources Management*, 25 (11), pp. 2683-2703.

Zhang, W.F., and Yan, H., 2012. Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences. *Pattern Recognition*, 45, pp. 947-955.

Zhang, X., Lai, K.K., and Wang, S.Y., 2008. A new approach for crude oil price analysis based on empirical mode decomposition. *Energy Economics*, 30, pp. 905-918.

Zhang, X., Yu, L., Wang, S., and Lai, K.K., 2009. Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method. *Energy Economics*, 31, pp. 768-778.

Zhu, S., Zhou, J., Ye, L. and Meng, C., 2016. Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. *Environmental Earth* Sciences, 75 (531).

APPENDICES

APPENDIX A

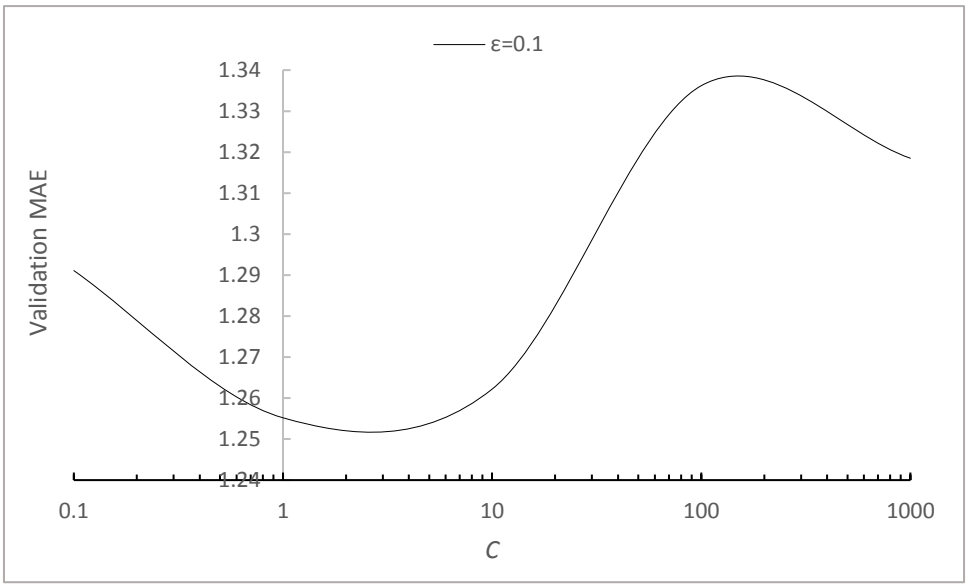


Figure A.1: Validation MAE vs. *C* of Ori-SVR (ε = 0.1)

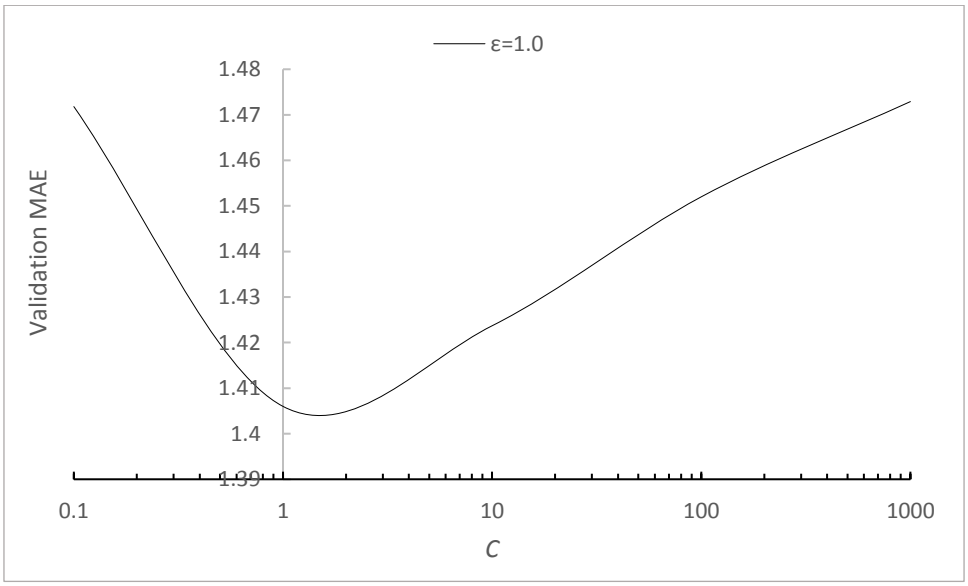


Figure A.2: Validation MAE vs. *C* of Ori-SVR (ε = 1)

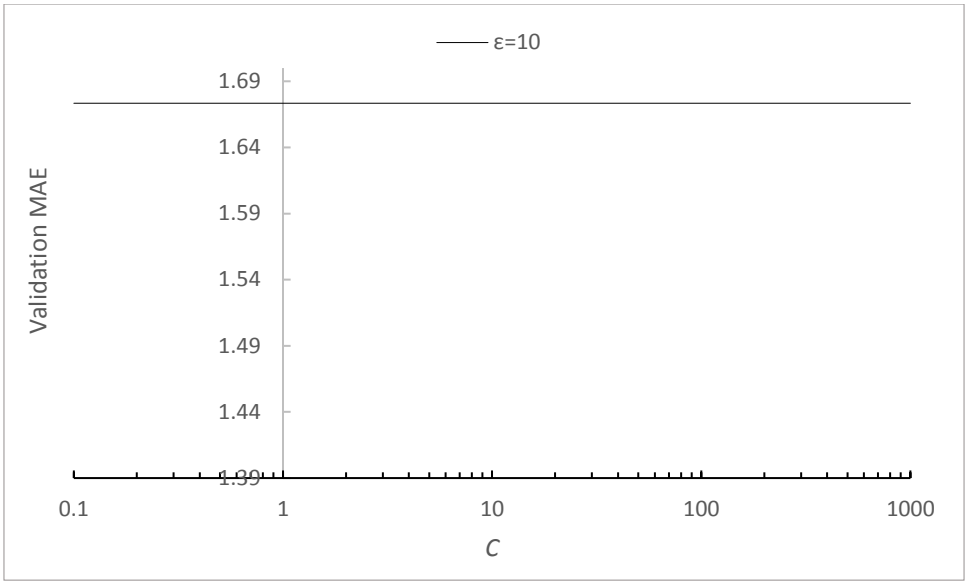


Figure A.3: Validation MAE vs. *C* of Ori-SVR (*ε* = 10), VMD-SVR (*a* = 10, 20 ,30 ,40 ,50), B-SVR and B-VMD-SVR (NLearn = 1, 10, 100, 1,000, 10,000, 100,000)

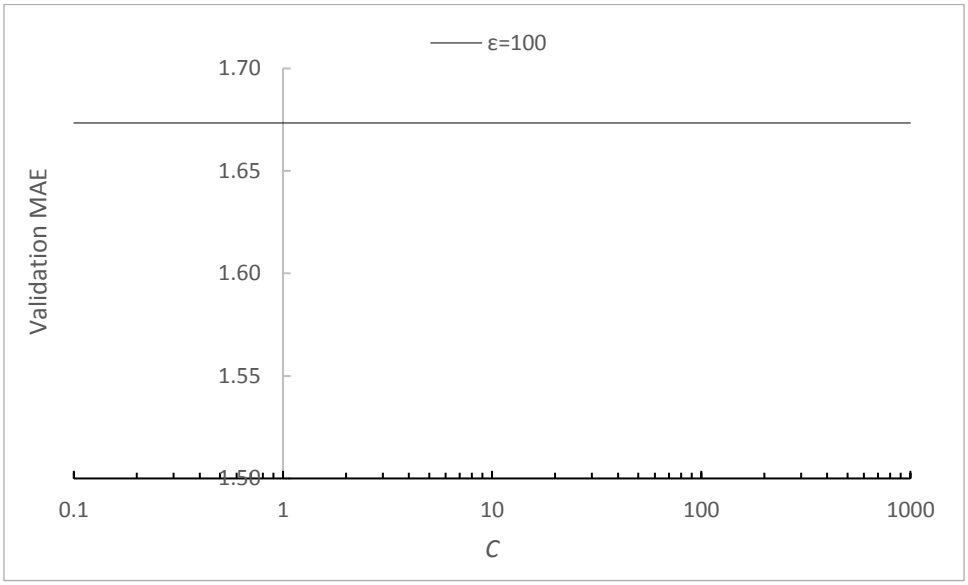


Figure A.4: Validation MAE vs. *C* of Ori-SVR (*ε* = 100), VMD-SVR (*a* = 10, 20 ,30 ,40 ,50), B-SVR and B-VMD-SVR (NLearn = 1, 10, 100, 1,000, 10,000, 100,000)

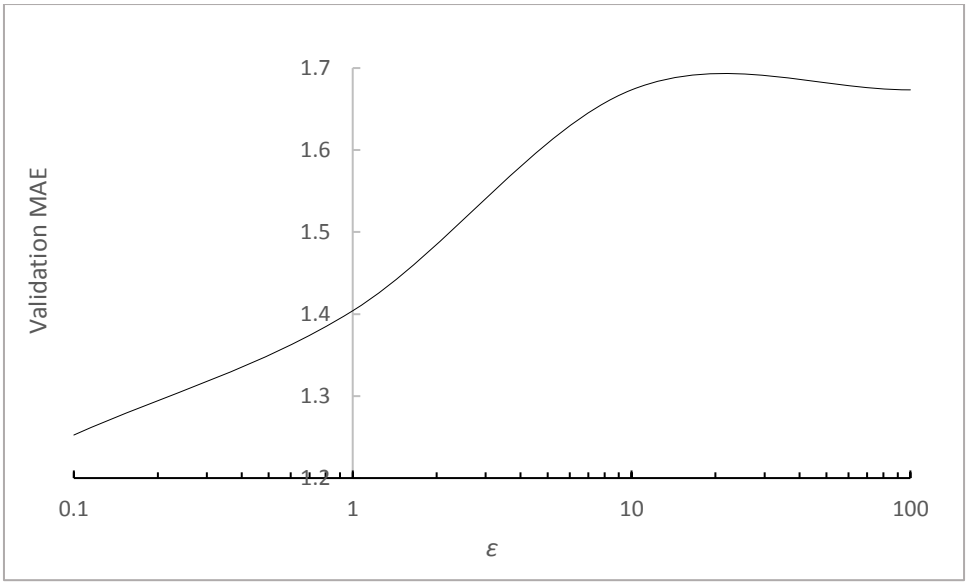


Figure A.5: Validation MAE vs. ε of Ori-SVR

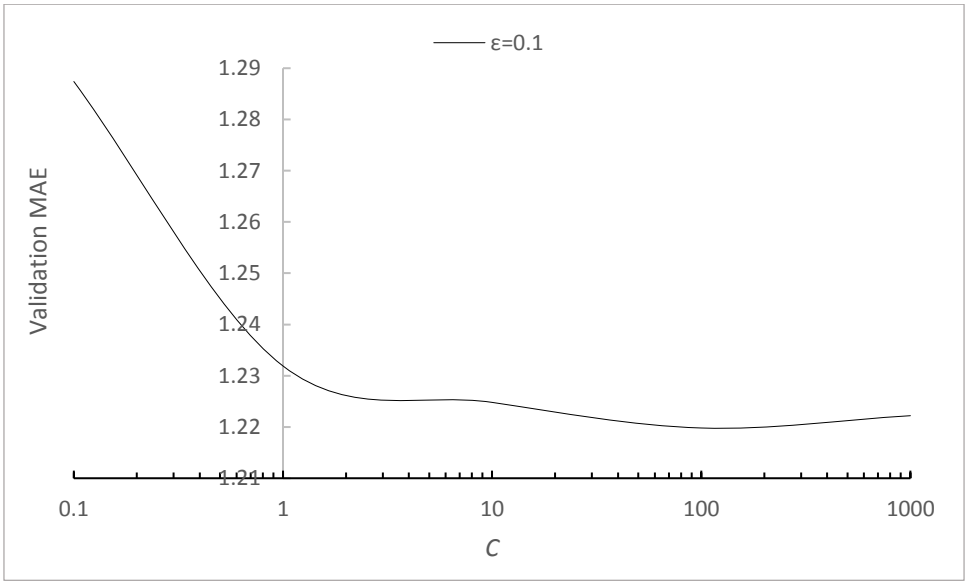


Figure A.6: Validation MAE vs. *C* of VMD-SVR (a = 10, ε = 0.1)

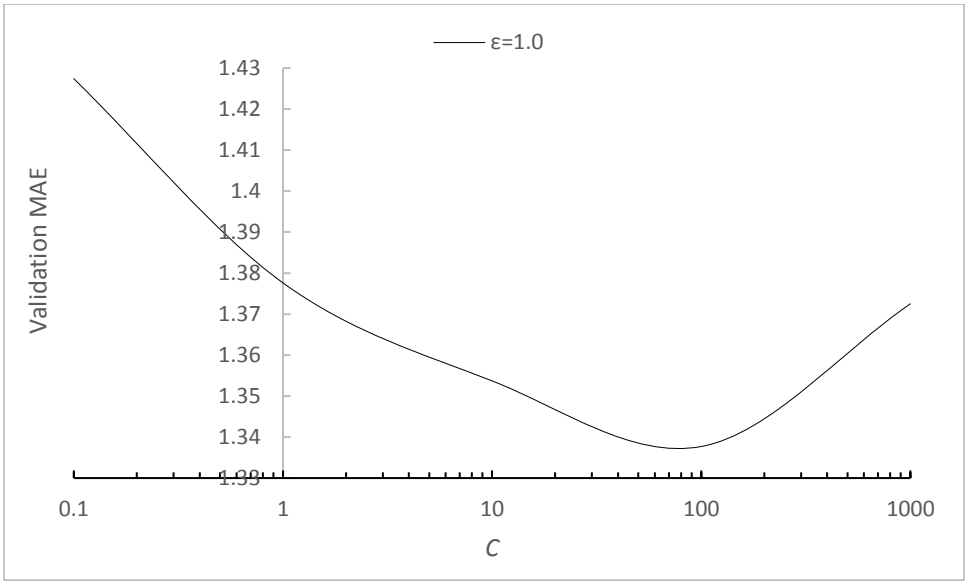


Figure A.7: Validation MAE vs. *C* of VMD-SVR (*a* = 10, *ε* = 1.0)

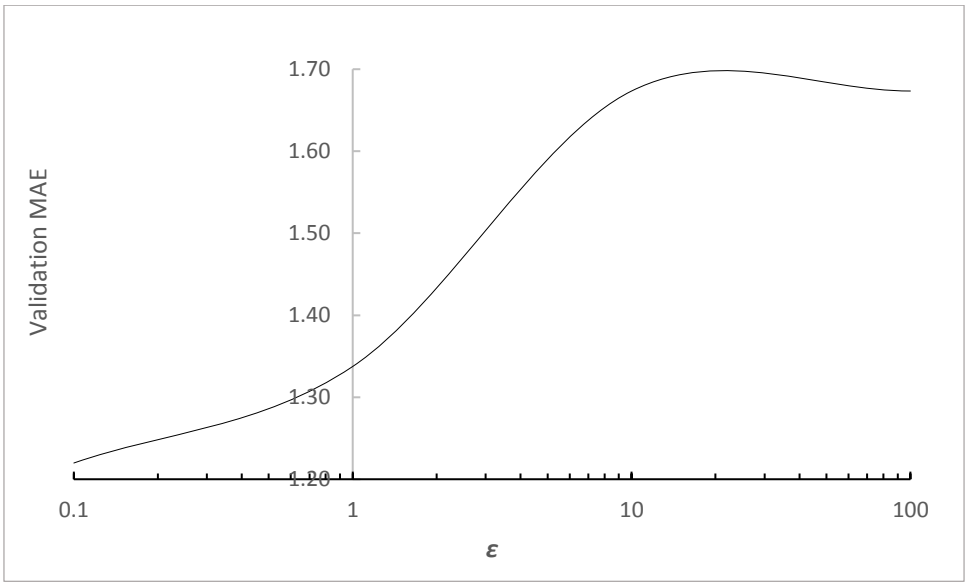


Figure A.8: Validation MAE vs. *ε* of VMD-SVR (*a* = 10)

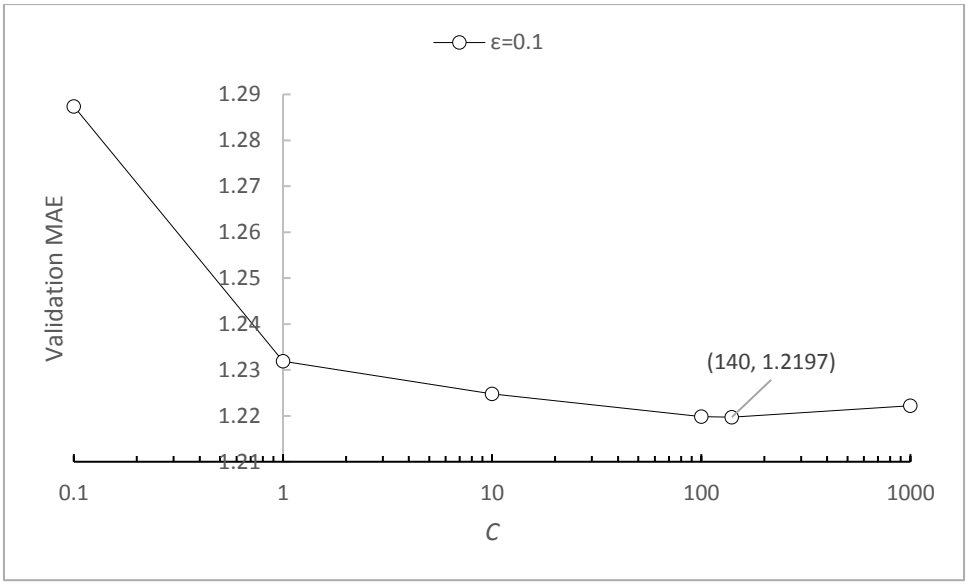


Figure A.9: Validation MAE vs. *C* of VMD-SVR (*a* = 10)

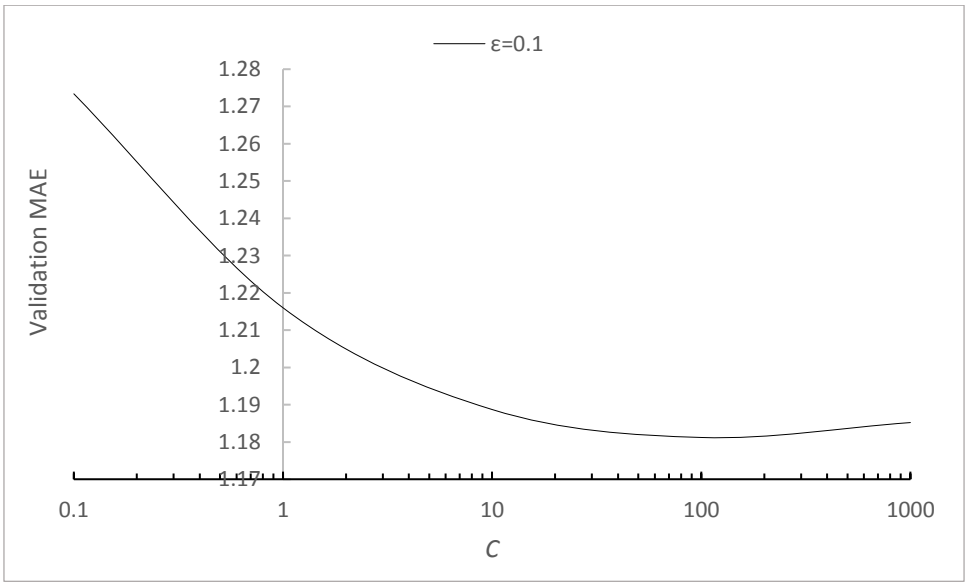


Figure A.10: Validation MAE vs. *C* of VMD-SVR (*a* = 20, *ε* = 0.1)

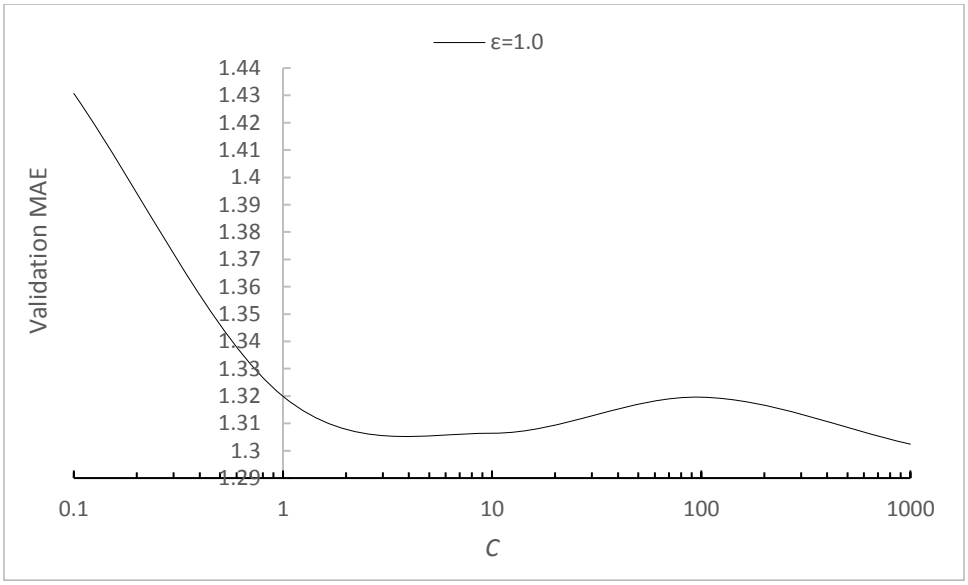


Figure A.11: Validation MAE vs. *C* of VMD-SVR (*a* = 20, *ε* = 1.0)

