STREAM FLOW ANALYSIS AND MODELLING USING ARTIFICIAL INTELLIGENCE TECHNIQUES

MOHAMMED A. B. SEYAM

FACULTY OF ENGINEERING

UNIVERSITY OF MALAYA

KUALA LUMPUR

2016

STREAM FLOW ANALYSIS AND MODELLING USING ARTIFICIAL INTELLIGENCE TECHNIQUES

MOHAMMED A. B. SEYAM

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF ENGINEERING
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION Name

of Candidate: **MOHAMMED A. B. SEYAM**

Matrix No: **KHA110050**

Name of Degree: **DOCTOR OF PHILOSOPHY**

Title of Thesis: **STREAM FLOW ANALYSIS AND MODELLING USING ARTIFICIAL INTELLIGENCE TECHNIQUES**

Field of Study: **WATER RESOURCES ENGINEERING**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature ……..……. Date: / / 2015

Subscribed and solemnly declared before,

Witness's Signature …… Date: / / 2015

Name: ……..

Designation: ……

ABSTRACT

The reliable prediction of stream flow (SF) is an important aspect in the planning, design and management of surface water and rivers systems. This prediction can be performed using either process-based or data driven-based models (DDMs). Several modelling approaches fall under DDMs, such as statistical and artificial intelligence (AI) techniques. AI includes artificial neural networks (ANNs), support vector machines (SVM) and other techniques. The main goal of this research is to develop and employ a group of efficient AI-based models for predicting the real-time hourly stream flow (Q) in the downstream area of the Selangor River basin, taken here as the paradigm of humid tropical rivers in Southeast Asia. The Q of this river is yet to be subjected to prediction using AI. Despite intensive applications of monthly and daily SF prediction using AI over the last two decades, the prediction of Q is rare, particularly in small rivers in humid tropical regions, such as the Selangor River. The significance of this research lies in the uniqueness of the considered process and the novelty of the applied methodology in the modelling process.

The performance of AI-based models can be improved through the integration of the hydrological description of SF in the modelling process through estimation of lag time (Lt) and analysis of the long-term changes of SF regimes which verified considerable changes may potentially result in increasing the probability of floods occurring in future. The integration process is essential to the selection of input and output variables of AI-based models and the lag intervals between them. The modelling process are performed in two phases to explore the possibility of improving the performance of AI-based models through the accurate timing of the model variables based on Lt estimation by two approaches, namely, the correlation coefficient and hydrological graphical approaches. Through the two modelling phases, four AI techniques, which include three types of ANNs, namely, the multi-layer perceptron network, radial basis function network, and

generalized regression neural networks, along with SVM, are employed to develop six AI-based models to predict the Q. Three scenarios were employed to achieve six combinations of input variables, the first adopts RF and the second adopts WL while the third adopts both WL and RF as input variables. A total of 8753 patterns of Q, water level, and rainfall hourly records representing a one-year period (2011) were utilized in the modelling process.

The performance evaluation of the developed AI-based models shows that high correlation coefficient (R) between the observed and predicted Q is achieved by most of the developed models. For example, R in SVM-M6 model is 0.992 and 0.953 for the training and testing data sets, respectively. The developed AI-based models were efficiently employed in some hydrological applications, such as Q prediction, analysis of the influence of both water level and rainfall on Q and estimation of the missing records of Q. They also were employed in flood early warning throughout the advanced detection of hydrological conditions that could lead to formations of floods.

Key words: Stream flow, surface water, hydrology, lag time, humid tropical rivers, hydrological modelling, artificial intelligence, artificial neural networks and support vector machines.

ABSTRAK

Ramalan aliran sungai (SF) yang tepat adalah satu aspek penting dalam perancangan, reka bentuk dan pengurusan permukaan air dan sistem sungai. Ramalan ini boleh dilakukan sama ada dengan menggunakan model berasaskan proses atau model berasaskan data (DDMs). Pendekatan model di bawah DDMs adalah seperti statistik dan teknik kecerdasan buatan (AI). AI termasuk rangkaian neural buatan (ANNs), mesin vektor sokongan (SVM), dan teknik-teknik lain. Matlamat utama kajian ini adalah untuk membangunkan dan menggunakan model berasaskan AI yang efisyen untuk meramal masa sebenar aliran sungai setiap sejam (Q) di kawasan hilir lembangan Sungai Selangor, di mana sungai ini mewakili sungai bertropika lembap di Asia Tenggara. Q sungai ini masih belum tertakluk kepada ramalan menggunakan AI. Walaupun aplikasi ramalan SF untuk bulanan dan harian dengan menggunakan AI telah digunakan sejak dua dekad yang lalu, ramalan Q jarang berlaku, terutamanya dalam sungai-sungai kecil seperti Sungai Selangor. Kepentingan kajian ini terletak pada keunikan proses yang telah diambil kira dan pembaharuan metodologi yang digunakan dalam proses pemodelan.

Prestasi model berasaskan AI dapat ditingkatkan melalui integrasi hidrologi SF dalam proses pemodelan melalui anggaran jarak masa (Lt) dan analisis perubahan rejim jangka panjang SF yang disahkan perubahan besar yang berpotensi boleh menyebabkan kebarangkalian banjir yang berlaku pada masa akan datang. Proses integrasi adalah penting untuk pemilihan input dan output pembolehubah model berasaskan AI dan jarak masa di antara mereka. Proses pemodelan dilaksanakan dalam dua fasa untuk meneroka kemungkinan meningkatkan prestasi model berasaskan AI melalui masa yang tepat bagi pembolehubah model berdasarkan Lt anggaran oleh dua kaedah: pekali korelasi dan pendekatan grafik baru. Melalui dua fasa pemodelan, empat teknik AI, termasuk tiga jenis ANN, iaitu, "*multi-layer perceptron network*", "*radial basis function network*", and

"*generalized regression neural networks*", bersama-sama dengan SVM, bekerja untuk membangunkan enam model berasaskan AI untuk meramalkan Q. Tiga senario telah digunakan untuk mencapai enam gabungan input pembolehubah, yang pertama menggunakan RF manakala yang kedua menggunakan WL dan ketiga menggunakan kedua-dua WL dan RF sebagai pembolehubah input. Sebanyak 8753 corak Q, paras air dan hujan direkodkan setiap sejam bagi mewakili tempoh satu tahun (2011) yang telah digunakan dalam proses pemodelan.

Penilaian prestasi model berasaskan AI menunjukkan bahawa ketepatan ramalan dicapai oleh kebanyakan model yang dibangunkan. Sebagai contoh, pekali korelasi Q yang diperhatikan dan Q ramalan menggunakan model SVM-M6 adalah 0.992 dan 0.953 untuk latihan dan ujian data set masing-masing. Model berasaskan AI yang berjaya digunakan dalam beberapa aplikasi hidrologi, seperti meramalkan Q, analisis kedua-dua pengaruh paras air dan hujan di Q dan anggaran rekod hilang Q. Model berasaskan AI juga berjaya digunakan dalam amaran banjir dan mampu mengesan keadaan hidrologi yang boleh membawa kepada banjir.

Kata kunci: aliran sungai, permukaan air, hidrologi, jarak masa, sungai bertropika lembap, pemodelan hidrologi, kecerdasan buatan, rangkaian neural buatan, rangkaian neural tiruan dan mesin vektor sokongan.

DEDICATION

I would like to dedicate my PhD thesis to my parents for I always feel their prayers in all aspects of my life, and to my dear brothers, sisters, friends, and colleagues. Special dedication goes to my beloved wife Aisha, and my lovely sons, Adnan, Ibrahim, and Yusuf.

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my god, ALLAH, who has given me the strength, enablement, knowledge, and required understanding to complete this thesis.

Next, I wish to express my unreserved gratitude to my supervisor, Assoc. Prof. Dr. Faridah Othman, for her help. Her constructive criticism and ideas have made this work worth reading. I would like to thank the University of Malaya and the Faculty of Engineering for providing me with the great opportunity of completing my PhD.

I extend my gratitude to the Ministry of Education for supporting me through the Malaysian International Scholarships (MIS). I would like to acknowledge the support of the Hydrology and Water Resources Division of the Department of Irrigation and Drainage, Malaysia, for making available the data used in this project.

I am most grateful to all those who have assisted, guided, and supported me in my studies leading to this thesis. Finally, I would like to extend my deepest gratitude to my parents and my wife, who have always given me unremitting support during the preparation of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol	Item
AI	Artificial intelligence
ANNs	Artificial neural networks
BP	Back propagation
CCA	Correlation coefficient approach
CV	Coefficient of variation
DDMs	Data Driven-based models
GA	Genetic algorithm
GRNN	Generalized regression neural networks
Lt	Lag time
Lt_e	Estimated lag time
Lt_o	Observed lag time
MAE	Mean absolute error
MLP	Multi-layer perceptron networks
NGA	Hydrological graphical approach
NN	Neural network
Q	hourly stream flow
Q_p	Peak hourly stream flow
Q_{48}	The average of previous 48 hour stream flow
R	Correlation coefficient
R^2	Coefficient of determination
RA	Range between maximum and minimum stream flow
RBF	Radial basis function networks

RF	Rainfall
Rf_p	peak rainfall intensity
Rf_{48}	The average of previous 48 hour rainfall
SD	Standard deviation
SF	Stream flow
SF1	Mean annual flow
SF2	Maximum annual flow
SF3	Minimum annual flow
SF4	Maximum monthly flow per year
SF5	Minimum monthly flow per year
SVM	Support vector machines
WL	Water level

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

Water resources are the fundamental requirement for human life and civilization. In tropical humid regions, surface water is the main resource for domestic, industrial, and agricultural water usage. Surface water is commonly represented by stream flow (SF), which represents the runoff stage of the hydrological cycle. SF is the response of a river basin to rainfall (RF), and other related hydrological factors under particular meteorological circumstances. SF is considered one of the most complicated hydrological processes. In the past decades, hydrologists have struggled to understand the formation process of SF to analyze and predict it.

The availability of an accurate method of SF analysis and modelling is of immense importance to the proper resolution of several challenges related to the planning, design and management of surface water recourses and river systems, such as the management of water supply, the optimum design of water storage and drainage networks, irrigation planning, improving power generation efficiency, planning of impending increase or decrease of basin capacities and water quality control (Cui & Singh, 2015; Dibike & Solomatine, 2001; Toro et al., 2013; Turan & Yurdusev, 2009).

SF analysis and modelling are very useful in the management of risky hydrological events, such as floods and droughts. They can provide early warning of upcoming floods and high SF events. They also help in regulating basin outflow during droughts and low SF periods (Hassan et al., 2014; Zhang et al., 2015).

The availability of an accurate method of SF prediction and analysis can also assist in the environmental protection of river basins, such as the prevention and comprehension of hydrologic hazards, such as erosion, mud and sediment movement over river basins (Tehrany et al., 2015; Toro et al., 2013). SF studies are not only significant in hydrological applications related to river systems but also in socioeconomic conditions and human activities related to river basins, such as recreation, fish and wild life propagation (Rakhshanehroo et al., 2010).

A variety of prediction techniques and huge efforts have been proposed and conducted to investigate a wide range of hydrological processes related to the SF, such as formation mechanism, relation with other components of hydrological cycle, long-term variations analysis, modelling and prediction. Reviewing SF applications reveals how such task is a challenge, given that surface water hydrological systems are complex and dynamic, characterized by a huge amount of temporal and spatial instability of input and output variables, and generally exhibit non-linear reactions to influencing parameters, which are interrelated in complicated way (Akhtar et al., 2009; Alfieri et al., 2014; Charron & Ouarda, 2015; Cui & Singh, 2015; Dai et al., 2015; Hatmoko et al., 2015; Meshgi et al., 2015; Nolan et al., 2015; Noori et al., 2011; Saber et al., 2015; Yucel et al., 2015; Zazo et al., 2015).

Several methods and models have been employed in SF analysis and prediction. They can generally be categorized into two main approaches: process-based models and data driven-based models (DDMs) (Hassan et al., 2014; Remesan & Mathew, 2015).

Process-based models are also known as physically-based or conceptual models, and are based on the actual physics of hydrological processes. These models have been designed to simulate interior sub-relationships included in physical mechanisms that rule the

hydrological process, making them too complex, demanding, and time consuming (Athira & Sudheer, 2015; Sahoo & Jha, 2015).

By contrast, DDMs depend directly on observed data without, but in consideration of, physical mechanisms that underlie the hydrological processes. These models can investigate the relationship between the dependent and independent variables of the hydrological processes and dispense with the mathematical formulation of the complex underlying process, making it efficient, less demanding and less time consuming (Clark et al., 2015; Jain & Kumar, 2007; Nilsson et al., 2006).

Each of these techniques has a specific set of advantages and disadvantages, based on data availability and modelling conditions. The lack of a full physical description of complex hydrological systems encourages hydrologists and researchers to find alternative modelling tools. However, whereas DDMs may lack the ability to demonstrate the physical process, it is more practical, more rapid and less demanding; at the same time, it exhibits superior performance relative to process-based models, especially when it is dependent on sufficient training data (Aqil et al., 2007; Bronstert et al., 2014; Kentel, 2009).

Many modelling techniques are categorized under DDMs, such as statistical methods and artificial intelligence techniques (AI). Statistical methods include methods, such as linear and nonlinear regression models, and autoregressive integrated moving average model, whereas AI techniques include advance modelling techniques, such as artificial neural networks (ANNs), support vector machines (SVM), fuzzy rule-based systems (FRBS), and genetic algorithms (GAs) (Daniel et al., 2011; Kisi et al., 2012; Solomatine et al., 2008). AI techniques are considered one of the most promising and efficient modelling tools in DDMs (Kalteh et al., 2008).

AI-based models are becoming adequate alternatives in many hydrological modelling applications, especially when the data required for process-based models are unavailable or limited. Various AI modelling techniques, such as ANNs and SVM, have recently been applied successfully to a wide range of hydrological systems. General literature reviews on the applications of ANNs in hydrology and water recourses have been debated in the Task Committee on Application of ANNs in Hydrology by the American Society of Civil Engineering. The importance of ANNs as a prediction tool has been recognized and proven by this society (ASCE, 2000b).

In this research, three types of ANNs, namely, multi-layer perceptron network (MLP), radial basis function network (RBF), and generalized regression neural networks (GRNN), along with SVM were employed in real-time hourly stream flow (Q) prediction in Selangor River basin - as a paradigm of humid tropical rivers in Southeast Asia - which have not been predicted utilizing AI techniques. Also, statistical methods were employed in the long-term changes analysis of SF regimes and lag time (Lt) estimation between the upstream and downstream stations, which is necessary to select the lag intervals between the input and output variables of AI-based models.

1.2 Problem Statement

In the past century, considerable variations in SF regimes have been verified in about a quarter of the world's rivers (Descroix et al., 2012; Walling & Fang, 2003; Yang et al., 2005; Yue et al., 2003; Zhang et al., 2000). Apparent variations in the SF regime of the Selangor River have also been verified through long-term variation analysis over a 50-year period (Seyam & Othman, 2014b). These changes may potentially result in the formation of hydrological circumstances that can raise the probability of high and low SF events, which draws attention to the necessity of improving prediction tools and early

warning systems of the SF process to sidestep the hazardous effects that may ensue from the variations in SF regime of the Selangor River.

One of the most important keys to improving the SF modelling and prediction process is to develop new efficient SF modelling and prediction techniques. Performance, simplicity, less-demanding usage, applicability, and cost effectiveness are the main required characteristics of efficient hydrological modelling techniques (Ammar et al., 2009; Bierkens, 2006; Harou et al., 2009). Consequently, current trends in SF prediction applications are motivated to develop new and efficient techniques using DDMs, particularly, AI-based models, which are considered one of the most promising techniques in SF prediction applications (Nourani, 2012).

SF variability can occur across many time scales that can vary from hourly to daily, seasonal to annual, and beyond (Hassan et al., 2014). Despite the existence of intensive applications of monthly and daily SF modelling and prediction through AI techniques in the last decade (as shown in Chapter 2), the prediction of Q, especially in humid tropical regions, is uncommon in the literature (Gopakumar et al., 2007; Nourani et al., 2014). However, such prediction is very necessary and required in many hydrological practical applications, such as the planning, design, and management of rivers and water resource systems, particularly in small river basins, such as the Selangor River basin. To the best of the researcher's knowledge, AI techniques are yet to be used to predict real-time Q in the Selangor River basin.

In small river basins, the time of concentration, which is the time taken by water to move from the hydraulically most distal part of the river basin to the outlet or reference point downstream, is usually less than one day. Daily or monthly SF cannot be considered a sufficient representative of the real-time SF and its variations over short time periods. The prediction of Q is more useful and practical than the prediction of daily or monthly SF

because Q in small river basins can change dramatically within a period of a few hours. Q is considered as a sufficient representative of the real-time Sf and its variations over short time periods, especially in small river basins (Besaw et al., 2010).

Reviewing the literature shows that the performance of Q prediction using AI techniques still requires more improvement, given the weak performance of developed AI-based models in many previous studies, as discussed in Chapter 2. In addition, many preliminary considerations in Q predictions using AI techniques, such as the determination of the input and output variables of AI-based models and the lag intervals between them have not been adequately investigated so far, despite their potential role in improving the performance of AI-based models (Crout et al., 2008).

The prediction accuracy of AI-based models can be significantly improved through the accurate selection of model variables and lag intervals between them, which are mainly based on the accurate timing of the input and output variables of AI-based models (Fang et al., 2008). The more accurately the effect of such considerations in Q prediction are investigated, the more precise the Q prediction process is (Damangir, 2001; Yao et al., 2014). To the best of the researcher's knowledge, no known ideal hydrological approach has been developed so far to explain the dilemma of the accurate timing of the input and output variables of AI-based models, thereby lending this research great significance in the field of Q prediction and hydrological modelling using AI techniques.

1.3 Goal and Objectives

The main goal of this research is to develop and employ a group of AI-based models to predict the real-time Q in the downstream area using the hourly records of the water level (WL) and RF stations of the upstream area in the Selangor River basin as paradigm of humid tropical rivers in Southeast Asia, which has not been modeled before using AI techniques.

This research aims to achieve the main goal through the following objectives:

1. To improve the hydrological description of the SF process by investigating the long-term variations of the SF regime and Lt estimation in the river basin.
2. To develop a hydrological graphical approach (NGA) and derive new empirical formulas for estimating the Lt between the upstream and downstream stations.
3. To develop a group of efficient AI-based models for predicting real-time Q by three types of ANNs, namely, MLP, RBF, and GRNN along with SVM.
4. To explore the ability of the accurate timing of the input and output variables of AI-based models to improve the prediction performance of Q.
5. To employ the developed AI-based models in several hydrological applications, such as prediction and analytical tools, estimation of the missing records of Q and early warning of high SF events.

1.4 Significance of the Research

The significance, particularly the original contribution, of this research, lies in the uniqueness of the considered process and the novelty of the applied methodology in the modelling process. The high performance and applicability of the developed AI-based models also have an immense role in enhancing the significance of the research.

Q is the hydrological process considered in this research. The prediction of the real-time Q, especially in humid tropical regions, is rare in the literature, although such prediction is necessary and required in many hydrological applications, particularly in small river basins, such as the Selangor River basin, thereby lending this research real novelty.

This research also integrates the hydrological description of SF in the modelling process using AI techniques, which is also rarely found in the related literature. The integration process and variety of the employed modelling techniques, such as the four AI techniques, lead to very high prediction accuracy. The performance evaluation of the results of the AI-based models shows that high correlation coefficient (R) between the observed and predicted Q was reached for most of the developed models

The research is also highly significant given the high applicability of the developed AI-based models. These models have been employed successfully in a wide range of hydrological applications, such as the prediction of a head Q, investigation of the influence of WL and RF on Q, and estimation of the missing records of Q. Furthermore, the models are beneficial in flood early warning and the advance detection of hydrological conditions that may lead to the formation of floods.

1.5 Scope of the Research

This study is concerned with the analysis and prediction of Q using four AI techniques, which include three types of ANNs -MLP, RBF and GRNN- along with SVM. The hourly records of WL and RF data from the upstream stations represent the input variables (independent) of the AI-based models, while the Q data in the downstream station represent the output variable (dependent) of the AI-based models. The records of the upstream stations were utilized to predict the real-time Q in the downstream station in an ahead period that is approximately equal to the estimated Lt between the upstream and downstream stations.

The study area of this research is the Selangor River basin, which is the main river in the state of Selangor in Malaysia. The Selangor River basin may be considered as a suitable paradigm of humid tropical rivers in Southeast Asia. Hydrological data were collected from hydrological stations located in the Selangor River basin.

The SF data from the 1960 to 2011 regime were extracted from the Rantau Panjang gauging station and were then employed to investigate the long-term variations in SF regime. The WL and RF data of three-year period from 2009 to 2011, which were utilized in the Lt estimation, whereas WL and RF data of one-year period (2011) were utilized in the development of the AI-based models, were sourced from four stations located in the upstream area of the Selangor River basin. The hydrological data were subjected to normality and homogeneity testing using Shapiro–Wilk and Pettitt's tests, respectively.

The accuracy of the AI modelling process was improved by enhancing the hydrological description and understanding of the SF process through the analysis of the long-term variations of the SF regime and Lt estimation. The Lt between the upstream and downstream stations was estimated using three methods: (1) empirical formulas, (2) the correlation coefficient approach (CCA) and (3) HGA based on the hydrological definition

of the Lt. This understanding and accurate Lt estimation are necessary to select the input and output variables of AI-based models and the lag intervals between them. Three performance evaluation criteria, namely correlation coefficient (R), coefficient of determination (R^2) and mean absolute error (MAE) were employed to assess the performance of the AI-based models.

The developed AI-based models were employed as prediction and analytical tools to investigate the influence of WL and RF on Q. Furthermore, they were applied to estimate the missing Q records. Finally, they were employed in floods early warning and the advance detection of the hydrological conditions, which could lead to the formation of floods through six hydrological scenarios.

1.6 Thesis Outline

The thesis is organized in five chapters, beginning with the introduction, which includes a general background of the research topic, problem statement, objectives, significance, and scope of the research.

The literature review is described in the second chapter, where many topics are reviewed, such as the general principles of SF modelling and AI techniques, such as ANNs and SVM, along with their related previous applications on SF prediction. The preliminary considerations in SF modelling using AI-based models, such as limitations of AI techniques, selection of the appropriate SF modelling technique, determination of the input and output variables of the AI-based models and improvement of the performance of the AI-based models, are included in the second chapter.

The third chapter presents the research methodology and briefly describes the study area, data collection and preliminary data analysis. The research methodology includes the

hydrological description of the Selangor River basin which includes the overview of the Selangor River basin hydrology, the long-term variations in the SF regime and the Lt estimation between the upstream and downstream stations. This chapter also includes a detailed description of the modelling process and development of AI-based models to predict the real-time Q which includes several steps such as selection of model variables, the lag intervals between the input and output variables, the modelling patterns, model structure, model training and the performance evaluation criteria.

The results and discussion are found in the fourth chapter. This chapter presents a detailed hydrological description of the Selangor River basin, including an analysis of the long-term changes in SF regimes over a 50-year period from 1960 to 2010 and the results of Lt estimation between upstream and downstream stations, which is required in the selection of the lag intervals between the input and output variables of the AI-based models. This chapter also presents the results and discussion of the two phases of the modelling process and the six AI-based models, which were trained and developed by the four AI techniques: MLP, RBF, GRNN, and SVM. The results include the description of the developed AI-based models, the performance evaluation criteria, such as R, R^2, and MAE, of the AI-based models, and a comparison between the observed and predicted Q by the AI-based models. This chapter also includes description of some hydrological applications of the developed AI-based models.

Finally, the conclusions and recommendations are presented in the fifth chapter, where various conclusions and recommendations derived from the research results are presented, as well as the proposed future research works related to the research topic.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Chapter 2 presents a general overview of the principles of SF modelling using AI techniques, followed by a brief background of the modelling approaches that have been applied in SF prediction and analysis, especially AI techniques, along with their previous applications in SF prediction and analysis. A brief summary of preliminary considerations in SF modelling using AI techniques is included as well. The preliminary considerations include an explanation of the advantages and disadvantages of AI techniques, selection of the appropriate SF modelling technique, determination of the input and output variables of the AI-based models and how to improve the performance of AI-based models.

2.2 General Overview of Stream Flow Modelling using AI Techniques

The SF prediction process exhibits a high amount of temporal and spatial variability and is overwhelmed by the complexity of physical processes and uncertainty in parameter estimations. Therefore, SF modelling using traditional process-based models requires a considerable amount of effort and significant quantity of data, whereas DDMs can offer an efficient modelling and prediction tool. AI techniques are considered one of the most promising DDMs techniques in SF modelling and prediction applications (Maier et al., 2010; Nourani, 2012).

SF modelling and prediction using AI techniques depends mainly on previous records of WL, RF and SF records in predicting the ahead SF. These data are usually gauged in monthly, daily, hourly, or shorter time steps. In a large river basin, the monthly or daily time step may be adequate for SF prediction applications, and the spatial variation of the model's variables in large river basins is usually more significant than the temporal variation. In a small river basin, the monthly and daily time step are usually longer than the response time of the river basin to RF event. Therefore, an hourly time step is mandatory for accurate and real-time SF prediction. In a small river basin, such as the case study of this research, the Selangor River Basin, hourly records of the hydrological variables are necessary to develop reliable models for SF prediction (Besaw et al., 2010; Talei & Chua, 2012).

In humid tropical rivers, SF is generally perennial and fluctuates depending mainly on the intensity and duration of the RF. Most of the SF in the downstream area is sourced from the RF in upstream areas, which requires time to arrive downstream. The concept of travel time is used to estimate the time needed by the water to move from any location within the river basin to another. This conception is frequently employed in many hydrological applications. Various expressions have been adopted and used to describe the concept of travel time, such as concentration time and Lt, as a result of developments in hydrological models and applications (Banasik et al., 2005; Green & Nelson, 2002; Grimaldi et al., 2012; Honarbakhsh et al., 2012; Talei & Chua, 2012). The performance of AI-based models can be significantly improved by the accurate selection of model variables and the lag intervals between them, which mainly depend on the accurate timing of the input and output variables of AI-based models (Fang et al., 2008).

The uncertainty in the hydrological modelling using AI techniques is mainly related to three main categories (i.e. data, models and human). The uncertainty is considered to be caused mainly by the following elements:

- Structural uncertainty: caused by processes those are not accounted for in the model; that is results from the simplification of the processes simulated in the model.
- Variables uncertainty: caused by inaccurate measurements or mistakes in selecting model variables and related to a number of unrelated variables, which may be inserted in the model.
- Modelling uncertainty: related to modelling technique applied in the modelling process.
- Human uncertainty: related to the knowledge, experience and expertise of the modeller (Campoli et al., 2014; Maier et al., 2008; Piotrowski, 2014).

SF modelling and prediction using AI techniques have the advantage of minimizing the bad role of most uncertainties, given that they depend directly on investigating the hydrological data, particularly exploring the relations between the variables (i.e., input and output variables), without full description and understanding of the hydrological and physical behavior of the river basin systems.

2.3 Modelling Approaches of Stream Flow

Modelling approaches of hydrological processes can be categorized into two main approaches: process-based models and DDMs. Process-based models are also known as hydrological models (including conceptual and physically based models). DDMs are also known as empirical and black-box models. In contrast to process-based models, DDMs depend directly on mathematical equations, which are not derived from hydrological processes in the river basin but from the direct investigation of the hydrological data. Recent advancements in computer sciences and technology have significantly enhanced

the abilities of DDMs, especially those of AI (Remesan & Mathew, 2015; Solomatine et al., 2008).

Given that SF analysis and prediction are two of the most common hydrological problems, several models based on different areas of knowledge under many approaches have been proposed and developed, leading to various levels of accuracy and modelling output (Toro et al., 2013). Similar to modelling approaches of hydrological processes, SF modelling approaches can also be generally categorized into two main approaches: process-based models and DDMs. These approaches have their specific set of advantages and disadvantages, based on data availability and modelling conditions (Aqil et al., 2007; Kentel, 2009).

2.3.1 Process-Based Modelling Approach

Process-based modelling approach includes the modelling techniques by which the specifications of the model are derived from a group of functional elements and their relations with one another and the system environment, through physical processes. The functional elements are selected at an identified level of hierarchy, commonly one level under the level of the whole system. Therefore, the model system can be considered an suitable equivalent of the actual system at the identified level of hierarchy (Mäkelä et al., 2000).

In hydrological modelling applications, process-based models can be recognized by three sides: (i) demonstration level of hydrological processes, (ii) spatial illustration of the model and (iii) temporal range of the model. Two main types are considered in the classification of process-based models depending on the demonstration of hydrological processes: physically-based models and conceptual models (Kokkonen & Jakeman, 2001; Toro et al., 2013).

Physically-based models depend on the mathematical simulation of interior sub-processes included in the prototype and physical mechanisms that rule the process (Chen & Chau, 2006). An example of the simple hydrological model (i.e., physically- based) is that in which the input is RF subdivided into its constituents and routed over the sub-processes to the basin outlet as SF to the surface and deep storage or to the atmosphere as evapotranspiration (Toro et al., 2008). Figure 2.1 displays a physically-based model that includes some hydrological parameters together with their interactions (Toro et al., 2008).

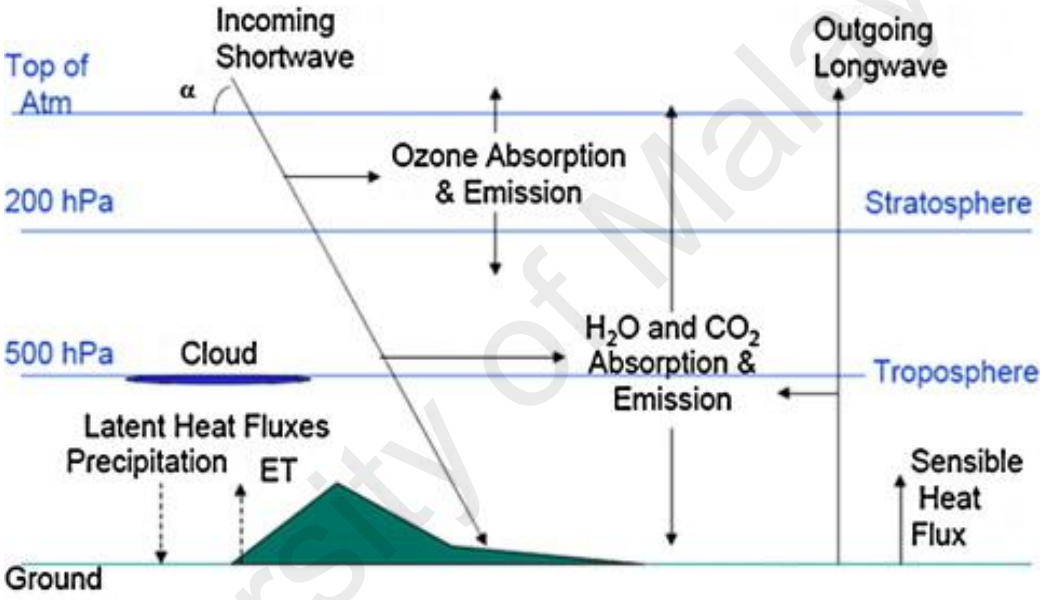


Figure 2.1: Simple hydrological climate model (physically-based model)

(Toro et al., 2013)

In conceptual models, elements should be calculated from fitting the model to hydrological description and historical records. Conceptual models employ a mathematical structure based on the full description of river basin features, such as RF specifications (i.e., intensity and duration of RF events), basin specifications (i.e., area, shape, slope and land use patterns, and vegetation and soil types), and climatic specifications (i.e., temperature, humidity, and wind speed) to predict SF (Jain & Kumar,

2007; Nilsson et al., 2006). Figure 2.2 shows a classic structure of a simple conceptual model (Francés et al., 2007).

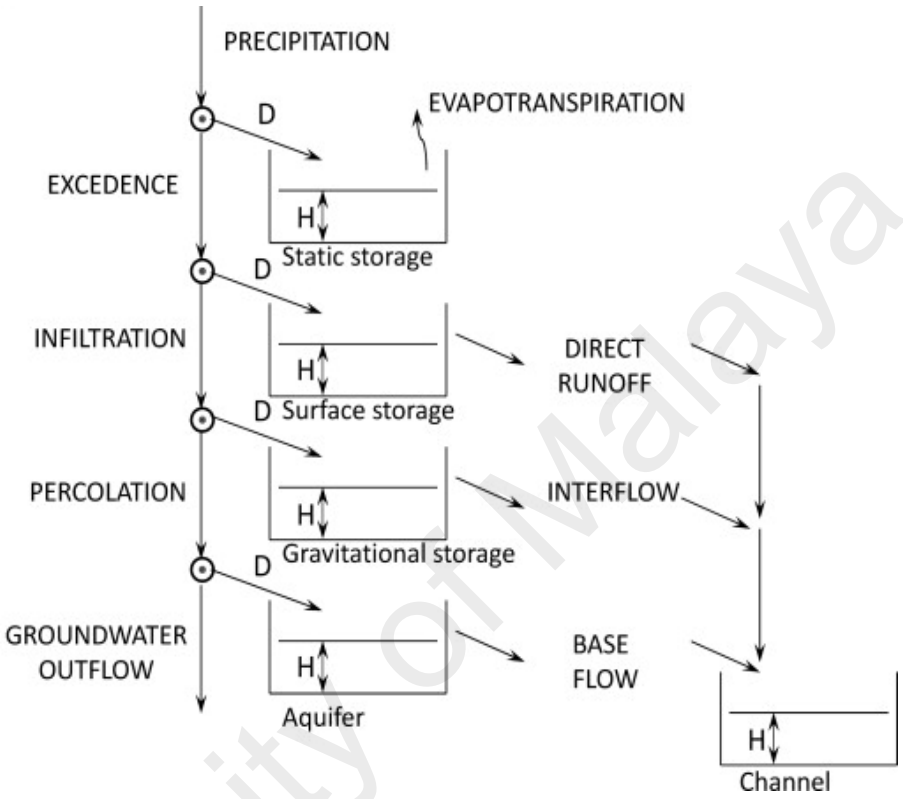


Figure 2.2: Conceptual watershed hydrological model

Adopt from (Francés et al., 2007)

where *D* represents the flow from one tank to another and *H* represents the existing water in each tank.