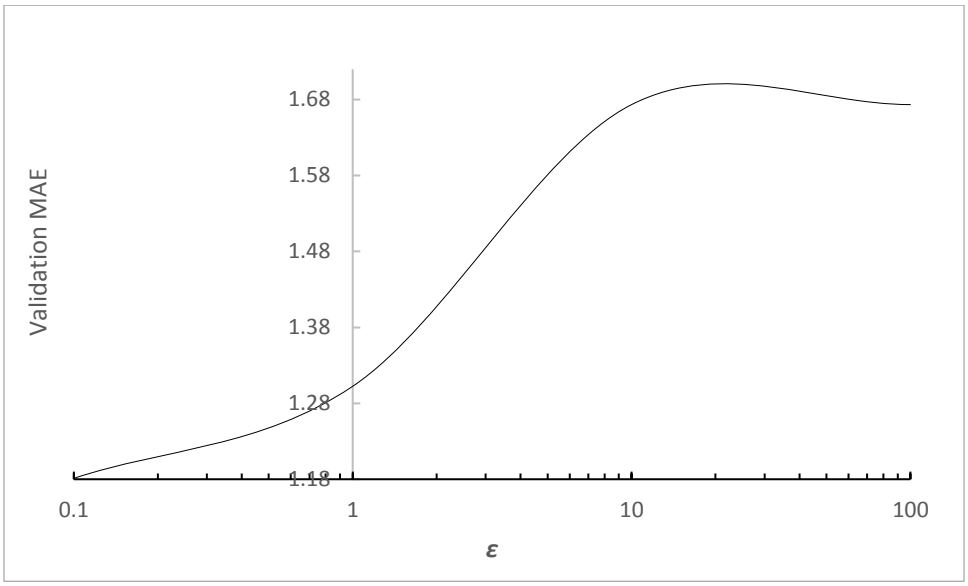


Figure A.12: Validation MAE vs. *ε* of VMD-SVR (*a* = 20)

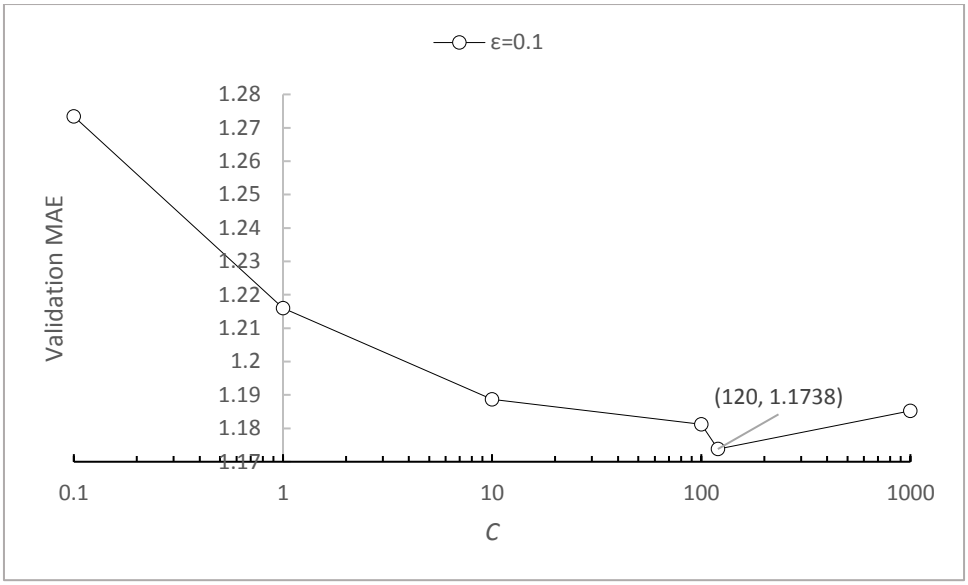


Figure A.13: Validation MAE vs. *C* of VMD-SVR (*a* = 20)

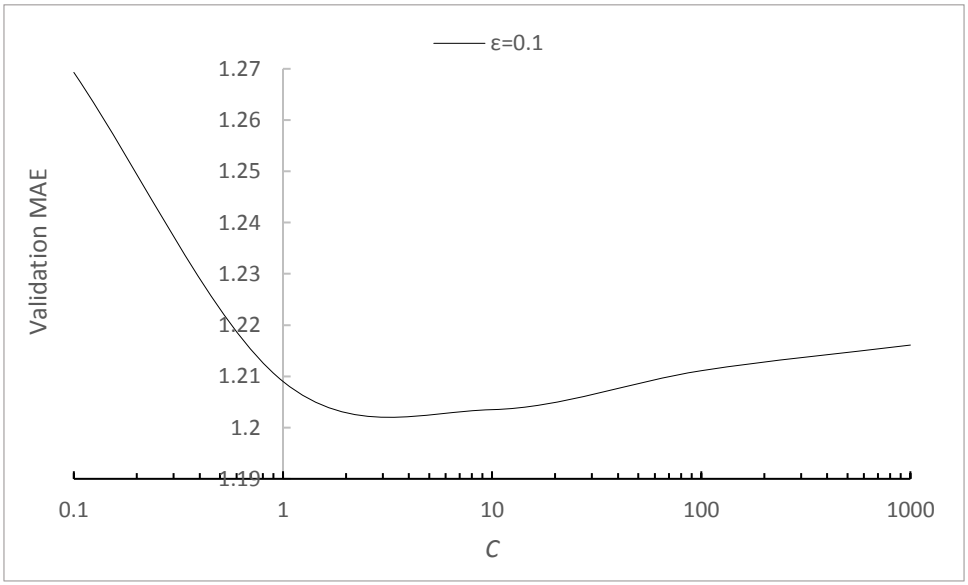


Figure A.14: Validation MAE vs. *C* of VMD-SVR (*a* = 30, *ε* = 0.1)

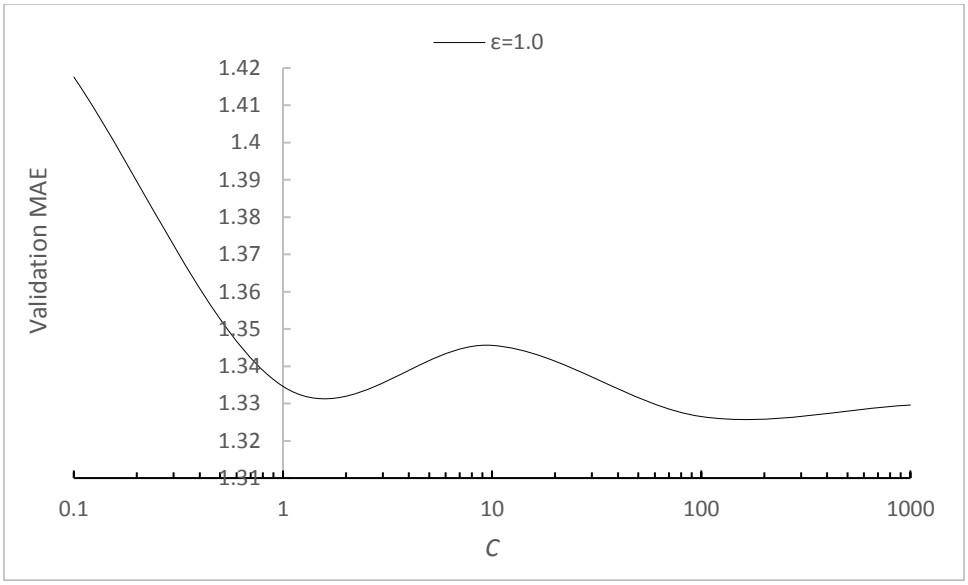


Figure A.15: Validation MAE vs. *C* of VMD-SVR (*a* = 30, *ε* = 1.0)

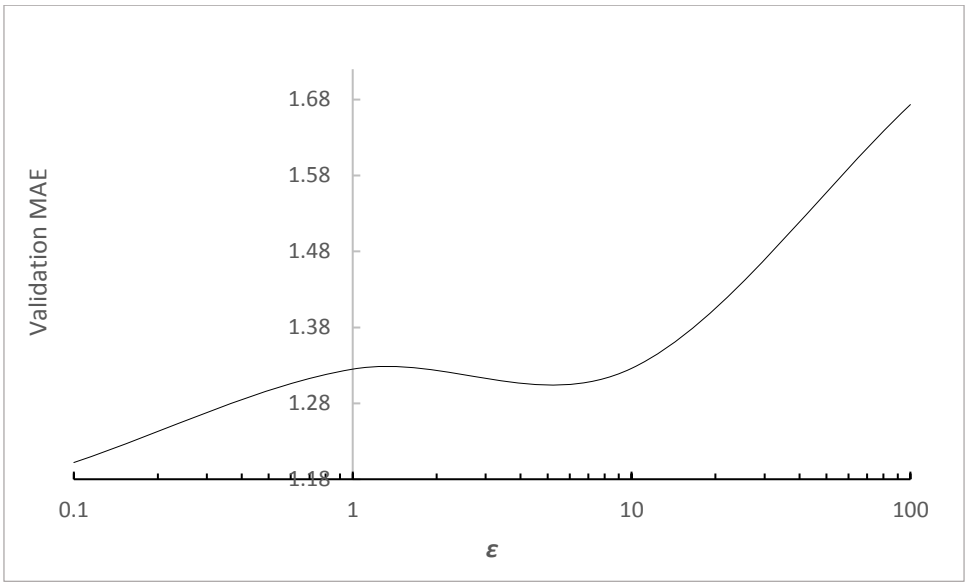


Figure A.16: Validation MAE vs. *ε* of VMD-SVR (*a* = 30)

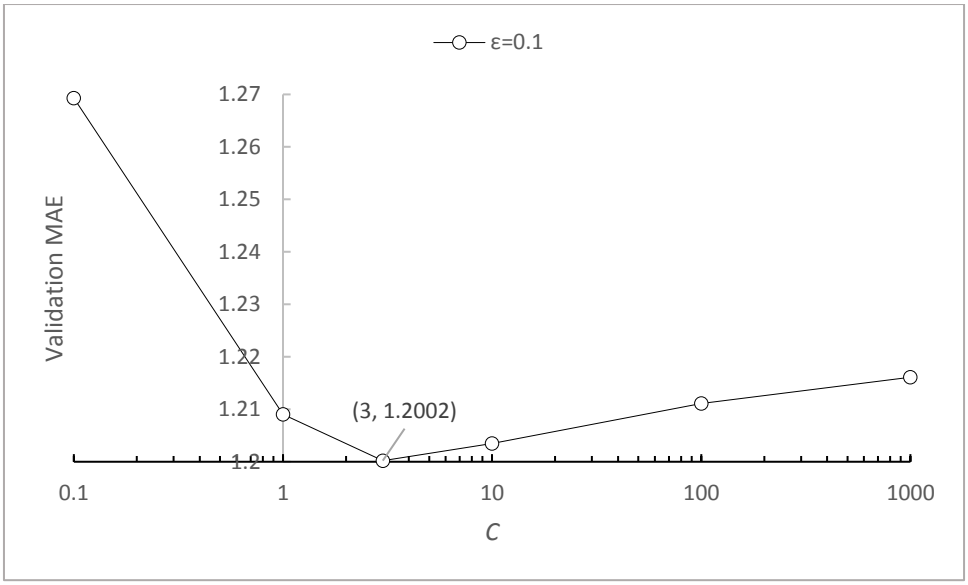


Figure A.17: Validation MAE vs. *C* of VMD-SVR (*a* = 30)

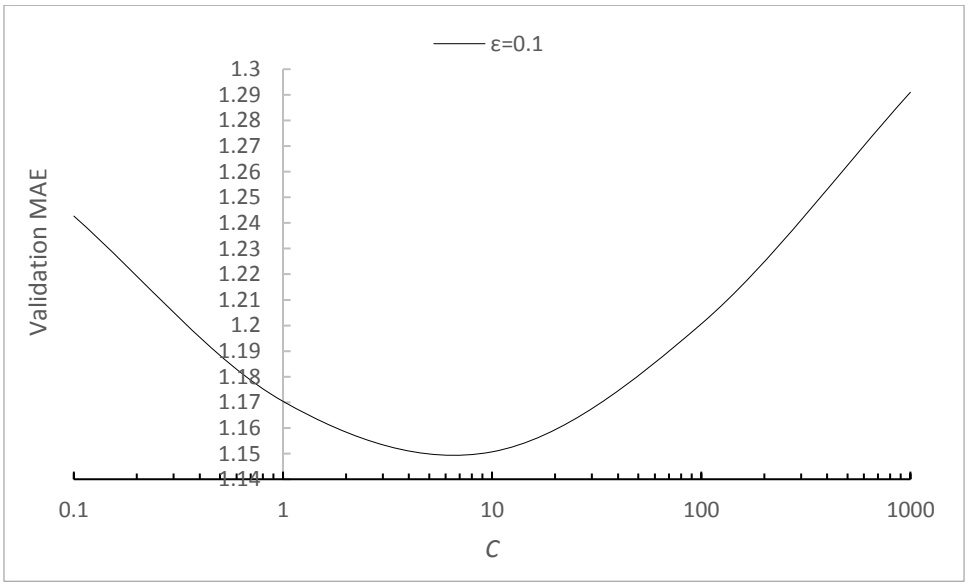


Figure A.18: Validation MAE vs. *C* of VMD-SVR (*a* = 40, *ε* = 0.1)

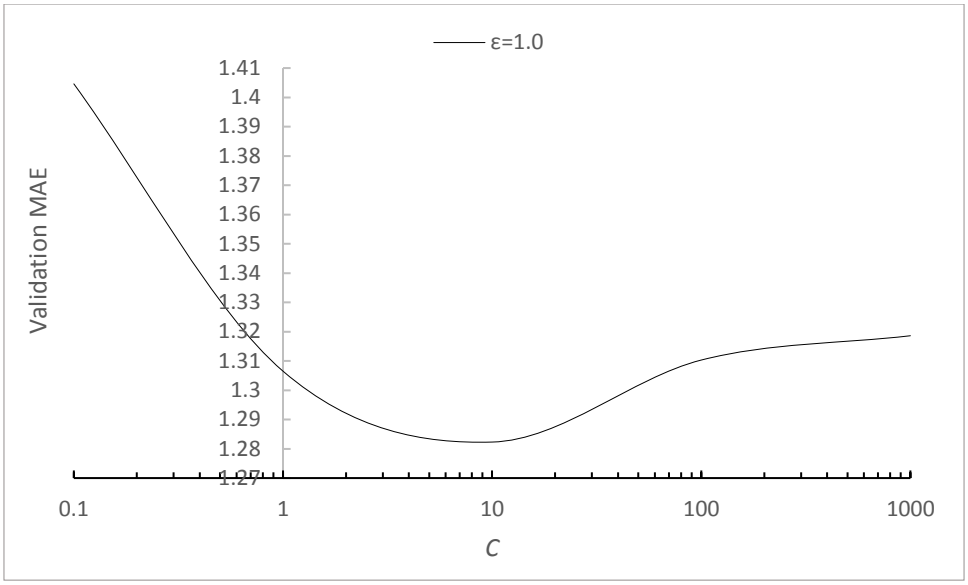


Figure A.19: Validation MAE vs. *C* of VMD-SVR (*a* = 40, *ε* = 1.0)

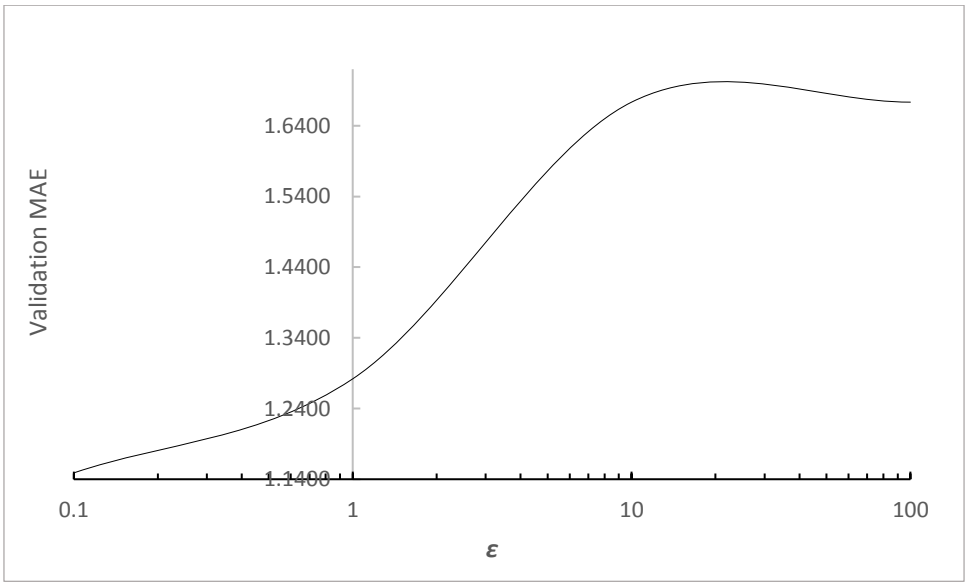


Figure A.20: Validation MAE vs. *ε* of VMD-SVR (*a* = 40)

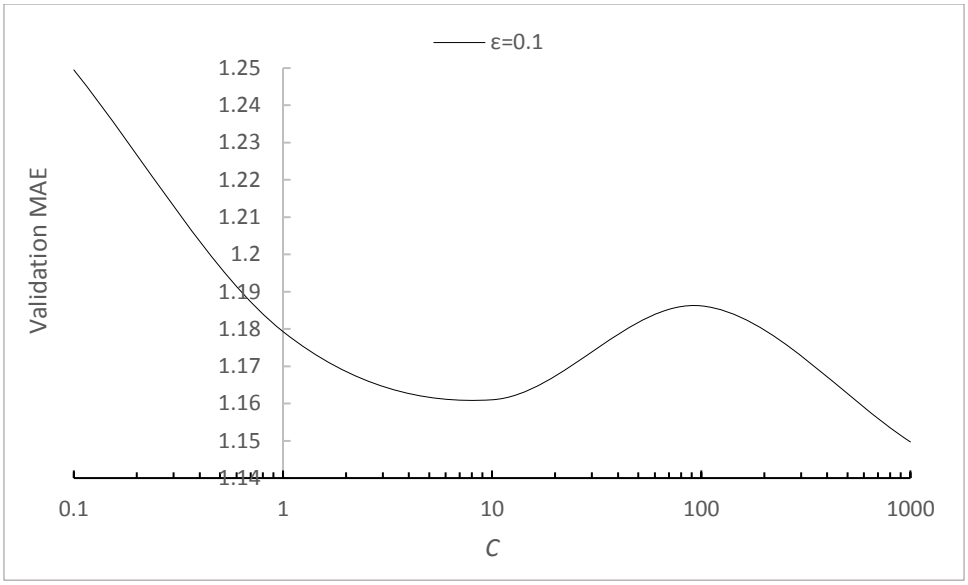


Figure A.21: Validation MAE vs. *C* of VMD-SVR (*a* = 50, *ε* = 0.1)

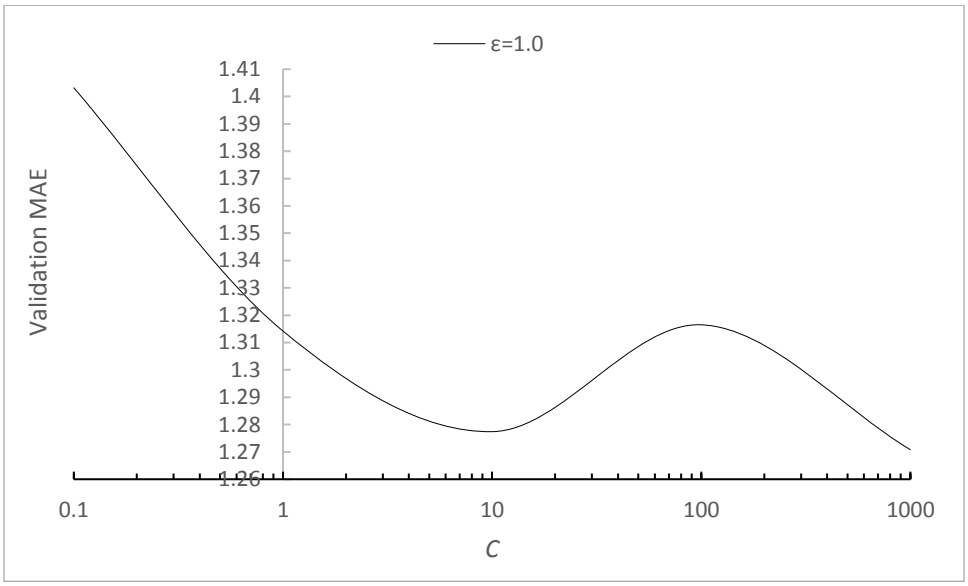


Figure A.22: Validation MAE vs. *C* of VMD-SVR (*a* = 50, *ε* = 1.0)

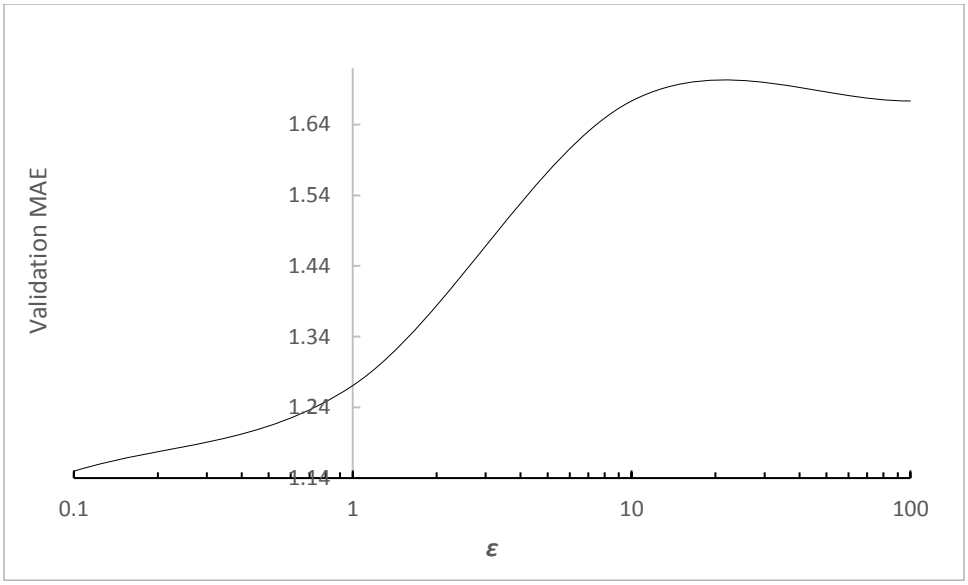


Figure A.23: Validation MAE vs. ε of VMD-SVR (a = 50)

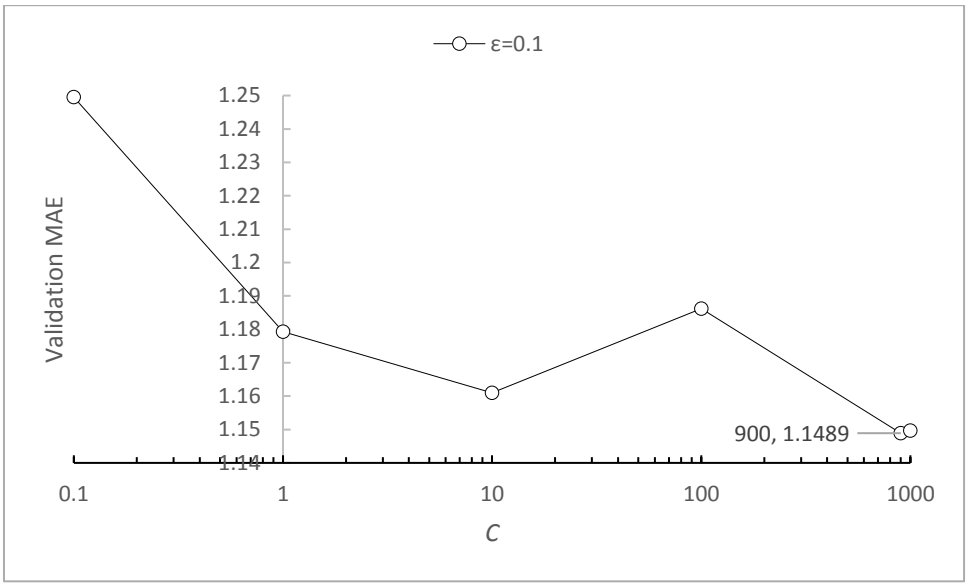


Figure A.24: Validation MAE vs. C of VMD-SVR (a = 50)

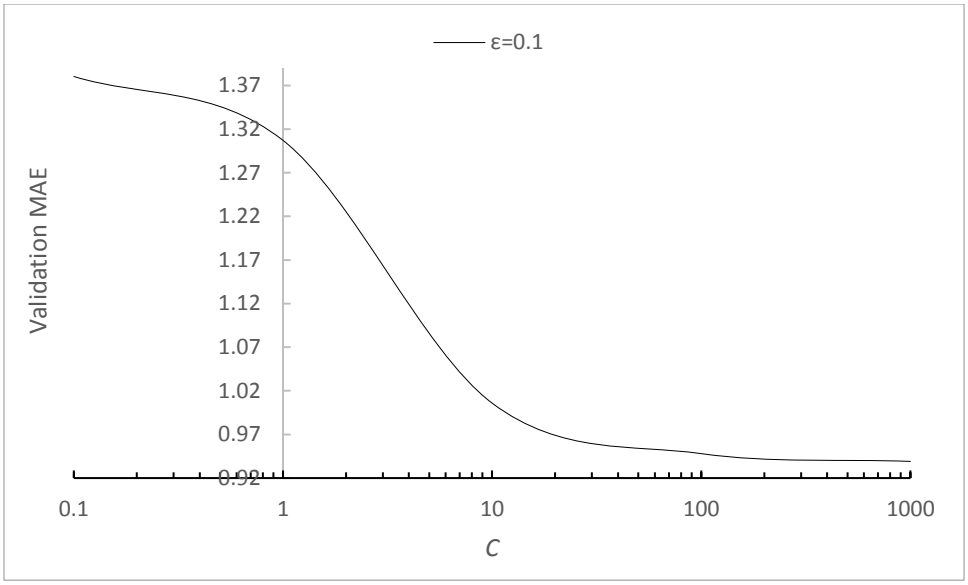


Figure A.25: Validation MAE vs. *C* of B-SVR (NLearn = 1, ε = 0.1)

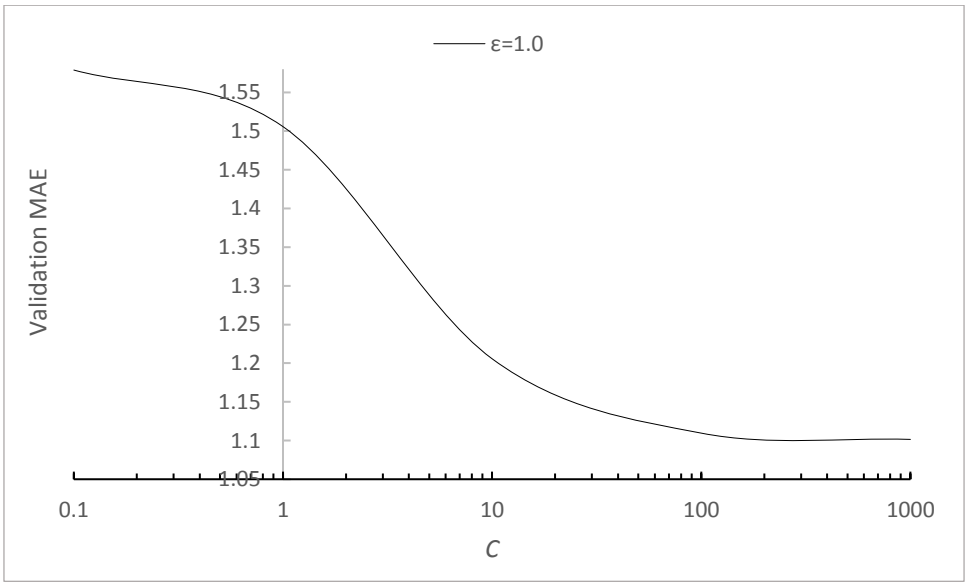


Figure A.26: Validation MAE vs. *C* of B-SVR (NLearn = 1, ε = 1)

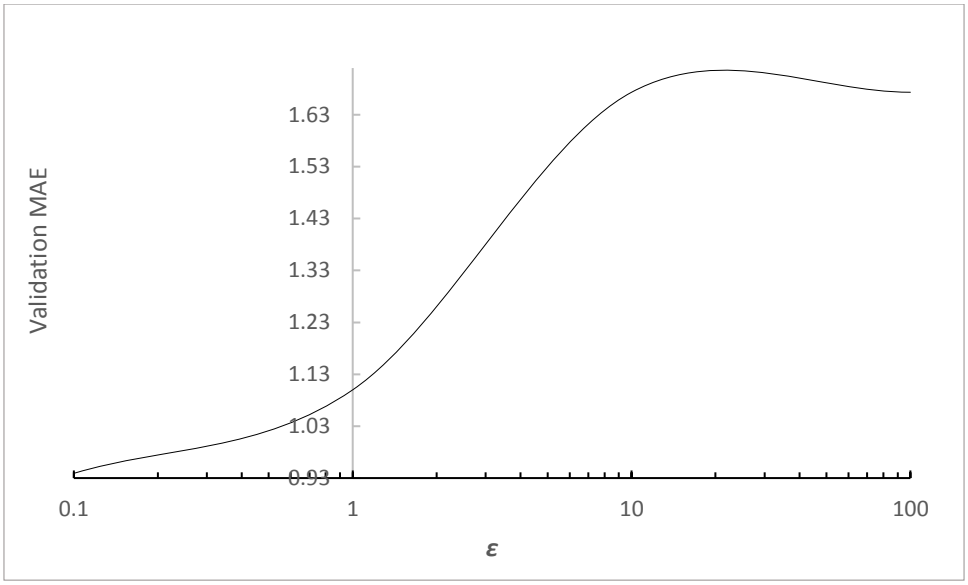


Figure A.27: Validation MAE vs. ε of B-SVR (NLearn = 1)

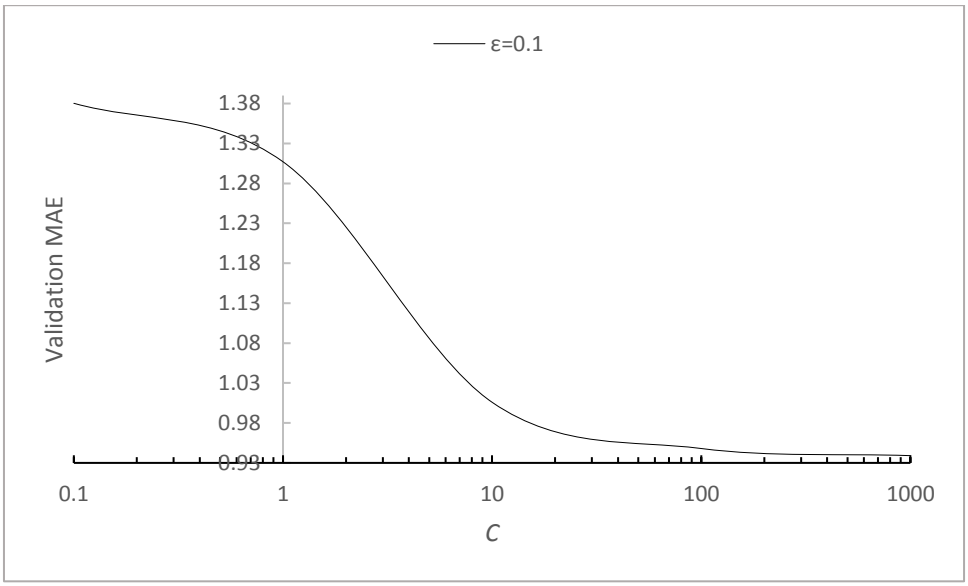


Figure A.28: Validation MAE vs. *C* of B-SVR (NLearn = 1)

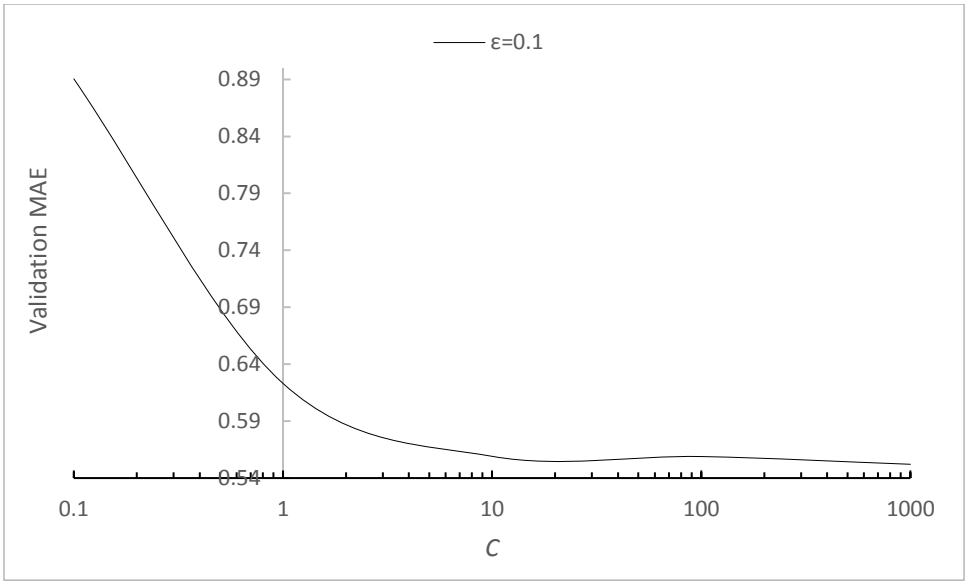


Figure A.29: Validation MAE vs. *C* of B-SVR (NLearn = 10, ε = 0.1)

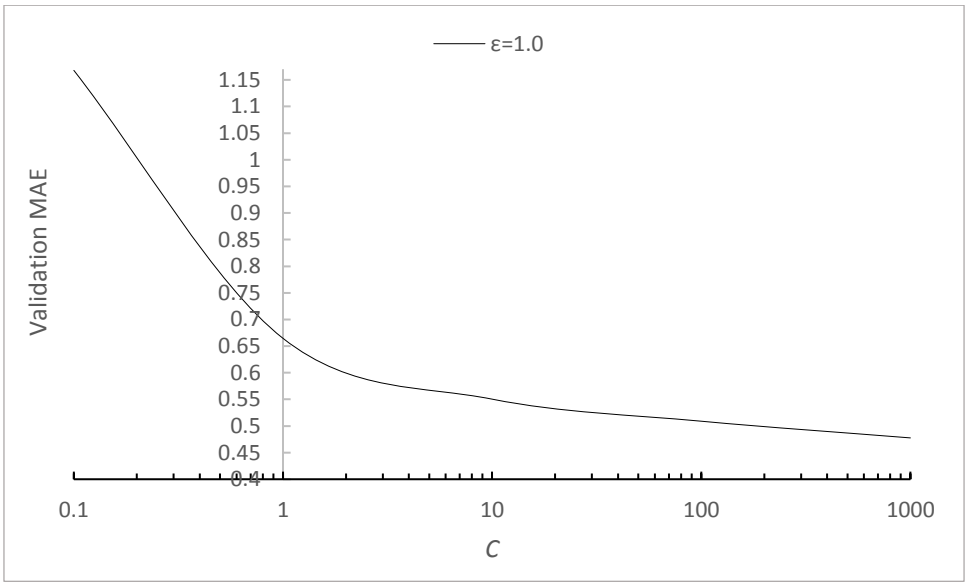


Figure A.30: Validation MAE vs. *C* of B-SVR (NLearn = 10, ε = 1)

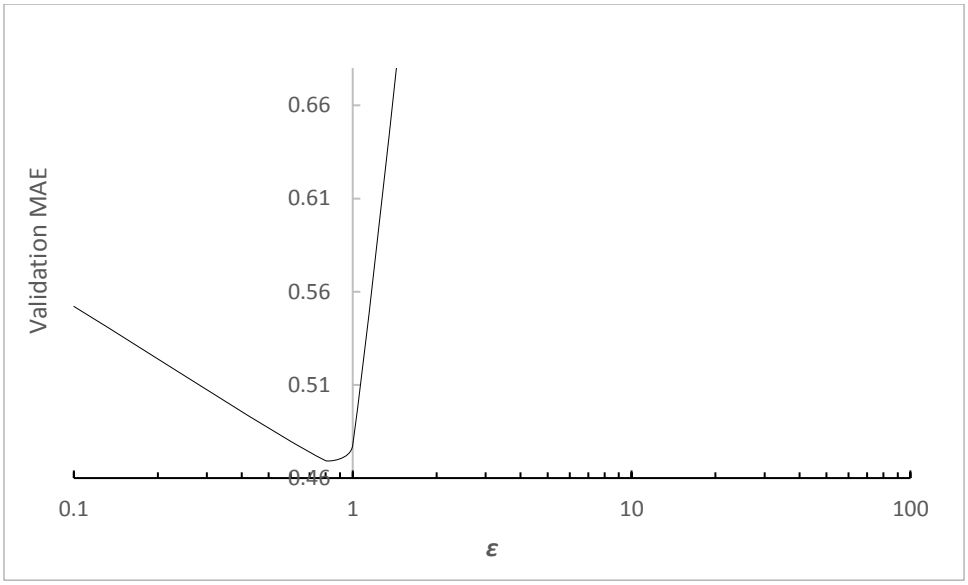


Figure A.31: Validation MAE vs. ε of B-SVR (NLearn = 10)

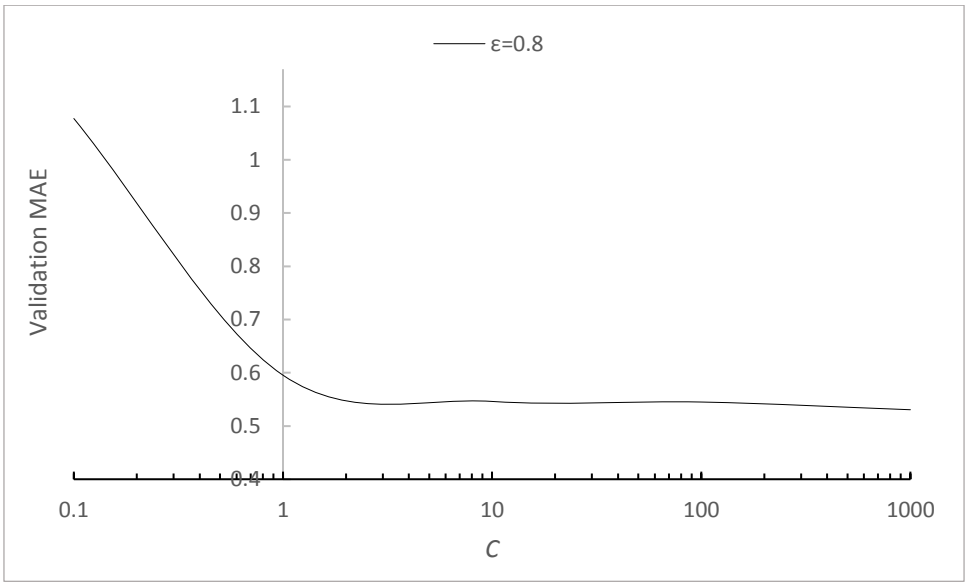


Figure A.32: Validation MAE vs. *C* of B-SVR (NLearn = 10, ε = 0.8)

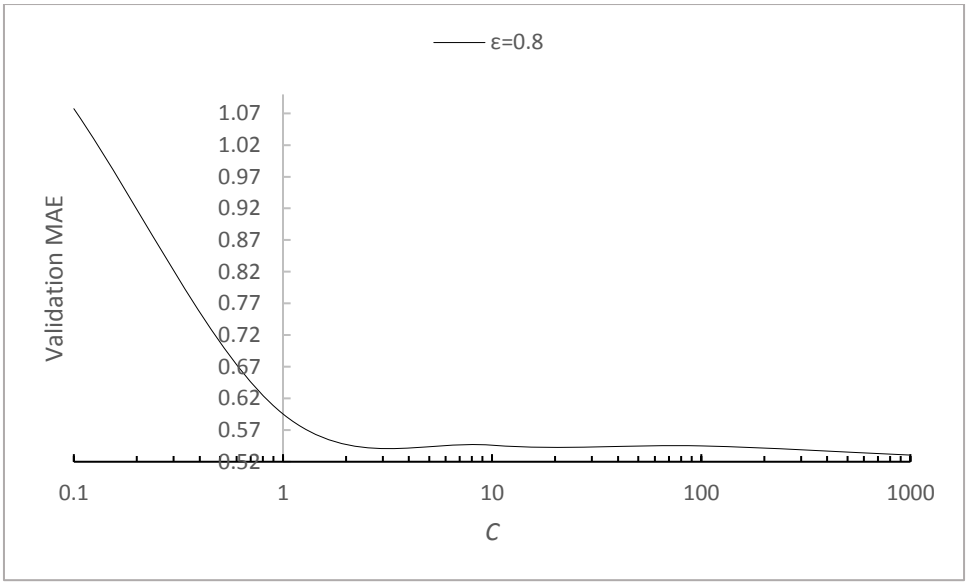


Figure A.33: Validation MAE vs. *C* of B-SVR (NLearn = 10)

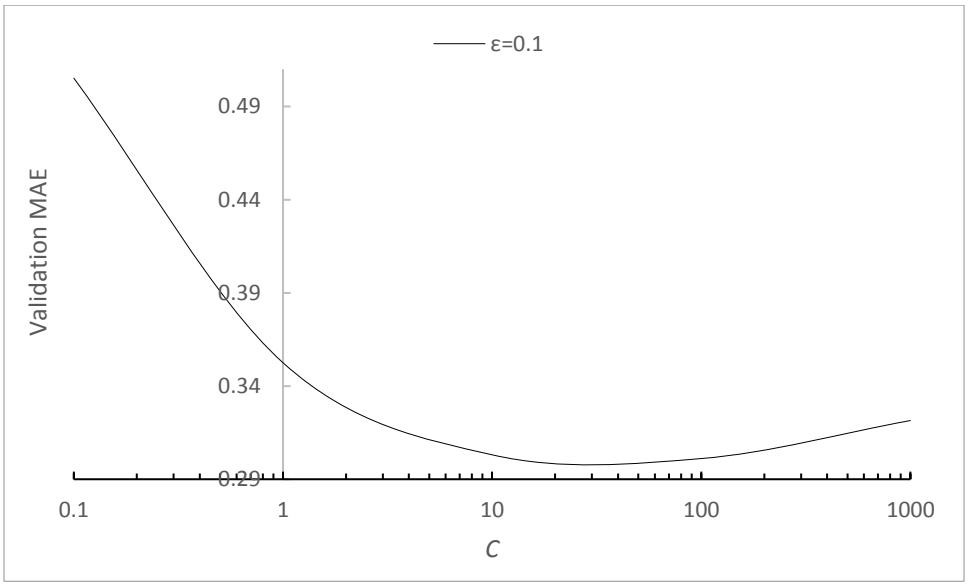


Figure A.34: Validation MAE vs. *C* of B-SVR (NLearn = 100, ε = 0.1)

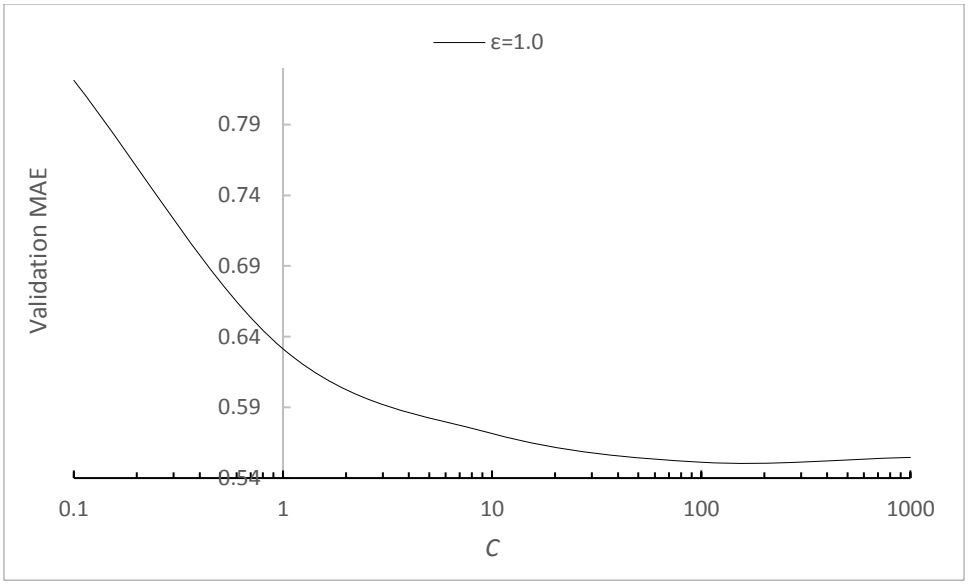


Figure A.35: Validation MAE vs. *C* of B-SVR (NLearn = 100, ε = 1)