For many reasons, such as the unavailability of the required data, especially in developing countries (Samsudin et al., 2011), and the complexity of the physical process of surface hydrological systems, which are mainly caused by the data gathering of multiple parameters and variables that vary in space and time, the process-based modelling approach is incapable of an adequately precise and reliable performance in the SF prediction process (Akhtar et al., 2009; Firat & Turan, 2010).

2.3.2 Data Driven-Based Modelling Approach

DDMs depend on investigating the row data of a system, particularly on studying the relations between system parameters (i.e., input and output variables) without exploring the physical behavior of the process. DDMs characterize large advancements in classic empirical models.

Figure 2.3 displays a general explanation of the modelling concept of DDMs. The main function of DDMs is to investigate the relation between the inputs and outputs variables of a system using training data, which are assumed to be demonstrative of the system's behaviors. Once the model is trained, it should be verified and tested to evaluate the modelling accuracy and error and to determine how well the model can generalize new data and cases.

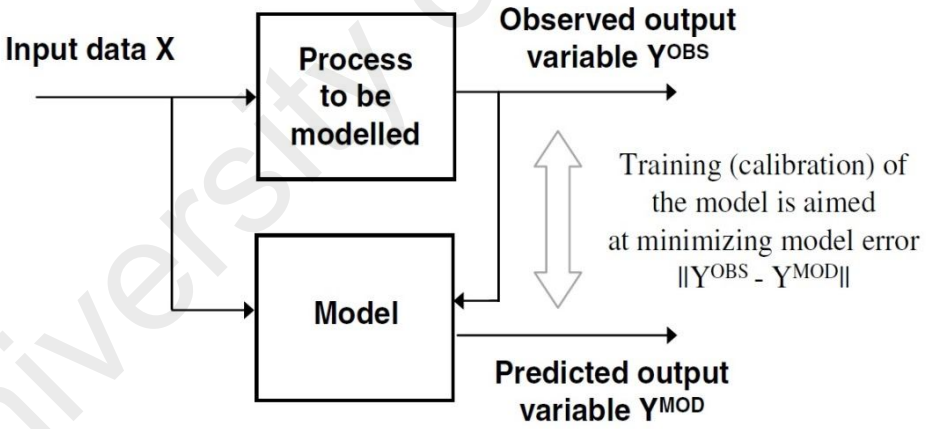


Figure 2.3: General description of the modelling concept of DDMs

(Solomatine et al., 2008)

In SF modelling, DDMs are based directly on observed hydrological data, such as RF, WL, and SF without, but in consideration of, the physical mechanisms that underlie the SF processes. SF modelling and prediction using traditional process-based models require much effort and a significant amount of data, whereas DDMs provide less-demanding efficient SF prediction tools.

However, DDMs may lack the full capability to describe the hydrological mechanism of the SF process. It is more practical, more rapid, and less demanding in terms of data usage, and it exhibits superior performance relative to process-based models. The latest progress in computational intelligence has improved the abilities of DDMs in hydrological applications (Aqil et al., 2007; Kentel, 2009). AI is considered one of the common types of DDMs. It includes numerous modelling techniques, such as ANNs, SVM, FRBSs, and GAs (Daniel et al., 2011; Kisi et al., 2012; Solomatine et al., 2008).

Considering the complexity in SF prediction and modelling, simple empirical models, such as statistical methods, are inadequate to model the complex hydrological systems, such as SF, because they are completely nonlinear, complex, and dynamic systems. Applying more advanced techniques, such as ANNs and SVM, which can analyze and investigate the complexity of the SF process, even without ensuring the whole physical description of the SF system, is significant for high accurate SF prediction (Aqil et al., 2007; Toro et al., 2013).

Given the aforementioned reasons, statistical techniques have been applied widely in analysis of SF long-term change, whereas AI techniques, such as ANNs and SVM, have been widely applied in many complex hydrological applications, including SF prediction, as discussed in Section 2.7 (Ch et al., 2013; Kisi et al., 2012; Machado et al., 2011; Sahu et al., 2011; Shabri & Suhartono, 2012; Wei et al., 2013). ANNs and SVM have recently become adequate and promising alternative tools in SF prediction, especially when the

physical description and data required for the process-based models are unavailable or incomplete (Nilsson et al., 2006; Samsudin et al., 2011).

In this research, three types of ANNs (i.e. MLP, RBF and GRNN) along with SVM were applied in the prediction of Q in the Selangor River basin. A brief description of these methods is described in the following sections.

2.4 Artificial Neural Networks

2.4.1 Introduction to Artificial Neural Networks

ANNs is an advanced data-driven modelling technique with a flexible mathematical structure making it proficient in modelling the non-linear and complex relations among the observed data sets without the need to fully physically recognize the natural systems (Adamowski & Sun, 2010). ANNs have the capability to learn and generalize from historical and previous data to create expressive explanations even when modelling data contain some errors or shortage (Jain et al., 2004; May & Sivakumar, 2009).

The fundamental premise of ANNs is inspired by the human brain's learning systems. ANNs are a simplified mathematical example of natural neural networks (Armaghani et al., 2015; Ziaee et al., 2015). The human brain contains a huge number of neurons, linked together by synapses to form networks of neurons, which are named. The peculiarity of the brain is its operative usage of huge parallelism, the parallel computing arrangement, and the vague information-processing ability. Each neuron includes a cell that utilizes biochemical reactions to collect signal and convey output reactions (Abraham, 2005). Figure 2.4 presented a schematic diagram of mammalian neuron.

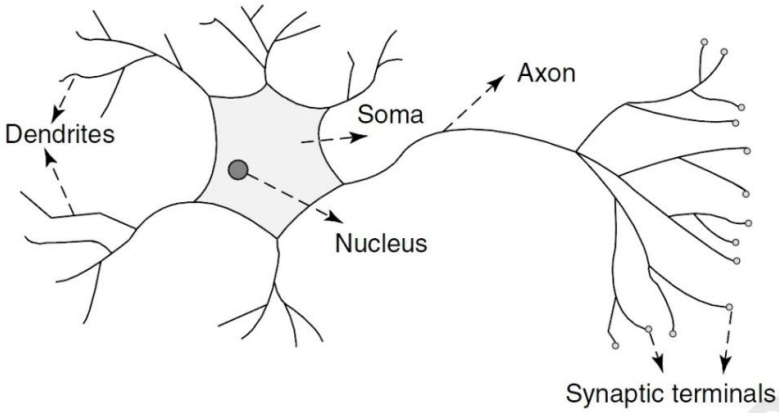


Figure 2.4: Schematic diagram of mammalian neuron (Abraham, 2005)

Treelike networks of nerve fibers named dendrites are linked to the cell body (soma), where the cell nucleus is placed. Single fiber named axon are spreading from the soma are linked to other neurons through synaptic terminals and dendrites of the next cell. The transformation of signals between the neurons is a composite chemical process (Abraham, 2005).

Similar to the mechanism of mammalian neuron, the artificial neurons joined together to form NN. The construction of ANNs is, as rule, layered. Three training processes are performed in the ANNs by three steps i.e. the Neural network (NN) receives data from the external source and transforming them to the neuron which process data then the neurons produce the output of the networks.

An artificial neuron example is presented in the Figure 2.5. The model consists of N inputs, one output, a summation block and an activation block (Hoła & Schabowicz, 2005).

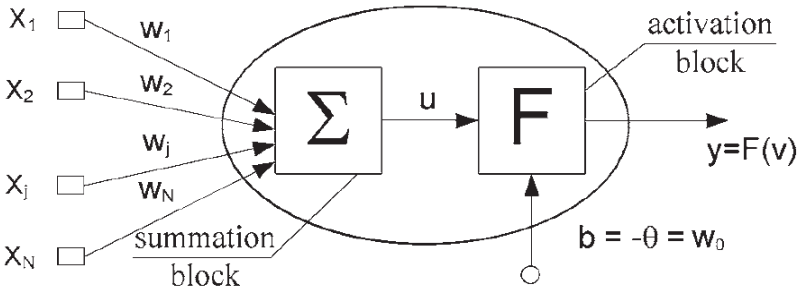


Figure 2.5: Schematic diagram of artificial neuron (Hoła & Schabowicz, 2005)

ANNs are comprised of a vast, interconnected structure of processing elements. The computational power of these processing elements is minimal when in isolation. However, within large networks, the computational power is massive, providing ANNs with the capacity to model complex, nonlinear, and interrelated systems such as hydrological systems, even without prior physical and geometric description of the hydrological system (May & Sivakumar, 2009).

The structure of ANNs entails three or more layers: input layer, hidden layer(s) and output layer. The role of the input layer is to send the input data pattern to the hidden layer. The output layer produces an output of the NN to a particular input. The intermediate hidden layers, which may be only one hidden layer, receive the input data from the first layer. These act as a collection of feature detectors in many ways based on the activation function and network architecture.

Selecting suitable neural network architecture is the most essential and challenging task in the ANNs-based model building process. The modeler should choose an efficient testing means applicable to a large number of options to keep the model within manageable scales. The main assumptions to be defined are network topology, training algorithm, and input selection (Anctil et al., 2004; Sudheer et al., 2002).

A random sample of neural networks that contain of three layers is shown in Figure 2.6

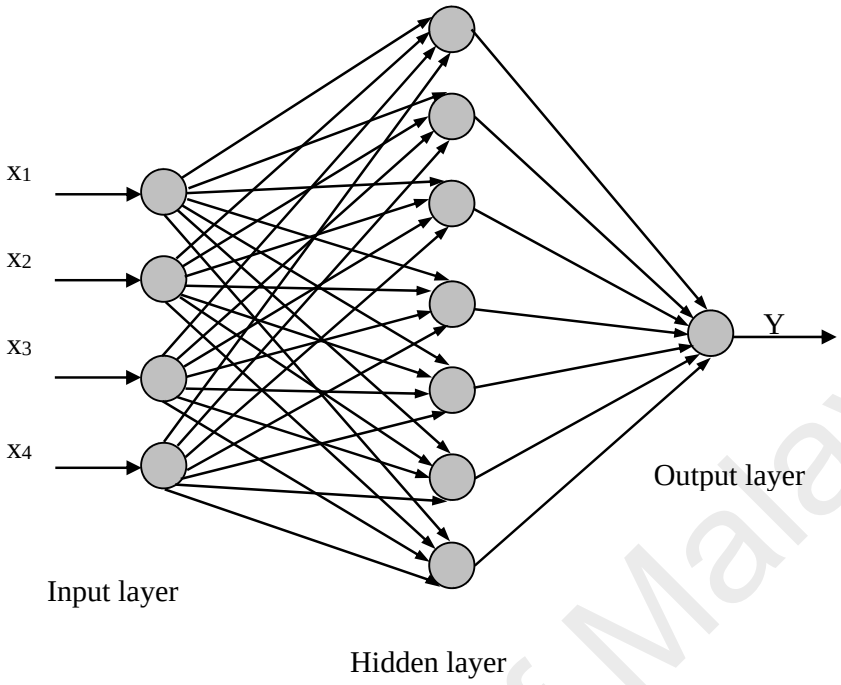


Figure 2.6: General schematic diagram of ANNs with three layers

2.4.2 History of Artificial Neural Networks

A first attention in ANNs was appeared after the development of neurons by McCulloch and Pitts in 1943 which offered as empirical models that could execute the computational processes (Krose & Smagt, 1996). Hebb (1949) developed the first description of biological learning rule for synaptic adjustment. He proposed that the connectivity of the brain is repeatedly varying as an organism learns differing functional operations, and that neural gatherings are created by such variations.

Rosenblatt (1958) introduced a theory for a hypothetical nervous system called a perceptron which is a novel method of supervised learning. The theory is used to present how information about the physical process is recognized, remembered, and how does

information reserved in memory influence recognition and behavior. The theory works as a channel between biophysics and psychology.

When Minsky and Seymour (1969) presented the shortages of perceptron models in a book entitled "Perceptrons". Several NN researches were being conveyed and many academics and scientists left the neural network field. Only a few academics and scientists continued their efforts in NN field, such as Teuvo Kohonen, Stephen Grossberg, James Anderson, and Kunihiko Fukushima (Krose & Smagt, 1996).

Hopfield (1982) discovered how to store data in dynamically steady networks. His efforts help the scientists to use neural networks in their physical applications. The attention in NN re-arisen after many significant theoretical discoveries were attained in the 1980s, such as the error back-propagation by (Rumelhart et al., 1986), the most widespread learning algorithm for training of multilayer-perceptrons NN and also by the new hardware progresses which improved the processing capabilities.

The new attention of neural networks is reflected in the occurrence of huge number of researchers, experts, funding, applications, conferences, and journals associated with neural networks (Krose & Smagt, 1996). Since the late 1980s, ANNs have been used successfully to model a variety of different process in many fields. Recently, ANNs have become gradually common in numerous applications as a modelling tool since it has the capability to explain the really complex processes. The flexible topology of ANNs makes it proficient in modelling nearly all input-output relationships especially the prediction applications (ASCE, 2000b).

2.4.3 Training Process of Artificial Neural Networks

When ANNs are trained, a specific input results in a target output. This process is presented in Figure 2.7. The NN training process includes modifying the connections (weights) between neurons, depending on comparing the NN output with the target (observed data), until the NN output closes to the observed data with minimum global error (E).

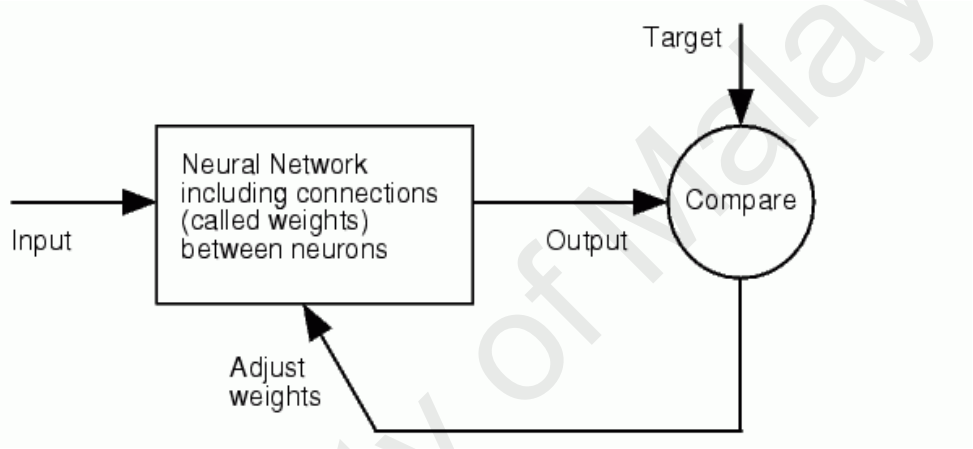


Figure 2.7: Neural Network Mechanism

Studying the training processes of ANNs mathematically is like estimating a real multivariable function F(X) by another formula F(W,X), where $X = (X_1, X_2, \dots, X_p)$ is the input vector and $W = (W_1, W_2, \dots, W_m)$ is the weight vector. In the training process, the aim is to select the vector arranged for the optimum calculation of f(x) based on observed data (Dibike & Solomatine, 2001).

During the training process, consider an input pattern x_p from the training set $\{(x_1, z_1), (x_2, z_2), \dots \dots, (x_p, z_p)\}$ is inserted to the NN to provide an output y_p which varies from the target Z_p.

The goal of the training process is to make y_p and y_p having same value for, $p = 1,2, \ldots \ldots \ldots, P$. The function of the training process is to decrease E, based on the squared error between the predicted and observed data, as can be seen in Equation 2.1

$$E = \sum_{p=1}^{p} E_p = \sum_{p=1}^{p} \sum_{k=1}^{K} (z_{pk} - y_{pk})^2 \quad (2.1)$$

where y_{pk} is the output of the k_{th} node in the output layer for the p_{th} data pattern and E_p is the total output error from all the nodes for the p_{th} (Mustafa et al., 2015).

As the training process continues, the NN structure is frequently changed, includes modifying the neurons number and the weights of the links between neurons. The changes continue until the ANNs output gain acceptable similarity with the observed data (Basheer & Hajmeer, 2000; Palani et al., 2008; Singh & Datta, 2007).

After minimizing E for the training set, new input patterns, those have not employed in the training process are interred to the NN to produce new outputs to test NN. NN should be able to predict new input patterns with comparable accuracy to learned patterns (Mustafa et al., 2015).

Many types and architectures of ANNs have been developed and used in regression and classification applications such MLP, RBF, GRNN and Kohonen self-organizing networks (Figueroa-García et al., 2015), Recurrent neural networks (Rather et al., 2015), self-organized neural networks (Zhang, 2000), adaptive self-organizing map neural network (Kiumarsi et al., 2015), Hierarchical markovian radial basis function neural network classifier (Kokkinos & Margaritis, 2015), Fuzzy delayed neural network (Wang et al., 2015), adaptive neuro fuzzy inference system (ANFIS) (Kisi et al., 2012), Hopfield neural networks (Bai et al., 2015), Cellular neural networks (Chua & Yang,

1988), Cohen–Grossberg neural networks (Liu et al., 2015) and bidirectional associative memory (BAM) neural networks (Qi et al., 2015).

In addition to the developed types and architectures of ANNs, there are many algorithms available for training the NN, such as gradient descent backpropagation algorithm, gradient descent backpropagation with momentum algorithm, conjugate gradient backpropagation algorithm, Quasi-Newton algorithm, and Levenberg-Marquardt training algorithm (Chang et al., 2014; Piotrowski, 2014; Schmidhuber, 2015).

In this research three ANNs types namely, multi-layer perceptron networks (MLP), radial basis function networks (RBF) and generalized regression neural networks (GRNN) along with SVM were applied in SF modelling and prediction.

2.4.3.1 Multi-Layer Perceptron Network (MLP)

MLP are the widely employed, feed-forward networks with unlimited numbers of hidden layers. The back propagation learning algorithm is the common learning rule for MLP. In MLP, the neurons are arranged in layers as illustrated in Figure 2.6.

The Figure presents a random sample of MLP containing three layers. Each neuron in the hidden and output layers receives weighted inputs from all neurons in the previous layer. The active incoming vector is then forwarded through an activation function such as the sigmoid, linear, or cubic polynomial function, to the next layer.

This means that each single neuron performs two actions. Initially, data from an external source is assimilated for the input layer, or from neurons in a previous layer for the hidden and output layers. Then, it creates an output dependent upon a prearranged activation function and sends it to the neuron in the next layer. This process in one neuron is comparatively non-complex; complications with MLP are eventually achieved through

contact and combinations between neurons in networks layers (Adamowski & Sun, 2010).

As an example of a training process, consider the net input to neuron in the hidden is the summation of the weighted inputs from the neurons in the input layer and is denoted by $I_{(in)}$, $I_{(in)} = w_1x_1 + w_2x_1 + w_3x_3 + \cdots \quad w_nx_n$

where, x_i is the input vector, and N is the total number of data patterns. Then $I_{(in)}$ is proceeded by an activation function to produce the output V, V= f ($I_{(in)}$). For instance, the most common activation function is a sigmoid function, which is denoted as follows:

$$f(x) = \frac{1}{1 + exp(-x)} \tag{2.2}$$

This process is reiterated for all input vectors. At the end of a pass, via the whole training data set all the neurons modify the weights depending on the difference between the observed and simulated data regarding each weight. These variations then change weight in order to make errors decay rapidly.

Considering w_m represents the value after iteration m of a weight w, then:

$$W_m = w_{m-1} + \Delta w_m \tag{2.3}$$

Where Δw_m is the variation in weight w at the end of iteration m and is computed as follows:

$$\Delta w_m = -\varepsilon\, d_m \tag{2.4}$$

where ε is the factor guiding the rate of change in weights. d_m is given by:

$$d_m = \sum_{n=1}^{N} \left(\frac{\partial E}{\partial w_m}\right)_n \tag{2.5}$$

where N is the total number of data patterns, and E is the training output error (Dibike & Solomatine, 2001).

The back propagation learning algorithm (BP) is the common learning rule for MLP. BP is trained using the Levenberg–Marquardt optimization technique. Throughout all BP simulations, the weights are modified .Training process contains two stages:

1- Forward pass: The outputs of NN are calculated and the difference between the observed data and network outputs (the error) also calculated.

2- Backward pass: The error is used to change weights between the neurons in networks layers. Each data pattern to be trained using this process.

This process is still repeated time and time again until one of stopping conditions reached such as a specified number of trials elapse, or when the error falls in suitable level, or when the error can't do more improvement.

2.4.3.2 Radial Basis Function network (RBF)

RBF is a feed forward NN like MLP but with a different training process as presented in Figure 2.8. It includes only one hidden layer with a number of neurons that are completely connected to the output layer (Wu et al., 2008). The mapping function of RBF is mostly built based on the Gaussian activation function. The training process in RBF runs in two stages: the first is in the hidden layer, and the second in the output layer (Dibike & Solomatine, 2001).

In RBF, connections from the input layer to the hidden layer require unit weights without a training process. The hidden layer executes a fixed nonlinear transformation with constant parameters. The hidden layer also contains neurons and a parameter vector called the center that can be recognized as the hidden layer's weight vector. The standard

Euclidean distance is used to determine the distance between the center and the input vector (Haddadnia et al., 2003; Sahoo & Ray, 2006).

Each neuron in the hidden layer is represented by an activation function that receives and transforms the input. Euclidean distance is the input to the activation function while the output of the RBF neuron φ_i is described by Equation 2.6

$$\varphi_i(x) = \omega\|x - c_i\| \tag{2.6}$$

where c_j is the center of the ith RBF neuron, ω is the activation function, x is the input vector, and $\|x - c_i\|$ indicates a norm that is commonly the Euclidean distance.

While there are several options for ω, the Gaussian function is generally used as the activation function. As such, the output of the RBF neuron with the Gaussian activation function φ_i is described by Equation 2.7

$$\varphi_i(x) = -exp\left(-\frac{\|x - c_i\|^2}{2\alpha_i^2}\right) \tag{2.7}$$

where αi is the spread or the radial distance from the center of the i^{th} RBF neuron. The function value ω is maximum at the center, c, and decrease as the x, goes away from the center. Thus, the neurons in the RBF have localized receptive fields. The weighted sum of the inputs is sent to output using a activation function. The output y of the RBF is calculated using the Equation 2.8

$$y_m = \sum_{i=1}^{m} \beta_i \varphi_i(x) + b \tag{2.8}$$

where β_i is the joint weighted value of the ith basis function, b is the bias and m is the number of RBF centers. Since each RBF center must replay at least one input pattern, m

is less than or equal to the total number of input patterns. RBF accuracy is highly influenced by c, α, and m selection (Firat, 2008; Haddadnia et al., 2003; Iliyas et al., 2013).

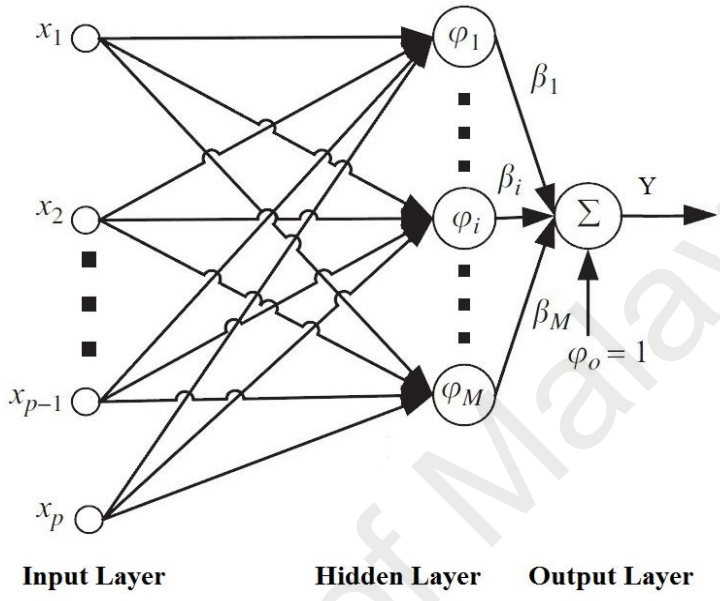


Figure 2.8: Schematic diagram of RBF architecture

(Iliyas et al., 2013)

2.4.3.3 Generalized Regression Neural networks (GRNN)

GRNN is new ANNs technique developed by Specht (1991). Unlike the feed forward neural 1986network, the training data are propagated through the network just one time as one-pass learning algorithms. Accordingly, it does not need an iterative training process and so training is accomplished in a short time (Firat, 2008).

MLP performance is greatly affected by the initial weight of the training process. Conversely, this issue is not challenged in GRNN training, nor is the local minima issue (Cigizoglu, 2005). Generally, GRNN performs comparably to other common ANNs techniques that use a smaller training dataset (KiŞI, 2006).

The GRNN structure is presented in Figure 2.9. GRNN contains four layers: the input, pattern, summation and output layers. The GRNN structure resembles that of MLP and RBF with two hidden layers. However, its training process differs vastly, since it depends on kernel regression networks with one-pass learning algorithms (Cigizoglu, 2005).

The number of neurons in the first layer is the number of input variables. The first layer is joined to the pattern layer which includes one neuron for each training pattern. The pattern layer is joined to the summation layer which consists of two dissimilar kinds of summation, namely a single division unit and summation units (Firat, 2008).

In GRNN training, the radial basis and linear activation functions are used in the pattern and output layers. The neurons of the pattern layer are linked to the S and D summation neurons in the summation layer. S represents the summation unit utilized to calculate the sum of weighted replies of the pattern layer. D represents the summation neuron used to compute un-weighted outputs of pattern neurons. The output layer only divides the output of each S by that of each D, yielding the output for the input pattern (Kim et al., 2004).

Considering a set of training data $\{x_i, d_i\}_{i=1}^{N}$ where x_i is the input vector, d_i is the corresponding output value and N is the total number of data patterns. GRNN is based on the following formulas:

$$Y_i' = \frac{\sum_{i=1}^{N} y_i . exp[-G(x, x_i)]}{\sum_{i=1}^{N} exp\,[-G(x, x_i)]} \quad (2.9)$$

$$G(x, x_i) = \sum_{k=1}^{m} \left(\frac{x_k - x_{ik}}{\sigma}\right)^2 \quad (2.10)$$

where y_i is the weight connection between the i^{th} neuron in the pattern layer and the S-summation neuron, G is the Gaussian function, m is the number of variables of an input vector, x_k and x_{ik} are the i^{th} element of x and x_i, respectively, and σ is the spread parameter whose optimum value is selected using trial and error (Firat, 2008).

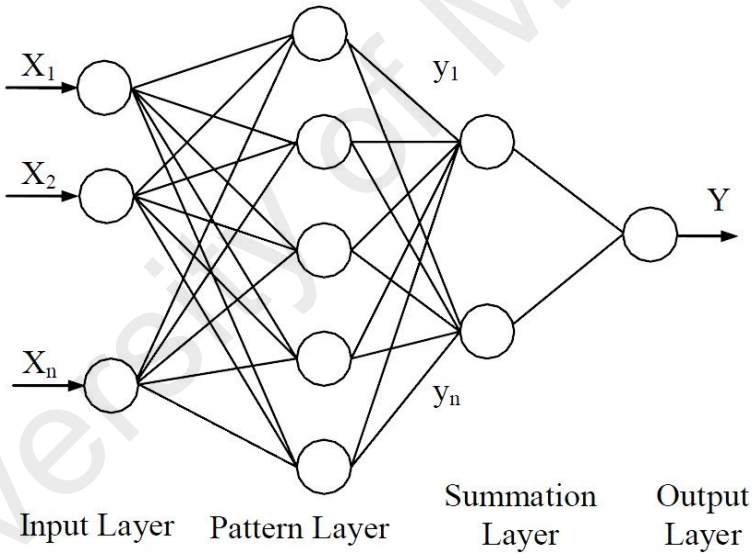


Figure 2.9: Schematic diagram of GRNN architecture

2.5 Support Vector Machine

2.5.1 Introduction to Support Vector Machine

SVM is the state-of-the-art neural network technology based on statistical learning theory introduced as a classification tool by Vapnik in the 1970s. SVM is a tool for investigating learning problem that offers adequate performance using linear or nonlinear function. It is a type of neural network that automatically determines the structural components. SVM has been applied successfully in various classification and clustering applications. Recently, it has been extended to apply regression and prediction applications (Khader & McKee, 2014; Li et al., 2013; Lou et al., 2013; Shi & Xu, 2012; Wei, 2015; Wieland et al., 2010; Zakaria & Shabri, 2012; Zhou et al., 2013).

In SVM, the linear model is primarily employed to set nonlinear class boundaries by nonlinearly mapping the input data into a high-dimensional feature space. In the new domain, the linear model characterizes a nonlinear decision boundary in the original domain (Solomatine et al., 2008; Wang et al., 2009). In other words, SVM constructs an optimal separating hyperplane in the new hypothetical field. This hyperplane may be either a line, a plane, or a surface that divides the data into two classes. When the data are split linearly, linear machines are trained for an optimal hyperplane that separates the data with minimum error (Chen & Yu, 2007; Solomatine et al., 2008).

SVM is advantageous because it follows the structural risk minimization principle, which aims to limit errors in both the training data set and the generalized model. With this feature, SVM can effectively generalize results even with limited input patterns (T. Asefa et al., 2006; Ding et al., 2014).

2.5.2 History of Support Vector Machine

SVM is a comparatively new AI modelling technique based on statistical learning theory introduced by Vapnik in the 1970s. SVM has been developed as a classification tool and it was applied successfully in a wide range of classification and clustering applications in. Recently, SVM have been successfully extended to apply in regression and prediction applications (Solomatine et al., 2008; Wu et al., 2008; Yu et al., 2006).

In the last few years, it has become a commonly used as modelling technique, due to the high performance of SVM, and have become an actual challenger to ANNs in regression and prediction applications. Since then there have been increasing SVM applications in the wide range fields such, civil engineering, water resources and other engineering applications (Tirusew Asefa et al., 2006; Behzad et al., 2009; Han et al., 2007; Misra et al., 2009; Zakaria & Shabri, 2012).

2.5.3 Training Process of Support Vector Machines

Consider a set of training data $\{x_i, d_i\}_{i=1}^{N}$ (x_i is the input vector, d_i is the corresponding output and N is the number of data patterns), the linear regression function of SVM can be expressed as follows:

$$f(x, w) = w \cdot \emptyset(x) + b \tag{2.11}$$

where w is the weight vector; b is the bias; and $\emptyset$ is nonlinear mapping function.

The two factors (*w* and *b*) are computed by minimizing the following function:

$$L_{\varepsilon}(y, f(X, \omega)) = |y - f(X, \omega)|_{\varepsilon}$$

$$L_{\varepsilon}(d_i, y_i) = \begin{cases} |d - (w \cdot \emptyset(x) + b)| - \varepsilon & if\ |(w \cdot \emptyset(x) + b) - y| > \varepsilon \\ 0 & otherwise \end{cases} \quad (2.12)$$

where *y* represents observed value. ε denotes the tube size and corresponds to the approximate accuracy of the training data points. Within the extent of the ε-tube and penalized losses L_{ε}, the loss function describes the tolerated errors when data are located external of the tube.

The nonlinear SVR problem can be expressed as the following optimization problem:

$$R_{w,\xi_i,\xi_i^*} = C \sum_{i=1}^{N} (\xi_i + \xi_i^*) + \frac{1}{2}\|\omega\|^2 \quad (2.13)$$

$C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$ is the first term in Equation (2.13) and represents training error (risk).

where ξ and ξ^* are slack variables represent the upper and lower training errors, respectively, subject to error tolerance ε. These variables describe the difference between the observed data and the related boundary values of the ε-tube.

It is zero when the predicted data are within the ε-tube, as shown in Figure 2.10. $\frac{1}{2}\|\omega\|^2$ is the second term and denotes the generalization term. It is a measure of function flatness. C is a positive constant that represents the regularized constant and regulates the trade-off between empirical risk and the regularization term. By maximizing the value of *C*, we can enhance the significance of empirical risk relative to the regularization term.

Equation (2.13) can then be solved using Equations (2.14) and (2.15) according to the following convex optimization problem:

Minimize:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2.14}$$

Subject to $\left.\begin{matrix} w_i\,\emptyset_i(x) + b - d_i \le \varepsilon + \xi_i \\ d_i - (w_i\,\emptyset_i(x) + b) \le \varepsilon + \xi_i^* \\ \xi_i\,,\xi_i^* \ge 0 \end{matrix}\right\}\ i = 1,2,3 \tag{2.15}$

Figure 2.10 shows the main concept of SVM based on Equation (2.13-2.15). In this regression problem, most data patterns are presumably within the ε-tube. If the data pattern (*xi, di*) is outside the ε-tube, errors are induced in ξ and ξ^*. These variables are thus reduced in the objective function. By limiting both the regularization $\frac{1}{2}\|\omega\|^2$ + and the training error $\sum_{i=1}^{N}(\xi_i + \xi_i^*)$, in order to alleviate under- and over-fitting (Chen & Yu, 2007; Noori et al., 2011; Wu et al., 2008).