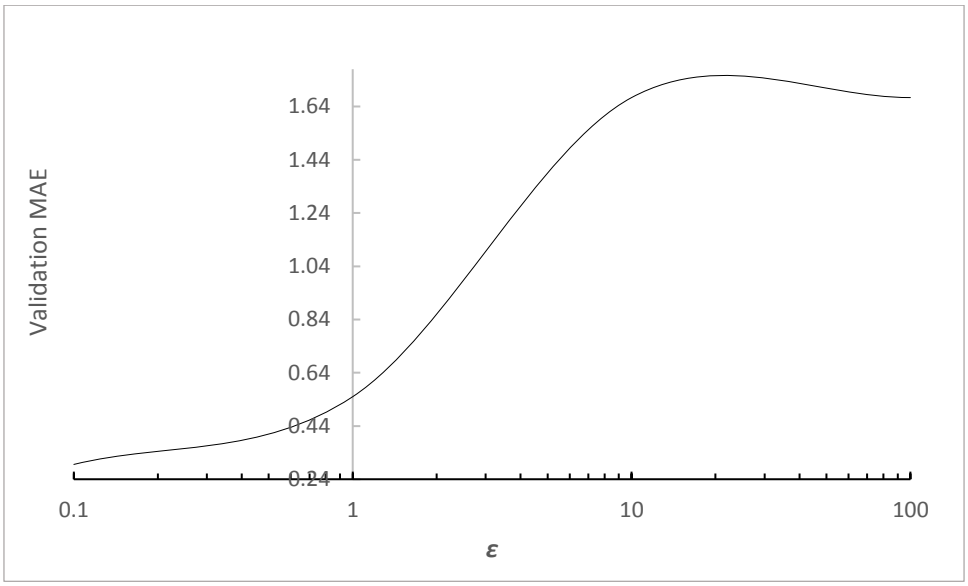


Figure A.36: Validation MAE vs. ε of B-SVR (NLearn = 100)

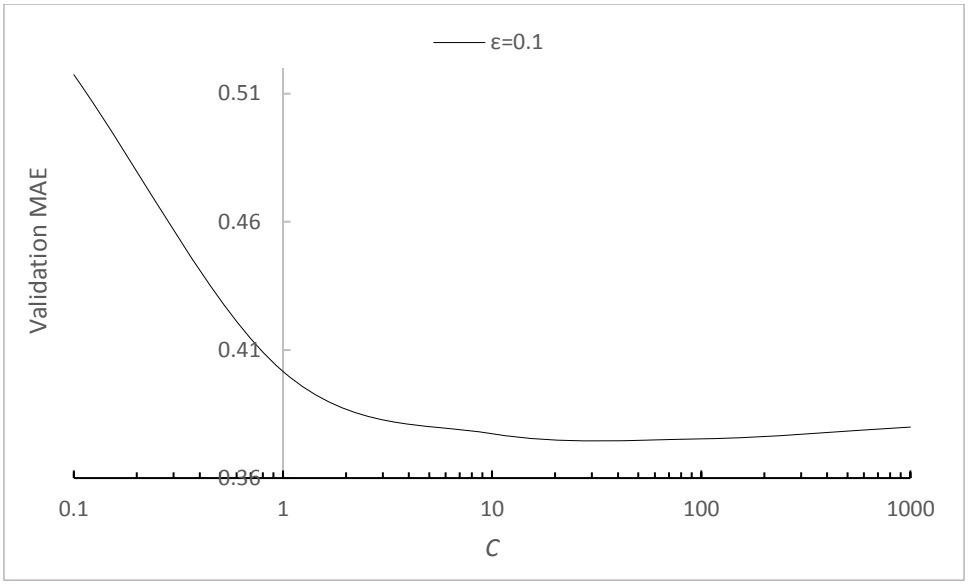


Figure A.37: Validation MAE vs. *C* of B-SVR (NLearn = 1,000, ε = 0.1)

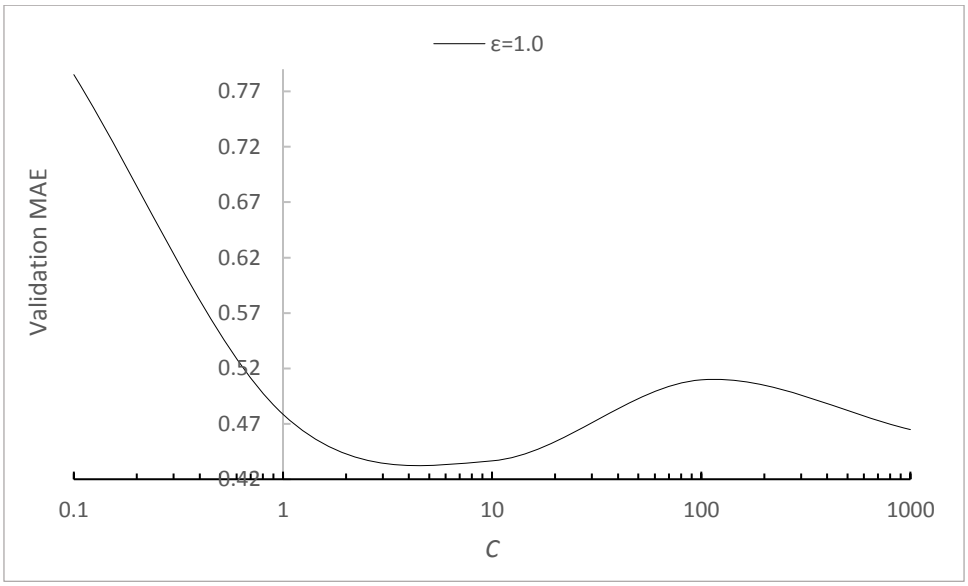


Figure A.38: Validation MAE vs. *C* of B-SVR (NLearn = 1,000, ε = 1)

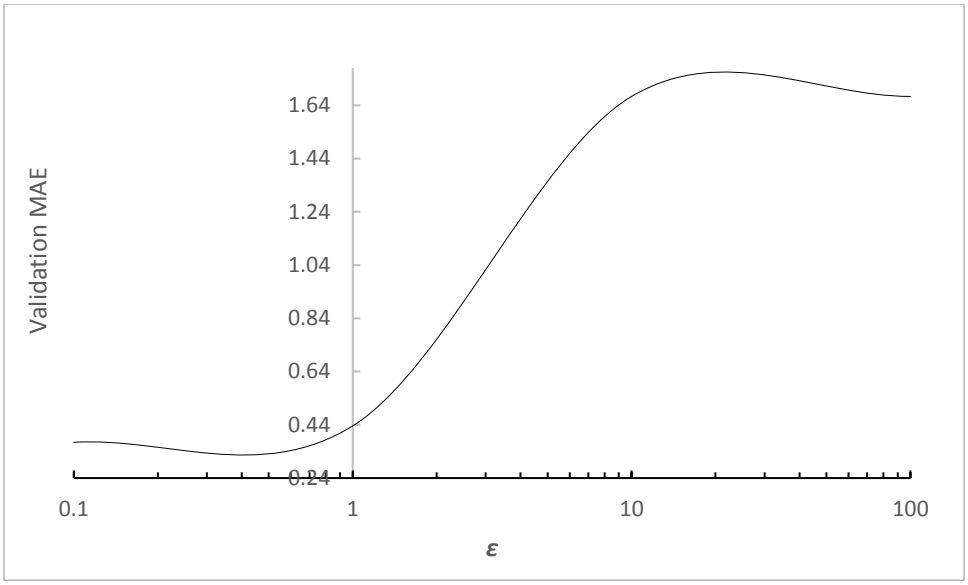


Figure A.39: Validation MAE vs. ε of B-SVR (NLearn = 1,000)

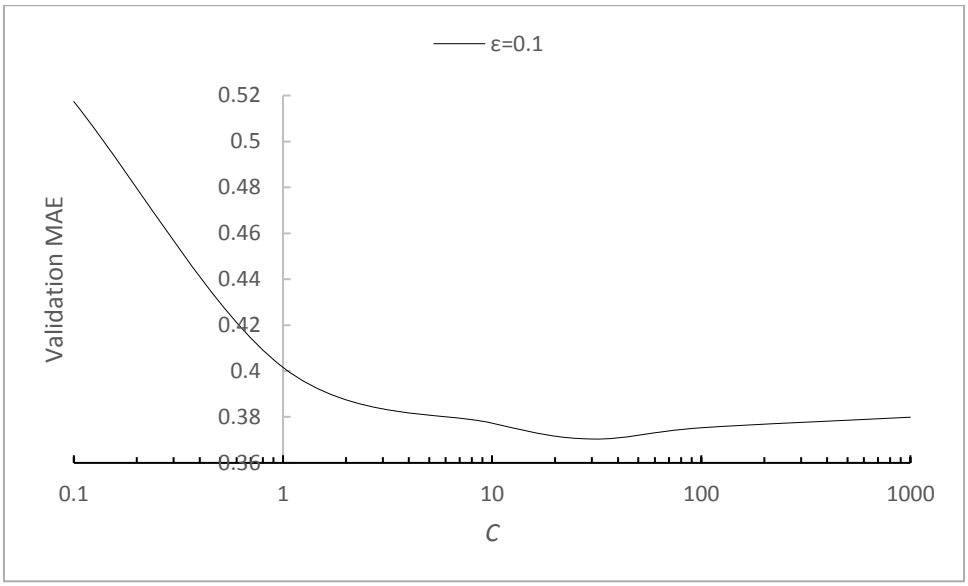


Figure A.40: Validation MAE vs. *C* of B-SVR (NLearn = 1,000)

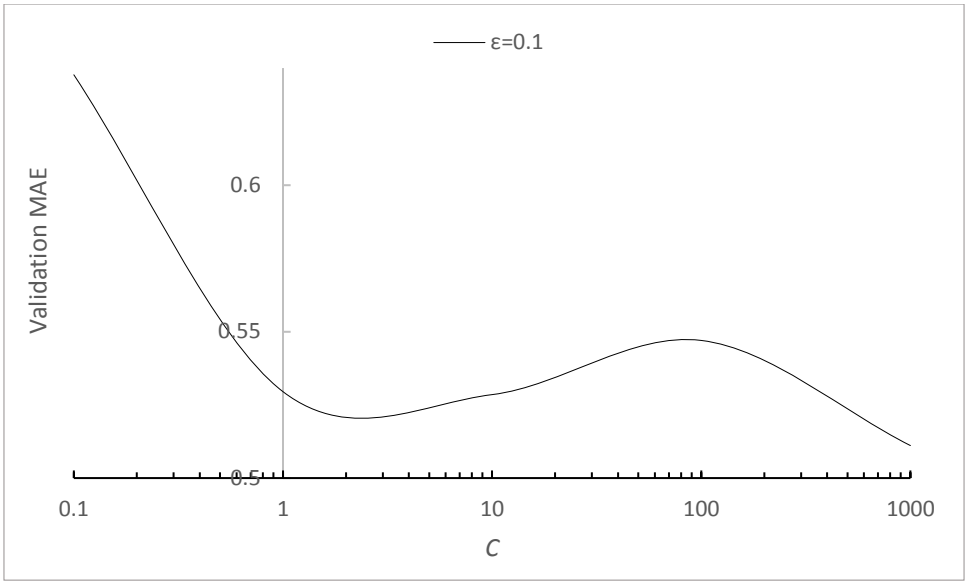


Figure A.41: Validation MAE vs. *C* of B-SVR (NLearn = 10,000, ε = 0.1)

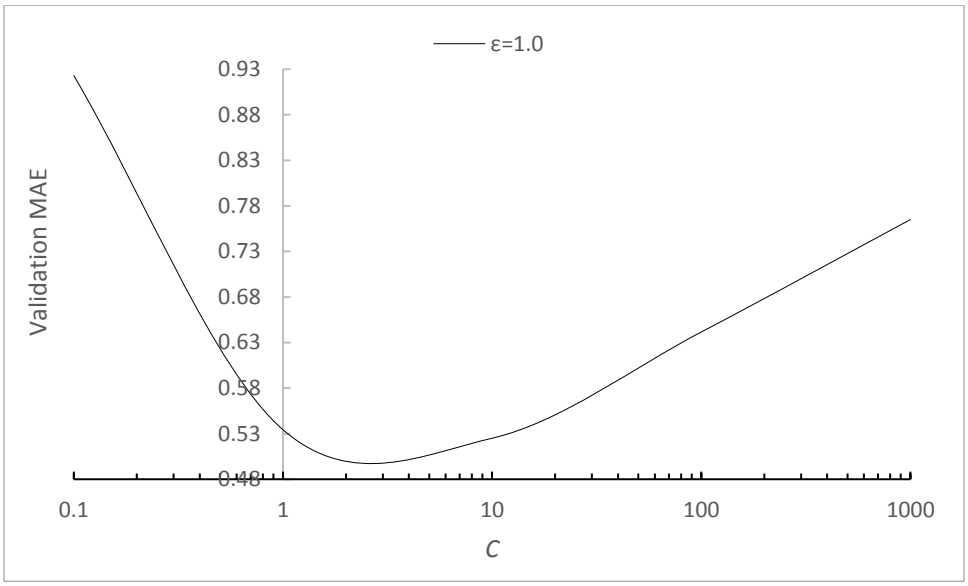


Figure A.42: Validation MAE vs. *C* of B-SVR (NLearn = 10,000, ε = 1)

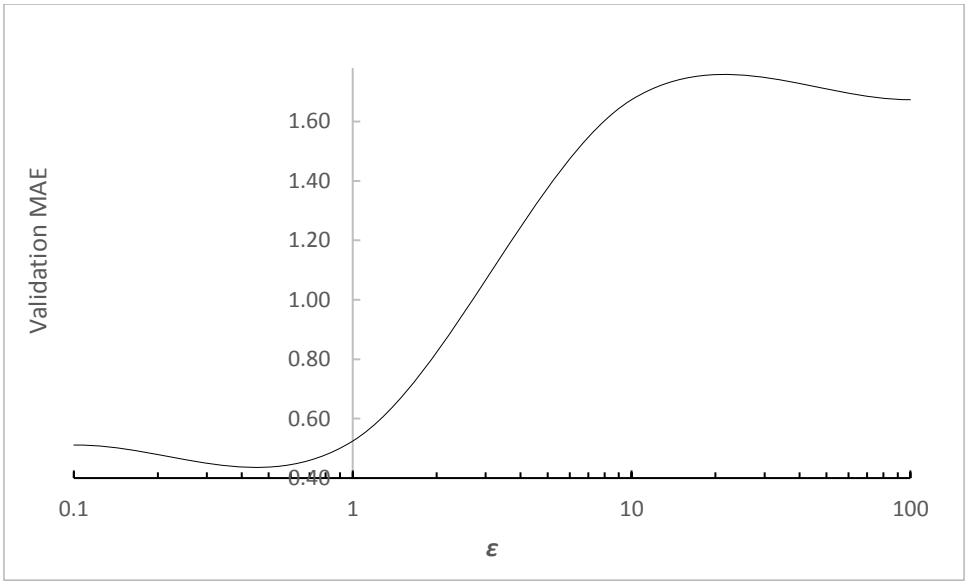


Figure A.43: Validation MAE vs. ε of B-SVR (NLearn = 10,000)

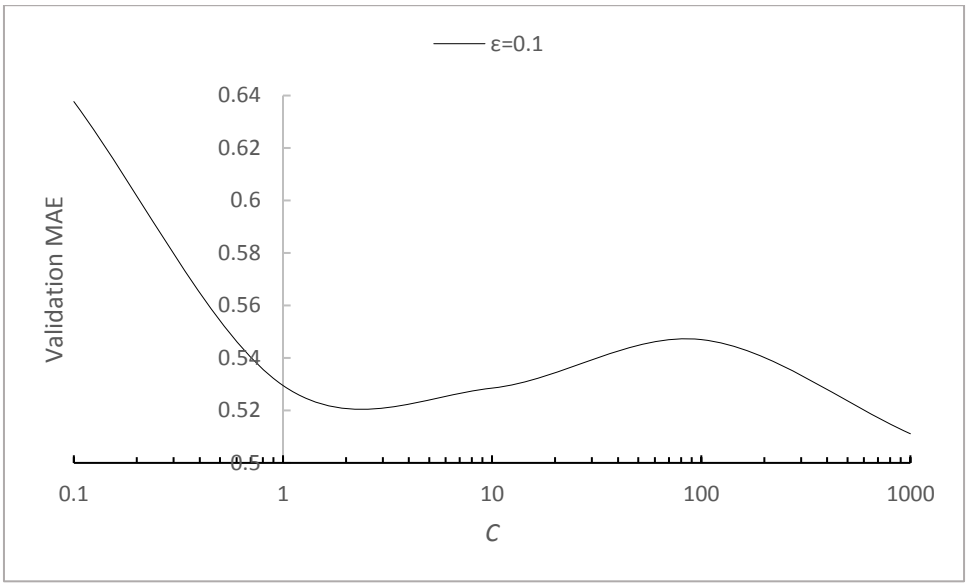


Figure A.44: Validation MAE vs. *C* of B-SVR (NLearn = 10,000)

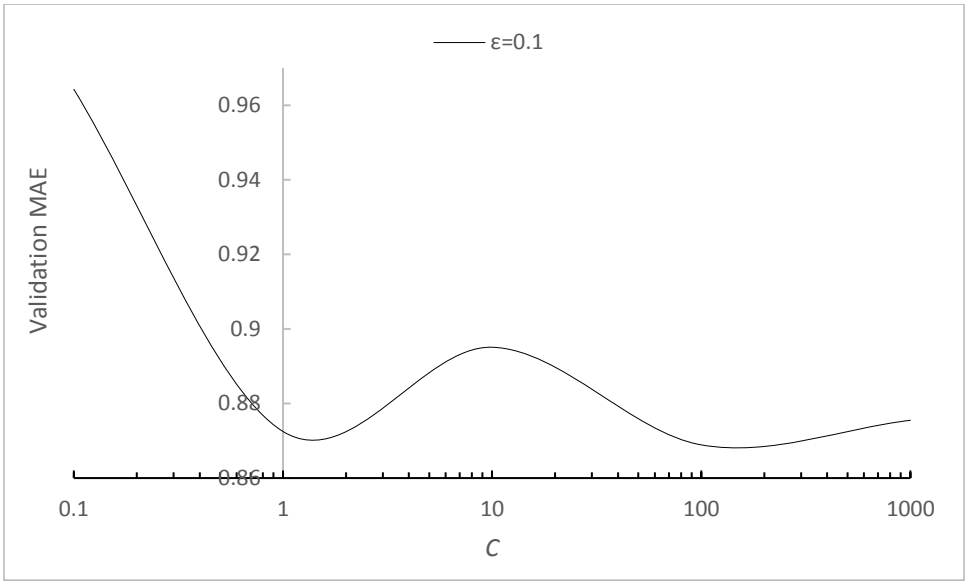


Figure A.45: Validation MAE vs. *C* of B-SVR (NLearn = 100,000, ε = 0.1)

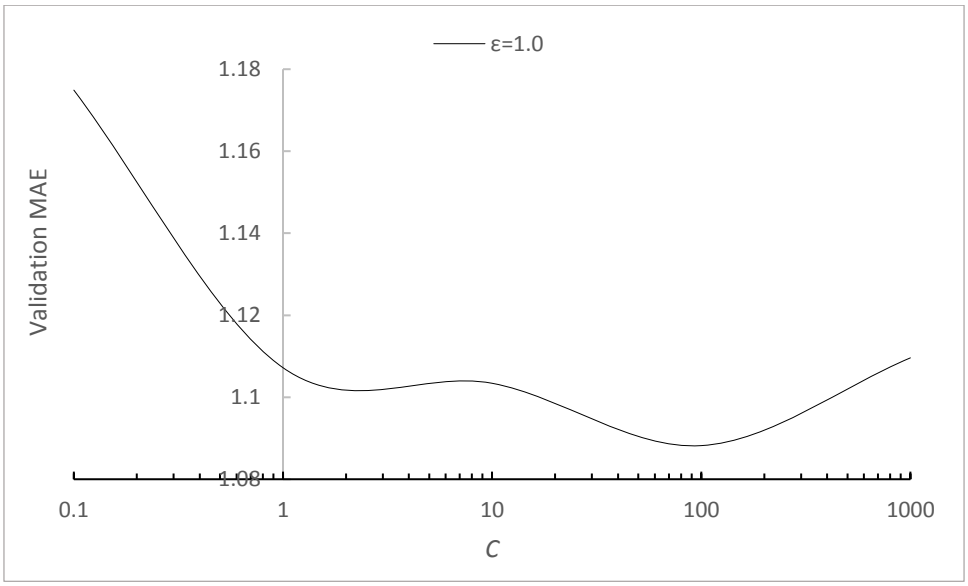


Figure A.46: Validation MAE vs. *C* of B-SVR (NLearn = 100,000, ε = 1)