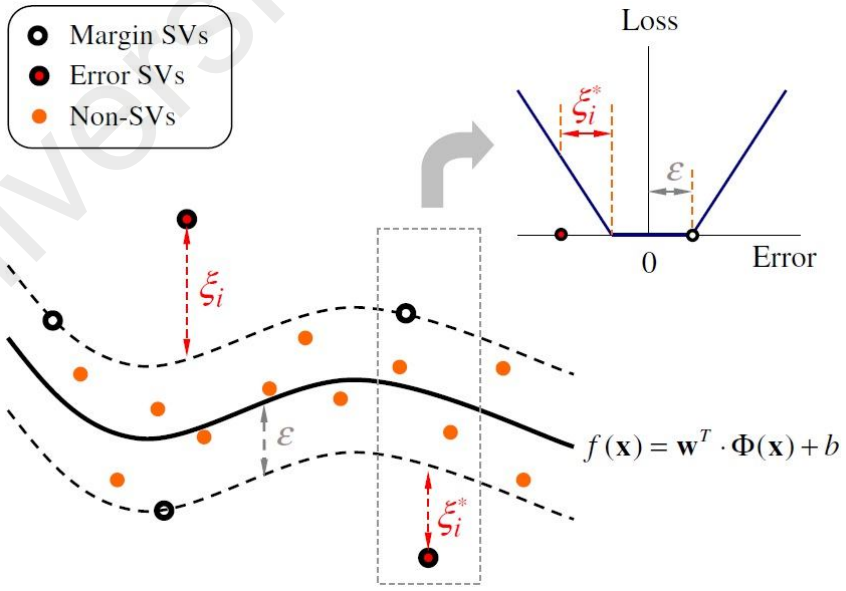


Figure 2.10: Nonlinear SVM with Vapnik's e-insensitive loss function (Chen & Yu, 2007)

Figure 2.11 presents the Schematic diagram of SVM, where the K(xi,x) is the output of the i_{th} hidden node for input vector x, it is a mapping of the input x and the support vector x_i by selecting the kernel function (Chen & Yu, 2007).

Some mostly used kernel functions in SVM are as follows:

- Linear $K(x_i, x) = x_i \cdot x$
- Polynomial $K(x_i, x) = [\gamma(x_i \cdot x) + c]^d$
- Sigmoid $K(x_i, x) = \tanh[\gamma(x_i \cdot x) + c]$
- Radial basis function $K(x_i, x) = \exp(-\gamma|x_i - x|^2)$

Many applications in hydrological modelling have proved the efficiency of the radial basis function in SVM. The results of the SVM model to be stated as Equation (2.16),

$$f(x) = \sum_{k=1}^{m} \overline{\alpha}_k \cdot K(x_k, x) + b \qquad (2.16)$$

Where, x_k represents the support vector, and m represents the number of support vectors. The SVM model employed herein has three interdependent parameters (C, ε, γ) to be valued. The near optimal values of these parameters are obtained by a trial and error method. The Lagrange coefficients $\overline{\alpha}_k$ and the bias term b can be solved analytically, and the best structure is thus achieved.

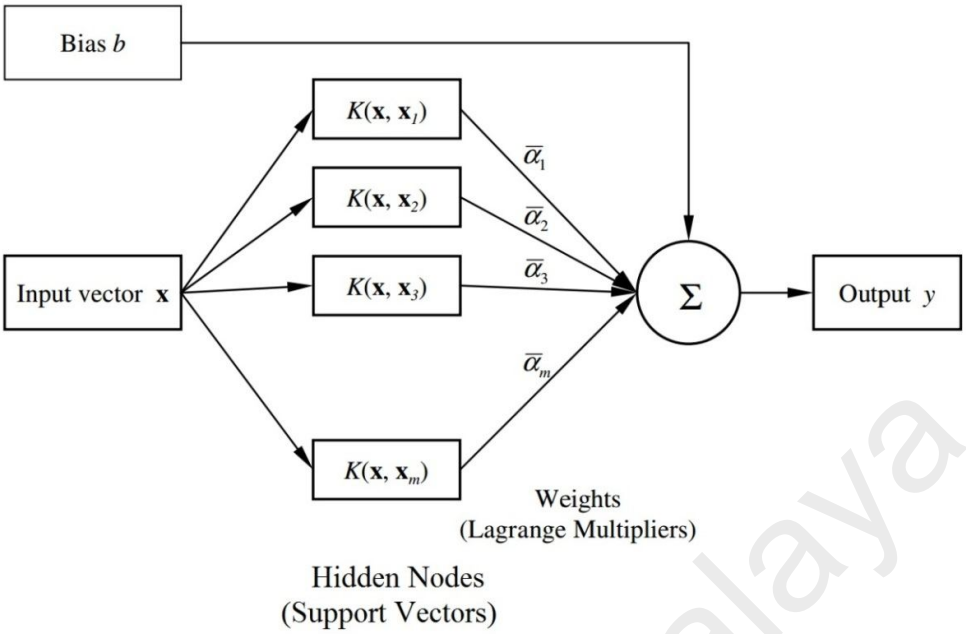


Figure 2.11: Schematic diagram of SVM architecture (Chen & Yu, 2007)

2.6 Previous Studies

Numerous applications of SF prediction and analysis have recently been implemented worldwide. They are now in the mainstream of hydrological science in terms of the number of approaches, techniques, models, applications, and publications. Hundreds of applications can be found in the literature on this field. A few examples of related applications are presented and evaluated in the following subsections.

2.6.1 Long-Term Variations Analysis of the Stream Flow

Investigations of the long-term variations of SF regime have increasing attention over the last decade. Several applications have been performed worldwide to explore the probable long-term variations in SF regime.

Burn et al. (2004) investigated the hydrologic trends of the Liard River basin in Canada. They reached the conclusion that the river's hydrologic trends are related to variations in both climatological and large-scale atmospheric progression.

Xiong and Guo (2004) studied the changes in the annual SF of the Yangtze River in China, using a trend test. According to results, there was no important trend in the annual maximum flood, whereas a declining trend was noticed in the annual minimum and mean SF.

Birsan et al. (2005) evaluated the trends in the daily SF data of 48 basins in Switzerland using the Mann–Kendall test over 3 sub-periods (1931–2000, 1961–2000, 1971–2000). An increment in annual SF was noted in approximately 60% of the stations.

Ma et al. (2008) examined the variations in annual SF for 8 sub-basins in the Shiyang River basin in China over a 50-year period. They employed the Kendall and Pettitt tests to detect changes in SF. According to results, significant decline was verified in the annual SF in 5 of the 8 sub-basins.

Kumar et al. (2009) investigated the SF trend in Indiana state which is located in Midwest region of USA using 31 gauging SF stations over about 50-year period. Trends were estimated using the Mann–Kendall test for SF characteristics including low, mean, and high annual SF. A declining trend was verified in low and mean SF in about 70% of stations.

Korhonen and Kuusisto (2010) analyzed the variations of SF regime in Finland over the 1912–2004 period. Neglected changes were noticed in mean annual SF, but significant changes in seasonal means of SF were detected. The mean monthly SF in winter and spring where increased, whereas no variations in the mean monthly SF in autumn were noticed.

Burn et al. (2010) studied the probable influence of climate change in extreme hydrological events trends' in Canada. Both high and low SF events of 68 gauging stations were investigated over about 50-year. They utilized the Mann-Kendall test to evaluate the trends. The results verified a decline in the maximum annual SF and maximum spring SF with declining trend in event timing (earlier events).

Morán-Tejeda et al. (2011) examined the changeability in the Duero River basin's SF regime (northern Spain) over the 1961–2006 period. A moderate decrease in SF was witnessed in winter and spring. Variations of SF regime appeared in both the timing and magnitude asspects.

Gautam and Acharya (2012) applied a widespread trend investigation system of SF in Nepal. According to results, 24% of variables had trends, of which 41% were declining and 59% increasing.

Miao et al. (2012) explored the explanations behind SF changes in the Yellow River in China. They employed SF data from 23 stations over about 50-year period from 1956 to 2008 in the investigation process. According to results, SF presented a declining trend.

Hannaford and Buys (2012) investigated the trends in SF regime in UK over about 40-year period from 1969–2008. Trends were inspected depending on seasonal scale. According to results, SF and high SF event increased in winter and autumn, whereas SF declined in spring and no variations appeared in summer.

Sun et al. (2013) quantified the variations in annual SF using climate elasticity in Poyang Lake Basin, China. They utilized SF datasets at four hydrological stations over 40-year period from 1961 to 2000. Slight variations have been noticed in the timing of the mass center of the SF with minor increase of annual SF.

Z. Chen et al. (2013) evaluated the impacts of climate change and human activities on SF regime of Kaidu River Basin in China over 50-year period from 1960 to 2009. According to results, noticeable increases have been verified in annual SF.

Jena et al. (2014) investigated the reasons for recent floods in the Mahanadi river basin in eastern India. Trends of extreme floods were analyzed related to trends of extreme RF. The analysis proved that the latest occurrences of floods are because of an increment in extreme rainfall in the basin.

Mediero et al. (2014) analyzed the floods trends' in magnitude, frequency and timing aspects in Spain. Annual maximum SF data were extracted in three periods: 1942 to 2009, 1949 to 2009 and 1959 to 2009. Mann–Kendall test was chosen to explore trends. An overall declining trend in magnitude and frequency of floods was noticed in the three periods, with more remarkable occur in the third period. A trend in the timing (i.e. towards later floods) was also noticed in the northwest of Spain. Most such trends in flood series could be elucidated by the rising trends in evapotranspiration.

2.6.2 ANNs Applications in Stream Flow Modelling

ANNs have been widely applied in SF modelling and prediction for a multiplicity of objectives in the two last decades. Nilsson et al. (2006) investigated the opportunity of modelling monthly Sf for two Norwegian river basins using ANNs and conceptual runoff modelling (CM). ANNs offered the best predictions of monthly SF for both basins with R^2 of 0.82 and 0.71, respectively. Thereafter, they used a combination of ANNs and CM improve the modelling performance. The R^2 for both basins was improved to 0.86 and 0.75, respectively.

Nayebi et al. (2006) employed ANNs-based models for daily SF prediction in the upper sub-watershed of the Kor watershed in southwestern region of Iran. They investigated the

effect of minimum air temperature on the modelling performance. They developed four ANNs-based models with deferent combination of input variables for SF prediction. The models with minimum air temperature as input variables achieved the best performance in SF prediction.

Sahoo and Ray (2006) applied ANNs for SF prediction of a Hawaii stream in Hawaii island, USA. The predicted SF by ANNs-based models were compared to SF estimated by conventional rating curves which estimated by the United States Geological Survey, the results verified that ANNs-based models outclass the conventional rating curves in SF prediction.

Sahoo et al. (2006) applied ANNs to evaluate flash floods and their associated with water quality specifications using observed data of a Hawaii stream in Hawaii island, USA. They demonstrated that ANNs can predict SF, turbidity and specific conductance with high R although they didn't reached good results in prediction of dissolved oxygen, PH and water temperature.

Pulido-Calvo and Portela (2007) employed ANNs for one-step ahead daily SF forecasting in some Portuguese basins. They applied many ANNs-based models with several inputs combinations such the flow in previous days. The models with inputs of three previous days flow combined verified very high performance. Generally the work proved that it is potential to reach accurate daily SF predictions using ANNs, even with inadequate data.

Aqil et al. (2007) utilized two types of ANNs namely, feed forward and recurrent NN with three types of training algorithm in real-time SF prediction. ANNs-based models were developed and evaluated based on results for 1 to 5-h ahead prediction in the Cilalawi River in Indonesia. According to results, high performance was reached for most of models for 1-h ahead with R around 0.91. However, the model performance declines

with increasing the lead-time, the results suggested that recurrent and feed forward network are capable to predict the SF up to 5 hr in advance with high accuracy.

Jain and Kumar (2007) developed a new hybrid time series NN model contains an overall modelling framework. It integrates between the conventional and ANNs, it was tested using the monthly SF data in Colorado River in Lees Ferry, USA. The results suggested that the proposed approach provided a strong modelling tool able to capture the non-linearity of the hydrological time series and accordingly generating more accurate predictions.

Ahmed and Sarma (2007) generated SF data of the Pagladia River in Assam, India by ANNs. They compared ANNs with other existing models such as autoregressive moving average (ARMA) model and Thomas-Fiering model. The comparison conducted based on five different statistics of the historical data and synthetically generated data. ANNs demonstrated the highest performance in generating SF data.

Gopakumar et al. (2007) explored the applicability of ANNs for prediction of daily SF in the Achencoil River basin un India. Although the developed ANNs model revealed high performance for rainy period, the performance for the low flow period didn't revealed same performance. To improve ANNs-based models performance', the modelling data were analyzed using Self Organizing Maps (SOM). The new approach for SF modelling utilizing the result of SOM analysis enhanced the model performance of daily SF prediction.

Singh and Kumar (2007) proposed the application of ANNs to estimate the missing mean monthly SF of Narmada River in India. The performance of ANNs-based model was compared with the Langbein's log deviation method and provided an adequate alternate to this method.

Turan and Yurdusev (2009) employed both of FFBP, GRNN and fuzzy logic to estimate missing SF records by the records of the four SF gauge upstream stations in the Birs River in Switzerland. The performances of these models were evaluated to select the best fit model. Based on the performance evaluation, the three modelling techniques demonstrated acceptable performance. However, FFBP outperforms other models.

Kentel (2009) applied ANNs for estimation of monthly SF. They used previous RF, SF, and the related month as input variables to forecast SF of Guvenc River in Turkey. They studied the influences of input vectors, number of training trails, and initial weights of the connections of NN on the ANNs-based models performance. ANNs achieved promising results in SF prediction, making it an adequate alternative SF prediction technique.

Rakhshanehroo et al. (2010) applied ANNs for flood prediction in similar basins in Iran. ANNs was trained as an event-based modelling tool utilizing data from only 2 of the basins. The prediction performance then evaluated for all basins, high prediction accuracy was reached. They came to conclusion that the ANNs model may be utilized for flood prediction in similar basin with accepted accuracy.

Besaw et al. (2010) developed ANNs-based model to predict SF in ungauged basins from sub-basins in Northern Vermont, USA. They employed time-lagged records of RF and temperature as input variables of the ANNs-based model. Time series analysis of the climate Q data offers an efficient method to decide the suitable steps number of time-lagged input variables. The results suggested that the proposed methods are appropriate to predict Q in ungagged basin and also it was shown that Q prediction was superior to those using daily records for the small river basins.

Sahu et al. (2011) applied ANNs for forecasting SF in open channel flow. They compared the performance of ANNs-based model with four widely used approaches. The results showed that the ANNs-based model is superior to other models in SF prediction.

Machado et al. (2011) explored the ability of ANNs-based model to predict the monthly RF-SF in the Jangada River basin, Paraná, Brazil. The prediction accuracy of ANNs-based model was compared to those of a conceptual-based model. The ANNs-based model achieved the higher performance based on some statistical performance evaluation criteria such as R and Nash-Sutcliffe statistics.

Tiwari et al. (2012) proposed a novel modelling approach for prediction of daily SF using neural units with higher-order synaptic operations (NU-HSO). They compared between ANNs-based models with NU-HSO and conventional ANNs-based models. The prediction process was performed using 1- to 5-day lead time prediction in the Mahanadi River basin at the Naraj gauging station. According to results, ANNs-based models with NU-HSO achieved higher performance than conventional ANNs-based models based on some statistical performance evaluation criteria such as R and Nash-Sutcliffe statistics. Thus, this results shows that ANNs-based models with NU-HSO can be an adequate alternative SF prediction technique.

Tiwari et al. (2013) employed self-organising maps (SOM) to homogeneously classify the data sets of four types of ANNs-based models developed for daily SF predictions. The four types are traditional ANNs, wavelet-based NN (WNN), bootstrap-based NN (BNN) and wavelet-bootstrap-based NN (WBNN). According to results of pridection process, the SOM's efficiency in clasfication of data into different clusters were noticed through enhancing the accuracy and reliability of daily SF prediction.

Wei et al. (2013) developed WNN hybrid modelling approach for monthly SF prediction in the Weihe River, China. Monthly SF records from three stations were employed to train and test the model for 48-month-ahead prediction. The prediction results using WNN achieved high enhancement in the model performance in comparison to the results of normal ANNs-based model.

Sahay and Srivastava (2014) developed a wavelet transform-genetic algorithm-neural network model (WAGANN) for prediction 1-day-ahead monsoon SF. Discrete wavelet transform (DWT) was used for preprocessing the time series and genetic algorithm (GA) for optimizing the initial parameters of an ANNs to the NN training. Four WAGANNs models with different combination of inputs variables are developed for prediction of SF in two Indian Rivers, the Kosi and the Gandak. According to results, WAGANNs models domestrated better pridection accuracy than autoregression models (ARs) and GA-optimized ANNs based models which use original SF time series for inputs variables.

Elsafi (2014) employed ANNs to predict SF in the Nile River at Dongola Station in Sudan. Readings from stations along the Blue Nile, White Nile, Main Nile, and River Atbara from the period between 1965 and 2003 were employed in the modelling process. The results showed that the ANNs model may be utilized for flood prediction in the Nile River with high accuracy.

2.6.3 SVM Applications in Stream Flow Modelling

Given that it is comparatively new, SVM is not as widely applied as ANNs, although it has been applied in several SF applications. However, recent literature provides some applications of SF modelling and prediction using two or more AI techniques (i.e., SVM and ANNs) to improve the performance of the modelling process.

Tirusew Asefa et al. (2006) applied SVM for SF predictions in arid regions based on two time scales: seasonal SF and Q. The results of these models revealed a good efficiency in explaining spatial and temporal SF process. SF was predicted using local-climatological data and demanding less input variables than process-based models. Seasonal SF prediction was also enhanced by integrating atmospheric circulation indicators in the modelling process.

Lin et al. (2006) applied SVM for long-term SF prediction. They used a shuffled complex evolution algorithm to detect the suitable specifications of the SVM-based model. The SVM prediction model was evaluated by the long-term monthly SF records of the Manwan Hydropower station, Lancang River in Yunnan Province, China. They compared the performance of SVM-based model with ARMA and ANNs-based models, the results verified that SVM could be considered as a very promising tool in long-term SF prediction.

Chen and Yu (2007) applied SVM in real-time flood prediction in Lan-Yang River, Taiwan. They used the cross-correlation technique to select the input variables of the SVM-based models. The real-time prediction performance was evaluated, the results indicated that the SVM is a probable prediction technique in SF.

Behzad et al. (2009) investigated the ability of SVM, ANNs, and ANNs integrated with GA (ANNs–GA) models to predict one-day lead SF of the Bakhtiyari River in Iran. They used local climate and RF records in the modelling process. The results proved that the SVM-based model outperforms the ANNs and ANNs–GA in predicting one-day lead SF.

Noori et al. (2011) explored the ability of some preprocessing techniques (i.e. principal component analysis (PCA), Gamma test (GT), and forward selection (FS) in improving the monthly SF prediction by SVM in the Sofichay River, 120 km to Tabriz southwest, Iran. They used 18 input variables, such as monthly RF, discharge, sun radiation, and temperature (as minimum, maximum and mean) with three temporal lags belong to t, t-1, and t-2. Consequently, PCA, GT, and FS techniques were applied to decrease the input variables from 18 to 5 by PCA and GT, and to 7 by FS. Results showed that the performance of the improved SVM-based model (i.e. PCA-SVM and GT-SVM) models outperform the conventional SVM-based model. R^2 between the observed and predicted

SF for PCA-SVM based model was equal to 0.92 and 0.88 in the training and testing data sets, respectively.

Guo et al. (2011) applied SVM for monthly SF prediction. They used adaptive insensitive factor to improve the performance of SVM-based model and the wavelet denoise method to minimize the noise in SF data. The performance of the SVM-based model is explored and compared with the performance of ANNs-based model. The results verified that the improved SVM-based model is of better generalization capability and prediction accuracy than ANNs-based models.

Samsudin et al. (2011) proposed a novel hybrid prediction model for monthly SF, which combines the group method of data handling (GMDH) and the least squares support vector machine (LSSVM). They applied GMDH to determine the useful input variables for the LSSVM model. Monthly SF data from two stations, the Selangor and Bernam Rivers in Selangor state of Peninsular Malaysia were employed in the modelling process. The performance of the new hybrid model was compared with ANNs, Autoregressive Integrated Moving Average (ARIMA). RMSE and R were used to evaluate the models' performances. The new hybrid model has been found to provide more accurate prediction compared to the other models.

Shabri and Suhartono (2012) applied SVM for monthly SF prediction in the Kinta River in Perak, Malaysia. They investigated the capability of a least-squares support vector machine LSSVM model to enhance the performance of SF prediction. They applied Cross-validation and grid-search techniques to determine the model variables. The accuracy of the LSSVM model was compared with the conventional statistical ARIMA, ANNs and conventional SVM models. According to results, the LSSVM model outclasses the other modelling techniques and it could be employed effectively in SF prediction.

Kisi et al. (2012) evaluated the performance of some AI techniques such as ANFIS, ANNs and SVM in prediction daily SF in two stations in north-western Turkey. They also compared the performance of the three models with two linear regression models. The results showed that the ANFIS, ANNs and SVM are superior in prediction of daily SF.

Kalteh (2013) applied two AI techniques (i.e., SVM and ANNs) in prediction of monthly SF of Kharjegil and Ponel stations in Iran. He also coupled the SVM and ANNs with the wavelet transform to improve the modelling performance. According to results, both ANNs and SVM coupled with wavelet transform, provided more precise prediction than the traditional ANNs and SVM. However, it is noticed that SVM coupled with wavelet transform provided better prediction than ANNs coupled with wavelet transform. The results also indicated that traditional SVM outperform slightly better than traditional ANNs.

Ch et al. (2013), investigated the ability of the hybrid model (support vector machine-quantum behaved particle swarm optimization) SVM-QPSO in forecasting monthly SF of Vijayawada and Polavaram stations of Andhra Pradesh in India. The results indicated that SVM-QPSO is an accurate and reliable prediction tool for monthly SF.

Tehrany et al. (2015) proposed a novel ensemble method by coupling SVM and frequency ratio (FR) to produce spatial modelling in flood formation assessment in the upper catchment of the Kelantan basin in Malaysia. They applied another machine learning algorithm (decision tree (DT)) to evalulate the performance of the proposed method. Around 155 flood sites were selected from several sources over the study area. The flood sites were accidentally separated into two dataset; (115 sites) for training and the remaining (40 sites) for testing. According to results, coupling SVM and FR domestrated higher prediction accuracy than DT as the prediction rate was 85.21% and 82.00 % for

the two methods respectively. The results demonstrated the efficiency of the proposed ensemble method in flood formation assessment.

Wei (2015) proposed a new method to predict river stages with a head prediction from 1 to 4 hr in the Tanshui River Basin, Taiwan throughout 50 historical typhoon events over 11-year period from 1996 to 2007. He employed both lazy and eager learning approaches. Two lazy learning models namely, the locally weighted regression (LWR) and the *k*-nearest neighbor (*k*NN) models and three eager learning models ANNs, SVR, and linear regression (REG) where employed in this study. According to results, in the eager learning models, ANNs and SVM produced more accurate prediction results than REG while in the lazy learning models, LWR outperformed more than *k*NN.

2.6.4 Accurate Time Applications in Stream Flow Modelling

Although the Lt estimation process has been extensively studied during the current and past decades (Allen, 1976; Askew, 1970; Banasik et al., 2005; Bhadra et al., 2010; Honarbakhsh et al., 2012; Reed et al., 1975; Toth, 2008), the accurate timing of input and output variables of SF prediction models is still a general problem in AI-based models; consequently, this issue is under ongoing investigation by hydrological modelers worldwide (Abrahart et al., 2007; Akhtar et al., 2009; Nourani et al., 2014).

Sudheer et al. (2002) developed a novel statistical method of minimizing the timing errors in SF predictions models. The technique is based on statistical analyses, such as cross-, auto- and partial-auto-correlation of the potential influencing variables that correspond to different time lags. The results obtained from these statistical analyses helped to identify the most adequate variables combinations. The approach of (Sudheer et al., 2002) has been also applied by (Aqil et al., 2007) in real-time SF modelling using ANNs.

de Vos and Rientjes (2005) investigated the constraints facing the application of ANNs for RF-runoff modelling. They found that timing errors is one of the main limitations as a result of a dominating autoregressive component introduced by using previous SF records as model input. Two probable solutions to the timing problem were proposed. The first solution is to try several alternatives in the determination of the variables of ANNs-based models. The second solution is to evaluate the performance of model through a combination of multiple indices during the modelling process.

Yu et al. (2006) employed the hydrological theory of response time in river basin in determination of the variables of SVM-based model which developed for real-time flood stage prediction in Lan-Yang River, Taiwan. They developed two models to predict multiple-hour-ahead stage. The results verified that the SVM-based models with accurate selection of variables can be applied successfully in prediction of the flood one-to-six-hours ahead.

Abrahart et al. (2007) integrated a time-error correction procedure into the optimization process of ANNs –based models to predict SF in both short and long forecasting periods. Their course of action produced adequate improvement over a shorter forecasting period, but only slight improvement for longer forecasting was reached.

Talei and Chua (2012) studied the influence of Lt on event-based RF-runoff modelling with some AI-based models (i.e. the adaptive network-based fuzzy inference system (ANFIS)) in a sub-basin within the Kranji Reservoir basin in Singapore. They observed that the models produced considerably more accurate results compared to other models in which Lt was not included in the modelling process.

Chang et al. (2014) applied one static ANNs and two dynamic ANNs to develop multi-step-ahead (Floodwater Storage Pond) FSP water level prediction models in the Yu–Cheng station in Taipei, Taiwan through two scenarios. Scenario 1 assumes RF and FSP

water level records as input variables of the model while scenario 2 assumes only RF records as input variables of the model. They calculated R in the recognition of the maximum correlations between FSP water level and RF at different lag intervals for each station then employed Gamma test (GT) to select the effective variables (RF stations) that pointedly influence the FSP water level. According to results, the GT can recognize the effective RF stations as inputs to the ANNs-based models. Coefficients of efficiency of prediction process is within 0.9–0.7 (scenario I) and 0.7–0.5 (scenario II) in the testing stages for 10–60-min., ahead predictions, respectively. According to results, it can be concluded that ANNs-based model is beneficial tool to the local authorities for flood control and awareness.

2.6.5 General Remarks on the Previous Studies

Over the last two decades or so, the use of AI techniques for the prediction of hydrologic processes has become a well-established research field. In the late 1990s, AI was a novel modelling method and, accordingly, considerable number of studies was directed to the employ the AI in several hydrological processes to judge their convenience as a new alternative modelling tool. A comprehensive literature review of the applications of AI techniques in hydrology and water resources engineering over that era is provided in the Task Committee on Application of ANNs in Hydrology by the ASCE (ASCE, 2000a; ASCE, 2000b) and some other review papers (Dawson & Wilby, 2001; Maier & Dandy, 2000).

Given the fast evolving of AI techniques in the hydrologic modelling during the last 15 years, several review papers was prepared to describe and evaluate the use of AI techniques in hydrological applications. Table 2.1 presents some review papers on the application AI-based models on hydrology and water resources over the last decade, from 2006 to 2015.

Table 2.1: List of some review papers on the application AI-based models on hydrology over the last decade

Author and Year	Paper title
Solomatine (2006)	DDMs and computational intelligence methods in hydrology
Kalteh et al. (2008)	Review of the SOM approach in water resources: Analysis, modelling and application
Solomatine and Ostfeld (2008)	DDMs: Some past experiences and new approaches
Maier et al. (2010)	Methods used for the development of NN for the prediction of water resource variables in river systems: Current status and future directions
Abrahart et al. (2012)	Two decades of anarchy? Emerging themes and outstanding challenges for NN river forecasting
Nourani et al. (2014)	Applications of hybrid wavelet–AI models in hydrology: A review
Seth (2015)	Use of ANNs and GA in Urban Water Management: A Brief Overview

Among the literature review, many remarks were obtained based on investigation of the methods and results included in the previous studies:

- The majority of SF prediction studies employed a daily or monthly time step, whereas the prediction of Q using AI techniques especially in humid tropical regions is uncommon in the literature. To the best of the researcher's knowledge, AI techniques are yet to be used to predict real-time Q in the Selangor River basin.

- The main objective of the Hydrologists and Neurohydrologists is to develop efficient SF models in terms accuracy, simplicity, less-demanding usage, applicability and cost

effectiveness. Therefore, considerable efforts are required to simplify the AI-based model along with enhancement of its performance particularly in complex process such as SF prediction. Developing models with high performance make them beneficial tool to the local authorities for the SF prediction and related applications.

- Despite the huge efforts in hydrological modelling by AI, so far, there is no recognized method for selecting the best input variables and optimum structure of AI-based model in Q prediction. Trial-and-error method was commonly employed which requires efforts, particularly in complex application.

- The performance of Q prediction using AI techniques still requires more improvement, given the low performance and limited applicability of AI-based models in many previous studies. In addition the methodologies for AI-model's development need to be more simplified with higher chance of applicability.

- The accurate timing of input and output variables of SF prediction models is still a general problem in AI-based models; consequently, this issue is under ongoing investigation by hydrological modelers worldwide.

- One of the significant concerns was investigating which AI techniques can best fit the hydrological description of SF. The weak awareness about this matter has resulted in the use of different AI techniques without adequate concern as to the suitability and applicability of the developed models.

- Most of the developed AI-based models for Q prediction did not provide enough investigation of the influence of the hydrological variables on Q which is the base for hydrological description. Despite of the black box nature of AI techniques, the recent progress in AI assists to produce general hydrological description of the SF in both time and spatial aspects as can be seen in few studies.

2.7 Preliminary Considerations in Stream Flow Modelling Using AI-based Models

Many preliminary considerations should be addressed and solved before the SF modelling process can be performed. These considerations include several topics, such as the advantages and disadvantages of AI techniques, the selection of appropriate modelling techniques, the determination of the most adequate variables of the AI-based models and the lag intervals between them and how to improve the performance of the modelling process. It could be achieved throughout improving the understanding of the SF process by some procedures such as the analysis of long-term changes in SF regimes and the Lt estimation.

2.7.1 Advantages and Disadvantages of Artificial Intelligence Techniques

Given the progress of computer technology, AI techniques are now capable of implementing fast, easy, and efficient simulations of complex processes and huge number of data patterns (Dawson et al., 2006). Hydrological modelling using AI techniques has many advantages over traditional modelling techniques, some of which are as follows:

- AI-based models have the capability to directly use field-recorded data without detailed inspection and simplification, in contrast to traditional regression analyses, which require it in advance. However, AI techniques require less inspection but there is need to check the quality of data in terms of normality, homogeny, etc.

- AI-based models have the capability to model and simulate any types of nonlinear systems and process that are difficult to model by traditional modelling techniques.

- Trained AI-based models can generate a predicted value for the output variable for any reasonable input pattern that fell within the range of their training data.

- AI-based models can investigate mathematical interactions between variables of the considered process and the effect of the independent variables on the dependant variable (Baxter et al., 2004; Kerh & Lee, 2006).

Despite the significant advantages of AI techniques, a number of disadvantages limit their operation and may adversely affect their performance. The effects of the disadvantages of AI techniques can be dramatically minimized by the better understanding and careful practice of AI-based models. The main limitations of AI techniques are as follows:

- The modelling capabilities of AI techniques are highly affected by the specifications of the model. Unfortunately, no ideal approach exists so far for determining the best specifications of AI-based models, such as the structure and training algorithm of ANNs and the kernel function and hyper-parameters (i.e., C and ε, respectively) of SVM. As an alternative, the specifications of AI-based models are generally selected based on the previous experience of the modeller and some guidelines and procedures listed in the literature (ASCE, 2000; T. Asefa et al., 2006).

- There is no recognized approach to selecting the optimum modelling technique. Neurohydrologists commonly employ a trial-and-error method, followed by a detailed performance evaluation analysis, which sometimes requires huge efforts, particularly in complex applications (Basheer & Hajmeer, 2000; Maier et al., 2010).

- The danger of over fitting sometimes occurs in ANNs because they behave like the human brain, which has the defect of over-memorizing (i.e., called “over

fitting" in ANNs techniques). ANNs-based models can be over trained; as a result, their generalizability in the prediction process diminishes. Over fitting commonly occurs when the training time is too long or the model includes too many hidden neurons (Abrahart et al., 2004).

- AI techniques are efficient interpolators but not efficient extrapolators. The capability of ANNs to perform well when deal with input data that fell within the range of their training data, but their performance declined when the input vectors were far from the range of the variables used for training purposes (Sivakumar & Berndtsson, 2010).

- There is no ideal approach to select the best number of dataset pattern or dividing them during the training process. This is a classic challenge that is still facing Neurohydrologists and commonly implementing trial and error method, of their own choosing. The more patterns lead to slower training process and could be expensive or not easy to collect whereas an inadequate number of records decline the performance of AI-based model (Bowden et al., 2005; Sivakumar & Berndtsson, 2010).

2.7.2 Selection of the Appropriate Stream Flow Modelling Technique

No single data driven-based modelling technique can explore all the hydrological modelling and prediction processes, particularly those with complex processes such as SF. The wide range of modelling techniques are depending on different aspects of modelling mechanisms, and thus no single technique can fully investigate all the features of SF process (Sivakumar & Berndtsson, 2010).

The decision to choose any of the SF modelling techniques mainly depends on two elements: first, the available amount of understanding of the physical behavior of the SF

process and the river basin hydrology, and second, the availability of hydrological data that describe the SF and related variables. Figure 2.12 shows the appropriateness of SF modelling techniques based on two elements: understanding the physical process and availability of data.

Figure 2.12 shows that in the case of sufficient understanding both the physical process and the data, the process-based models are the suitable modelling technique. The model in this case is developed depending on the physical understanding of the process. The model may then be calibrated using the available data. In the case of sufficient data but insufficient understanding of the physical process, the AI-based models are the suitable modelling technique. If both data and understanding of the physical process are insufficient, statistical methods can be applied only as an initial tool to improve the general understanding of the SF process (Basheer & Hajmeer, 2000; Sivakumar & Berndtsson, 2010).

After the data collection process in the study area, Selangor River basin, it was noticed that the description of the physical characteristics of the SF process and the related parameters of the river basin hydrology have not been sufficiently addressed and described yet. However, enough data have been successfully collected for the SF and the related variables. Accordingly, AI-based models are the most suitable modelling technique for SF prediction and modelling based on the availability of sufficient modelling data.

Several types of modelling techniques can be classified as AI-based models. In this research, ANNs and SVM are applied in SF modelling and prediction, whereas statistical methods are applied in the analysis of the long-term variations of SF regimes.

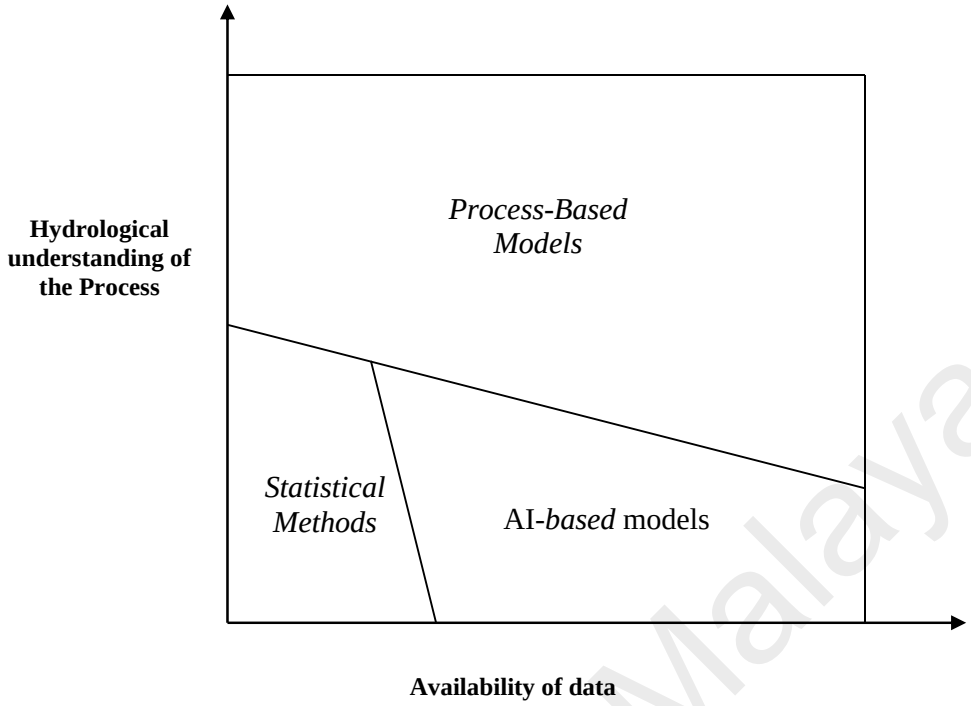


Figure 2.12: Appropriateness of stream flow modelling techniques

Adopted from (Basheer & Hajmeer, 2000)

2.7.3 Determination of the Input and Output Variables of the AI-based models

SF modelling and prediction using AI is a complex process and it is influenced by many parameters that are intricately interrelated (Akhtar et al., 2009). Determining the most adequate input and output variables is considered one of the main challenges in developing AI-based models. In SF prediction, it is commonly based on a priori and previous knowledge of the river basin hydrology, which provides an initial indication of the potential inputs and outputs (Kretzschmar et al., 2014; Maier & Dandy, 2000; Nourani et al., 2014).

Given that the main goal is to predict the ahead real-time Q in the downstream area of a tropical river from the upstream WL and RF records, the hourly records of the WL and RF upstream stations are used as input variables of the AI-based models, whereas the

hourly records of SF in the downstream station are used as output variable of the AI-based models.

The input and output variables of the AI-based models where selected through three scenarios, scenario (1) adopts only the RF data of upstream stations as input variables, scenario (2) adopts only the WL data of upstream stations as input variables, and scenario (3) adopts both the WL and RF data of upstream stations as input variables.

Equation 2.16 shows the relationship between Q and influential variables:

$$Q_{(t+Lt)} = f(X_{(t)}) + e \quad (2.16)$$

where $Q_{(t+Lt)}$ presents the ahead hourly stream flow; $X_{(t)}$ is the input vector, which includes the input variables RF and/or WL; and e is the random error.

Determining the variables of the input vector of the AI-based models includes also finding the lag intervals of antecedent records of the WL and RF records with the highest effect on the ahead Q, which can be detected through the Lt estimation (Sudheer et al., 2002).

2.7.4 Improvement of the Performance of the Modelling Process

The performance of SF models can be improved through a better understanding of the hydrological systems of a river basin such as analyzing the SF regime (Gautam & Acharya, 2012). The SF model performance can also be significantly improved by selecting the most adequate input and output variables, which mainly depends on Lt estimation; therefore, it is of great significance in several surface water hydrological analyses and models (Bowden et al., 2005; Fang et al., 2008; Yao et al., 2014).

In this research, WL and RF records of the upstream station are used as the input variables of the AI-based model, whereas the SF data from the downstream station are used as the output variables of the AI-based model. Therefore, estimating the Lt between the

upstream and downstream stations is a fundamental step in determining a potential combination of input and output variables for AI-based models. It is also important in exploring the sensitivity of various combinations of input and output variables to the prediction accuracy of AI-based models.

2.7.4.1 Analysis of Long-Term Variations in Stream Flow Regimes

SF regimes can be described by various parameters such as rate, magnitude, duration, timing and fluctuations over a varied scale of frequencies, including hourly, daily, monthly, yearly, decadal and multi- decadal (Krasovskaia & Gottschalk, 2002; Morán-Tejeda et al., 2011). Investigation of these parameters assists in understanding the whole SF regime and related hydrological phenomena, such as low and high SF events. Furthermore, the long-term variations in a SF regime can be recognized clearly based on the description of SF through these parameters (Poff et al., 1997; Richter, 1996; Yang et al., 2005).

Due to the variations of SF, long-period records of SF are essential to investigate and describe the SF regime. The statistical analyses used in long-term variations in SF regimes should be performed using long-period records (i.e. 50 years or more) as trends resulted from short observations may be part of weather fluctuations or just temporary changes (Gautam & Acharya, 2012; Kundzewicz & Robson, 2004; Opitz-Stapleton & Gangopadhyay, 2011).

There are many reasons why variations in SF appear, for example, climate change, human activities and geomorphic variations, which are possibly the main sources of SF change (Chang et al., 2014; Sang et al., 2013). Usually the changes in SF grow slowly; over the last 100-year period, an apparent decline in yearly SF has been verified in about 25% of the world's rivers (Descroix et al., 2012; Walling & Fang, 2003; Yang et al., 2005; Yue et al., 2003; Zhang et al., 2000).

Studying the changes in the SF regime is essential to enhance understanding the river hydrologic system which is necessary for the improvement of accuracy of SF and flood prediction (Gautam & Acharya, 2012; Xu et al., 2012). It is not only important from hydrologic aspects but also from both socioeconomic and natural aspects. As example, ecosystems are highly influenced by variations in SF regime because they are dependent on SF to protect their composition and continuity (Richter, 1996).

In the view of the above discussion, the better analysis of SF regimes is considered a very important step to improve the researches knowledge about the SF process which leading to improve the performance of SF modelling and prediction processes.

2.7.4.2 Lag Time Estimation

The travel time concept is used to estimate the time needed by the flow to move from any location to another within the river basin. This notion is frequently employed in many hydrological applications. Due to developments in hydrologic models and applications, various expressions of travel time have been adopted and often used, such as concentration time and Lt (Green & Nelson, 2002; Honarbakhsh et al., 2012; Thomas et al., 2015).

The travel time concept could be described by two ways: the hydrological (operational) and conceptual (theoretical) definitions. The hydrological definitions are applicable when hydrological data is available (Fang et al., 2005). The conceptual definition for time of concentration is the period taken by a water to move from the hydraulically most distal part of the basin to the outlet or reference point downstream (Fang et al., 2005; Kirpich, 1940; McCuen et al., 1984). The hydraulically most distant point is the point with the longest travel time to the basin outlet, and not necessarily with the longest flow path. The theoretical definition of Lt is the time a water drop takes to travel from an upstream location to a downstream location within river basin (Woodward, 2010).

Travel time reflects the speed at which the river basin responds to RF events (Pavlovic & Moglen, 2008) and is influenced by several parameters including the slope and length of the flow path, flow path roughness, flow depth, initial soil moisture, and duration and intensity of the effective RF (Green & Nelson, 2002; McCuen, 2009; Singh, 1988). These parameters are very complex and thus render estimation difficult and time consuming. Due to the complexity of describing all physical and hydrological specifications of the entire flow path and other basic parameters affecting travel time, many empirical equations for Lt estimation have been derived based on flow path and basin average parameters to simplify travel time estimation (Green & Nelson, 2002; Singh, 1976).

Hydrologically, a perfect estimation of travel time cannot be achieved, as it requires infinite, steady and continuous RF over the river basin, which is an impossible condition in reality (Saghafian & Julien, 1995).

- **Definition of Lag Time**

In hydrological modelling applications, travel time is generally represented by Lt. The main hydrological definition of Lt is the difference in time between the center of mass of effective RF and the center of mass of direct runoff (i.e. hydrograph) produced by the effective RF (Banasik et al., 2005; Viessman & Lewis, 2003).

Several other hydrological definitions of Lt between the WL upstream station and SF downstream station can be handled, which are reported by (Viessman & Lewis, 2003), (Fang et al., 2005), (Honarbakhsh et al., 2012), (Grimaldi et al., 2012), (Talei & Chua, 2012) and referenced herein:

(1) The time interval from the time of maximum WL rate to the time of the peak of hydrograph of SF station;

(2) The time interval from the time of the centroid of actual WL excess to the time of the peak of hydrograph;

(3) The time from the end of WL excess to the inflection point on hydrograph; and

(4) The time interval from the beginning of WL excess to the centroid of hydrograph.

The hydrological definitions of Lt between the RF upstream station and SF downstream station are similar with the hydrological definitions between the WL upstream station and SF downstream station and they are based on the same hydrological concepts. It should be noted that for these definitions to be applicable to estimating the Lt between two locations or between two stations within the river basin, they must be hydraulically connected without any significant non-natural barriers (Viessman & Lewis, 2003).

Figure 2.13 describes two of definitions of Lt. (1) Time interval from the centroid of the RF to the centroid of hydrograph (t_1); and (2) time interval from the centroid of RF to the peak of hydrograph (t_2) (Talei & Chua, 2012).

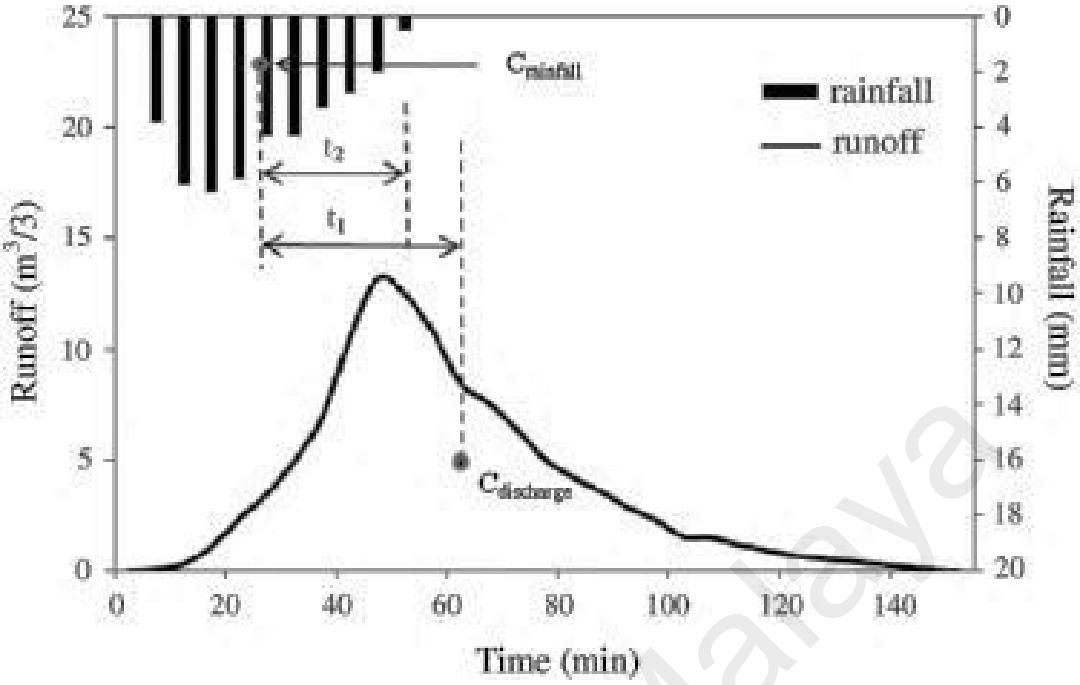


Figure 2.13: Schematic illustration of lag time estimation based on two different definitions (Talei & Chua, 2012)

- **Lag Time Estimation Methods**

In literature, three main approaches have been applied for estimating the Lt, first of which is the estimation using empirical formulas. Huge number of empirical formulas have been used to estimate Lt, Li and Chibber (2008) presented and evaluated about fourteen empirical formulas with different data requirements. In the current research, four empirical formulas were employed to estimate the Lt between upstream stations, and a downstream station.

The second approach is to estimate Lt by calculating the R between WL or RF hourly records of upstream stations and Q records of downstream station. The Lt is thus defined as the lag interval that required in providing the highest R between the upstream and downstream records. This method is not completely in high agreement with the classic definitions of Lt, but can be measured to provide an approximation of the Lt.

The third approach is to estimate the Lt based on the observed data of WL or EF and SF through one of the hydrological (operational) definitions of Lt. In the current research, the first hydrological definition of Lt was employed to estimate the Lt between upstream stations, and a downstream station.

- **Influence of the hydrological parameters on the Lag Time**

Lt between two places among the river basin could be different due to many hydrological parameters such as basin, flow path and RF characteristics. The basin parameters that affect the Lt are areal extent, surface topographic, vegetation, and land use. The flow path characteristics that affect the Lt are slope, length, roughness, flow depth and antecedent soil moisture. RF characteristics that affect the Lt are intensity and duration. There are other parameters may be slightly affect the Lt, such as wind speed, relative humidity and climate conditions (Green & Nelson, 2002; McCuen, 2009; Sabzevari et al., 2010; Singh, 1988). These parameters are very complex, thus making it difficult and time consuming to study.

Due to the complexity of description all physical and hydrological characteristics of the entire flow path and other parameters influencing the Lt; many empirical equations and estimation approaches have been derived based on the flow path and basin average parameters to simplify the estimation of the Lt (Green & Nelson, 2002; Singh, 1976). Although the availability of empirical equations to estimate the Lt (Grimaldi et al., 2012; Li & Chibber, 2008), the influence of the hydrological parameters that are likely affecting the Lt such as RF and SF have not been studied intensively.

The investigation of the influence of the hydrological parameters that are likely affecting the Lt is very important key in SF modelling and detection the times of high SF events. Mostly, the Lt reflects the speed at which the river basin responds to RF events (Pavlovic & Moglen, 2008).

The RF and SF are considered as the main variables affecting the Lt. To investigate the effect of these parameters on the Lt, RF intensity was represented by two variables, peak rainfall intensity (Rf_p) and the average of previous 48 hour rainfall (Rf_{48}) while the SF was represented by two variables, peak hourly stream flow (Qp) and the average of previous 48 hour stream flow (Q_{48}). Rf_{48} and Q_{48} are used to represent the degree of saturation in the river basin (Simas, 1996).

CHAPTER 3: METHODOLOGY

3.1 Introduction

Chapter 3 presents the methodology and briefly describes the study area, data collection and preliminary data analysis. The research methodology includes the hydrological description of the Selangor River basin and the development of AI-based models. The hydrological description of the Selangor River basin includes several procedures, such as the overview of the Selangor River basin hydrology, analyses of the long-term variations in the SF regime, and the Lt estimation between the upstream and downstream stations, which is essential in selecting the variables of AI-based models.

The development of AI-based models includes many steps such as the selection of the AI-based model variables and modelling patterns, identification of AI-based model structures and general description of the training processes. The main procedures of the research methodology are presented in detail through this chapter. The research methodology is briefly presented as follows:

- Review of the related literature, such as books, scientific reports, and journal papers.

- Selection of the appropriate case study area. The Selangor River basin was selected as the study area.

- Data collection from the hydrological stations located in the Selangor River basin. The SF, RF and WL hourly records of a one-year period (2011) were utilized in the development of the AI-based models, whereas the SF, WL and RF hourly

records of the high SF events over a three-year period (2009, 2010, and 2011) were utilized in the Lt estimation. The SF records over a 50-year period from 1961 to 2011 were used in the analysis of long-term variations in the SF regimes.

- Preliminary data analysis, including the basic statistical analysis, check for normality and homogeneity tests.

- The improvement of the hydrological description of the Selangor River basin through the long-term variation analysis of the SF regime over a 50-year period and the development of HGA for estimation of the Lt between upstream and downstream stations.

- Determination of the input and output variables of AI-based models and lag intervals between them depending on the estimation of the Lt between upstream and downstream stations.

- Modelling process and development of AI-based models to predict real-time Q.

- Evaluation of the developed AI-based models performance by multi-evaluation criteria, namely, R, R^2 and MAE.

- Application of the AI-based models in many applications, such as prediction and analytical tools to investigate the influence of the hydrological variables on SF. They are also employed in estimation of the missing records of Q and the flood early warning throughout the advance detection of the hydrological conditions that may lead to formations of floods through six hydrological scenarios.

The main steps of the research methodology are briefly illustrated in the following flowchart.

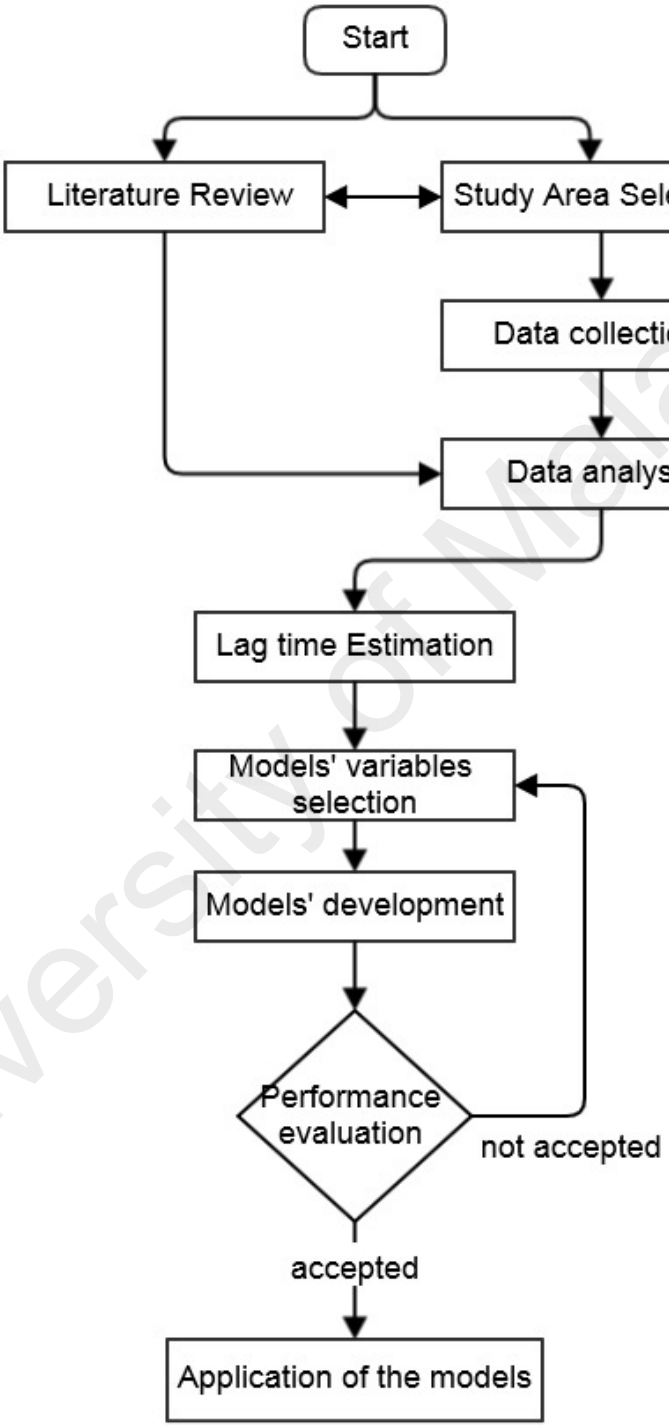


Figure 3.1: Main steps of the research methodology

3.2 Case Study Description

The study area of this research is Selangor River basin, which is one of the main rivers in Malaysia. It is located in northern part of the Selangor state and has an approximate area of 1960 km^2 (Hassan et al., 2004; Lee, 2002; Samsudin et al., 2011). The Selangor River Basin provides approximately 50% of the water consumed in Selangor and Kuala Lumpur states (Subramaniam, 2004).

3.2.1 General Description of Malaysia

Malaysia covers an area of about 330,000 km^2 including the Peninsular Malaysia, which is located in the southern east of the Asia, and the States of Sabah and Sarawak in the northwestern part of Borneo Island. Peninsular Malaysia, covering 131,598 km^2, has its land frontier with Thailand to the north, and is connected to Singapore by a causeway in the south. Malaysia locates near the Equator between latitude 1° and 7° North and between longitude 100° and 119° East (Shafie, 2009). Figure 3.2 displays regional map of Malaysia.

Peninsular Malaysia contains hills and mountains ranges, covering around 30% of the Peninsula area and track nearly parallel to the long axis of the country. The rolling to undulating land is seen mainly at the seaward flanks and the intervening zones among the mountain ranges. Figure 3.3 displays Political map of Malaysia

Figure 3.2: Regional map of Malaysia

Figure 3.3: Political map of Malaysia

3.2.2 Location and Topography of Selangor River Basin

The Selangor River basin is one of the main rivers in the Malaysian Peninsula's west coast. It is located in the Selangor state and has an approximate area of 1960 km^2, approximately a quarter of Selangor state. It is the main river in the Selangor state. Selangor River starts at the border between the states of Selangor and Pahang at an elevation of 1700 m and it streams nearly 110 km from the northeast to the southwest (Hassan et al., 2004; Lee, 2002; Samsudin et al., 2011). Figure 3.4 shows the location map and Figure 3.5 shows the topography map of the Selangor River basin.

3.2.3 Climate and Rainfall of Selangor River Basin

The study area can be considered a paradigm of the humid tropical rivers in Southeast Asia. This river basin is under a humid tropical climate, and the temperature varies slightly throughout the year. On average, the temperature reaches 32 °C in the daytime and 23 °C at night. Average annual rainfall is between 2000 mm and 3000 mm, and evaporation ranges from 1600 mm to 1800 mm. The annual average relative humidity is approximately 80% (Breemen, 2008; Shafie, 2009; Zin et al., 2013). The average flow of this river is 57 m^3/s. Approximately 10% of the time, the flow either exceeds 122 m^3/s or drops to below 23 m^3/s as a result of seasonal variations in rainfall (Green & Nelson, 2002).

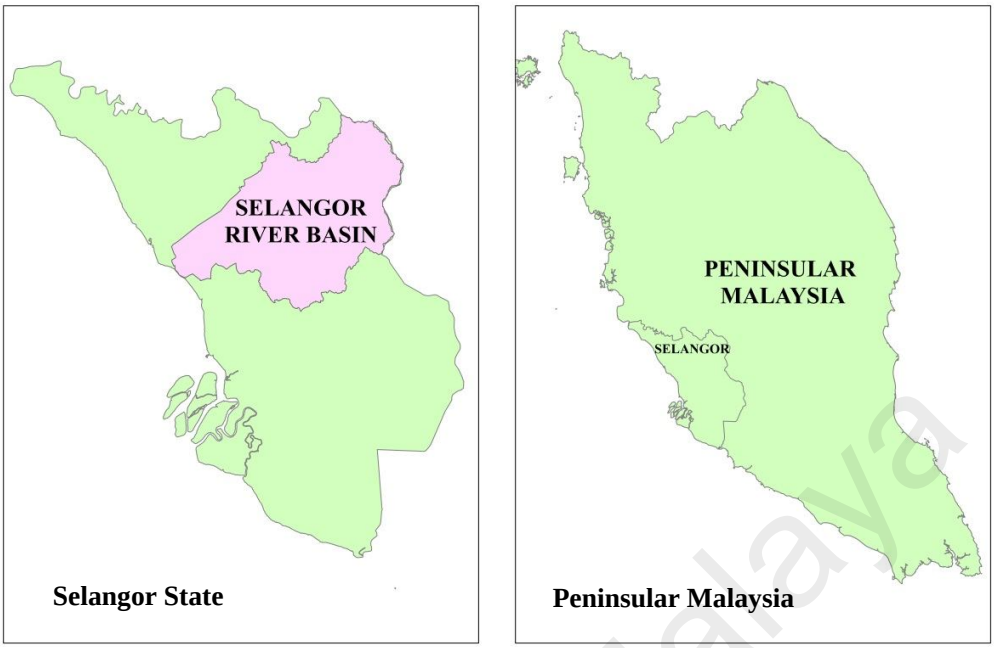


Figure 3.4: The location map of Selangor River basin

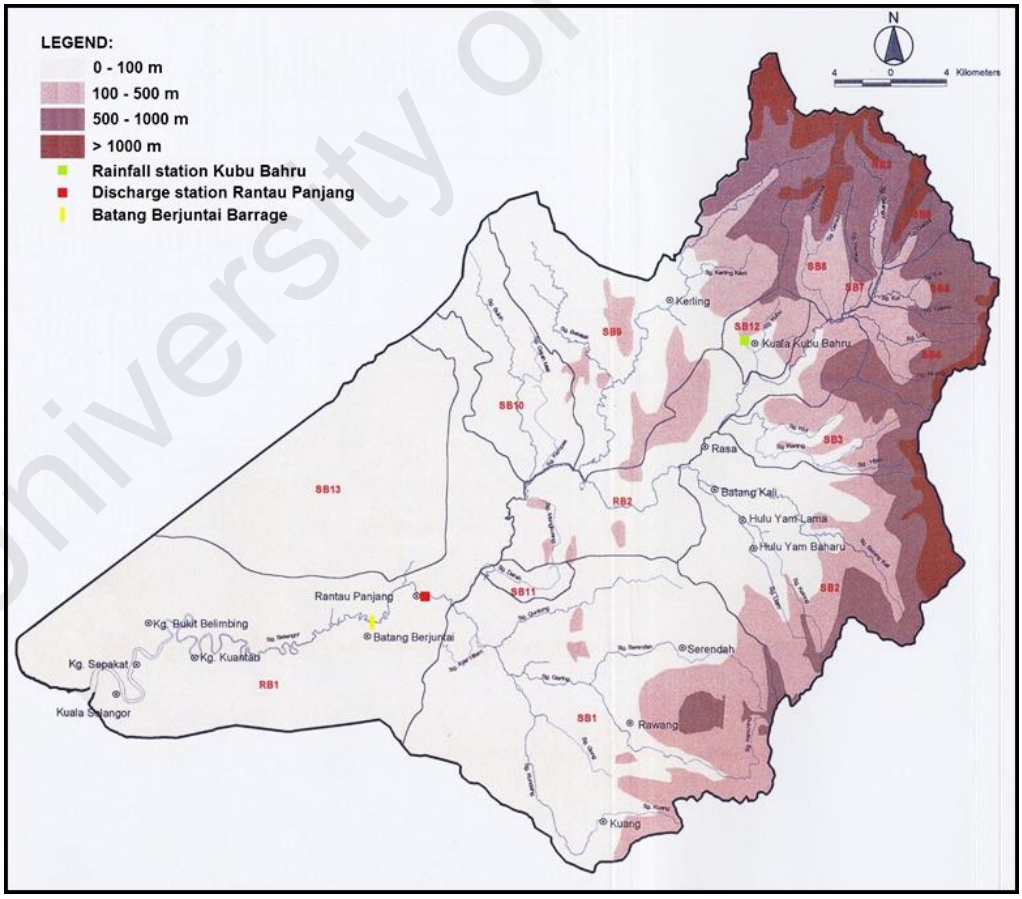


Figure 3.5: The topography map of Selangor River basin

3.3 Research Data

Data is considered the backbone of any SF modelling applications, thus, high quality data is vital to get high accurate results in modelling, prediction and analysis process. The hydrological data were sourced from the hydrological stations located in the Selangor River Basin. Two of these stations gauge SF, seven gauge water level, and more than twenty gauge rainfall.

Unfortunately, only the Rantau Panjang gauging station has enough data of SF data from 1961 to 2011. This station is located in the downstream area of the River basin. Before this station, all of the major tributaries of this river converge. Thus, the SF at the Rantau Panjang Station is suitable representative of the SF at the study area.

The WL and RF records were extracted from four upstream stations as shown in Table 3.1. The SF, WL and RF hourly records of 1-year period 2011 were utilized in development of the AI-based models, while the SF, WL and RF hourly records of the high SF events over three-year period (2009, 2010 and 2011) were utilized in the Lt estimation.

Figure 3.6 presents the location of the utilized hydrological stations and main tributaries in the Selangor River basin.

Table 3.1: Hydrological stations in the Selangor River basin

Station Name	Function	Latitude	Longitude
Rantau Panjang	SF	03 24 10.0	101 26 35.0
Ulu Yam	WL & RF	03 27 38.4	101 38 14.4
Batang Kali	WL & RF	03 28 11.7	101 38 23.3
Kerling	WL & RF	03 35 18.1	101 36 22.8
Ampang Pecah	WL & RF	03 32 25.4	101 39 48.3

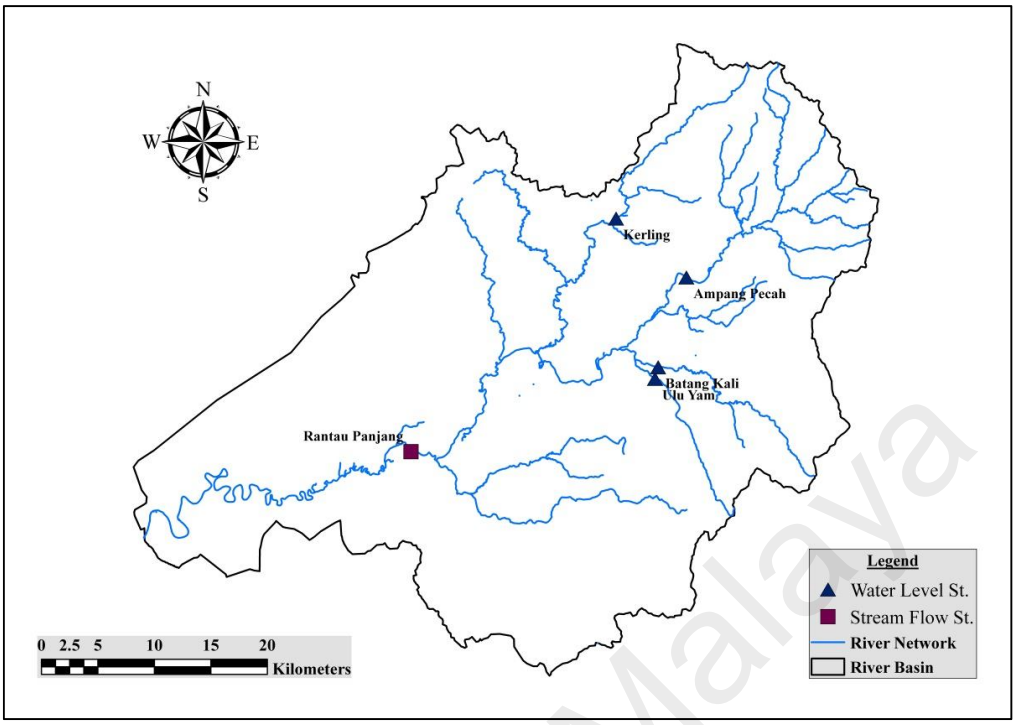


Figure 3.6: Locations of the main hydrological stations and tributaries in the Selangor River basin

3.4 Preliminary Data Check

The preliminary data check analysis includes the basic statistical analysis of data, normality and homogeneity tests. The outcomes of these tests verified that the SF data could be considered normally distributed and homogeneous.

3.4.1 The Basic Statistical Analysis

The research data were statistically analyzed to briefly demonstrate its quality and reliability. About 8753 patterns of Q, WL and RF hourly records representing one year period (2011), used for modelling process. The data basic statistical characteristics of data, such as minimum, maximum, mean and standard deviation (SD) of hourly records of all stations are shown in Table 3.2.

The Q, WL and RF hourly records of first three days of January 2011 were presented in Appendix A as example of the full records of data.