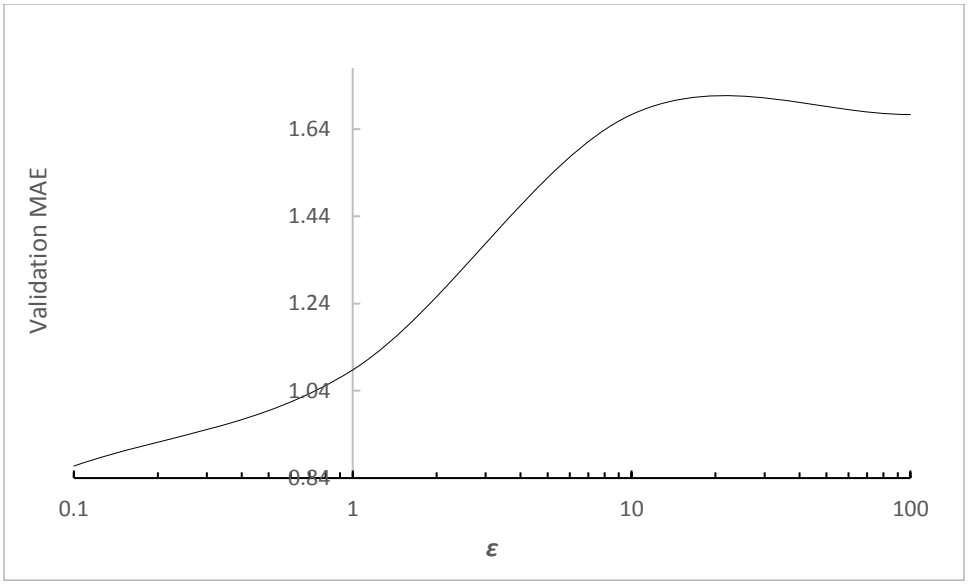


Figure A.47: Validation MAE vs. ε of B-SVR (NLearn = 100,000)

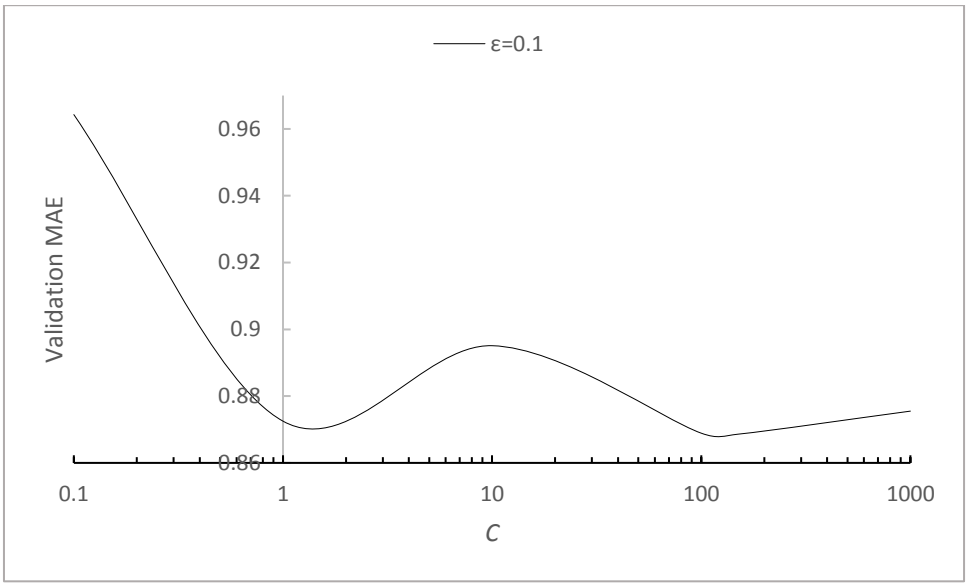


Figure A.48: Validation MAE vs. *C* of B-SVR (NLearn = 100,000)

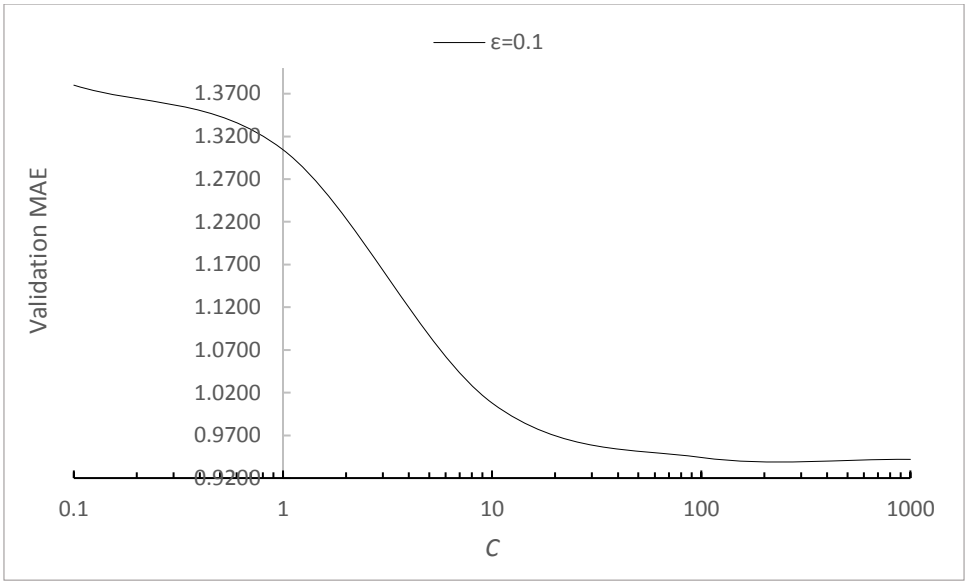


Figure A.49: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1, ε = 0.1)

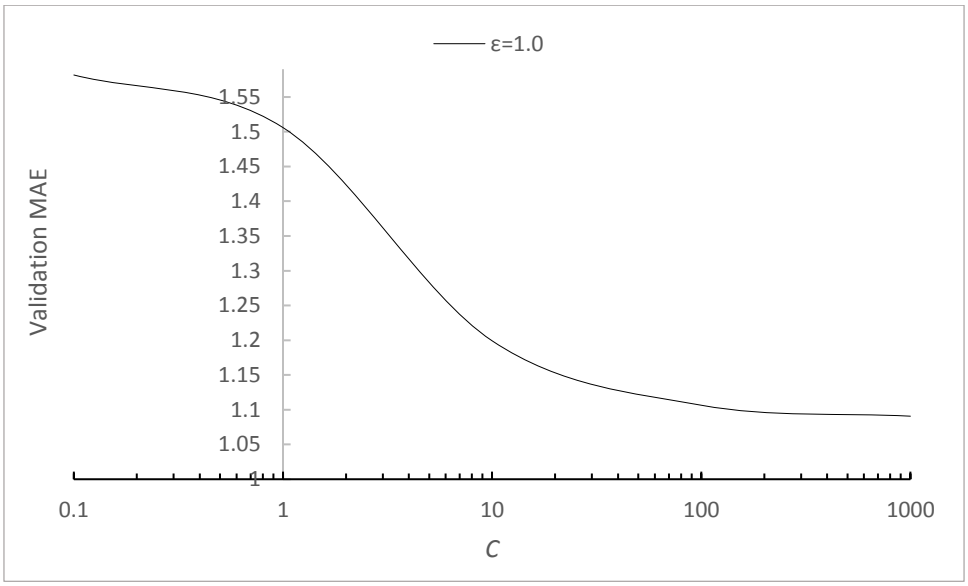


Figure A.50: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1, ε = 1)

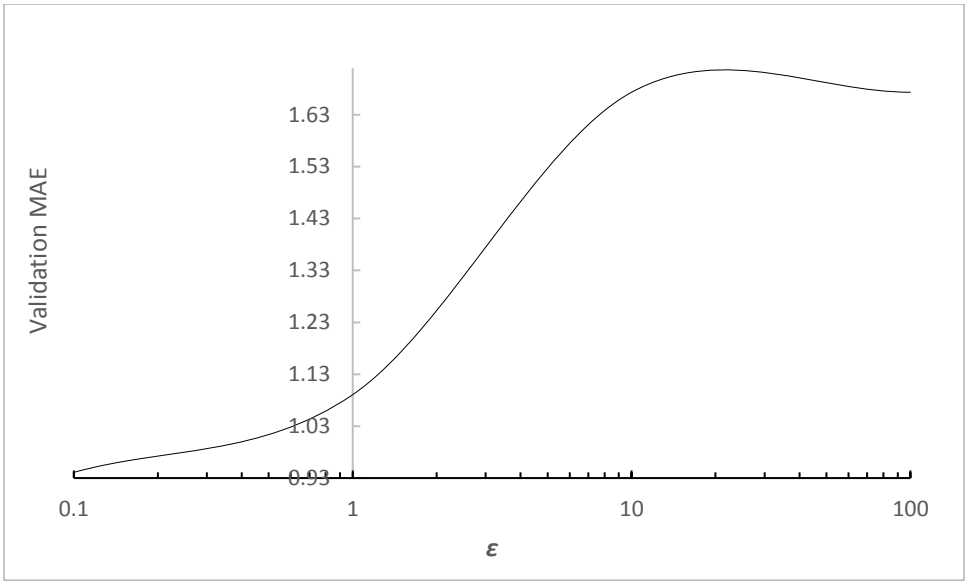


Figure A.51: Validation MAE vs. ε of B-VMD-SVR (NLearn = 1)

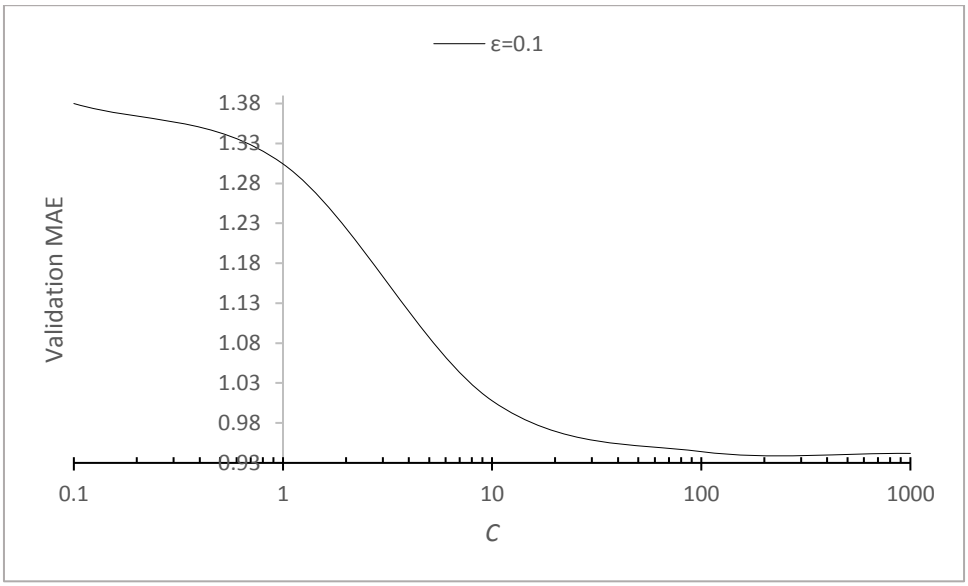


Figure A.52: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1)

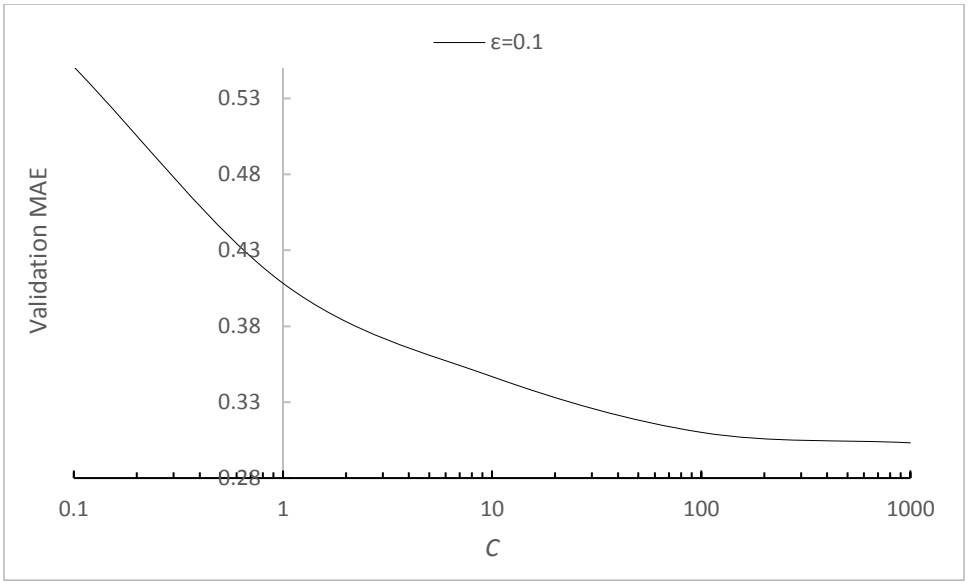


Figure A.53: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10, ε = 0.1)

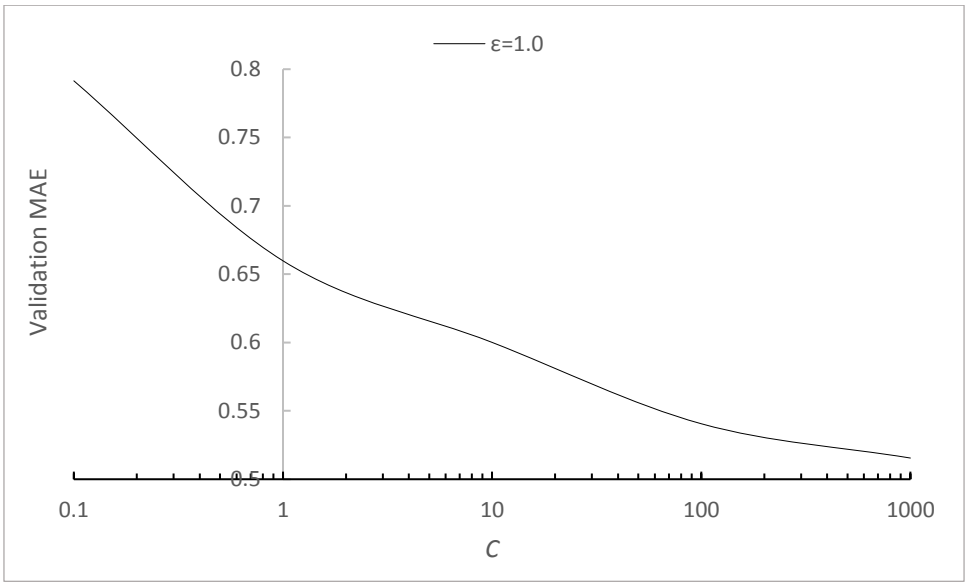


Figure A.54: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10, ε = 1)

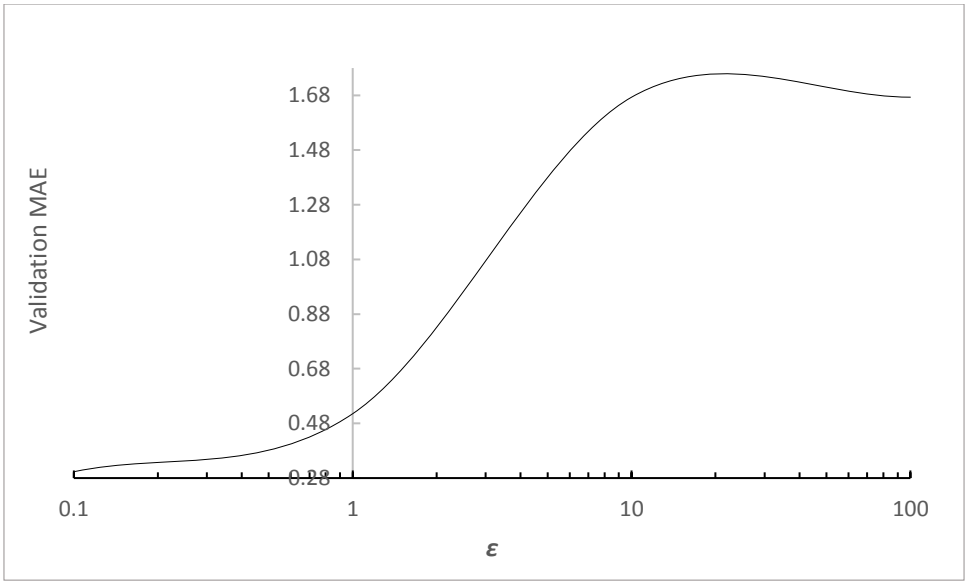


Figure A.55: Validation MAE vs. ε of B-VMD-SVR (NLearn = 10)

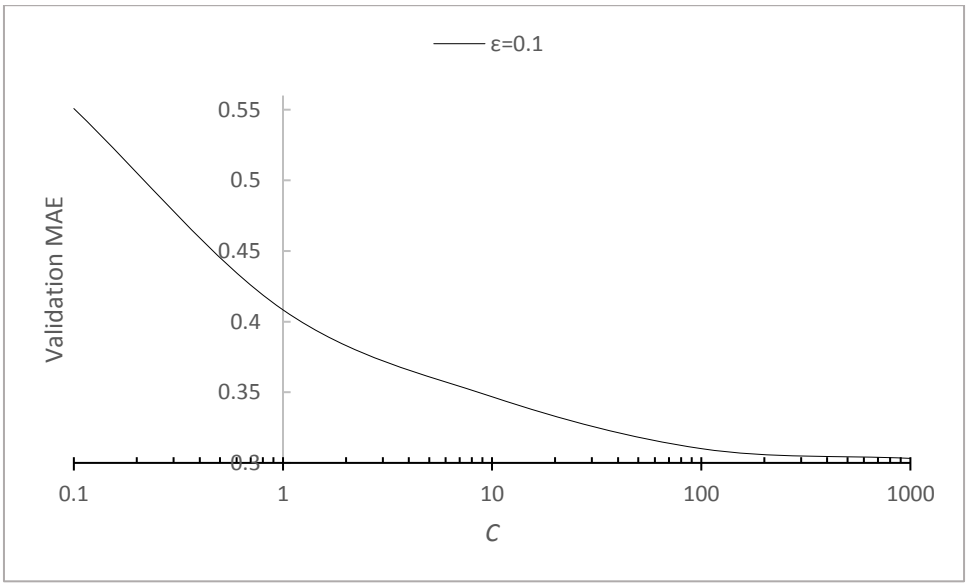


Figure A.56: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10)

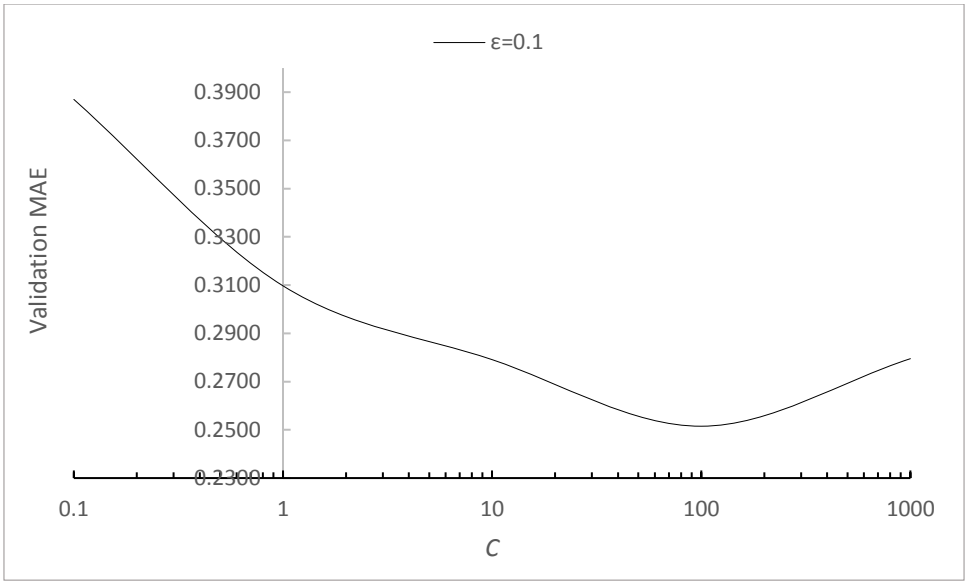


Figure A.57: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 100, ε = 0.1)

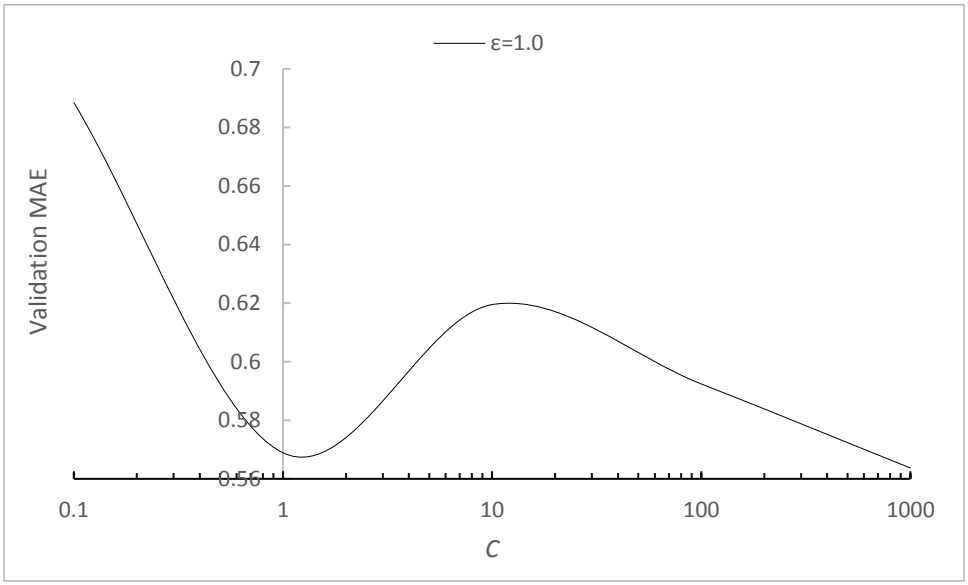


Figure A.58: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 100, ε = 1)

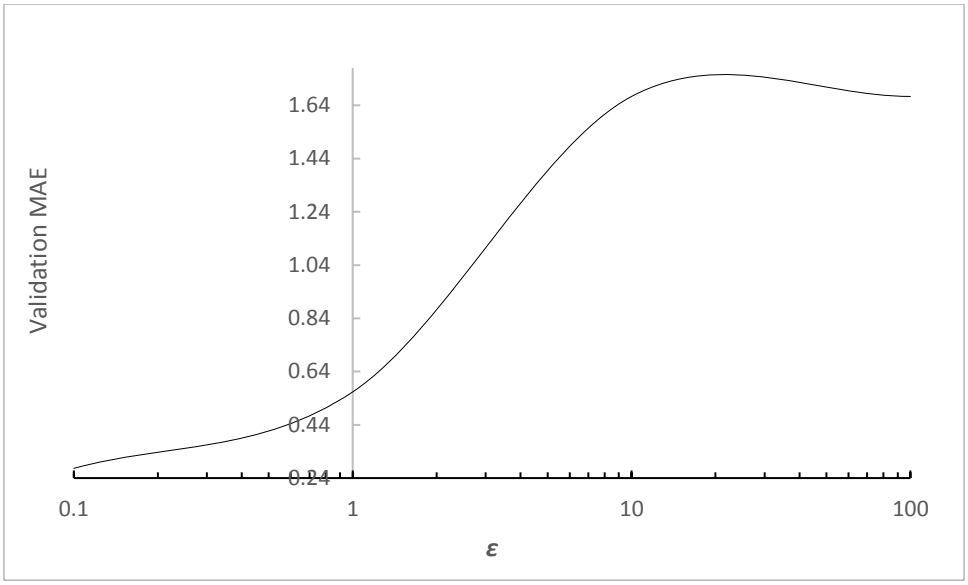


Figure A.59: Validation MAE vs. ε of B-VMD-SVR (NLearn = 100)

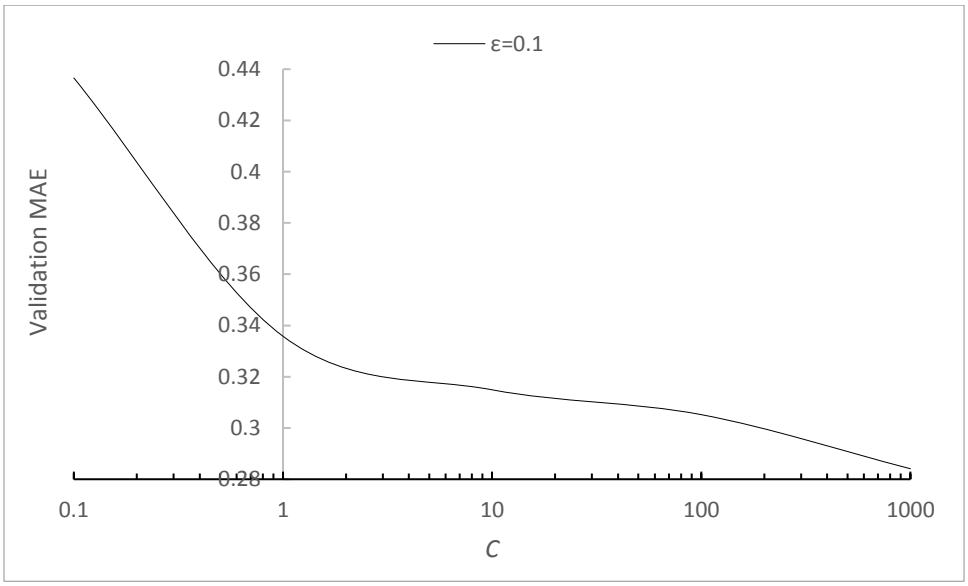


Figure A.60: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1,000, ε = 0.1)

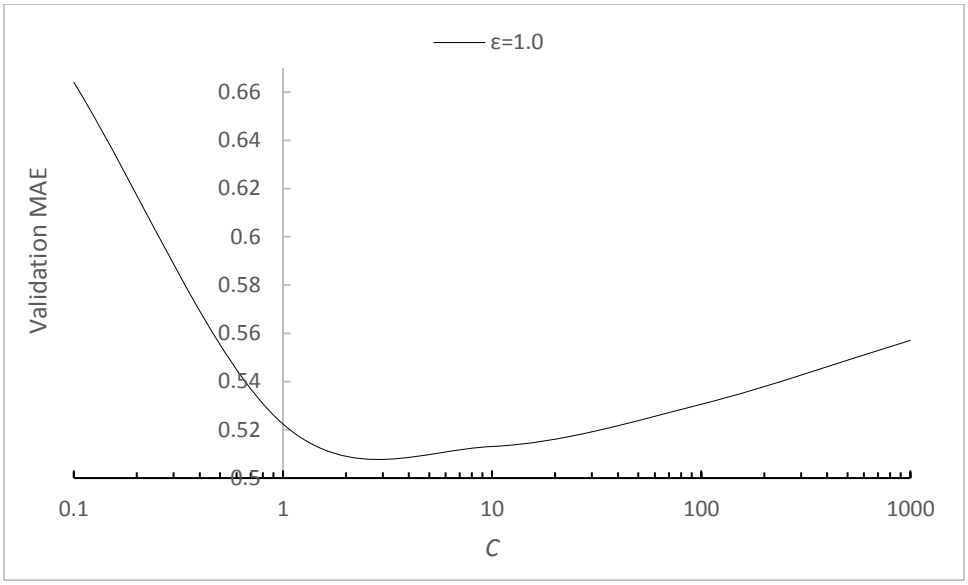


Figure A.61: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1,000, ε = 1)

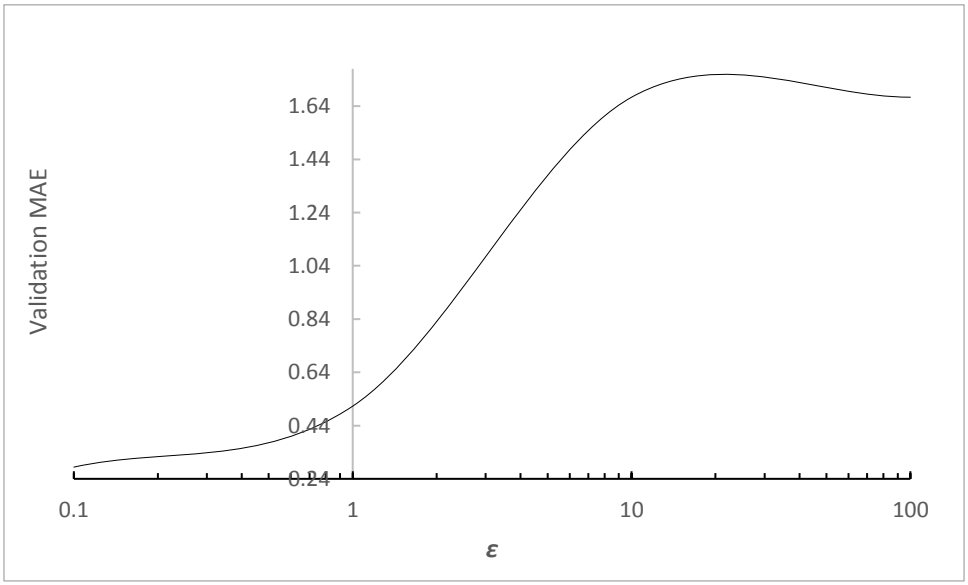


Figure A.62: Validation MAE vs. ε of B-VMD-SVR (NLearn = 1,000)

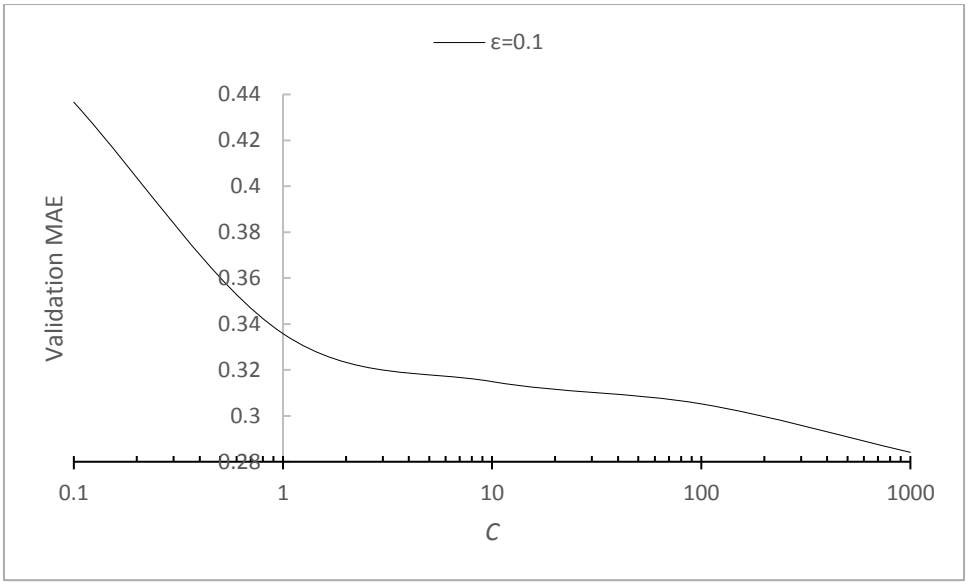


Figure A.63: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 1,000)

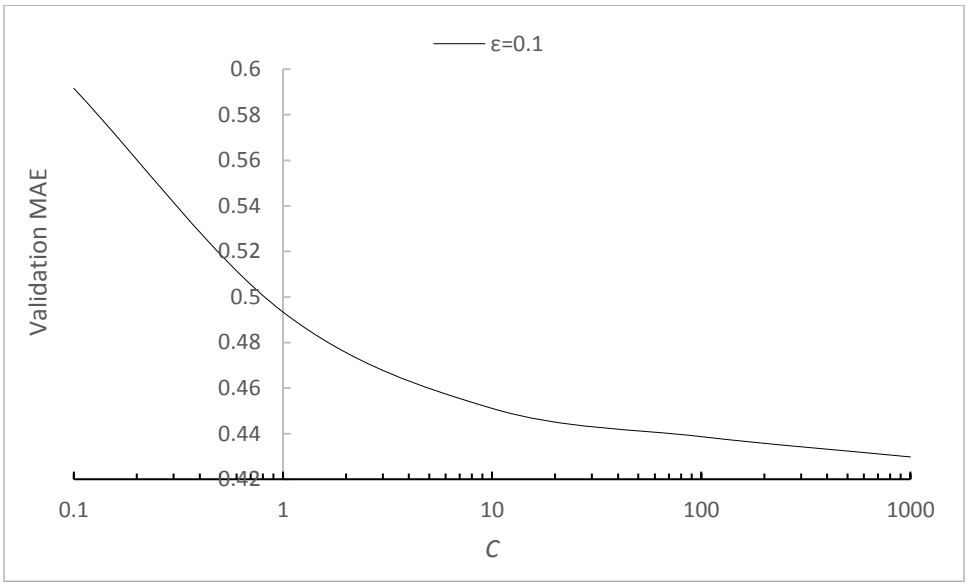


Figure A.64: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10,000, ε = 0.1)

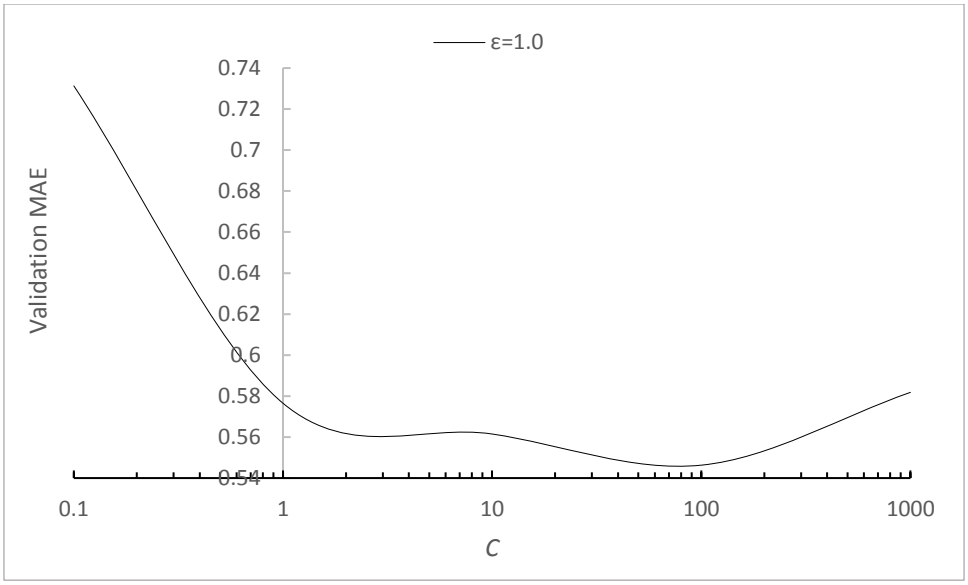


Figure A.65: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10,000, *ε* = 1)

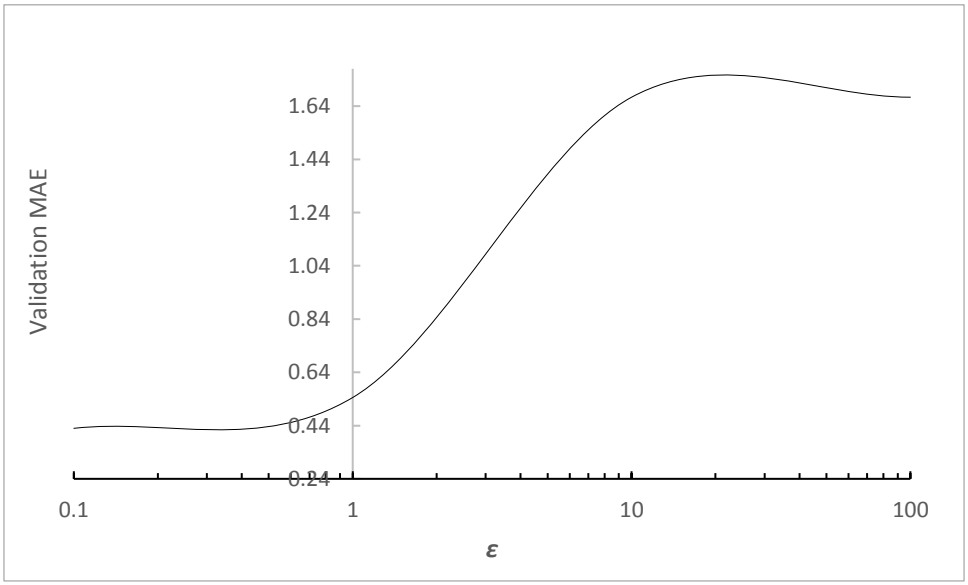


Figure A.66: Validation MAE vs. *ε* of B-VMD-SVR (NLearn = 10,000)

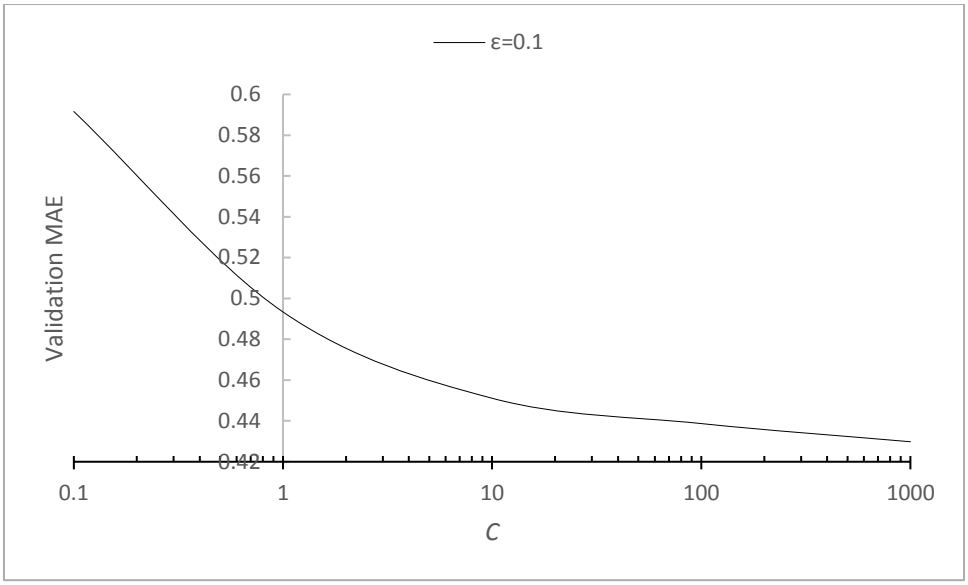


Figure A.67: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 10,000)

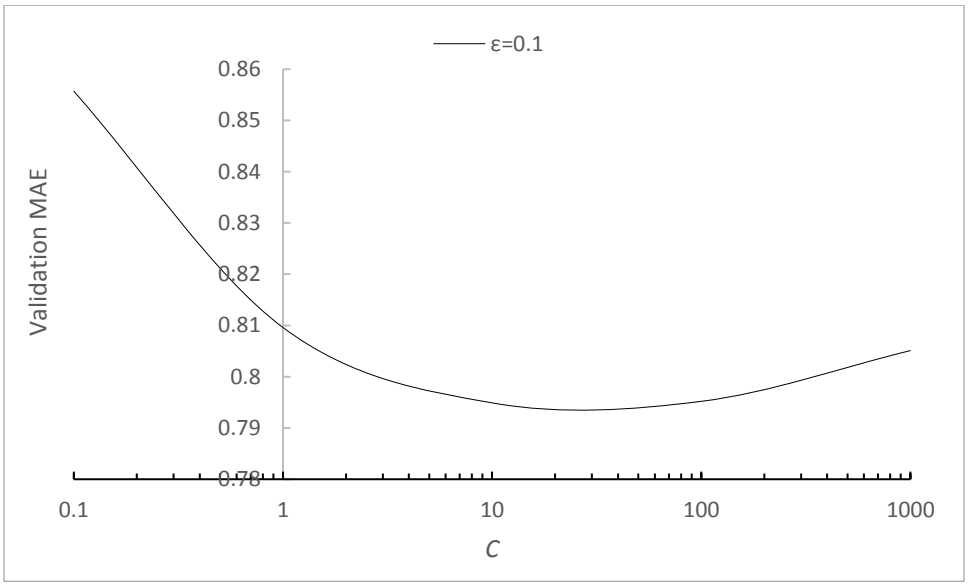


Figure A.68: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 100,000, *ε* = 0.1)

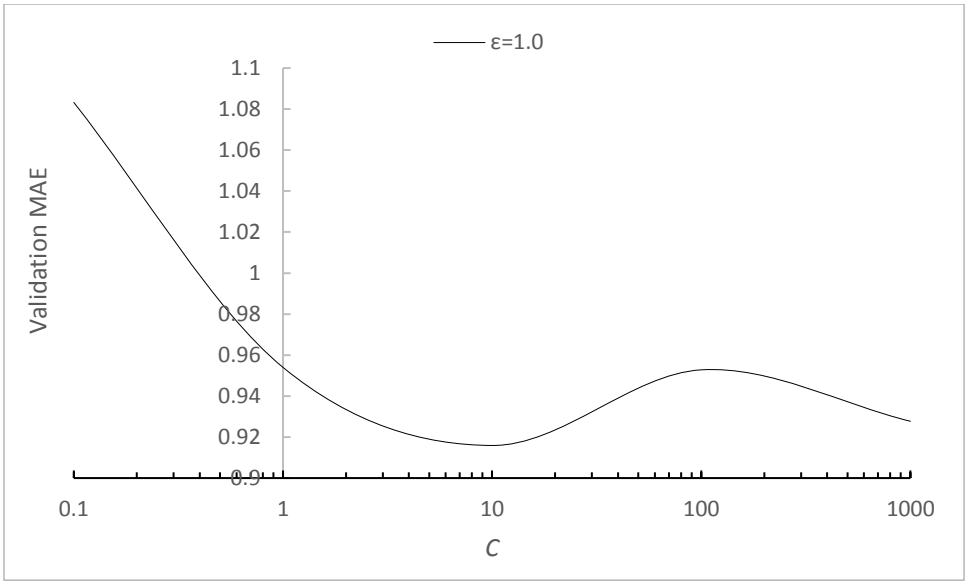


Figure A.69: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 100,000, *ε* = 1)

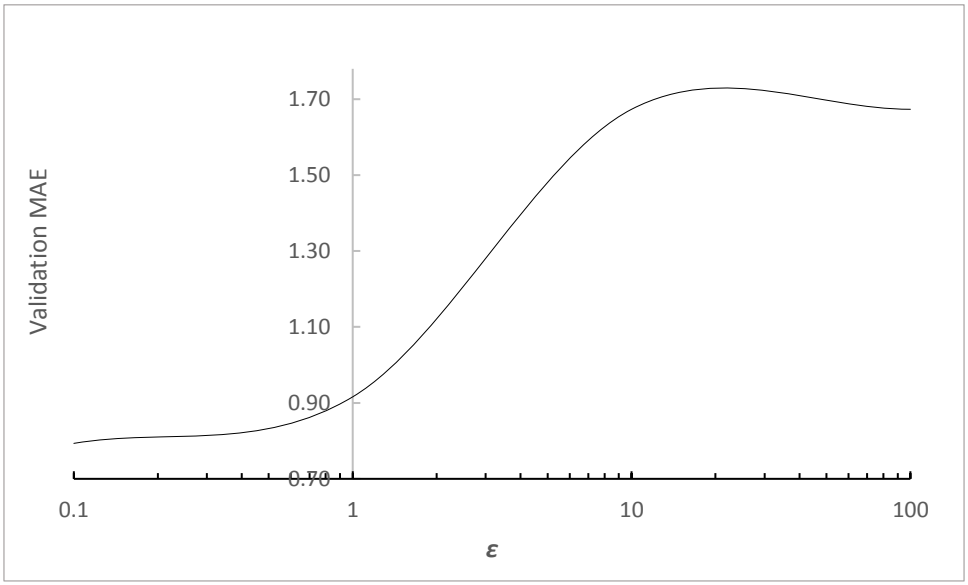


Figure A.70: Validation MAE vs. *ε* of B-VMD-SVR (NLearn = 100,000)