Table 3.2: Statistical basic analysis of the data used

Station	Function	Latitude	Longitude	Mean	Min.	Max.	SD
Rantau	Q (m^3/s)	03° 24' 10.0"	101° 26' 35.0"	60.35	23.94	294.6	39.00
Ulu Yam	WL (m)	03° 27' 38.4"	101° 38' 14.4"	32.24	30.56	35.49	0.49
Batang Kali	WL (m)	03° 28' 11.7"	101° 38' 23.3"	32.42	27.03	34.71	0.78
Kerling	WL (m)	03° 35' 18.1"	101° 36' 22.8"	44.18	43.93	45.61	0.12
Ampang	WL (m)	03° 32' 25.4"	101° 39' 48.3"	50.16	49.61	50.89	0.15
Ulu Yam	RF (mm/hr)	03° 27' 38.4"	101° 38' 14.4"	0.16	0.00	19.33	0.73
Batang Kali	RF (mm/hr)	03° 28' 11.7"	101° 38' 23.3"	0.24	0.00	22.67	0.91
Kerling	RF (mm/hr)	03° 35' 18.1"	101° 36' 22.8"	0.25	0.00	25.33	1.06
Ampang	RF (mm/hr)	03° 32' 29.1"	101° 39' 44.4"	0.24	0.00	28.00	1.08

3.4.2 The Normality Test

The normality test is a statistical test to inspect whether the data is fine-modelled by a normal distribution or not (Coin, 2008; Tenreiro, 2011). It is widely employed in statistical analyses. The Shapiro–Wilk test is the common normality test, particularly, in hydrological applications. It was derived in 1965 by Samuel Shapiro and Martin Wilk, then adjusted by Royston in 1992 and again in 1995. It utilizes the null hypothesis principle to check whether a sample came from a normally distributed population or not. The null-hypothesis of this test is that the population is normally distributed. Thus if the p-value is less than the selected significance level, then the null hypothesis is rejected and there is evidence that the data tested are not from a normally distributed population (Razali & Wah, 2011; Royston, 1992).

The results of the Shapiro–Wilk test contain two values: W and p-value. W lies between 0 and 1. High W values lead to acceptance of normality, whereas small W values lead to rejection of it. When W is equal to 1, it indicates complete data normality. For the p-value value, if it is higher than the selected significance level, the normality will be accepted (Razali & Wah, 2011).

The normality of average annual SF at Rantau Panjang station over 50-year period from 1961 to 2010 was tested with the Shapiro-Wilk test. Additional graphical technique of checking the normality was employed; it is based on comparison between the frequency distribution (histogram) of average annual SF records and the normal probability curve of the data. The data is considered normally distributed if the histogram of the data is in high agreement with the normal distribution curve of data (Mecklin & Mundfrom, 2004; Shahla Ramzan et al., 2013).

The Shapiro–Wilk test was applied on annual SF data. Based on the test results W (0.976) and P (0.411), the normality of annual SF are verified. The normality of annual SF was also inspected by graphical technique. Figure 3.7 demonstrates a very high similarity between the frequency distribution and normal probability curve of average annual SF data. Thus, it is normally distributed.

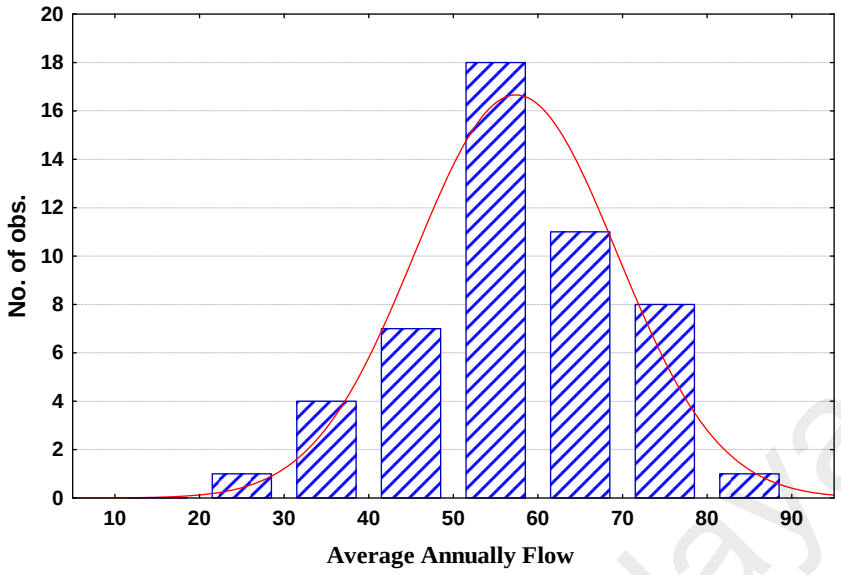


Figure 3.7: Frequency distribution and normal probability curve of average annual stream flow

3.4.3 The Homogeneity Test

The homogeneity test is a statistical test for discovering data variability. It is used to check whether data have been obtained from homogeneous or heterogeneous source. Literature offers many homogeneity tests of hydrological time series data, for example, the standard normal homogeneity test, Buishand's test and Pettitt's test (Buishand, 1982; Kang & Yusof, 2012; Pettitt, 1979).

In this research, Pettitt's test was applied to verify the homogeneity of annual average SF data from the Rantau Panjang station over 50 year-period from 1961 to 2010.

The p-value (Two-tailed) - calculated using 10000 Monte Carlo simulations, is equal to 0.161. It is higher than the significance level p-value of 0.05, meaning that the SF data is homogeneous data.

3.5 The Hydrological Description of Selangor River Basin

The performance of SF modelling process can be greatly improved by the amelioration of the hydrological description of the river basin. Therefore, it is of great significance to achieve suitable level of hydrological description before carrying out the modelling process. In this research, the hydrological description have been achieved via some procedures such as the over view of Selangor River basin hydrology, the long-term changes analyses in SF regime and the Lt estimation between the upstream and downstream stations which is very essential to select the variables of AI-based models and lag intervals between them.

3.5.1 Hydrological Overview of the Selangor River Basin

The main tributaries of the Selangor River are Rening, Kerling, Batang Kali and Guntong. Many minor branches joint the main tributaries or the main stream of the river itself. Due to high slope of flow paths in the upstream area, the branches are fast flowing through mountain with granite and sedimentary bedrock. In the downstream area, the river enters the fluvial plain and becomes a low-gradient, meandering river. The river bed slope in the last 30 km is around zero (Breemen, 2008).

The mean SF over a 50-year period from 1961 to 2010 is around 57 m^3/s. Seasonal variations in RF make SF to exceed 122 m^3/s or to drop under 23 m^3/s in about 10% over the time (Nelson, 2002). Figure 3.8 presents the mean annual SF of the Selangor River over a 50-year period between 1961 and 2011.

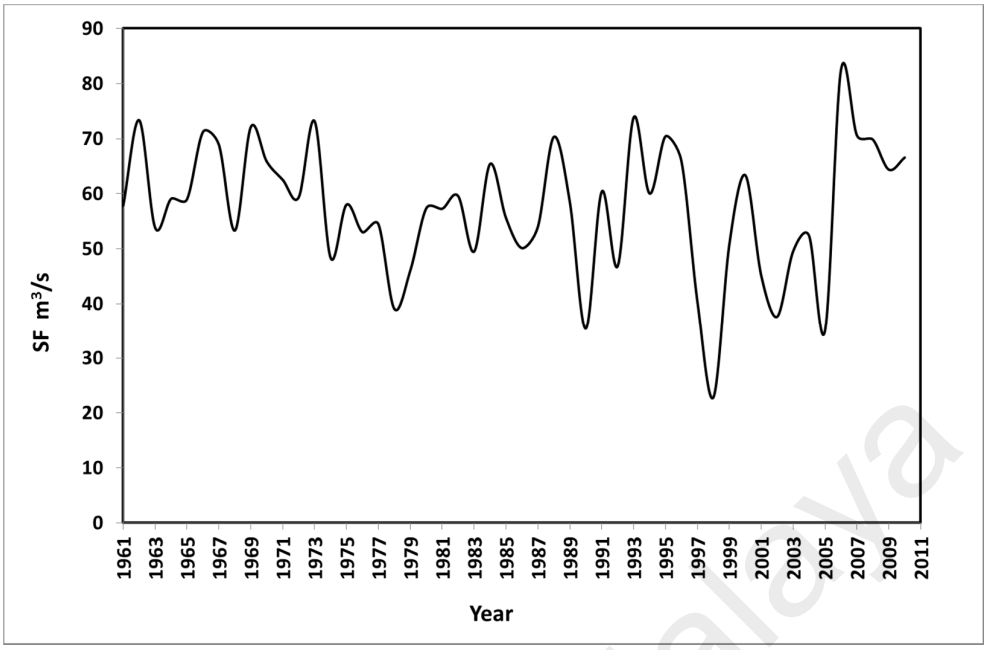


Figure 3.8: Mean annual stream flow in the Selangor River over a 50-year period from 1961 to 2010

3.5.2 Analysis of the Long-Term Variations of the Stream Flow Regime

To achieve accurate investigations of the long-term changes in the SF regime, the statistical analyses should be performed using lengthy periods (i.e. 50 years or more) (Kundzewicz & Robson, 2004; Walling & Fang, 2003).

The long-term variations analyses include an investigation of the changes in the hydrological variables describing the annual SF over the 50-year period from 1961 to 2010 along with testing their changes' trend. Analyses were performed based on two time scales. The first is yearly and the second is the sub-periodic changes. The sub periods were reached by segmentation of the 50 years into 7 sub-periods via two methods: the direct segmentation and change-point test.

The work also includes an exploration of the variations in the monthly SF regime and the yearly duration of high and low SF from 1961 to 2010 and inspecting the trend of changes.

For the high and low SF, the assessment was performed with respect to the duration and magnitude as both parameters play an important role to understand the variations of SF regime (Mirabbasi et al., 2012).

Investigation of yearly duration of high SF comprises three levels, danger level, warning level and alert level while investigation of the yearly duration of low SF analysis was conducted at a single level, that when the SF drop below 14.5 m^3/s, which is around 25% of the average SF over the study period.

3.5.2.1 Determination of Representative Hydrological Variables

The long-term variations in SF regime could be explored via SF features like magnitude, rate, frequency, duration, timing and rate of change. These features are generally applied in three circumstances: average, low and high flow. Several hydrological variables can be employed to investigate the variations in the features of SF (Moliere et al., 2009).

In this research, the variations in SF were investigated depending on SF rate (discharge), which is the quantity of water passing through an identified location per of time (Poff et al., 1997; Richter, 1996; Yang et al., 2005).

Nine hydrological variables describing SF were selected to investigate the long-term changes of the Selangor River's SF regime. The variables are mean annual stream flow (SF1), maximum annual stream flow over the sub-period (SF2), minimum annual stream flow over the sub-period (SF3), maximum monthly flow over a single year (SF4), minimum monthly stream flow over a single year (SF5), the deference between maximum and minimum stream flow (RA), SD, coefficient of variation (CV) and the Pluviometric Ratio (PR).

SD and CV are statistical measures of dispersion in a data series around its average; and the CV denotes the ratio of standard deviation to the SF1. The CV is employed in

matching the amount of dispersion and variation among data series (Albrecher et al., 2010; Boik & Shirvani, 2009).

PR corresponds to the ratio between maximum and minimum SF, and it is indication of seasonal variability. When the PR value is close to 1, seasonal variability is minor, but when the value is above 1, seasonal variability rises directly (Laraque et al., 2007). The mathematical formula for RA and PR are as follow.

$$RA = SF4 - SF5$$

$$PR = SF4/SF5$$

The values of nine variables were calculated from the Q records at the Rantau Panjang station over a 50-year period from 1961 to 2010. Table 3.3 shows the nine variables and their measurement units.

Table 3.3: Hydrological variables utilized to describe the annual stream flow

#	Var.	Definition	Unit
1	**SF1**	Mean annual stream flow	m^3/s
2	**SF2**	Maximum annual stream flow over the sub-period	m^3/s
3	**SF3**	Minimum annual stream flow over the sub-period	m^3/s
4	**SF4**	Maximum monthly stream flow over a single year	m^3/s
5	**SF5**	Minimum monthly stream flow over a single year	m^3/s
6	**RA**	The deference between maximum and minimum annual stream flow	m^3/s
7	**SD**	Standard Deviation	m^3/s
8	**CV**	Coefficient of variation	ratio
9	**PR**	The Pluviometric Ratio	ratio

3.5.2.2 Segmentation of the Study Period

To study and analyze the long-term changes in SF regime over long periods i.e. 50 years, the yearly variations may be not adequate to investigate the trend of long-term variations. For this cause, long periods could be segmented into short sub-periods, such as 7 or 10 years. The short sub-periods are commonly include consecutive years with comparable hydrologic features (Descroix et al., 2012).

The segmentation process could be carried out by many methods such as the Hidden Markov model, the Hubert model and the change-point test. The change-point method is a statistics test employed to find the date(s) at which a big change happens in a data series. The selected dates demonstrate a change in the mean or variance (Beaulieu et al., 2009; Rougé et al., 2012; Villarini et al., 2011). In literature, many approaches have been employed to check for the existence of change points in the data of long periods such as Bayesian inference, moving t-test and Pettit's test (Descroix et al., 2012; Ma et al., 2008; Rougé et al., 2012; Xiong & Guo, 2004; Zheng et al., 2007).

In this research, the 50-year period from 1961 to 2010 was segmented into seven sub-periods by two techniques. The first is the change points using Pettit's test, while the second method is direct segmentation method. The first technique entails selecting multiple change-points, as calculated using Pettitt's test. This technique leads to the segmentation of the study period into seven, non-identical sub-periods. The second method is direct segmentation in which the study period was divided into seven identical 7-year sub-periods. The sub-periods obtained in the two ways are presented in Table 3.4.

Table 3.4: Sub-periods obtained via two segmentation techniques

Sub-period No.	Segmentation technique	
	Change-point	Direct
1	1961-1972	1961-1967
2	1973-1978	1968-1974
3	1979-1986	1975-1981
4	1987-1991	1982-1988
5	1992-1995	1989-1995
6	1996-2004	1996-2002
7	2005-2010	2003-2009

3.5.3 Lag Time Estimation

Three approaches have been applied to estimate the Lt between the upstream and downstream stations. In the first approach, the Lt is estimated through four available empirical formulas. In the second approach, the Lt is estimated by CCA, whereas, the Lt is estimated by calculating the R to lag intervals between the antecedent hourly records of WL and RF records of the upstream stations and the Q in the downstream station. In the third approach, the Lt is estimated using HGA based on the hourly records of WL, RF and Q through the hydrological (operational) definition of Lt.

The first approach is performed only to provide an initial approximation of the Lt between the upstream and downstream stations. The results of both the second and third approaches are employed in the Q modelling process, particularly in the selection of the variables of AI-based models and lag intervals between them.

3.5.3.1 Lag Time Estimation Methods

- **Lag Time Estimation Using Empirical Formulas**

To estimate the Lt between the upstream and downstream stations using empirical formulas, it is necessary to determine the basic morphometric specifications of the flow path between the mentioned stations, such as flow path length, average flow path slope and mean flow of channel velocity.

By referring to the hydrological and topographical maps of the Selangor River basin, the basic morphometric specifications of the hydrological stations and the flow paths between them were determined; moreover, Lt was estimated between the upstream and downstream stations by carrying out the following subsequent procedures:

1. Determining the coordinates (x, y, z) of the downstream and upstream stations;
2. Determining the flow paths between the upstream and downstream stations and measuring the its' lengths;
3. Measuring the difference in elevation between the upstream and downstream stations;
4. Calculating the flow path slopes between the upstream and downstream stations; and
5. Applying the empirical formulas that require only morphometric parameters to estimate the Lt.

Lt estimation was performed by using the four empirical formulas listed in Table 3.5. The topographic specifications are shown in Table 3.6.

Table 3.5: Empirical formulas used to estimate the Lt in hours (hr)

Name	**formula**
Kirpich (1940)	$Lt = 0.00325\ L^{0.77}\ S^{-0.385}$
Johnstone and Cross (1949)	$Lt = 0.0017153\ L^{0.5}\ S^{-0.5}$
Carter (1961)	$Lt = 0.001628\ L^{0.6}\ S^{-0.3}$
Viparelli (1961, 1963)	Lt = L/ V

where Lt is the lag time (hr), L is the flow path length (meters) between the upstream and downstream stations, S is the average slope of the flow path from the upstream to the downstream station, and V is the mean flow channel velocity (m/s) with recommended values between 1 and 1.5 m/s (Grimaldi et al., 2012). In the Selangor river basin, which is a very flat area with a slight slope between the upstream and downstream areas, the flow channel velocity is fixed at 1.0 m/s.

Table 3.6: Topographic specifications of the upstream and downstream stations and the flow paths between them

Station Name	**Location**	**Elevation**	**H**	**L**	**S**
		M	**m**	**km**	**ratio**
Rantau Panjang	Downstream	13	-	-	-
Ulu Yam	Upstream	40	27	38.5	0.070%
Batang Kali	Upstream	46	33	38.3	0.086%
Kerling	Upstream	56	43	44.9	0.096%
Ampang Pecah	Upstream	51	38	47.6	0.080%

- **Lag Time Estimation Using the Correlation Coefficient approach**

In CCA, the Lt was estimated by calculating the R that corresponds to 24 different time lag intervals between the antecedent hourly records of WL or RF records at the upstream stations and the Q at downstream station. Calculation of the R leads to detect which the antecedent records of WL or RF records that have the highest effect on the predicted stream. The correlations were analyzed based on the hourly antecedent records.

Lt was estimated by calculating the R corresponding to 24 different time lags from 0 h to 24 h between the antecedent hourly records of WL or RF of upstream stations. These records represent the input variables of AI-based models, whereas those of Q in the downstream station denote the output variable. The estimated Lt using R approach was utilized to determine the input variables of the AI-based models in first phase of the modelling process.

- **Lag Time Estimation Using the Hydrological Graphical Approach**

Lt was estimated using HGA through the hydrological definition of Lt (the time interval from the time of maximum WL or RF at the upstream stations to the time of the peak rate of runoff at the downstream station). HGA where employed to estimate the Lt between both of the WL and RF upstream stations, and the downstream station. The followings procedures were performed to estimate Lt between the WL upstream stations and downstream station:

a. Collecting hourly records of the WL and SF stations and check the quality, continuity and reliability of the collected data;
b. Determining the high SF events during the three-year period: 2009, 2010 and 2011;
c. Drawing the hydrograph of the downstream station and WL graphs of the upstream stations for each event;

d. Determining the peak time in the hydrograph and WL graphs;

e. Estimating the Lt by measuring the time interval between the peak times in the hydrograph and WL graphs;

f. Repeating these steps for all high-SF events. 134 WL-SF events were used to estimate the Lt. A sufficient number of events are necessary for better Lt estimation; and

g. Analysis of the estimated Lt for all events and calculating the mean of the estimated Lt. The mean value is deemed the best representative of Lt between WL and downstream stations.

Figure 3.9 describes HGA for estimating the Lt using the observed WL–SF event by applying the first hydrological definition of Lt, also, the procedures to estimate Lt are described in the flowchart as shown in Figure 3.10 (Seyam & Othman, 2014a).

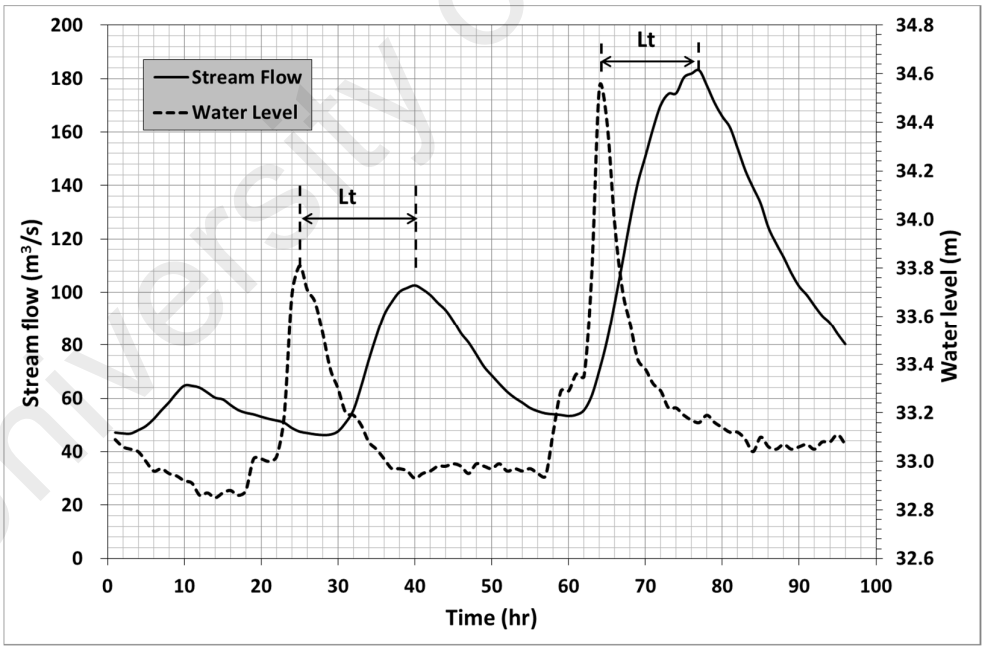


Figure 3.9: Hydrological graphical approach for estimation of Lag time between upstream water level stations and downstream station

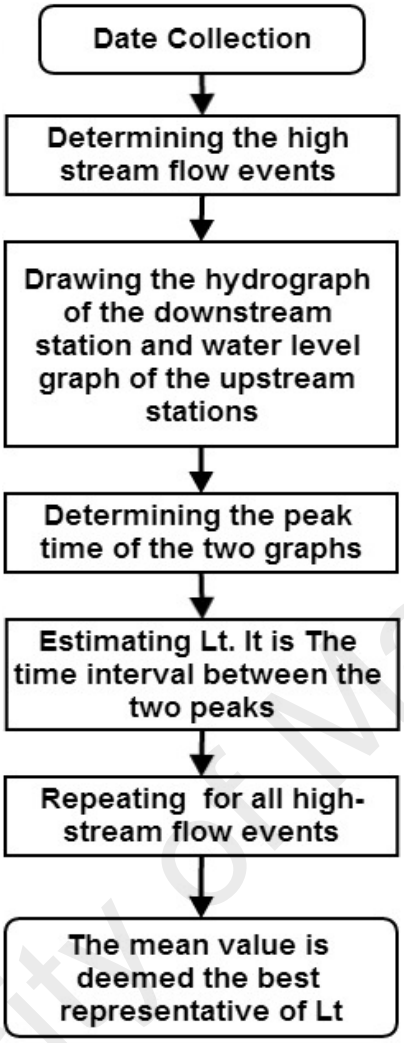


Figure 3.10: Flowchart of the hydrological graphical approach for estimation of Lag time based on the observed water level and stream flow

Same procedures were applied to estimate the Lt between the upstream RF stations and downstream station, the only difference is estimating time interval between the peak times in the hydrograph of downstream station and hyetograph of upstream stations as described in Figure 3.11 which describes the HGA of Lt estimation using the observed RF-SF event by applying the first hydrological definition of Lt. 100 RF-SF events were used to estimate the Lt.

The procedure of estimating Lt is additionally described in the flowchart in Figure 3.12. The estimated Lt using HGA between both the WL and RF upstream stations, and the

downstream station was utilized to select the lag intervals between the input and output variables of the AI-based models in second phase of the modelling process.

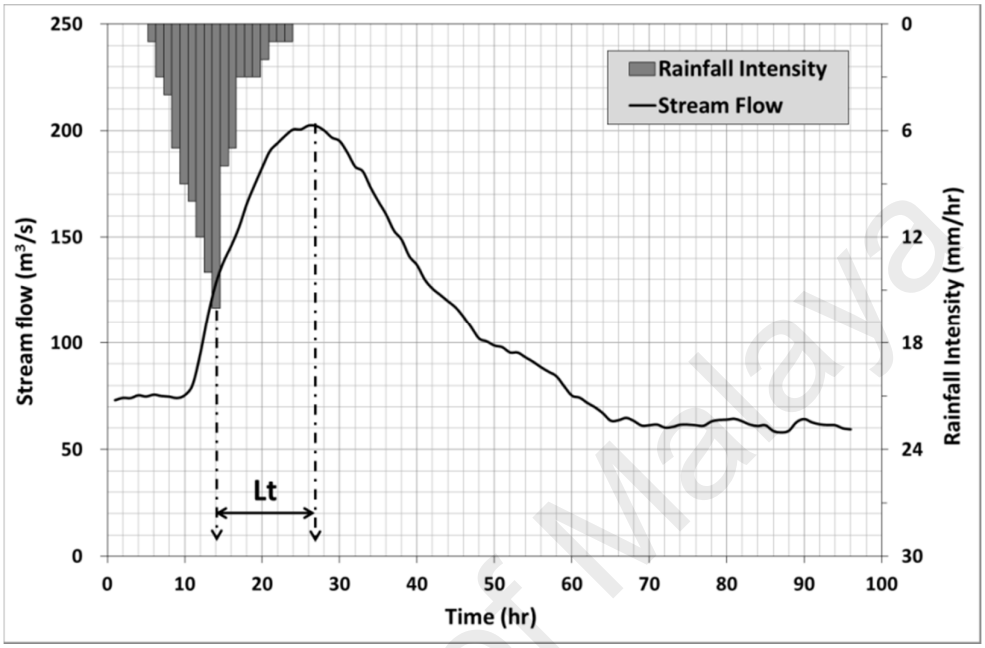


Figure 3.11: Hydrological graphical approach for estimation of Lag time between upstream rainfall stations and downstream station

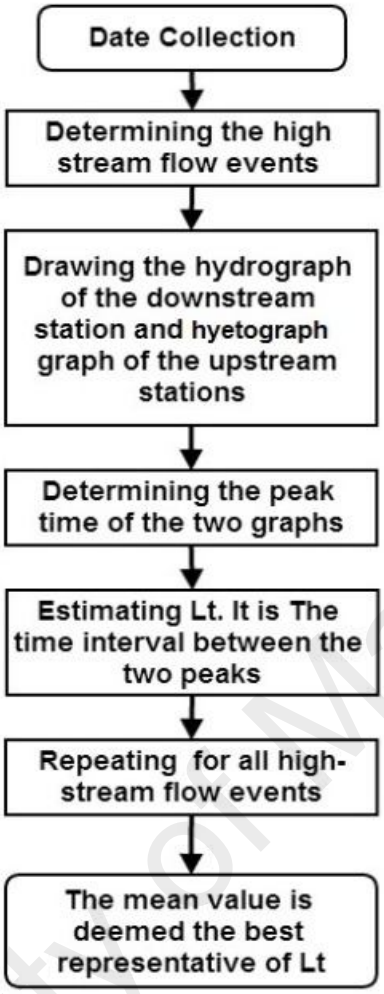


Figure 3.12: Flowchart of the hydrological graphical approach for estimation of Lag time based on the observed ranifall and stream flow

3.5.3.2 Derivation of new empirical formulas to estimate the Lt

The high complexity of surface water systems and interaction among the variables influencing the Lt, justify the necessity to derive new empirical formulas to estimate the Lt. Linear and nonlinear empirical formulas were both derived to estimate the Lt and study the influence of some related hydrological variables on the Lt.

The RF and SF are considered as the main variables affecting the Lt. To derive new empirical formulas to estimate the Lt, the RF was represented by two variables, peak

rainfall intensity (Rf_p) and the average of previous 48 hour rainfall (Rf_{48}) while the SF was represented by two variables, peak hourly stream flow (Q_p) and the average of previous 48 hour stream flow (Q_{48}). Rf_{48} and Q_{48} are used to represent the degree of saturation in the river basin (Simas, 1996).

The input (independent) variables of the empirical formulas are Rf_p, Rf_{48}, Q_p and Q_{48} while the output (dependent) variable is the Lt. Linear and non-linear empirical formulas were derived directly using the estimated Lt by HGA approach from 100 high RF-SF events.

3.6 Development of AI-based Models

The modelling process and development of AI-based models to predict the real-time Q includes several procedures and practical steps. After conducting the preliminary data check, the input and output variables of the model, the lag intervals between the input and output variables, and the modelling patterns should be selected followed by the selection of the model structure. Thereafter, the modelling process can be performed.

The selection of the AI-based models variables was performed in two phases. First, the results of the Lt estimated by CCA were employed to select the lag intervals between the input and output variables of the AI-based models. The results of the HGA were then employed in the second phase to select the lag intervals between the input and output variables of the AI-based models. The two phases of the variables selection and modelling process were used to explore the ability of improving the performance of AI-based models by the accurate timing of their variables based on the Lt estimation.

Finally, the developed models were evaluated and tested based on the performance evaluation criteria. These main steps should be performed in the development of any AI-based models, even used to predict the Q or any other similar hydrological process.

3.6.1 Variables Determination of AI-based Models

In the development of AI-based models, determining the adequate input and output variables is a key issue to achieve high performance models. In models of Q prediction, model variables are commonly selected based on a priori knowledge of river basin hydrology, which provides initial indications of potential inputs and outputs (Bowden et al., 2005; Maier & Dandy, 2000). The SF in tropical rivers can be characterized as the function of several influential variables, including RF, WL and the physical characteristics of the river (Firat, 2007).

In this research, the main objective is to predict Q of downstream area from the hourly WL and RF records of upstream station. Thus, the hourly records of WL and RF of upstream stations were employed as input variables (independent variables) while, those of SF data in downstream station was used as output variable (dependent variables). The Equation 3.1 presents the relationship between the Q and influential variables:

$$Q_{(t+Lt)} = f(X_{(t)}) + e \qquad (3.1)$$

where, $Q_{(t+Lt)}$ represents ahead hourly stream flow; Lt represents the lag time between upstream and downstream stations; $X_{(t)}$ is the input vector, which include the input variables i.e. RF and/or WL; e is the random error.

Three scenarios in selecting the input variables of the AI-based models were considered. In the first scenario, only the RF records of upstream stations were employed as input variables. In the second scenario, only the WL records of upstream stations were

employed as input variables while in the third scenario, both of RF and WL records of upstream stations were employed as input variables.

In these three scenarios, two input vectors were applied. In the first, the single antecedent record of upstream stations was used. In the second, the average of three antecedent records was used. Given six input vectors, every one of them includes deferent combinations of input and output variables.

The single antecedent record of Q in the downstream station is considered another input variable of AI-based model that needed to predicts the Q for a head period equal to the Lt between the upstream and downstream stations. The estimated Lt between the upstream and downstream stations determines the Lag intervals between the input and output variables for the six input vectors. Using these input vectors, six combinations of input and out variables has been generated as shown in Table 3.7.

Table 3.7: Input vectors of the AI-based models

No.	Input Vector $X_{(t)}$	Output
1	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$
2	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$
3	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$
4	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$
5	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$
6	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+Lt)}$

Where, $Rf_{u(t)}$ represents a single records of hourly rainfall intensity at Ulu Yam station, $Rf_{u(t)}$ represents the average of three antecedent records of hourly rainfall intensity at Ulu

Yam station, $Wl_{u(t)}$ represents a single records of water level at Ulu Yam station and $Wl_{u(\bar{t})}$ represents the average of three antecedent records of hourly rainfall at Ulu Yam station.

$Rf_{b(t)}$ represents a single records of hourly rainfall intensity at Batang Kali station, $Rf_{b(\bar{t})}$ represents the average of three antecedent records of hourly rainfall intensity at Batang Kali station, $Wl_{b(t)}$ represents a single records of water level at Batang Kali station and $Wl_{b(\bar{t})}$ represents the average of three antecedent records of hourly rainfall at Batang Kali station.

$Rf_{k(t)}$ represents a single records of hourly rainfall intensity at Kerling station, $Rf_{k(\bar{t})}$ represents the average of three antecedent records of hourly rainfall intensity at Kerling station, $Wl_{k(t)}$ represents a single records of water level at Kerling station and $Wl_{k(\bar{t})}$ represents the average of three antecedent records of hourly rainfall at Kerling station.

$Rf_{a(t)}$ represents a single records of hourly rainfall intensity at Ampang Pecah station, $Rf_{a(\bar{t})}$ represents the average of three antecedent records of hourly rainfall intensity at Ampang Pecah station, $Wl_{a(t)}$ represents a single records of water level at Ampang Pecah station and $Wl_{a(\bar{t})}$ represents the average of three antecedent records of hourly rainfall at Ampang Pecah station.

$Q_{(t)}$ represents hourly stream flow at Rantau Panjang station and $Q_{(t+Lt)}$ represents ahead hourly stream flow at Rantau Panjang station with prediction time equal to Lt.

3.6.2 Estimation of the Lag Intervals between the Input and Output Variables

In determining the input variables of AI-based models to predict Q, the antecedent records of WL and RF that significantly affect the predicted Q should be estimated to select the most accurate lag intervals between the input and output variables (Sudheer et al., 2002).

These records could be accurately selected based on the results of Lt estimation between upstream and downstream stations as the hourly records of WL and RF at the upstream station represent the input variables of the AI-based model, whereas Q from the downstream station represent the output variables of the AI-based model.

The Lt between the upstream and downstream station was estimated by three approaches i.e. empirical formulas, CCA and NGA. The results of Lt are indicative of the potential lag intervals between the input and output variables for the AI-based models to predict real-time Q in the downstream area. Both of second and third approaches to estimate Lt, were employed in the selection of the lag intervals between the input and output variables of AI-based models, while, the first approach is performed only to provide an initial approximation of Lt between the upstream and downstream stations.

3.6.3 Integration of the Lag Time Results in the Selection of Models Variables

The results of Lt estimation were employed to explore the ability of the accurate timing of the input and output variables of AI-based models to improve the prediction performance of Q. The variables selection of AI-based models and modelling process were performed in two phases. First, the results of the Lt estimated by CCA were employed to select the lag intervals between the input and output variables of the AI-based models. The results of the HGA were then employed in the second phase to select the lag intervals between the input and output variables of AI-based models. The two phases of the variable selection and modelling process were used to explore the ability of improving the performance of AI-based models by the accurate selection of their variables based on the Lt estimation.

3.6.4 Selection of the Modelling Patterns

About 8753 patterns of hourly records of Q, WL and RF representing a one-year period (2011), were used for modelling. The basic statistical characteristics of the data, such as minimum, maximum, mean and standard deviation of hourly records of all stations employed are shown in Table 3.2.

For each ANNs-based model, the modelling data was divided into three datasets: 50% for training (4387 patterns), 25% for validation (2193 patterns) and 25% for testing the models (2193 patterns) (Maier et al., 2010). For each SVM-based models, the modelling data was divided into two datasets as 75% for training (6580 patterns) and 25% for testing the models (2193 patterns). The modelling patterns should be arranged as matrix in a suitable format, such as a spreadsheet for modelling requirements. The modelling matrix includes the data combinations of input and output variables as shown in Appendix B which presents group of modelling cases for three days.

The training dataset is utilized to train the models while the validation dataset is used in the early stopping of training process to prevent over-fitting and overtraining during the training step. The testing dataset serves to assess the performances of the AI-based models (Tiwari & Chatterjee, 2010).

3.6.5 Identification of AI-based Models Structure

After selecting the appropriate combination of input and output variables for the AI-based model and the modelling patterns, the structure of the modelling technique should be determined to initiate in the modelling processes. Model structure defines the functional form of the connection between inputs and output(s) variables of the model(Maier et al., 2010). The optimal model structure should compromise between generalization capability

and model complexity. Similar procedures are applied for both ANNs- and SVM-based models.

3.6.5.1 ANNs-based Models

The NN specifications, including network structure, connection scheme and weight range, should be selected to develop the ANNs-based models using the techniques (i.e., MLP, RBF, and GRNN). The number of layers and number of neurons per layer often specify the network framework. Next, the neuron specifications, that is, the activation function and its range, should be determined, followed by system dynamics and training algorithm selection. The following procedures describe how to identify the structure of the ANNs-based models:

1. Define the NN specifications, such as the network structure, the types of connections, the order of connections, and the weight range;
2. Select the node properties, such as the activation range and the activation (transfer) function;
3. Select the system dynamics, such as the weight initialization scheme, the activation-calculating formula, and the learning rule; and
4. Determine the topology of a NN, including the number of layers and neurons per layer. Each neural network should include three layers: input, hidden, and output. The input layer represents the input data, and the output layer comprises the model output. The hidden layer includes the activation function to provide nonlinearities for the NN and may consist of one or more layers with an unlimited number of neurons (Maier et al., 2010).

So far, no scientific approach to selecting the ideal number of hidden layers and neurons exists. The optimal number of neurons is identified using a trial-and-error process by developing many ANNs-based models and evaluating them. In the case of the ANNs-

based model with few hidden layers and neurons, high training and error occurs because of underfitting and high statistical bias. In the case of the ANNs-based model with many hidden layers and neurons, low training error with high generalization error occurs because of overfitting and high variance (Abrahart et al., 2004).

The optimum number of hidden layers of the ANNs-based models is influenced in a complex manner by several elements:

- the number of input variables;
- the number of data patterns;
- the complexity of the process to be modelled;
- the amount of noise in the training data;
- the type of the ANNs-based model; and
- the type of training algorithm and activation function (Maier et al., 2010).

3.6.5.2 SVM-based Models

An appropriate model structure should be selected first to develop the SVM-based models. Usually the construction of the SVM-based model involves similar procedures to those of ANNs-based models with some changes based on the differences between the modelling mechanisms of the two approaches. The following procedures describe how to identify the structure of the SVM-based models:

1. Select the appropriate features of the SVM-based model. This step is critical and based on the goal of the model;
2. Select the training constant and capacity of the model. They should be scaled within an appropriate range based on the size and complexity of the training data;
3. Select a suitable kernel function. Some kernel functions should be tested before selecting the best one; and

4. Select the best parameters *C* and gamma. *C* is the penalty coefficient used to define how much error can be tolerated and gamma is related to the kernel function being chosen, influencing the final space distribution. These two parameters influence the performance of SVM-based model and have to be selected carefully (X. Chen et al., 2013).

3.6.6 Models Training

Once the structure of the AI-based model is identified, the conditions for stopping the training process should be fixed prior to beginning the training process. Some of the conditions that control the training are the maximum number of iterations, maximum time of training, target performance that specifies the tolerance between the observed and the predicted Q, and minimum learning rate.

The training process of the ANNs- and SVM-based model usually involves similar procedures with some changes based on the differences between the modelling mechanisms of the two approaches.

The training process in the ANNs-based models includes the following steps: The input variable records are inserted into the input layer and then weighted and forwarded to the hidden layer. The neurons in the hidden layer create outputs by applying an activation function to the sum of the weighted input values. Next, the outputs of the hidden layer are weighted by the connections between the hidden and output layers. The desired results are finally produced in the output layer. The ANNs-based models reach the optimum prediction performance by continuously modifying the interconnected weights until good accord is achieved between the observed and predicted Q with minimum residuals between them.

The training process in the SVM-based models includes the following steps: The input variable records are inserted into the input layer and then passed on to the hidden layer to create outputs by applying a kernel function to the sum of the input values. Next, the outputs of the hidden layer are weighted, and the desired results are finally produced in the output layer. The SVM-based models reach the optimum prediction performance by changing the kernel function and modifying the values of SVM parameters continuously until a good accord is achieved between the observed and the predicted Q with minimum residuals between them (X. Chen et al., 2013).

3.6.7 Models Calibration

The ANNs-based model is still under training to reach the best model performance. However, the NN may "memorize" the training set instead of learning it. Calibration is employed to prevent memorization from occurring. Calibration is used to indicate that the NN has trained enough, thereby ending the iteration process. This process can be performed based on best correlation or minimum error. The calibration process is not required in the SVM-based models.

3.6.8 Performance Evaluation Criteria

The performance of the models was assessed based on three criteria: R, R^2, and MAE. R is a statistical technique used to indicate the strength and direction of a linear relationship between two groups of data representing two variables or observed and predicted data (i.e. the observed and predicted Q) (Perugu et al., 2013). The most widely used is the Pearson correlation coefficient (R).

It is obtained by dividing the covariance of the two variables by the product of their standard deviations, as described in Equation 3.2

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3.2)$$

where n is the number of pairs of data, and x and y are two variables (i.e. the observed and predicted Q).

The R values fall between +1 and −1 and indicate the strength of the linear relationship between the variables; R = 0 signifies no linear relationship between the variables. R is +1 in a perfectly increasing linear relationship circumstance, whereas R is −1 in a perfectly decreasing linear relationship instance.

R^2 describes the variance between the two groups of data or two variables (i.e. observed and predicted Q). It indicates how well the data fit a model. It is a statistic parameter employed to measure of how well real-world data are simulated by the model. R^2 values between 0 and 1, A value of 0 indicates no correlation, while a value of 1 mean that the model can explain all of the observed variance (Besaw et al., 2010).

MAE is used to evaluate the residual or the differences between the two groups of data or two variables (i.e. the observed and predicted Q). Theoretically, the minimum value of MAE is zero, meaning the model represents a perfect fit, something that is not easy to achieve. MAE has no maximum value (Davydenko & Fildes, 2014; Schulmerich et al., 2015). The following equation represents MAE:

$$MAE = \frac{\sum_{i=1}^{n}\left|X_{o,i} - X_{m,i}\right|}{n} \quad (3.3)$$

3.6.9 Procedural Steps in Building AI-based Models

In this research, both ANNs- and SVM-based models were designed and built using the STATISTICA software. In addition to both basic and advanced statistics, STATISTICA offers specialized tools for developing ANNs- and SVM-based models. The procedural steps in building and applying for SVM-based models and ANNs-based models vary slightly according to the employed tool. Using STATISTICA, the procedural steps involves the following procedures:

3.6.9.1 Data Importation

The input modelling data matrix which includes the modelling patterns should be fed into the program so that the model can be trained by the data entry tool. The data must be in a suitable format, such as a spreadsheet. The modelling data matrix includes the cases that the model uses in the training and testing processes.

3.6.9.2 Problem Definition

The input (independent) and output (dependent) variables for the model should be defined before conducting the training. The modelling process is performed in two phases. In first phase, the results of the Lt estimated by CCA are applied in the selection of the lag intervals between the AI-based models variables. Thereafter, the results of HGA are applied in the second phase. The input and output variables of the first and second modelling phase are described in chapter four.

3.6.9.3 Data Division

The modelling data cases for the ANNs-based models are divided into three datasets: 50% for training (4387 patterns), 25% for validation (2193 patterns) and 25% for testing the models (2193 patterns). The modelling data for the SVM-based models are divided into two datasets: 75% for training (6580 patterns) and 25% for testing the models (2193

patterns). The patterns of theses sets are randomly selected, and the modeller can change these percentages.

3.6.9.4 Model Structure Design

The appropriate structure of the models should be selected for the employed techniques. Four AI techniques (i.e., MLP, RBF, GRNN, and SVM) are employed in the two modelling phases, resulting in 32 AI-based models to predict Q. This step was previously presented in Section 3.6.5.

3.6.9.5 Model Training

Once the structure of the models is designed, the software becomes ready to start training process. The model continuously trains until one of the stopping conditions is achieved, in which case the training process is stopped, as mentioned in Section 3.6.6.

3.6.9.6 Model Testing

After stopping the training process, the model is then tested using a group of cases were not used in the training session. Thereafter, the model performance is assessed based on three criteria: R, R^2, and MAE. The model is then ready to be applied to predict any other values of input variables.

3.6.10 Flowchart of the Modelling Process Using STATISTICA Program

The main procedural steps in building AI-based models using STATISTICA are described in Figure 3.13.

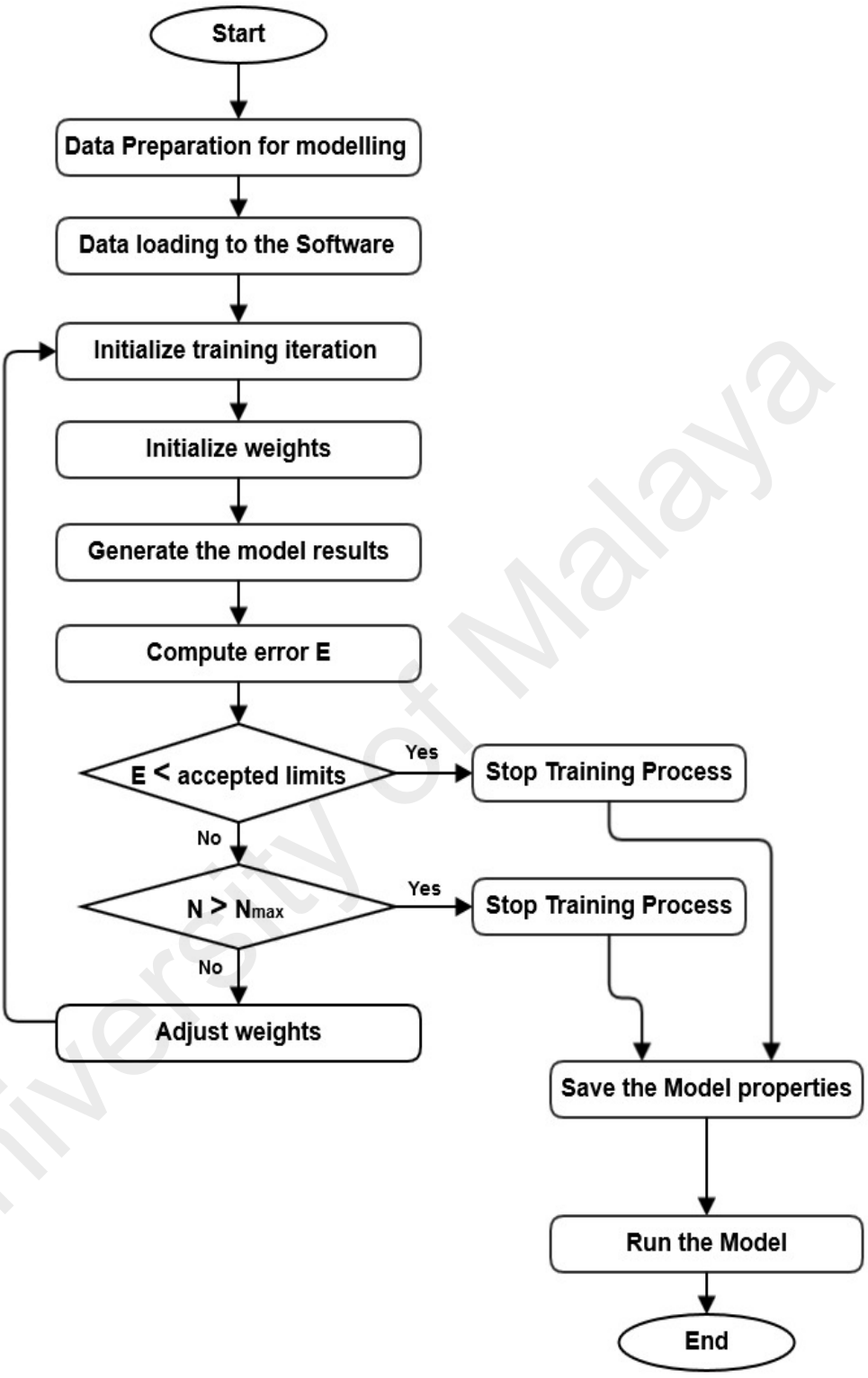


Figure 3.13: Flowchart of the procedural steps in building AI-based Models using STATISTICA program

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

Chapter 4 presents the results and discussion of the research. It contains a detailed hydrological description of the Selangor River basin, including an analysis of the long-term changes in the SF regimes of the Selangor River basin over a 50-year period from 1960 to 2010 and the results of the Lt estimation between the upstream and downstream stations, which are required in determining the input and output variables of the AI-based models and the lag intervals between them.

Chapter 4 also presents the results and discussion of the two phases of the Q modelling process, including the description of the six AI-based models that were trained and developed by four AI techniques, namely, MLP, RBF, GRNN and SVM, resulting in the development of 32 AI-based models to predict the real-time Q. The results include the specifications of the AI-based models, the performance evaluation criteria (i.e., R, R^2 and MAE) of the AI-based models and the comparison between the observed and the predicted Q via the AI-based models.

The results also include exploring the ability of improving the performance of AI-based models by the accurate selection of the lag intervals between the input and output variables of the model based on the estimation of the Lt. A description of some hydrological applications of the developed AI-based models are also included and discussed in Chapter 4.

4.2 Results of the Long-Term Changes of Stream Flow Regime

The results include the variations among nine variables describing the annual SF, the changes in monthly SF and changes in the annual duration of high and low SF events. The analysis was performed based on two time scales: yearly and sub-periodic. The sub-periods were estimated by segmentation of the study period into seven sub-periods using two methods, namely the change-point test and direct method.

Significant changes were observed in the nine variables as well as the monthly SF and the yearly duration of high and low SF with respect to time. The observed variations verified the presence of long-term variations in SF regime, which will raise the possibility of floods and droughts happening in future.

4.2.1 The Changes in the Hydrological Variables of Annual Stream Slow

4.2.1.1 The Yearly Changes

Investigation of the changes in the nine variables over 50-year period from 1961 to 2010 was performed firstly based on a yearly time scale.

Figure 4.1 presents three time series of SF1, SF2 and SF3. There is no apparent trend in the three variables. This figure demonstrates almost negligible variations and a minor trend in the three variables, although it is noticed that the values of the three variables get farther from the mean values with respect to time.

The time series of SD, CV, RA and PR over the 50-year period from 1961 to 2010 are shown in Figure 4.2. According to this figure, no clear trend was noticed in the SD, CV, RA and PR variables to describe the dispersion of annual SF data.

Generally, the outcomes of the investigation process based on a yearly time scale didn't provide clear trend and results, which justify the necessity to the investigation based on a sub-periodic time scale as reached via two segmentation methods described in Chapter 3.

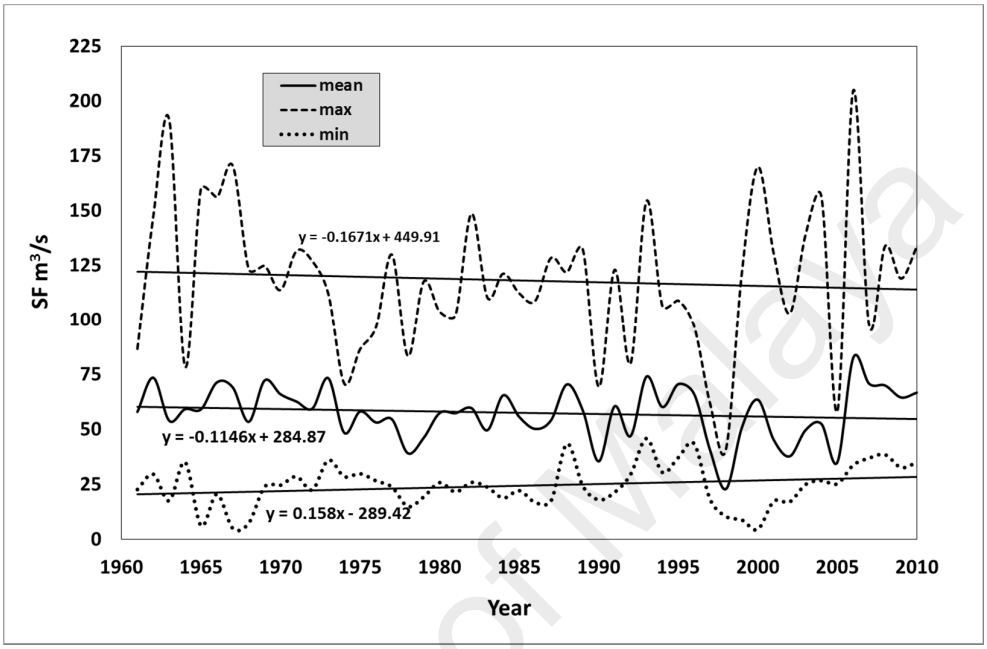
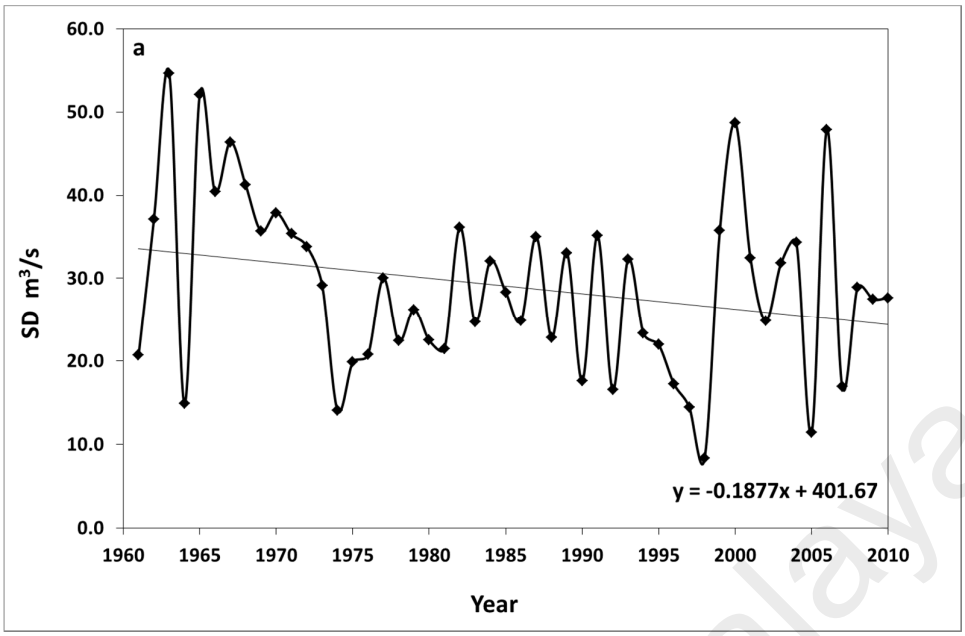
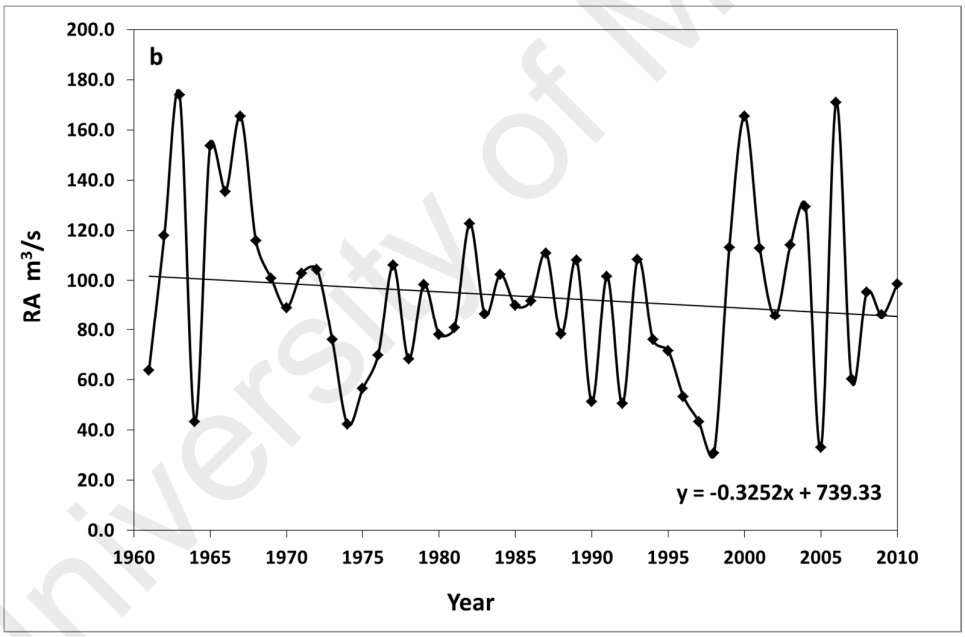
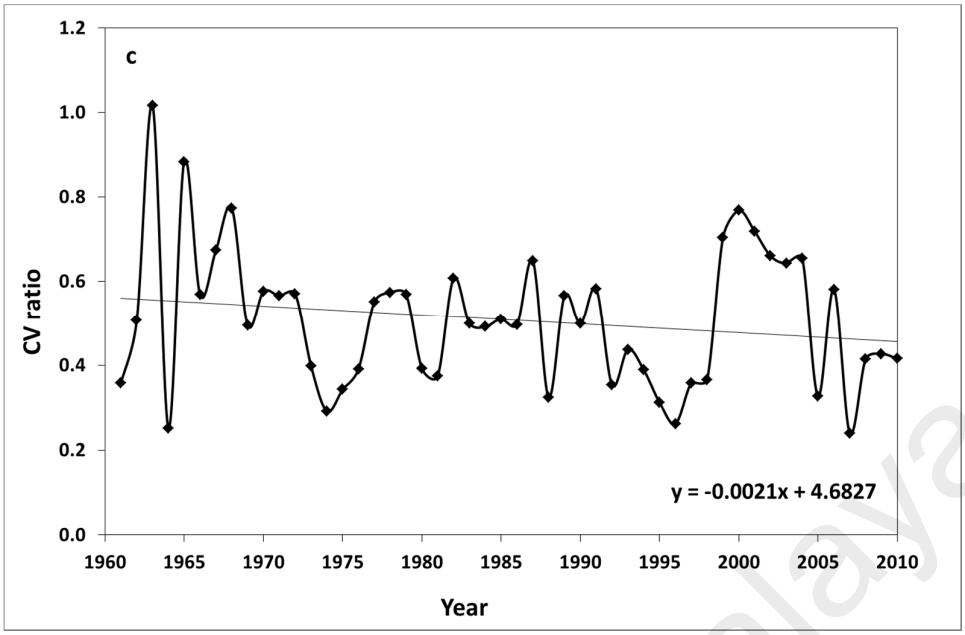


Figure 4.1: Changes in mean annual flow, maximum monthly stream flow per year and minimum monthly flow per year over the study period







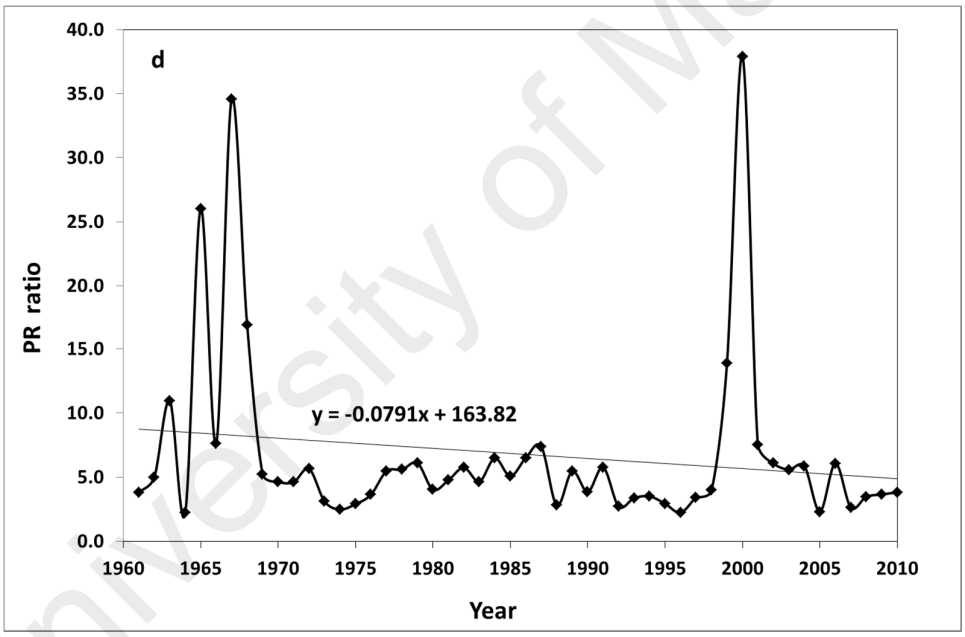


Figure 4.2: Changes in hydrological variables over the study period: (a) standard deviation (SD), (b) the range between maximum and minimum stream flow (RA), (c) coefficient of variation (CV) and (d) the Pluviometric Ratio (PR)

4.2.1.2 The Sub-Periodic Changes

- **The Changes over the Sub-Periods Obtained via the Change-Point Test**

The variations of the nine hydrological variables over the sub-periods reached by the change-point test are presented in Figure 4.3. Although the analysis shows almost negligible variations in SF1 over the sub-period as presented in Figure 4.3 (a), the SF2 gradually increased, particularly from the fourth sub-period.

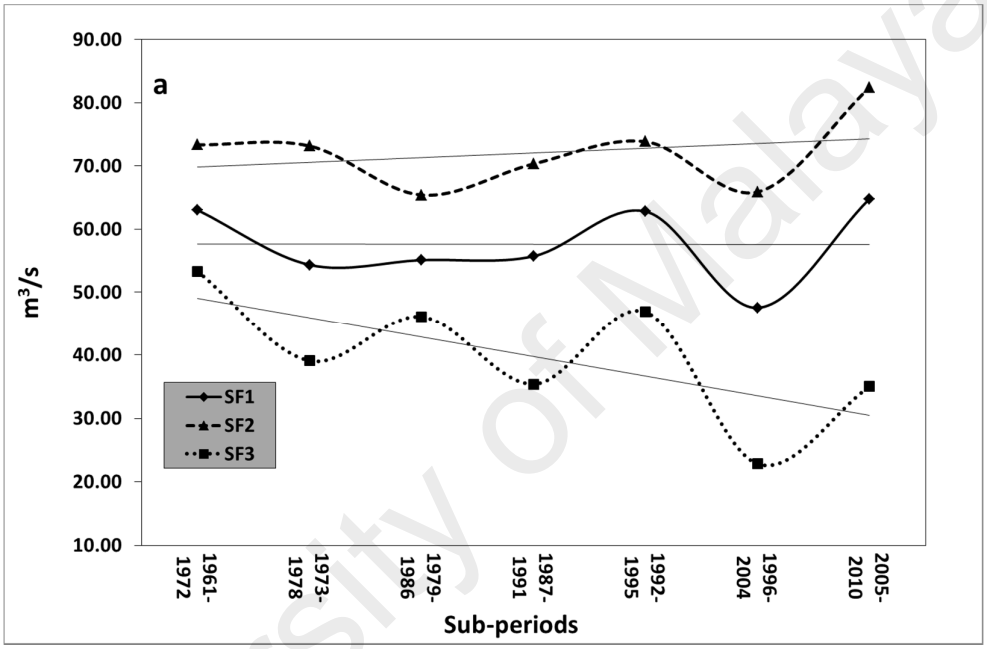
According to this result, the occurrence of high SF increased in the latest sub-periods. Such variations result in appropriate hydrological circumstances, whereas flood events will possibly take place more often in future.

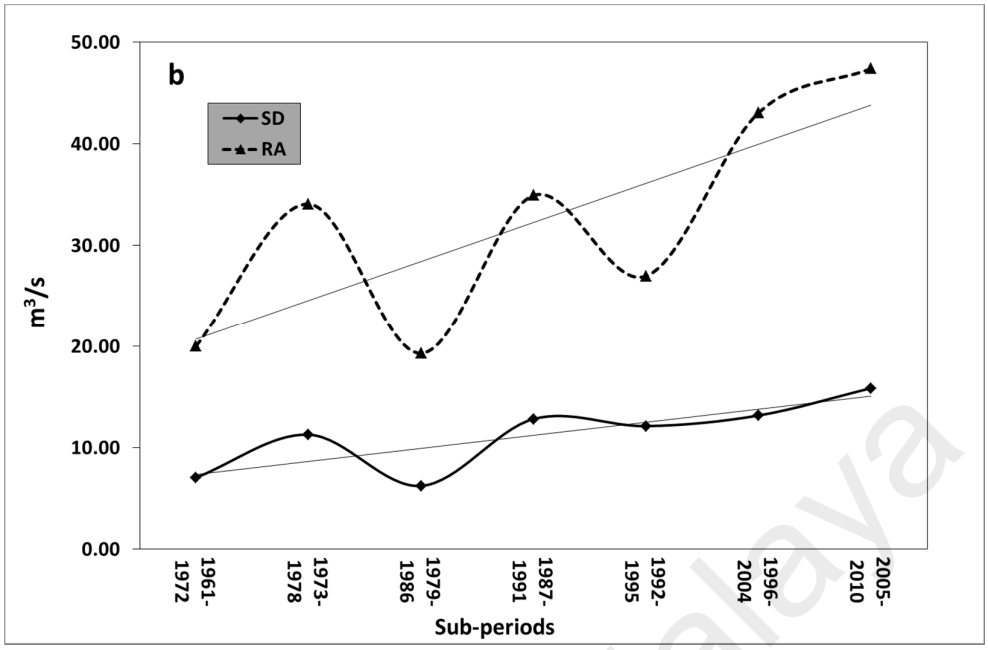
Although the SF2 generally increased, it is observed that the SF3 declined, especially from the fifth sub-period. Evidently, low SF occurrence increased in the latest sub-periods, leading to suitable hydrological conditions for droughts to occur. As such, drought periods may happen more frequently in future.

Figure 4.3 (b) indicates that the RA increases significantly with respect to time. Figure 4.3 (b) also presents an incessant increase of SD with respect to time. In addition, it is noted that the SF gets farther from the SF1, providing another indication of the increasing probability of high and low SF events occurring in future.

In the early sub-periods, the PR values are close to 1 as can be seen in Figure 4.3 (c), indicating that the dispersion in SF is negligible. In the later sub-periods, there is an incessant increment in PR. The PR is equal to 2.89 and 2.35 in the last two sub-periods, meaning that annual variability of SF is becoming very high. This is another evidence of increased probability of high and low SF events. Figure 4.3 (c) indicates a constant increase in the CV with respect to time. The CV in the last two sub-periods are doubled compared to its value in the first two sub-periods.

Obviously, the annual SF was stay around the mean in early sub-periods. Significant dispersion started to occur in the annual SF and the nine variables in the later sub-periods. These analyses therefore, verify the presence of considerable variations in the annual SF regime of the Selangor River basin along with the fact that these variations can produce appropriate hydrological circumstances for the increased probability of high and low Sf events occurring in future.