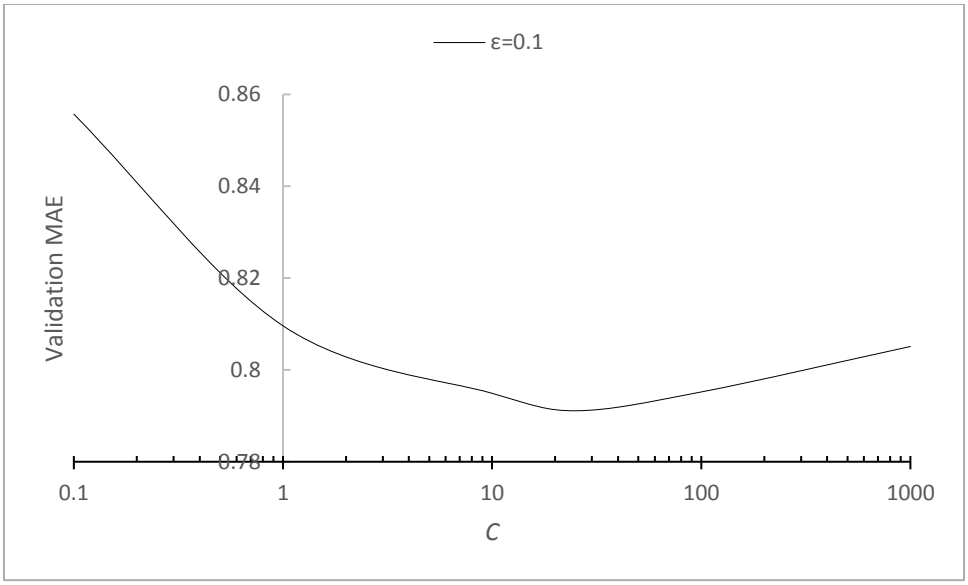


Figure A.71: Validation MAE vs. *C* of B-VMD-SVR (NLearn = 100,000)