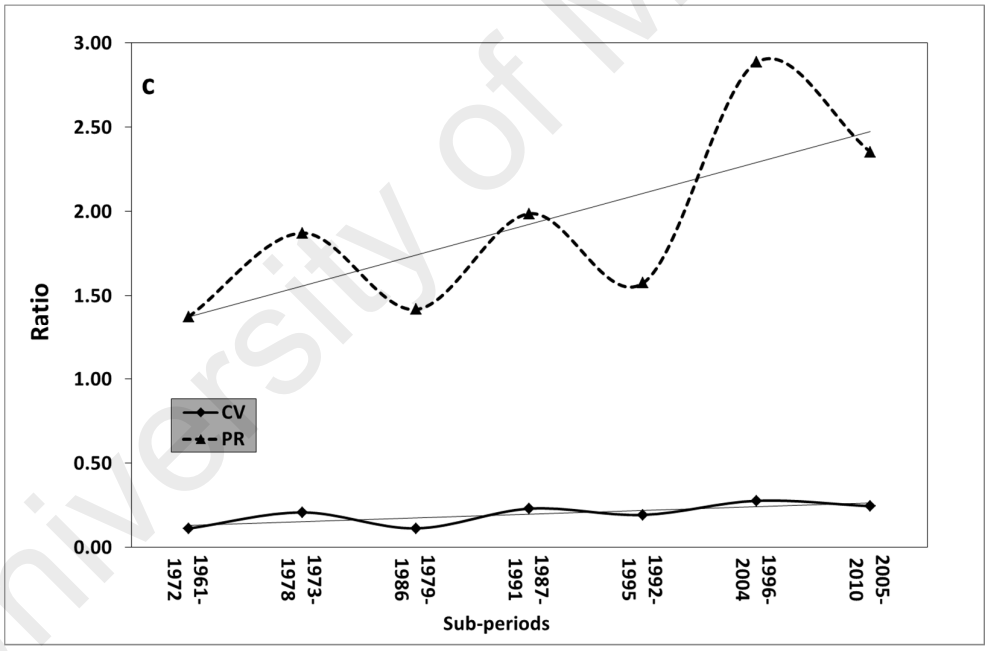
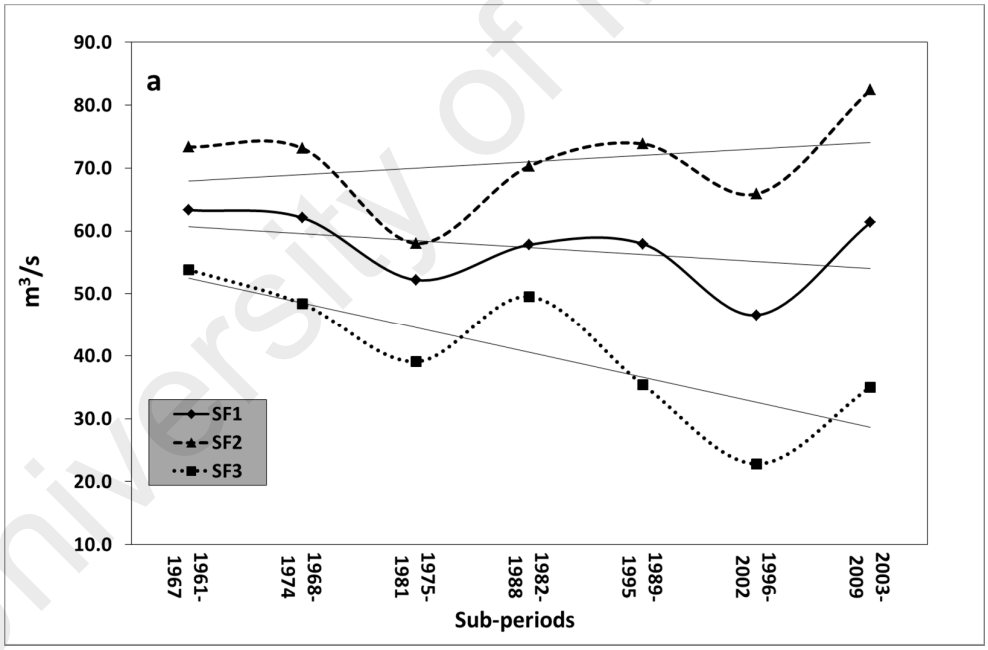


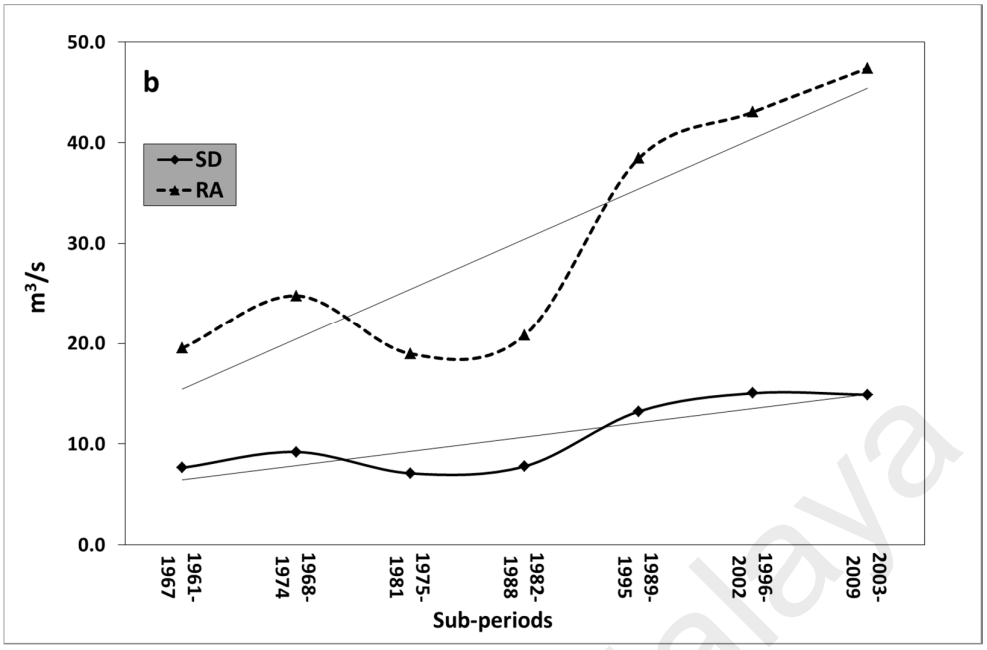
Figure 4.3: Variations in hydrological variables over the sub-periods obtained by change-point test: (a) mean, maximum and minimum annual flow, (b) standard deviation and range between maximum and minimum annual stream flow and (c) coefficient of variation and Pluviometric ratio

- **The Changes over the Sub-Periods Obtained via the Direct Technique**

The variations of hydrological variables over the sub-periods reached by direct technique are illustrated in Figure 4.4. Generally, the results of the hydrological variable variations' over the sub-periods obtained by direct method are comparable to those of change-point test.

This symmetry emphasizes the results regarding the long-term variations of annual SF and provides further evidence about formation appropriate hydrological circumstances for the increased probability of high and low SF events occurring in future.





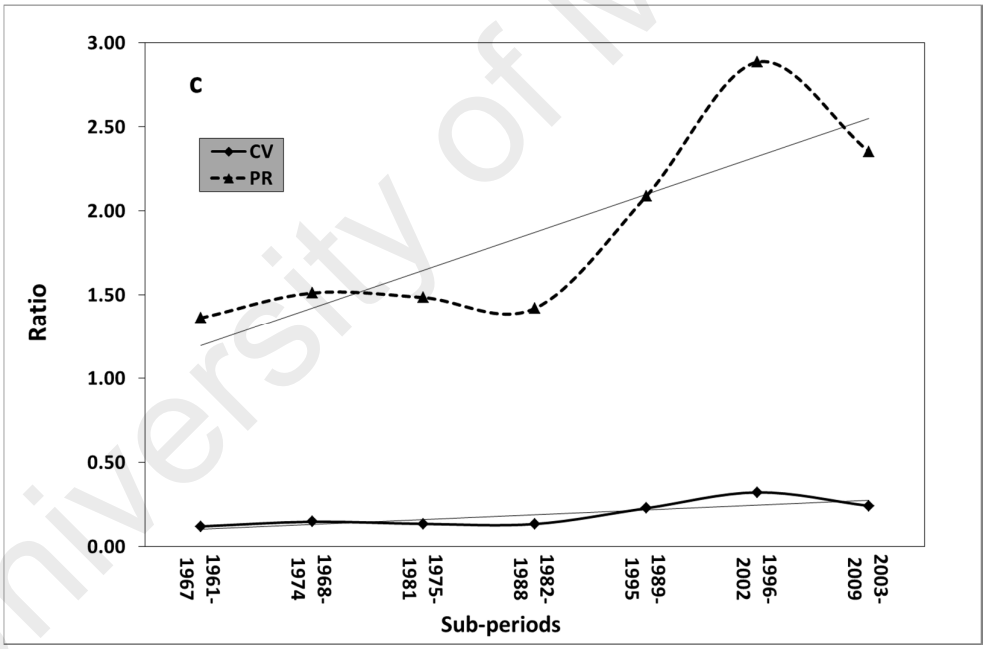


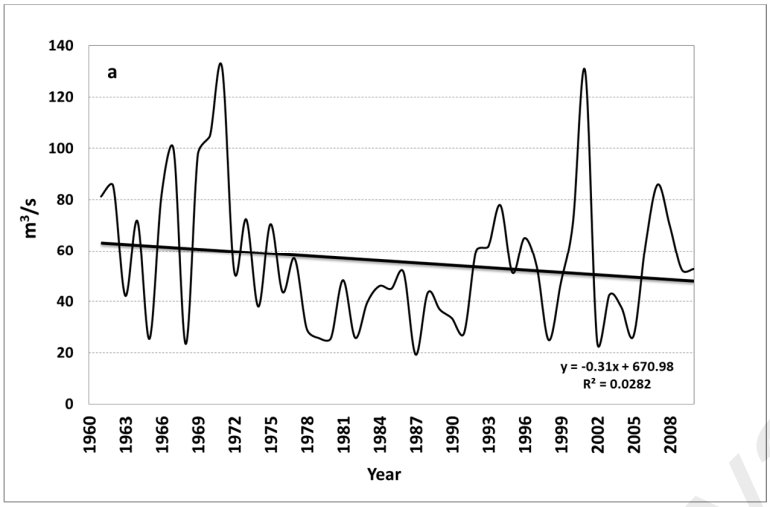
Figure 4.4: Changes in hydrological variables over the sub-periods obtained by direct technique: (a) mean, maximum and minimum annual stream flow, (b) standard deviation and range between maximum and minimum annual stream flow and (c) coefficient of variation and Pluviometric ratio.

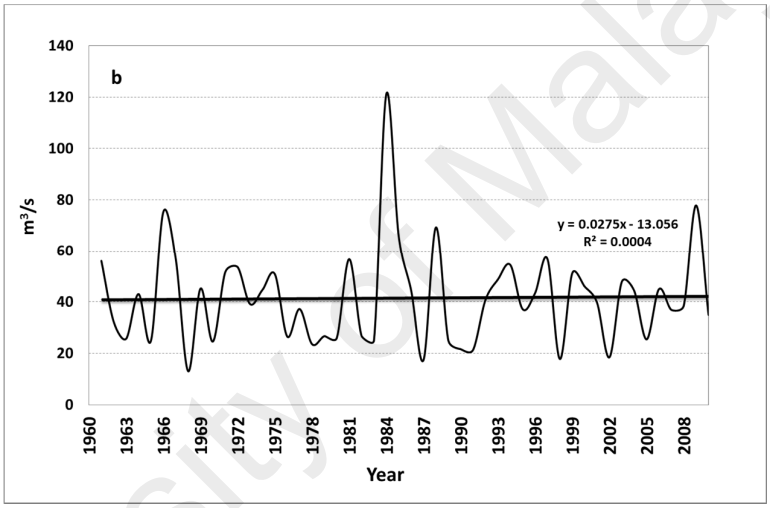
4.2.2 The Changes in the Monthly Stream Flow

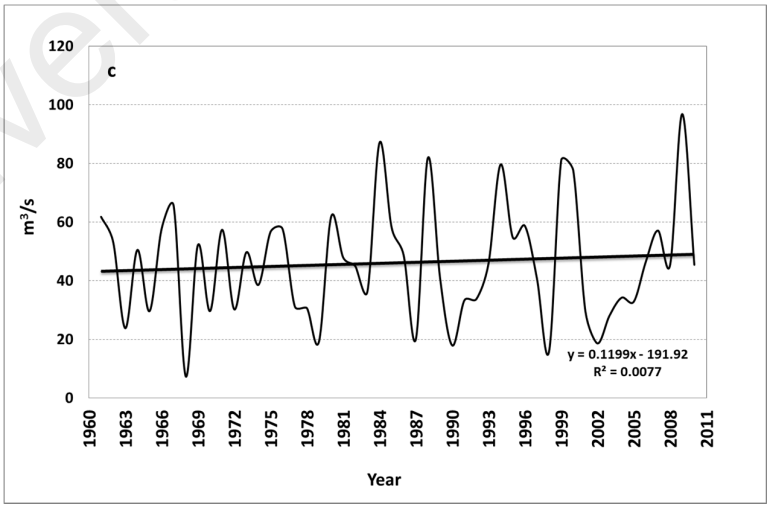
4.2.2.1 The Yearly Changes

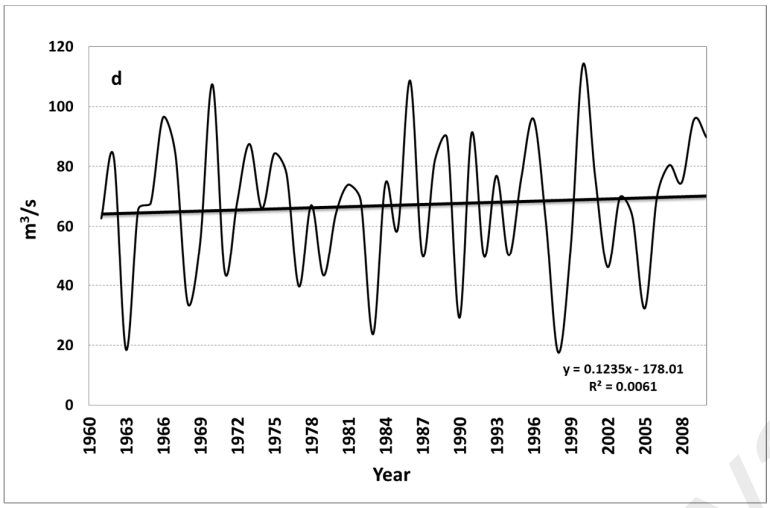
Investigation of the changes in the monthly Sf over 50-year period from 1961 to 2010 was performed firstly based on a yearly time scale. According to results, no significant change was observed in the monthly SF. However, slight decline was noticed in monthly SF for January, May, November and December, whereas slight increase was noticed in monthly SF for March, April, July, August and September. Almost no changes was noticed in monthly SF for the other months. Figure 4.5 presents changes of monthly SF of four months (January to April) over 50-year period from 1961 to 2010.

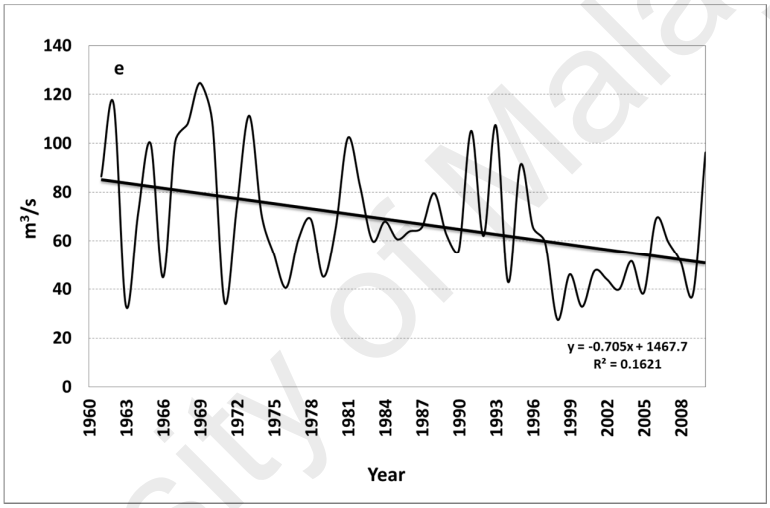
Generally, the results of the yearly investigation of monthly SF didn't provide clear trend and results, which justify the necessity to the investigation based on a sub-periodic time scale as reached via two segmentation methods described in Chapter 3.

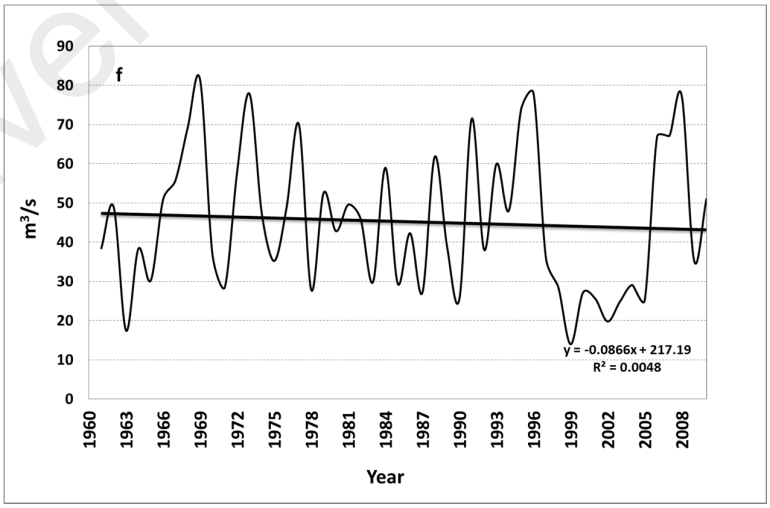


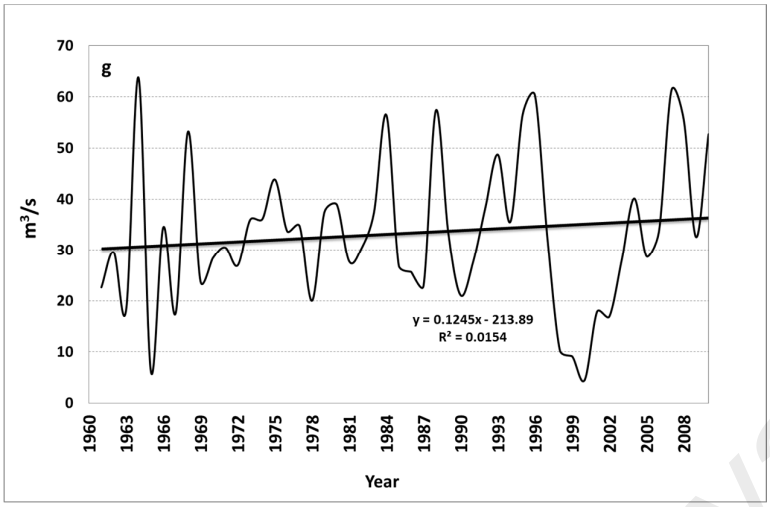


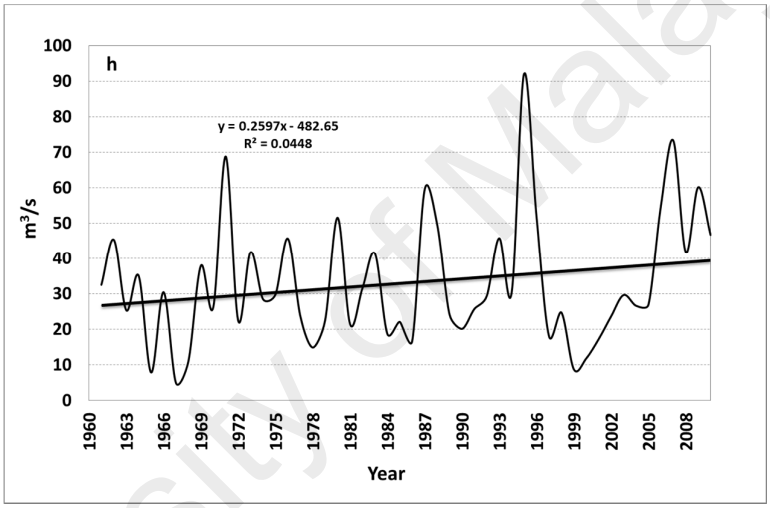


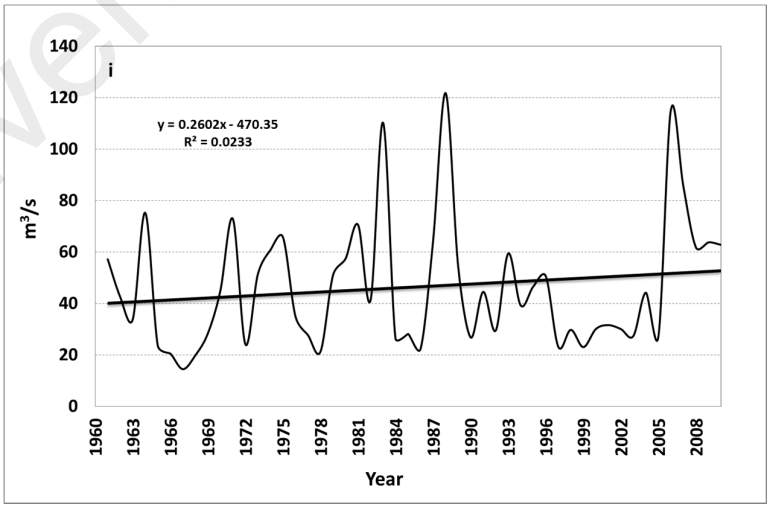


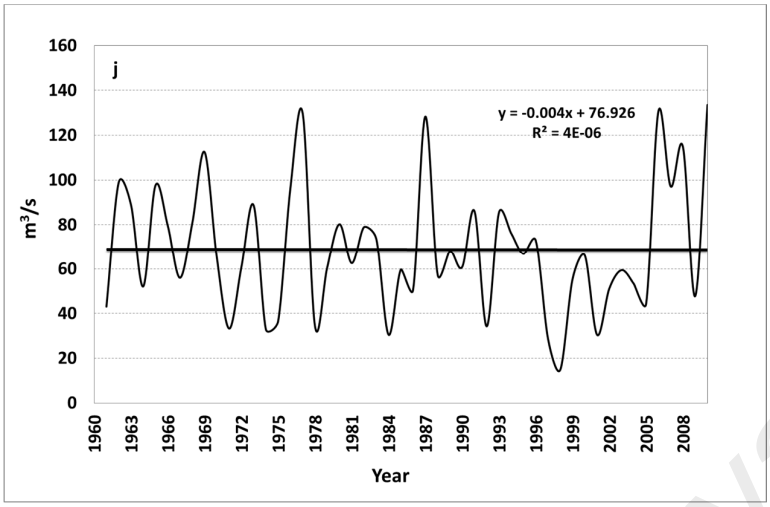


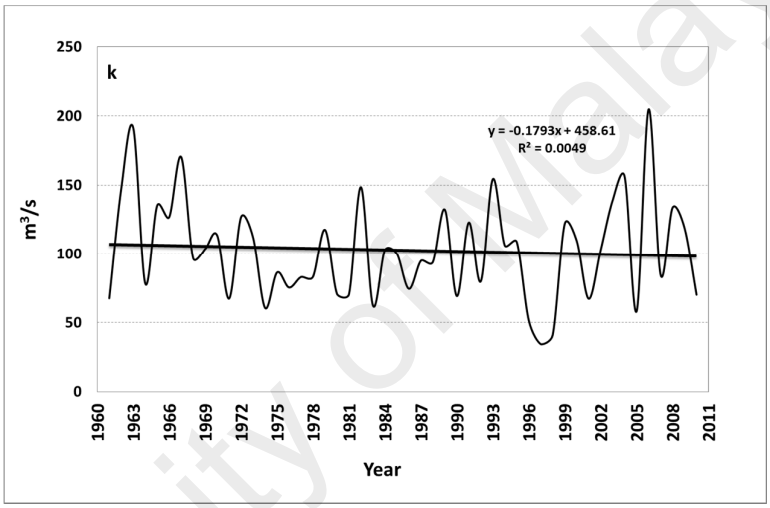












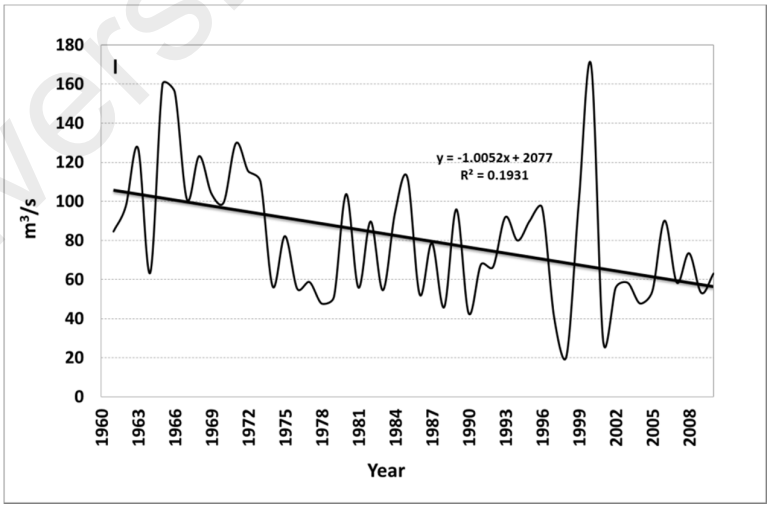


Figure 4.5: Changes in monthly stream flow over the study period

(a) January, (b) February, (c) March, (d) April, (e) May, (f) June, (g) July, (h) August, (i) September, (j) October (k) November and (l) December

4.2.2.2 The Sub-Periodic Variations

- **The Changes over the Sub-Periods Obtained via the Change-Point Test**

The Variations of the mean monthly SF of the sub-periods reached by the change-point test over the 50-year period from 1961 to 2010 are presented in Figure 4.6.

According to results, clear changes were observed in the mean monthly SF of the sub-periods over the study period particularly in January, May and from September to December, whereas in other months, no significant change was observed.

- **The Changes over the Sub-Periods Obtained via the Direct Technique**

The Variations of the mean monthly SF of the sub-periods reached by the direct technique over the 50-year period from 1961 to 2010 are presented in Figure 4.7. Generally, these changes are very similar to those changes perceived by means of the change-point test. This resemblance emphasizes the result regarding variations in monthly SF over the study periods.

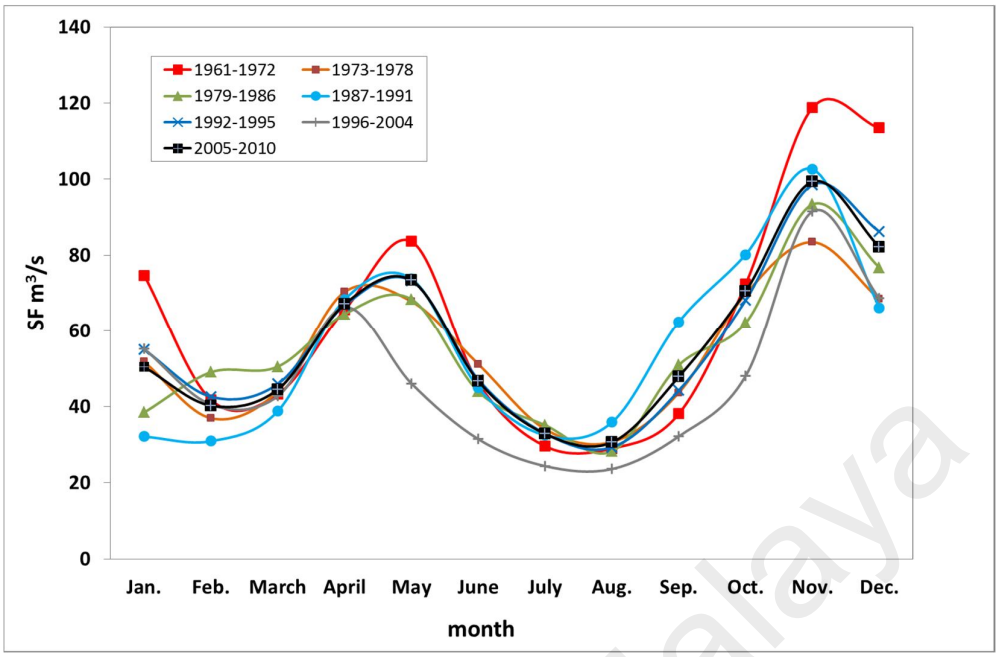


Figure 4.6: Changes in monthly stream flow over the sub-periods obtained by the change-point test

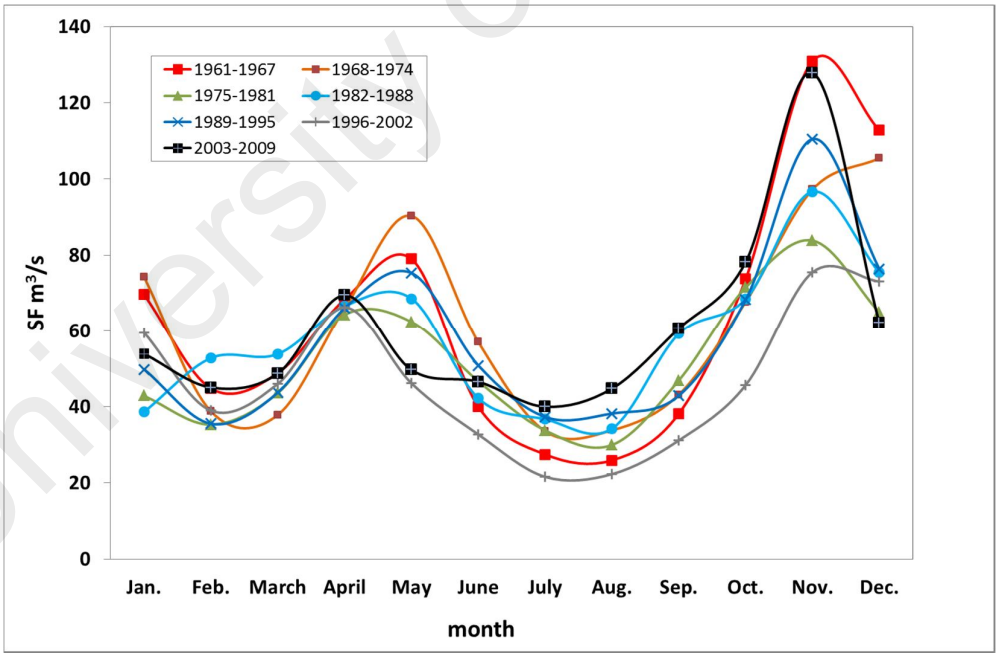


Figure 4.7: Changes in monthly stream flow over the sub-periods obtained by the direct technique

4.2.3 The Changes in High and Low Stream Flow Duration

4.2.3.1 The Yearly Changes

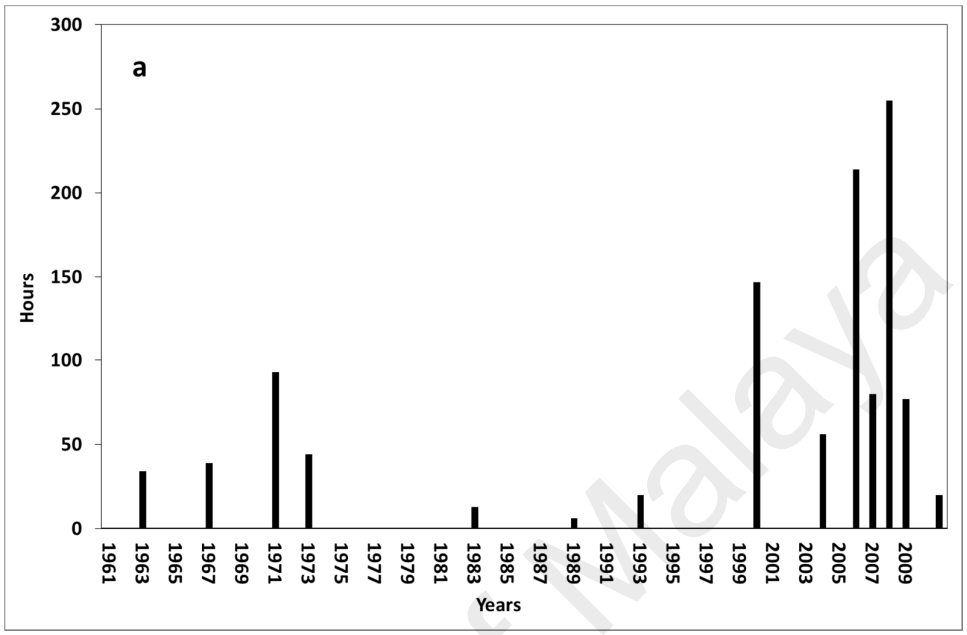
The yearly duration of high and low SF events were investigated over 50-year period from 1961 to 2010. The results include an evaluation of the changes in high and low SF duration and trend testing.

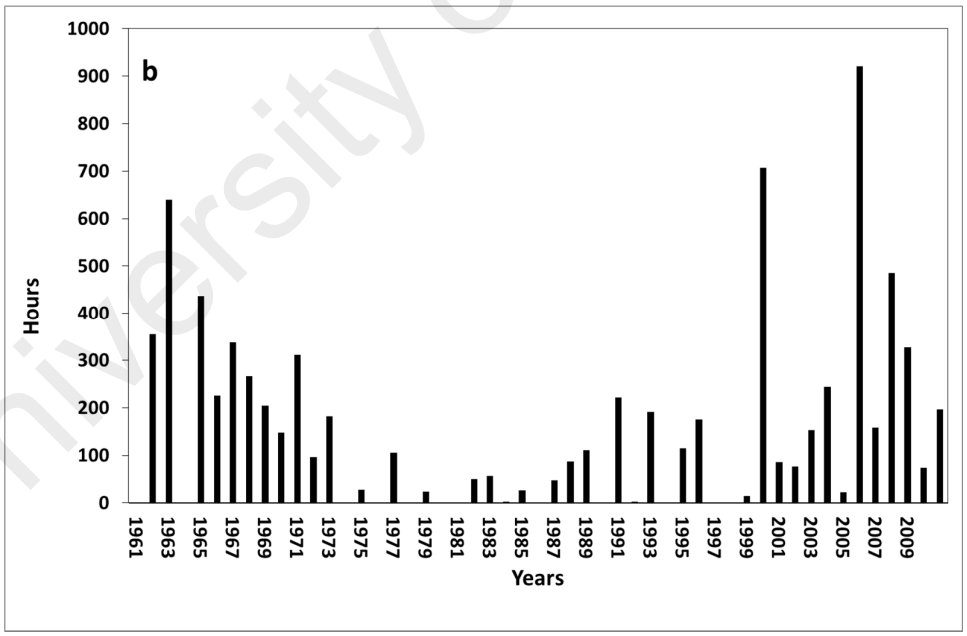
The high SF duration analysis was investigated based on three levels: danger level, when the SF is higher than 250 m^3/s; warning level, when the SF is above 180m^3/s; and alert level, when the SF is more than 160 m^3/s. The three levels are as determined by the Department of Irrigation and Drainage (DID). The low SF duration analysis was investigated in a single level, which is when the SF falls below 14.5 m^3/s which denotes about 25% of the average annual SF over 50-year period from 1961 to 2010.