APPENDIX B

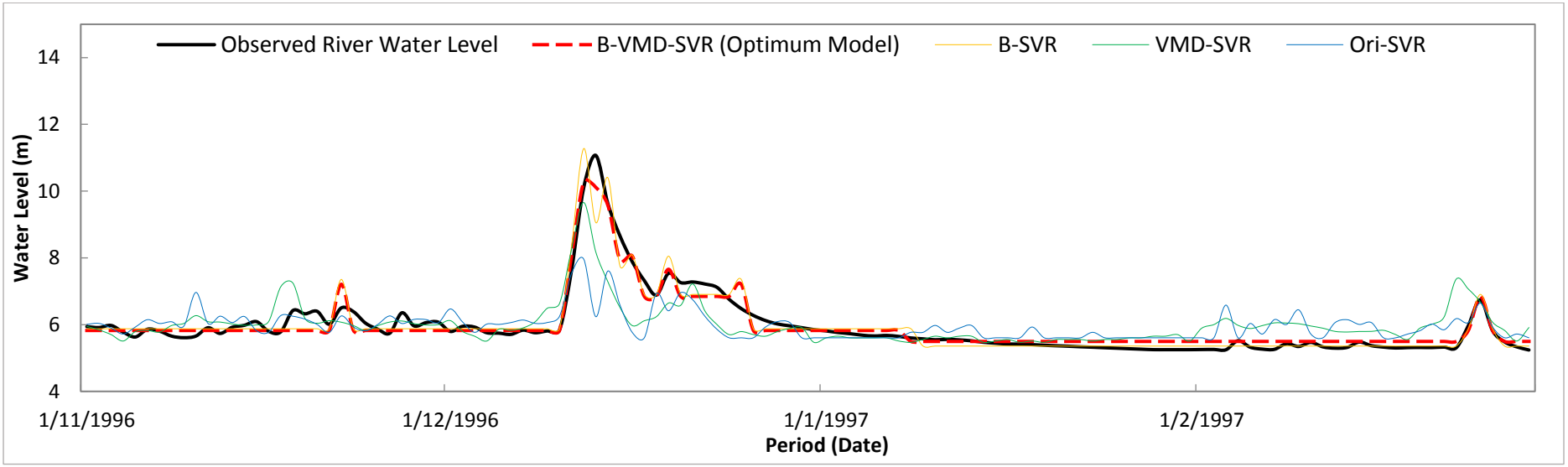


Figure B.1: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 1996-1997

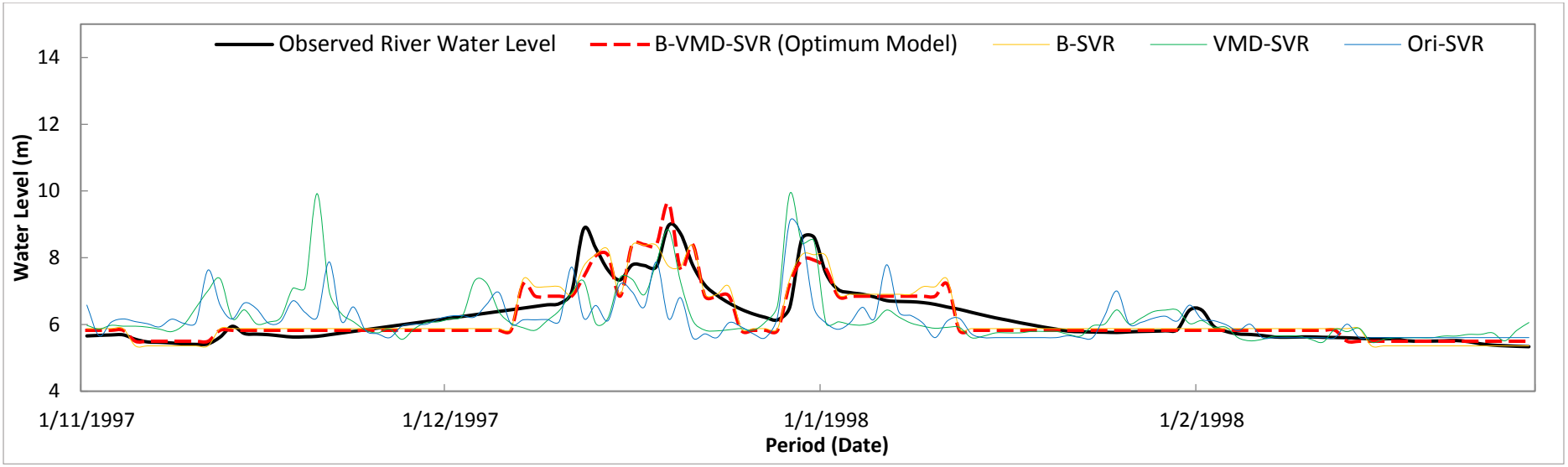


Figure B.2: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 1997-1998

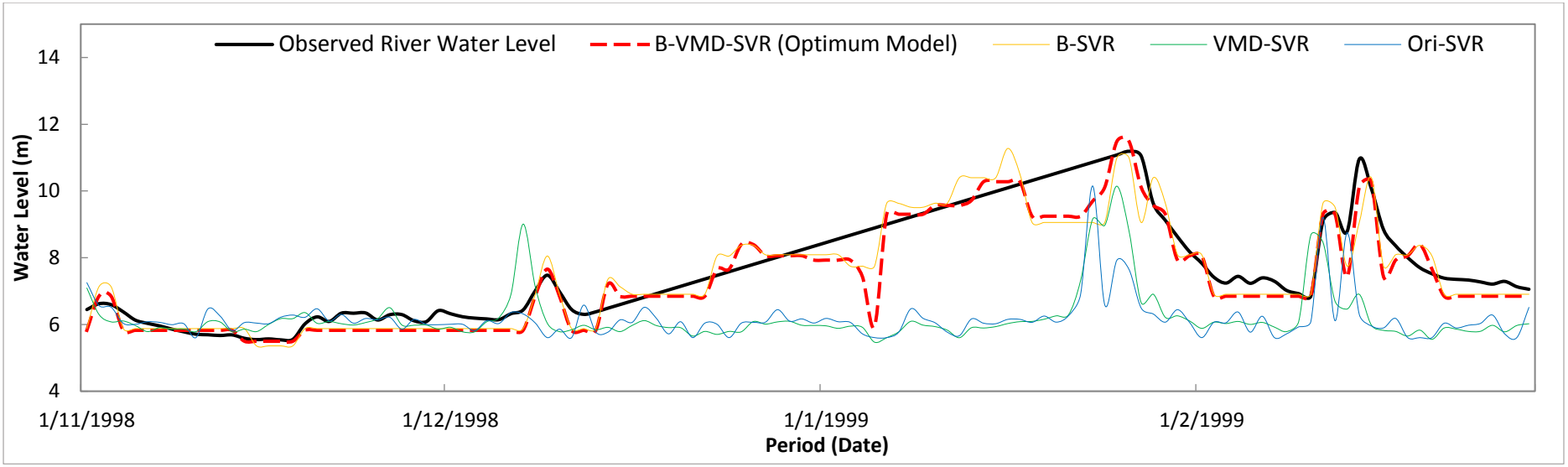


Figure B.3: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 1998-1999

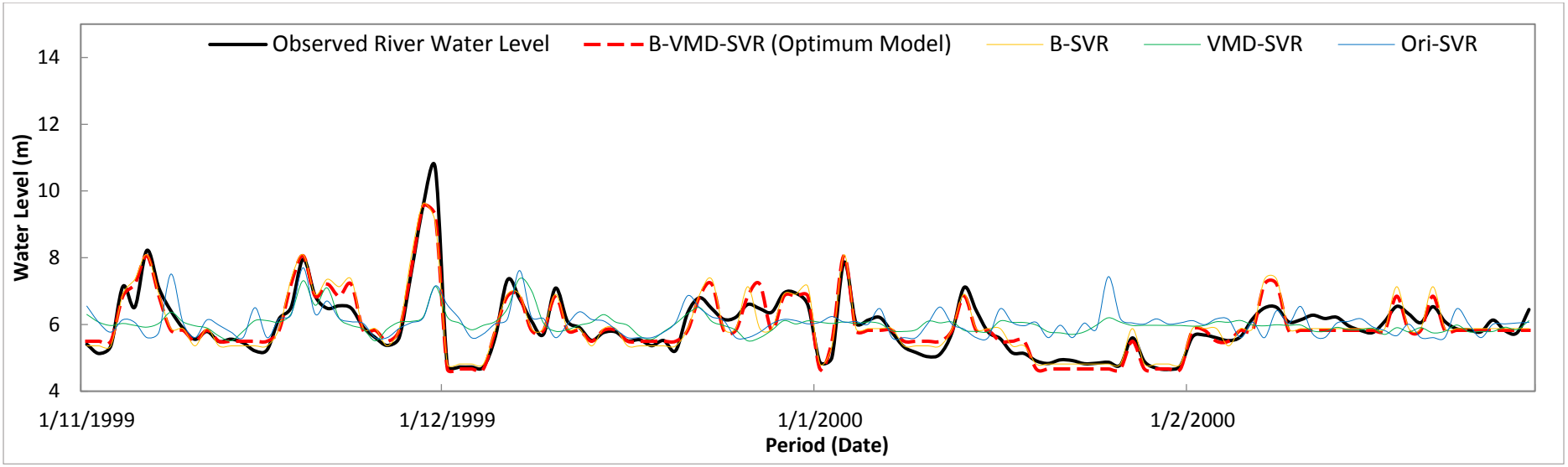


Figure B.4: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 1999-2000

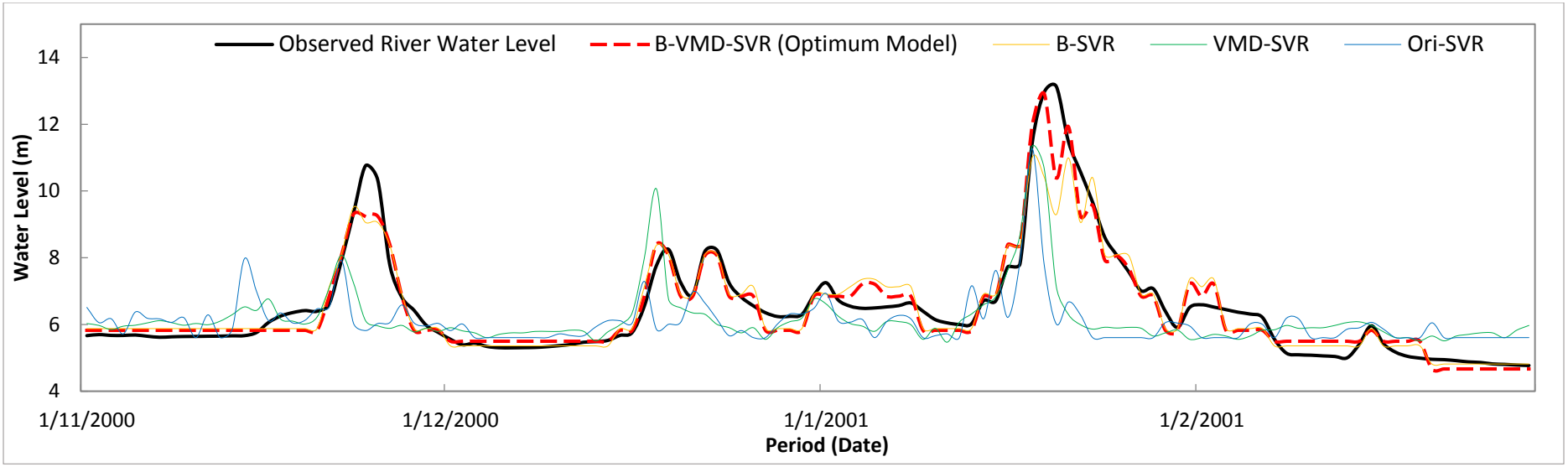


Figure B.5: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2000-2001

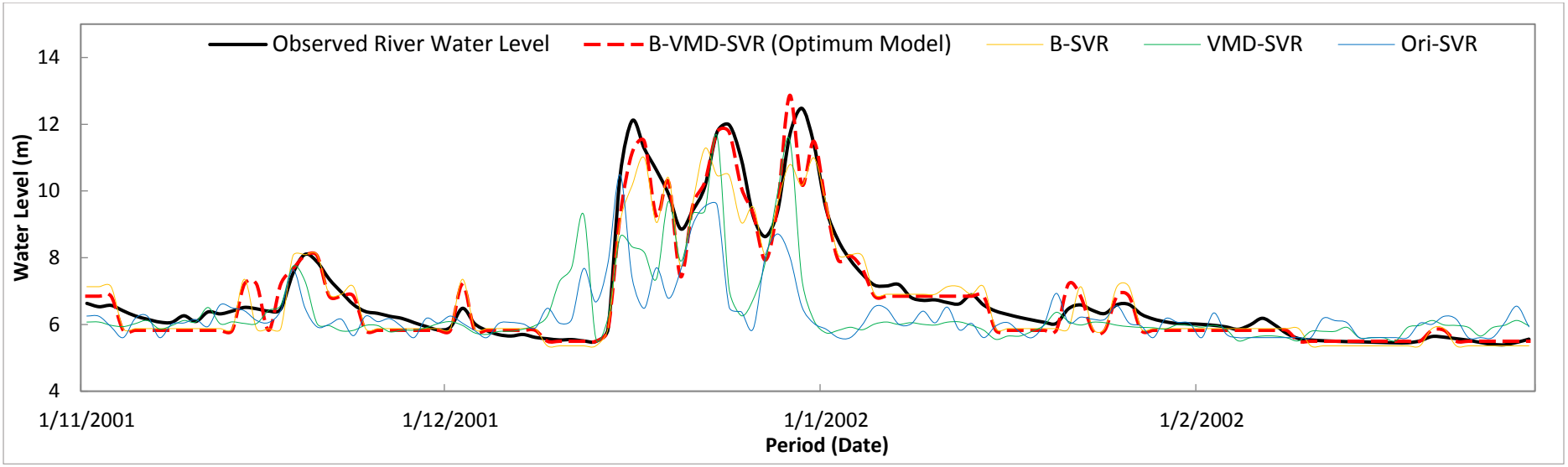


Figure B.6: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2001-2002

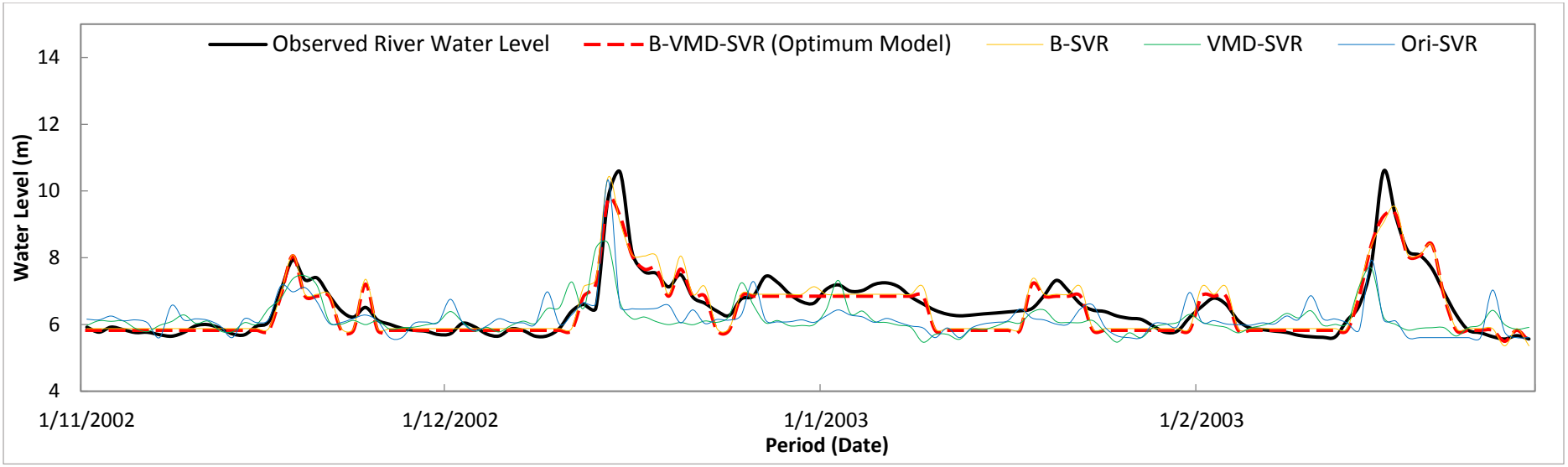


Figure B.7: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2002-2003

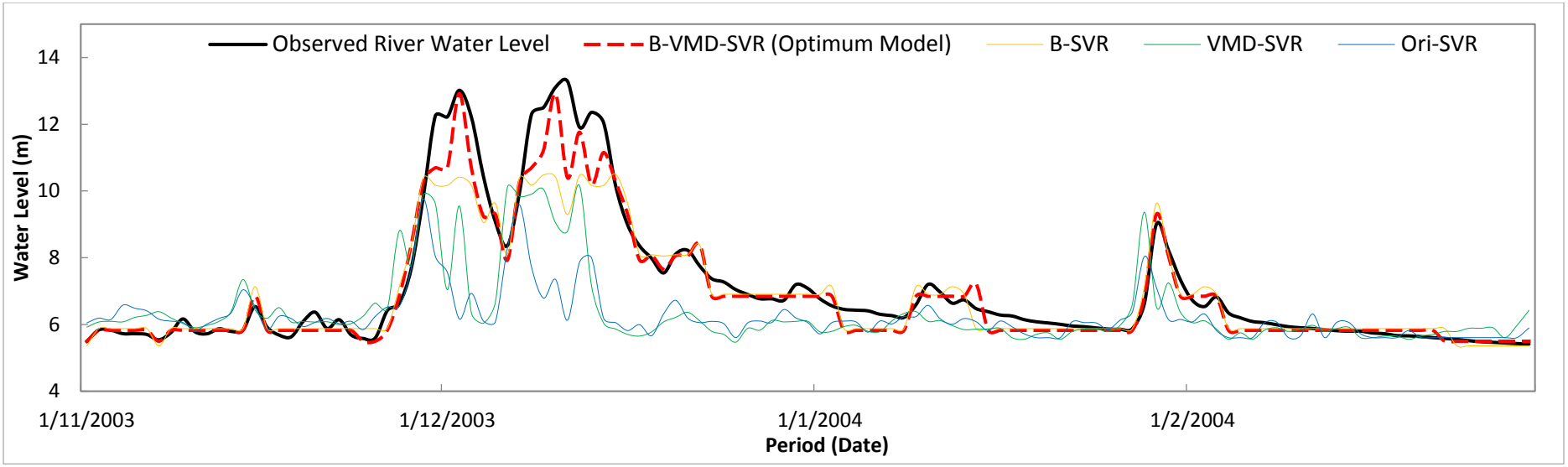


Figure B.8: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2003-2004

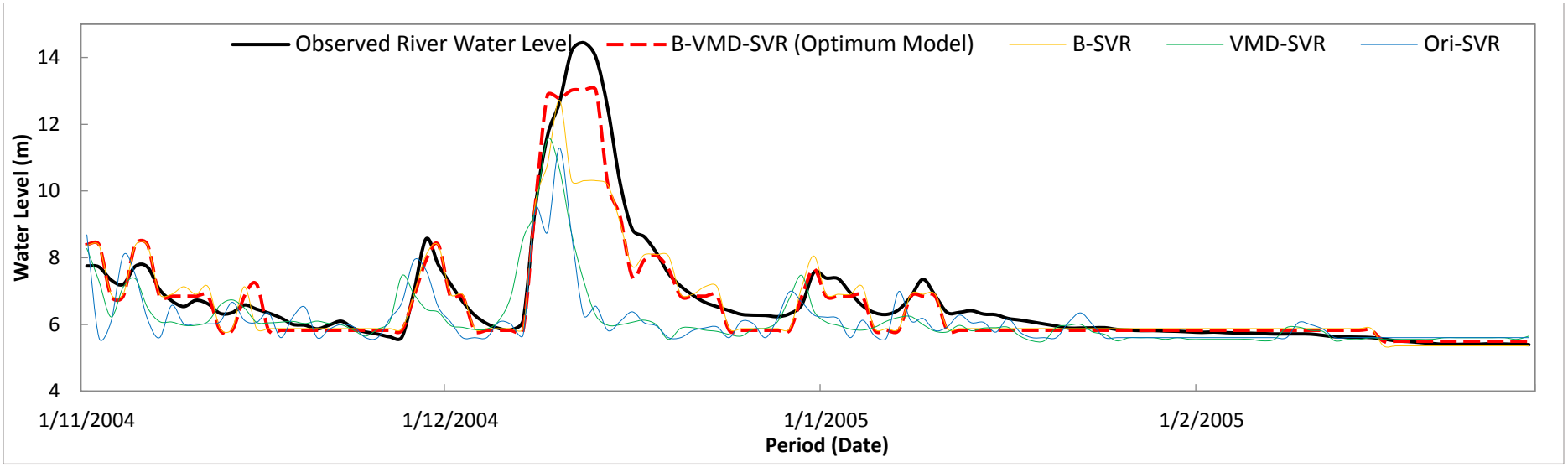


Figure B.9: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2004-2005

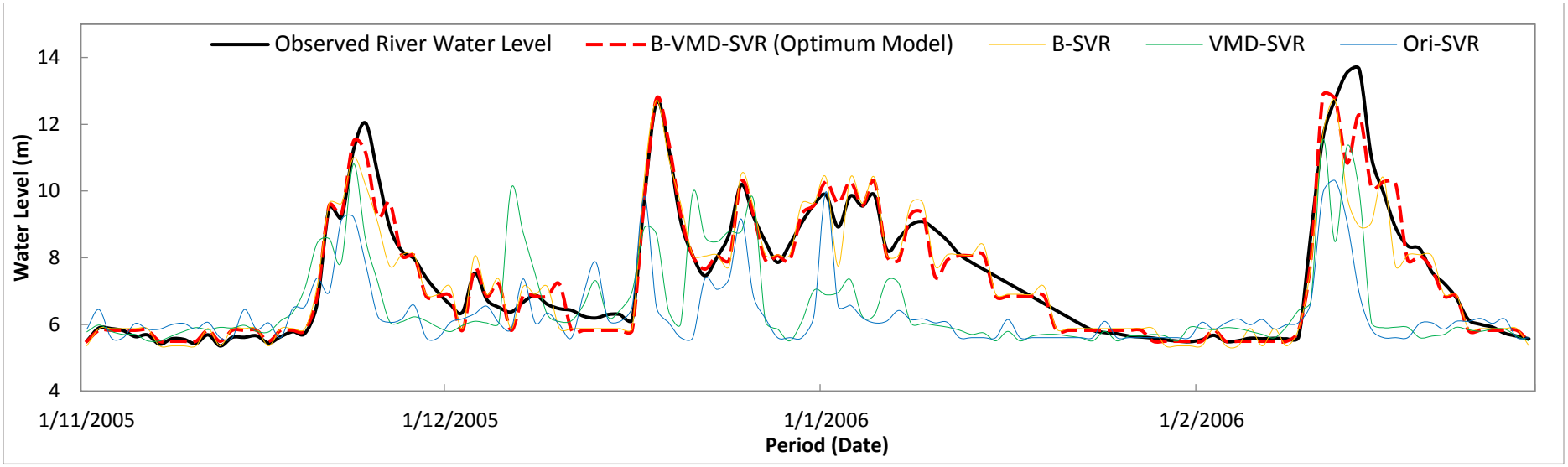


Figure B.10: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2005-2006

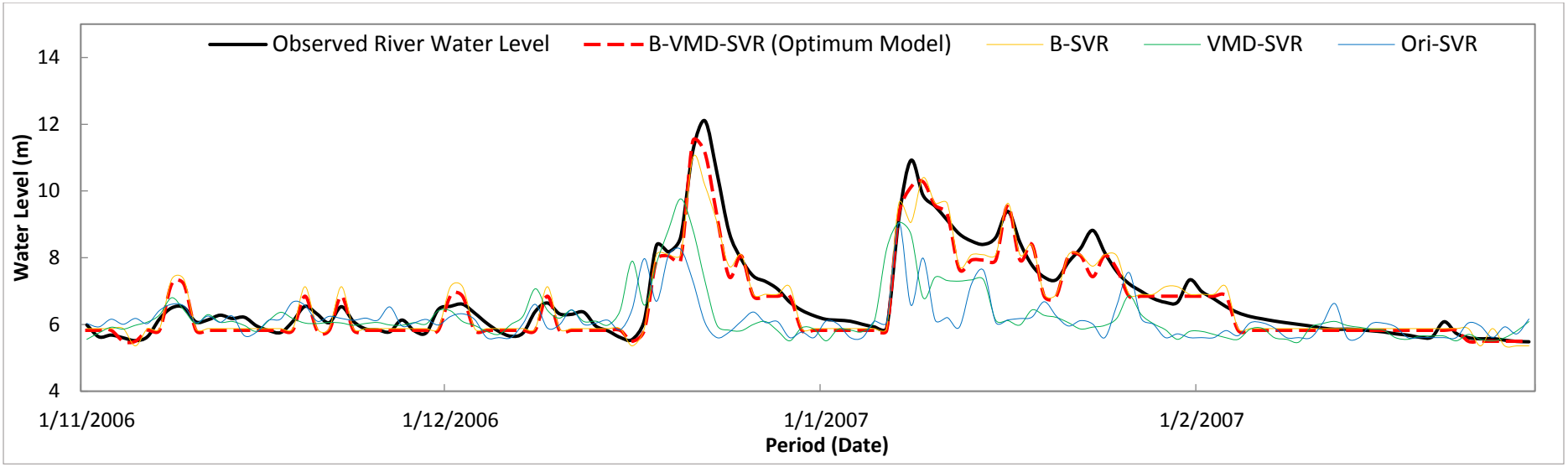


Figure B.11: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2006-2007

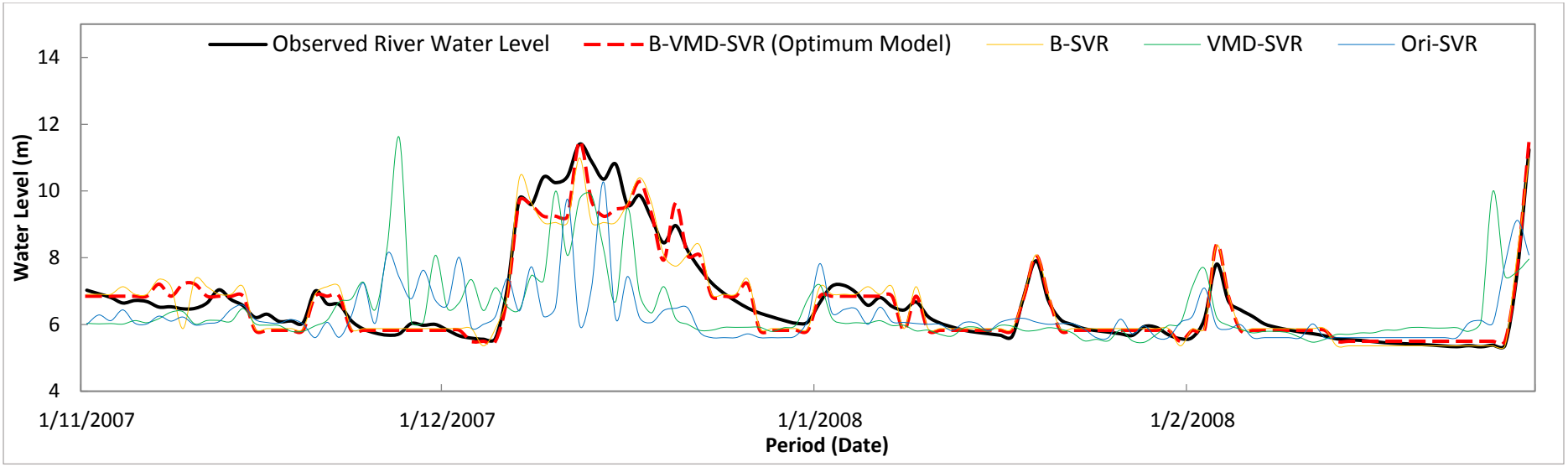


Figure B.12: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2007-2008

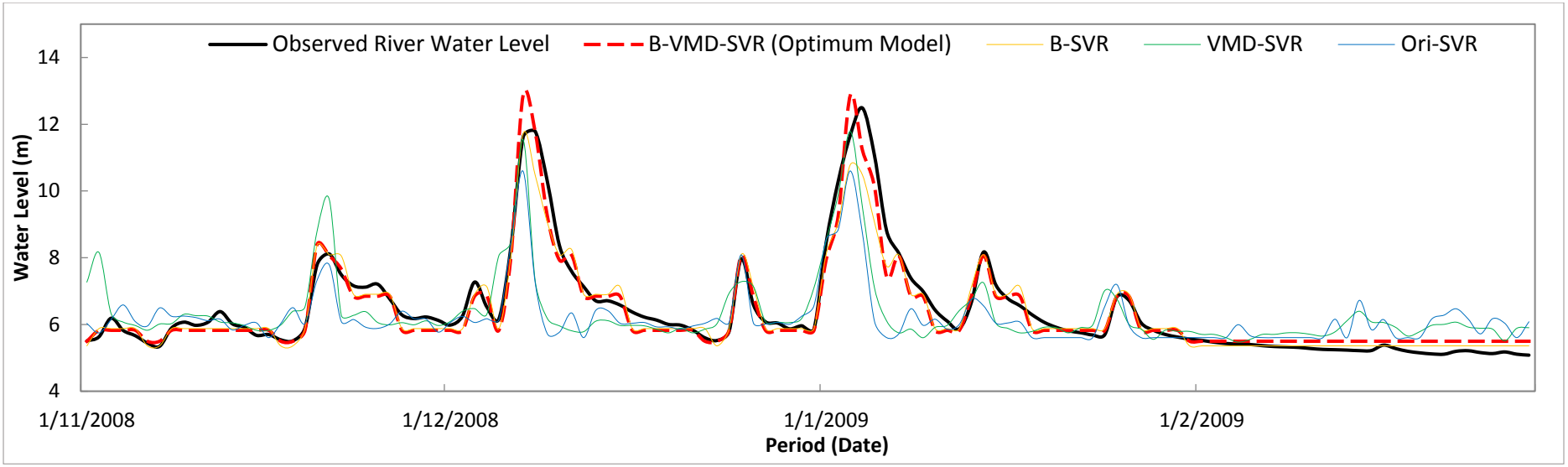


Figure B.13: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2008-2009

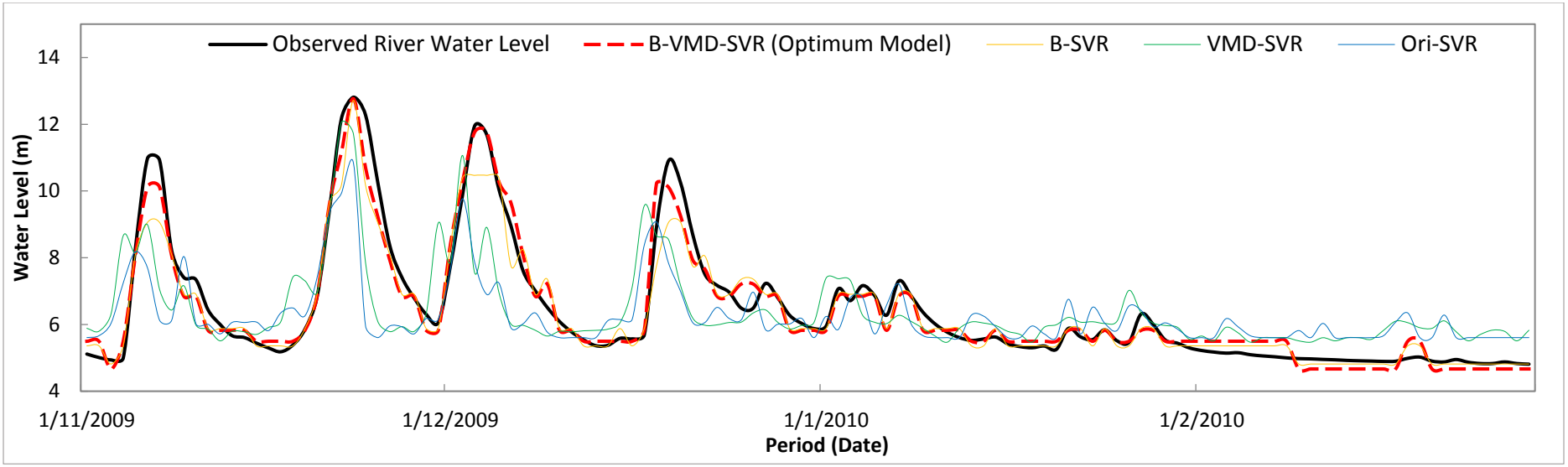


Figure B.14: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2009-2010

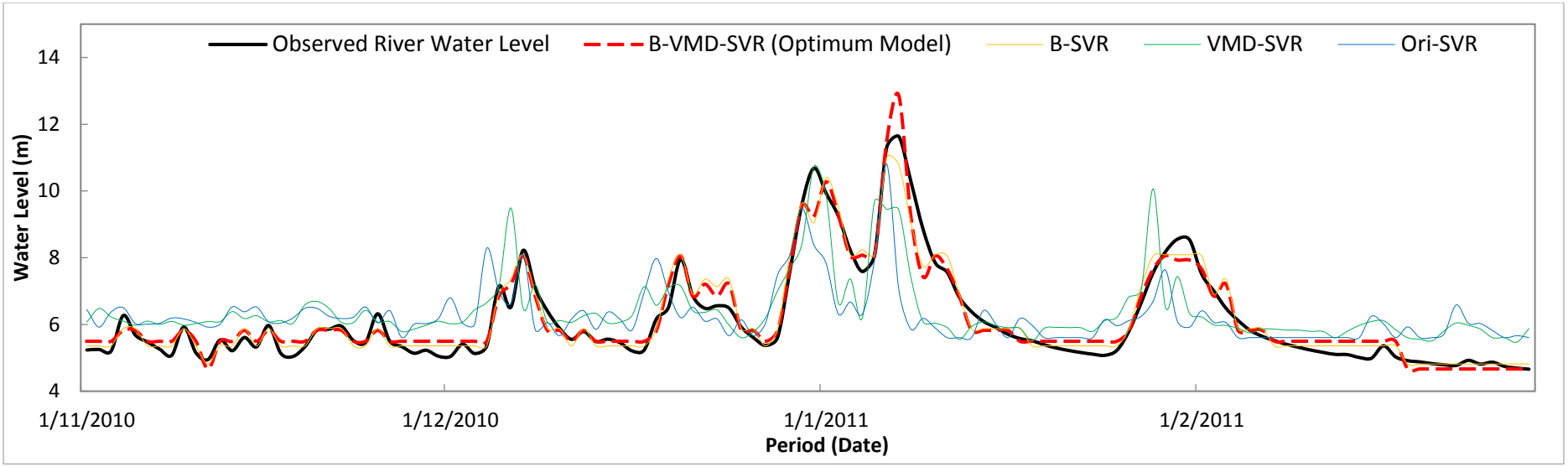


Figure B.15: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2010-2011

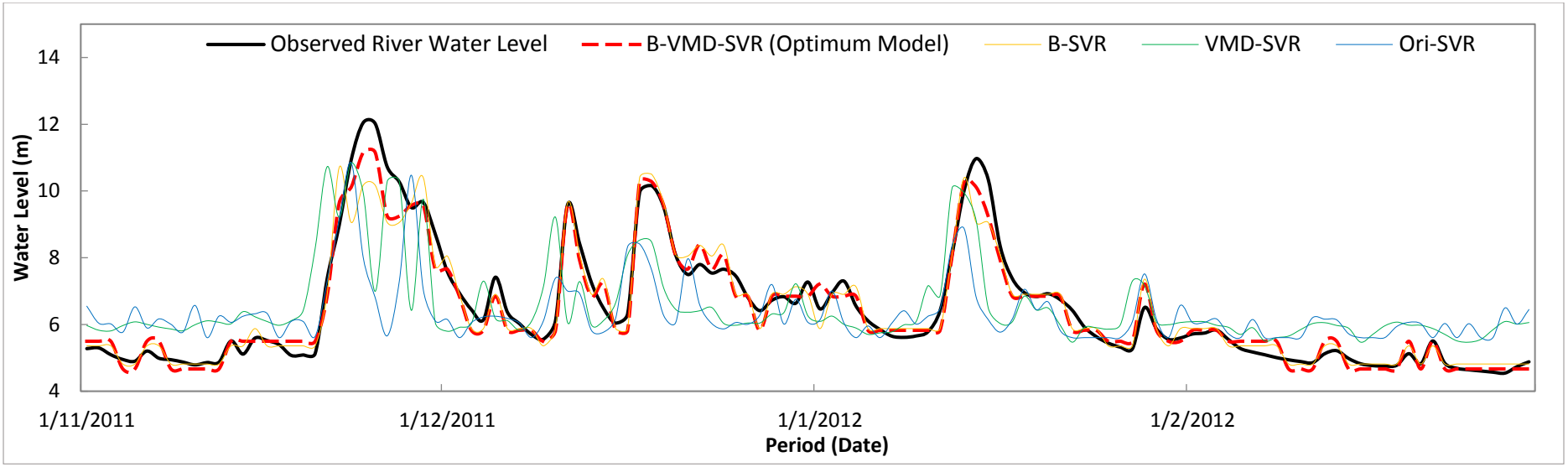


Figure B.16: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2011-2012

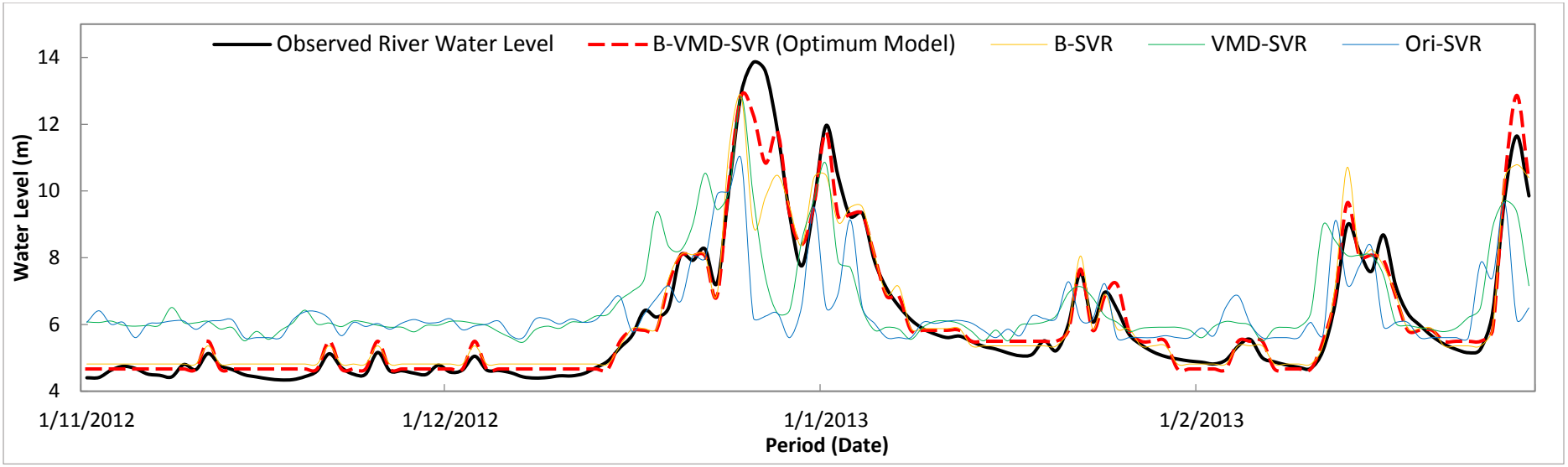


Figure B.17: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2012-2013

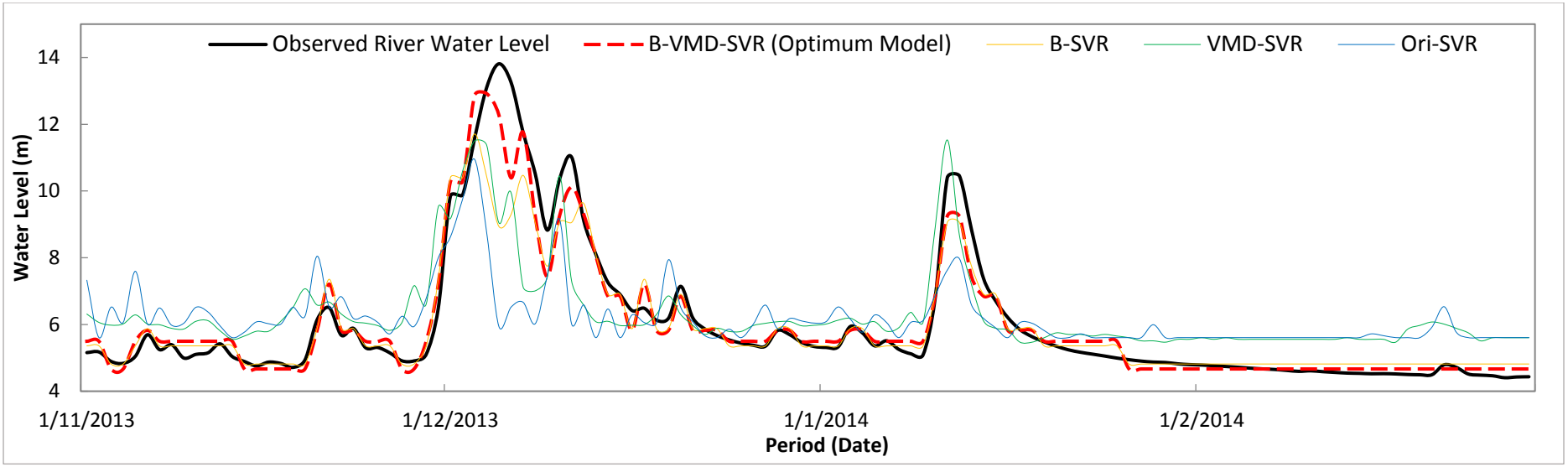


Figure B.18: Hydrograph of observed river water levels and predicted river water levels using processed and observed rainfall series for training sets of 2013-2014

APPENDIX C

1. Tiu, E.S.K., Huang, Y.F. and Ling, L., 2018. Improving the performance of streamflow forecasting model using data-preprocessing technique in Dungun River Basin. *International Conference on Civil & Environmental Engineering (CENVIRON 2017)*, 19 March 2018 Penang, Malaysia. E3S Web of Conferences, pp. 9. (Manuscript Published)