Figures 4.8 show the annual duration of the three levels of high SF while Figures 4.9 shows the annual duration of low SF. There is a noteworthy rise in the danger level duration, particularly in the last decade, while slight change occurs in both of duration of warning and alert levels. Minor change was also observed in the yearly duration of the low SF.

For more clear results, the three years moving average of annual duration of the high and low SF were employed to investigate the overall trend of changes over the 50-year period from 1961 to 2010. Figure 4.10 presents the three years moving average of the annual duration of the three levels of high while Figure 4.11 presents the three years moving average of the annual duration of low SF over 50 years. The three moving average also shows significant increment in the duration of danger level. Figure 4.9 also displays a minor increase in the duration of warning level, almost no variation in the duration of alert level and a minor decline in the duration of low SF. Such changes in the yearly

duration of SF danger level results in formation of suitable hydrological circumstances, whereas the flood events would probably take place more frequently in future.





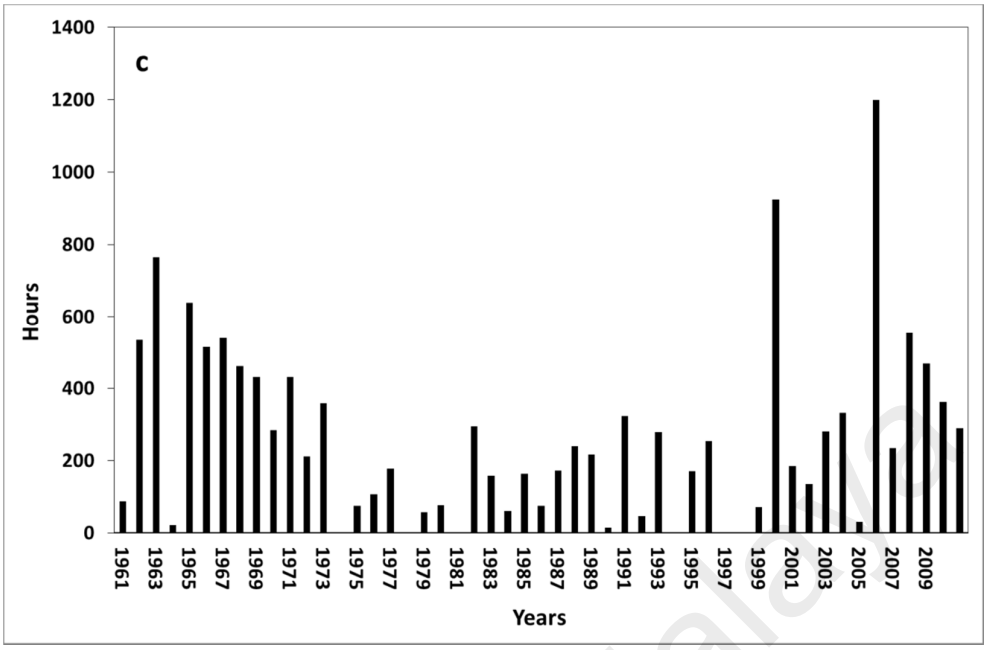


Figure 4.8: Yearly duration of high stream flow over 50 years: (a) danger level, (b) warning level and (c) alert level

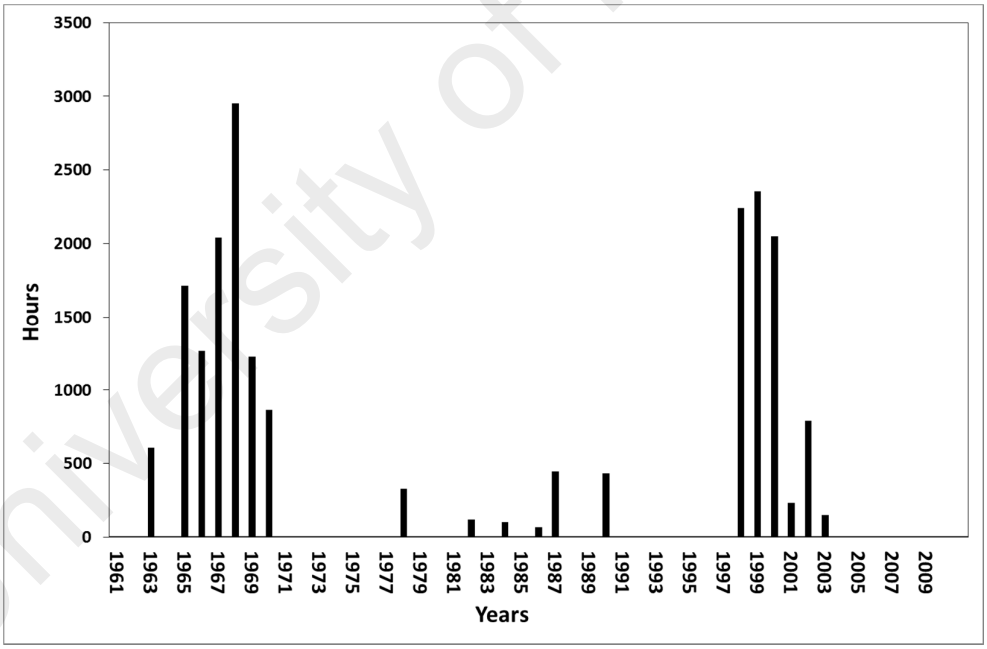
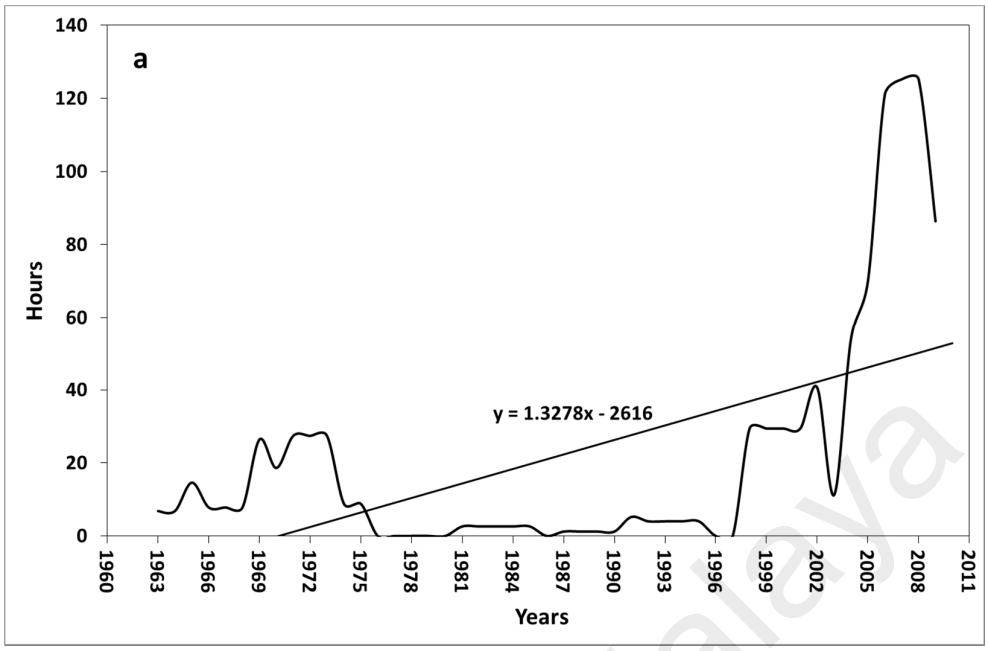
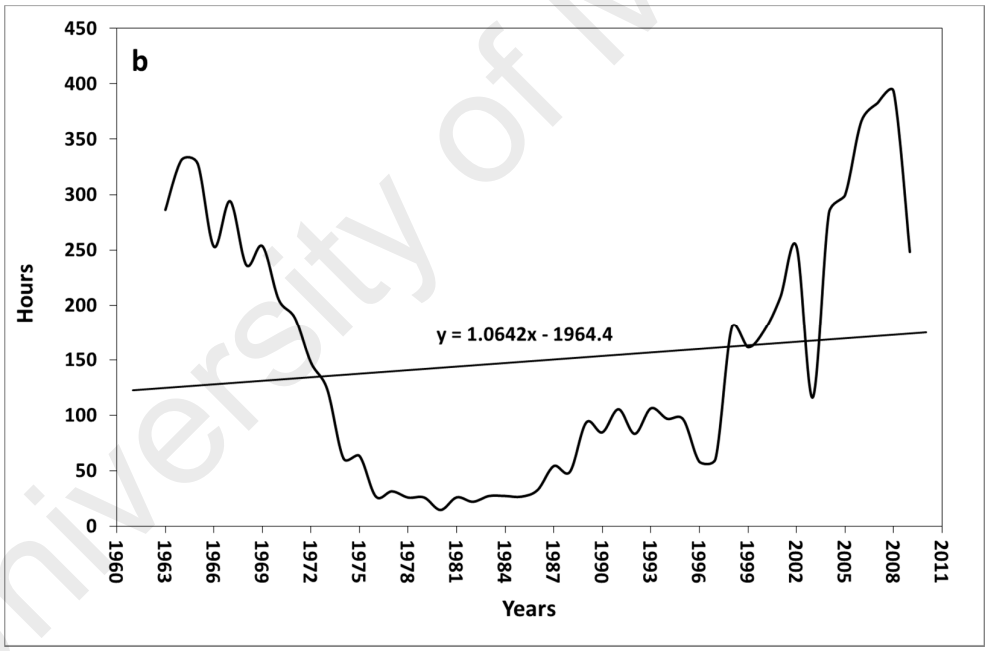


Figure 4.9: Yearly duration of low stream flow over 50 years





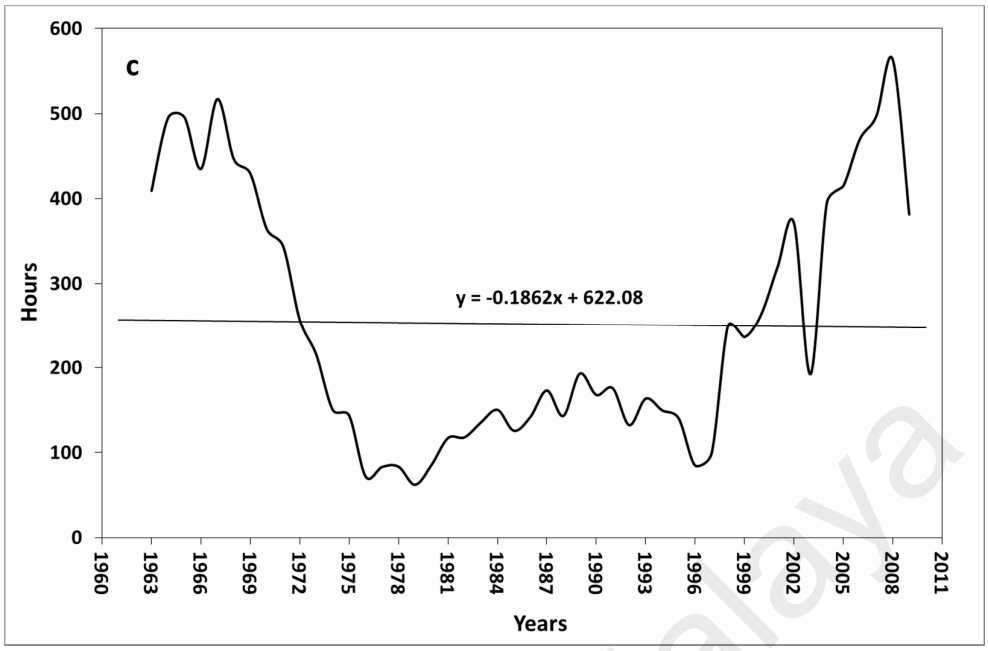


Figure 4.10: Three years moving average of the yearly duration of high and low stream flow: (a) danger level, (b) warning level and (c) alert level

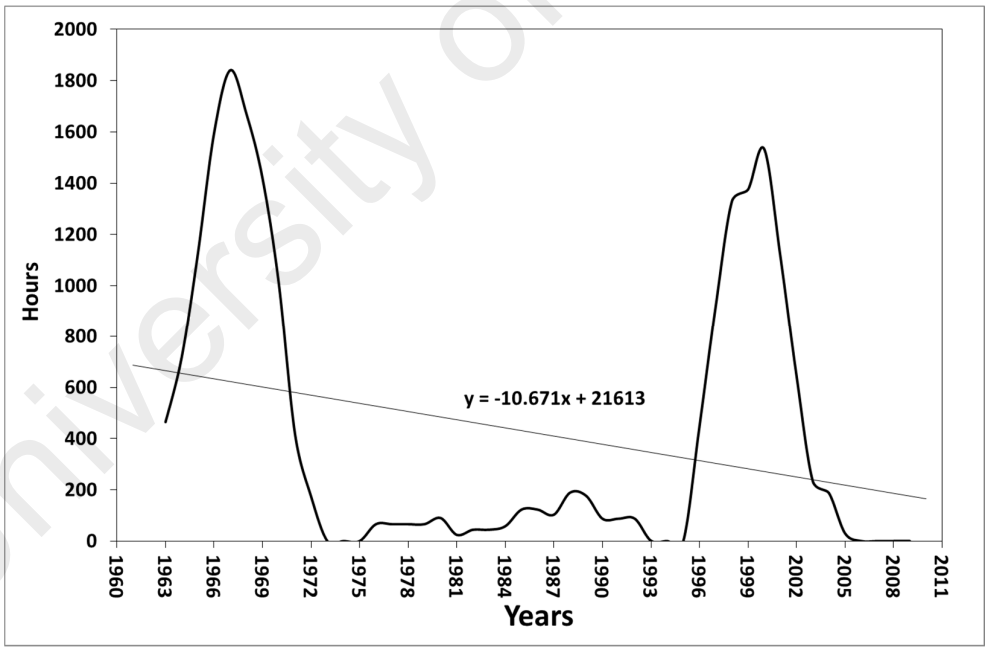


Figure 4.11: Three years moving average of the yearly duration of low stream flow

4.2.4 General Discussion about the Analysis of Stream Flow Regimes

The long-term changes analysis contains an investigation of the variations in nine variables describing annual SF. The changes in the monthly SF and annual duration of high and low SF were included in the investigation. The analyses were performed based on two time scales namely, yearly and sub-periodic changes. The sub-periods were achieved by segmentation of the study period into 7 sub-periods by two methods namely, change-point test and direct technique.

Even the results shown almost minor variations in SF1 over the study period, the SF2 generally increased while the SF3 declined with respect to time. It was also noticed that the variables describing the dispersion in SF data, incessantly increased with respect to time. No significant change was observed in the monthly SF, whereas an obvious increment was noticed in the annual duration of danger level of SF while a slight increment was noticed in the annual duration of both warning and alert levels. A slight decline was observed in the yearly duration of low SF. According to these outcomes, two main results were drawn as follows:

- Obvious variations were detected in the annual SF, monthly SF and annual duration of high SF over the 50-year periods from 1961 to 2010. These variations verified the existence of the long-term variations in the SF regime.
- The verified long-term variations in SF regime may potentially result in the formation of appropriate hydrological circumstances that can increase the probability of high and low SF events occurring in future.

4.3 Lag Time Estimation

Three approaches were applied to estimate the Lt: (1) empirical formulas, (2) CCA, and (3) NGA. The results of both the second and third approaches were employed in the SF modelling process, particularly in the selection of the lag intervals between the input and the output variables of the AI-based models. The first approach was performed only to provide an initial approximation of the Lt.

The results of the CCA were applied in the first phase of the modelling process, and those of HGA were applied in the second phase of the modelling process. The results of HGA were also employed in deriving new empirical formulas to estimate the Lt between the RF upstream station and the SF downstream station, which are presented in this section.

4.3.1 Lag Time Estimation using the empirical formulas

The estimated Lt between the upstream and downstream stations using the aforementioned empirical formulas in Chapter 3 is shown in Table 4.1. The mean estimated values (mean of the 4 empirical formulas) of the Lt between the downstream station and WL at Ulu Yam, Batang Kali, Kerling and Ampang Pecah stations are 12.4, 11.58, 12.6 and 13.90 hr, respectively.

The results of Lt estimation demonstrate a remarkable difference depending on the empirical formula Similar variations in the estimated Lt values were identified throughout several earlier studies (Grimaldi et al., 2012). For example Sharifi and Hosseini (2011) found that the value of Lt estimated by the different empirical formulas can vary up to 500%.

Table 4.1: The estimated Lag time with the four empirical formulas

Station Name	Estimated Lt by Empirical formulas (hr)			
	Kirpich	**Johnstone**	**Carter**	**Viparelli**
Ulu Yam	18.07	12.71	8.11	10.69
Batang Kali	16.63	11.44	7.60	10.64
Kerling	18.04	11.74	8.10	12.47
Ampang Pecah	20.24	13.25	8.86	13.22

4.3.2 Lag Time Estimation Using the Correlation Coefficient Approach

The results of the CCA for the WL and RF stations are presented in Figure 4.12. The highest R values for the WL and RF stations are around 12 and 17 hr, respectively. R is generally not high, and it can be explained by the high complexity of the relation between the WL, RF, and Q and by the influence of other hydrological parameters on SF (Chang et al., 2014).

Although R is generally weak, it is useful in selecting the lag intervals between the input and output variables of the AI-based models. The estimated Lt using this approach was utilized to select the lag intervals between the input and output variables of the AI-based models in the first phase of the modelling process.

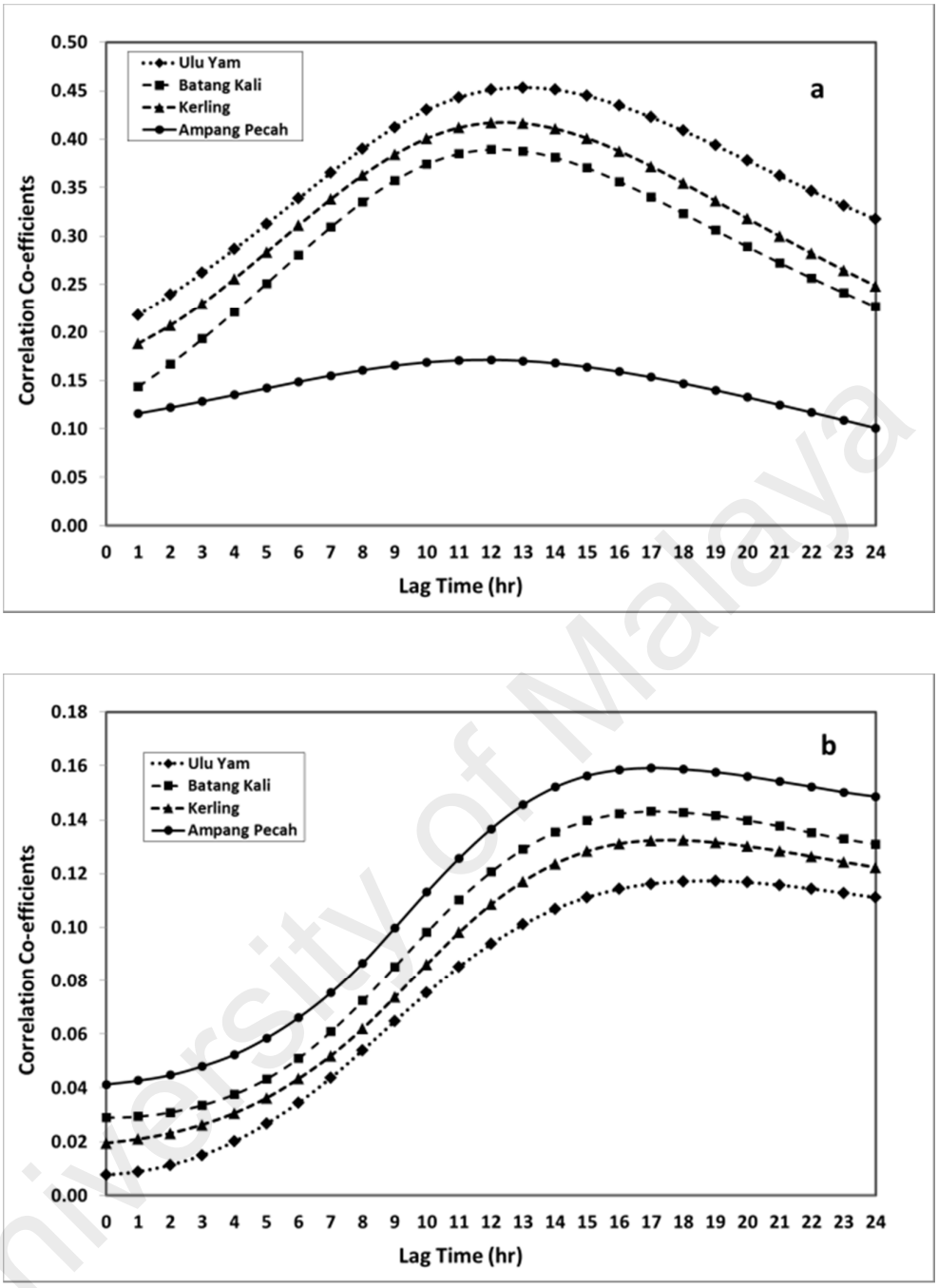


Figure 4.12: Correlation analysis between between hourly stream flow records in downstream station and the hourly records of upstream station in different time steps: (a) Water level stations and (b) Rainfall stations

4.3.3 Lag Time Estimation by the Hydrological Graphical Approach

The HGA was employed to estimate the Lt between the WL and RF upstream stations and the downstream station.

4.3.3.1 Lag time between Water Level Upstream Stations and the Downstream Station

A total of 134 WL–SF events were employed to estimate the Lt between four WL upstream station and the downstream SF stations by the NGA. Table 4.2 lists the estimated Lt of ten events of the four stations between the downstream SF and upstream WL stations. The estimated Lt of all the events (134) between the SF and WL stations is presented in Appendix C.

A basic statistical analysis of the results of the Lt between the SF and WL stations is presented in Table 4.3. It includes the mean, standard deviation (SD), coefficient of variation (CV), maximum and minimum.

The mean values of the Lt between the SF and WL stations at Ulu Yam, Batang Kali, Kerling and Ampang Pecah are 12.39, 12.38, 13.18 and 13.16 hr, respectively. The results suggest that the maximum estimated value of Lt that occurred at the Ampang Pecah station was 18 hr, whereas the minimum estimated value was 8 hr. Table 4.3 shows that the maximum Lt estimated values are 17, 17, 17 and 18 hr for the Ulu Yam, Batang Kali, Kerling and Ampang Pecah stations, whereas the minimum values are 9, 9, 9 and 8 hr for the same stations, respectively.

The SD of the estimated Lt for all the stations is similar, with the highest SD value (2.42) at Kerling and the lowest SD value (1.96) at Batang Kali. Likewise, the CV of the Lt for all the stations is similar, with the highest CV value (0.19) at Ulu Yam and Kerling and the lowest SV value (0.15) at Ampang Pecah.

Table 4.2: The estimated lag time of ten events between the downstream stream flow station and water level upstream stations

St. Ulu Yam			St. Batang Kali			St. Kerling			St. Ampang Pecah		
Q	**WL**	**Lt**	**Q**	**WL**	**Lt**	**Q**	**WL**	**Lt**	**Q**	**WL**	**Lt**
m^3/s	m	hr	m^3/s	m	hr	m^3/s	m	hr	m^3/s	m	hr
102.6	35.3	14.0	102.6	33.8	15.0	102.6	45.3	15.0	102.6	50.4	14.0
183.1	36.1	17.0	183.1	34.5	13.0	183.1	44.8	10.0	203.5	50.4	16.0
183.1	36.1	16.0	183.1	34.5	12.0	183.1	44.8	11.0	198.9	50.1	14.0
203.5	34.4	13.0	203.5	34.4	16.0	203.5	45.4	16.0	137.1	50.1	13.0
113.6	33.7	14.0	113.6	33.0	10.0	113.6	45.8	17.0	110.9	50.0	15.0
141.8	34.4	9.0	136.8	33.4	10.0	136.8	45.2	14.0	97.6	50.0	11.0
198.9	34.9	13.0	141.8	33.6	12.0	137.1	45.5	12.0	184.7	50.1	11.0
137.1	34.8	13.0	198.9	33.6	14.0	110.9	44.5	11.0	189.3	50.1	13.0
110.9	33.4	11.0	137.1	33.6	10.0	97.6	44.6	12.0	77.9	50.2	12.0
97.6	33.8	10.0	110.9	33.0	11.0	180.7	44.4	15.0	168.7	50.2	15.0

Table 4.3: Basic statistical analysis of the hydrological graphical approach results of the lag time between the downstream station and water level upstream stations

Station	Ulu Yam	Batang Kali	Kerling	Ampang Pecah
Mean	12.39	12.38	13.18	13.16
SD	2.38	1.96	2.42	2.03
CV	0.19	0.16	0.18	0.15
Mean + SD	14.77	14.35	15.60	15.19
Mean - SD	10.01	10.42	10.76	11.12
Maximum	17.00	17.00	17.00	18.00
Minimum	9.00	9.00	9.00	8.00

4.3.3.2 Lag time between Rainfall Upstream Stations and the Downstream Station

A total of 100 of RF-SF events were applied to estimate the Lt between the downstream SF station and four RF upstream stations by NGA. Table 4.4 presents the estimated Lt of ten events between the downstream SF and four upstream RF stations. The estimated Lt of all the events (100) between the SF and RF stations is presented in Appendix C.

A basic statistical analysis of the hydrological estimation of the Lt between the SF and RF stations is presented in Table 4.5. It includes the mean, standard deviation (SD), coefficient of variation (CV), maximum and minimum.

The mean values of the Lt between the downstream station and RF stations, Ulu Yam, Batang Kali, Kerling and Ampang Pecah are 14.44, 14.70, 15.05 and 14.74 hr, respectively. Table 4.5 shows that the maximum Lt estimated values are 20, 20, 21 and 21 hr for the Ulu Yam, Batang Kali, Kerling and Ampang Pecah stations, whereas the minimum values are 10, 10, 9 and 9 hr for the same stations, respectively.

The standard deviation (SD) of the estimated Lt for all the stations is similar, as the highest SD value for Ampang Pecah is 2.83 and the lowest for Kerling is 2.5. The coefficient of variation (CV) for Lt is also very similar for all stations, with the highest CV value at Ampang Pecah and Ulu Yam stations (0.19) and at Kerling (0.17).

Table 4.4: The estimated lag time of ten events between the downstream stream flow station and rainfall upstream stations

St. Ulu Yam			St. Batang Kali			St. Kerling			St. Ampang Pecah		
Q	**RF**	**Lt**	**Q**	**RF**	**Lt**	**Q**	**RF**	**Lt**	**Q**	**RF**	**Lt**
m^3/s	mm/hr	hr	m^3/s	mm/hr	hr	m^3/s	mm/hr	hr	m^3/s	mm/hr	hr
102.6	17.5	16.0	102.6	20.5	17.0	136.8	13.4	15.0	167.0	30.1	16.0
171.6	19.9	14.0	183.1	18.5	19.0	141.8	4.1	16.0	113.6	13.1	17.0
136.8	10.3	17.0	167.0	9.5	15.0	103.2	43.2	9.0	171.6	42.4	12.0
141.8	37.3	15.0	136.8	9.4	15.0	198.1	40.0	14.0	136.8	11.4	15.0
79.5	26.6	14.0	141.8	11.7	15.0	198.8	16.0	12.0	141.8	14.3	16.0
103.2	19.2	10.0	79.5	36.6	14.0	137.1	14.0	13.0	198.1	3.3	16.0
198.1	7.6	16.0	198.1	5.8	16.0	77.9	24.0	16.0	198.8	22.3	12.0
198.8	49.3	12.0	198.8	53.9	12.0	168.7	7.0	15.0	184.3	24.8	10.0
184.3	23.9	10.0	184.3	32.5	10.0	148.5	7.0	21.0	134.6	20.0	10.0
137.1	6.3	14.0	137.1	10.5	14.0	74.3	10.0	15.0	170.9	25.7	9.0

Table 4.5: Basic statistical analysis of the estimated Lag time between downstream station and rainfall upstream stations.

Station	**Ulu Yam**	**Batang Kali**	**Kerling**	**Ampang Pecah**
Mean	14.44	14.70	15.05	14.74
SD	2.76	2.58	2.50	2.83
CV	0.19	0.18	0.17	0.19
Maximum	20.00	20.00	21.00	21.00
Minimum	10.00	10.00	9.00	9.00
Mean + SD	17.21	17.28	17.55	17.57
Mean - SD	11.68	12.12	12.55	11.91

4.3.4 New Empirical Formulas to Estimate the Lag Time

The input (independent) variables of the empirical formulas are Rf_p, Rf_{48}, Q_p, and Q_{48}, and the output (dependent) variable is the Lt. The linear and nonlinear empirical formulas are directly derived using the estimated Lt by the HGA approach from 100 high SF-RF events.

Based on the estimated Lt between the upstream RF and downstream SF stations by the HGA approach from 100 high SF-RF events, the four hydrological variables were calculated to construct combination patterns of variables for every event to employ these combinations in deriving new empirical formulas.

The results of the combinations of ten events for Rf_p, Rf_{48}, Q_p, and Q_{48} and the estimated Lt between the Ulu Yam and Rantau Panjang stations are presented in Table 4.6. The results of the combinations of Rf_p, Rf_{48}, Q_p, and Q_{48} and the estimated Lt of all the events between the downstream SF and four RF upstream stations are presented in Appendix C.

Table 4.6: Results of Ten events: Peak rainfall intensity, previous 48 hour rainfall, peak stream flow, previous 48 hour stream flow and the Lt between the Ulu Yam station and Rantau Panjang station.

Q_p	Q_{48}	RF_p	RF_{48}	Lt
m^3/s	m^3/s	mm/hr	mm/hr	hr
102.6	58.3	17.5	36.0	16.0
171.6	93.3	19.9	27.0	14.0
136.8	73.4	10.3	43.8	17.0
141.8	110.8	37.3	100.0	15.0
198.1	123.8	7.6	79.8	16.0
198.8	99.9	49.3	92.0	12.0
184.3	146.0	23.9	131.1	10.0
137.1	84.8	6.3	16.0	14.0
110.2	76.9	18.2	44.9	12.0
120.7	92.1	14.9	27.0	13.0

4.3.4.1 Linear Empirical Formula

A new linear empirical formula was derived to estimate the Lt between the RF upstream and downstream stations (Equation 4.1). The independent variables of the formula are Rf_p, Rf_{48}, Q_p, and Q_{48}, and the dependent variable is Lt. A moderate agreement between the observed lag time (Lt_o) and estimated lag time (Lt_e), as shown in Figure 4.13

$$Lt = 15.95+0.0221*Q_p-0.024*Q_{48}-0.067*Rf_p-0.020*Rf_{48} \quad (4.1)$$

Table 4.7 shows the p-value and R between the hydrological variables and the Lt_e by the linear formula. The R value between Lt_o and Lt_e is 0.5194. The significance levels (p-values) of Q_p, Q_{48}, Rf_p and Rf_{48} are 0.0574, 0.0002, 0.00 and 0.00, respectively. It reveals the presence of a relationship between all the hydrological variables and Lt, excluding Q.

The regression analysis results between Lt_e and Q_p, Q_{48}, Rf_p, and Rf_{48} are −0.1957, −0.3678, −0.7944 and −0.6179, respectively.

Figure 4.14 shows the correlation between the hydrological variables and Lt_e. The regression analysis results between Lt_e and the hydrological variables signify that Lt_e is strongly inversely proportional to Rf_p and Rf_{48}, whereas it is moderately inversely proportional to Q_{48}. Based on the results, Lt_e is directly proportional to Q_p through a weak–strength relationship.

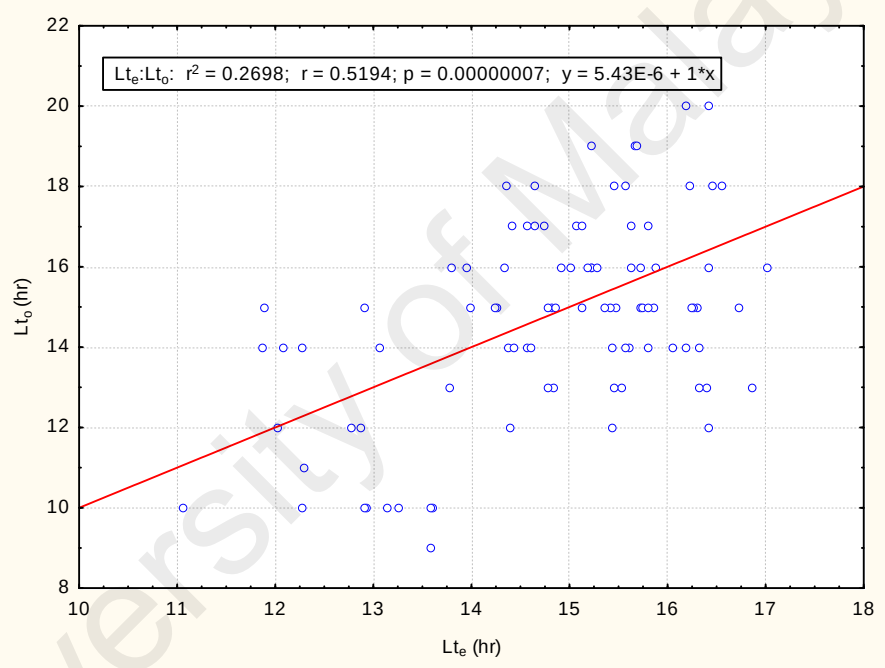
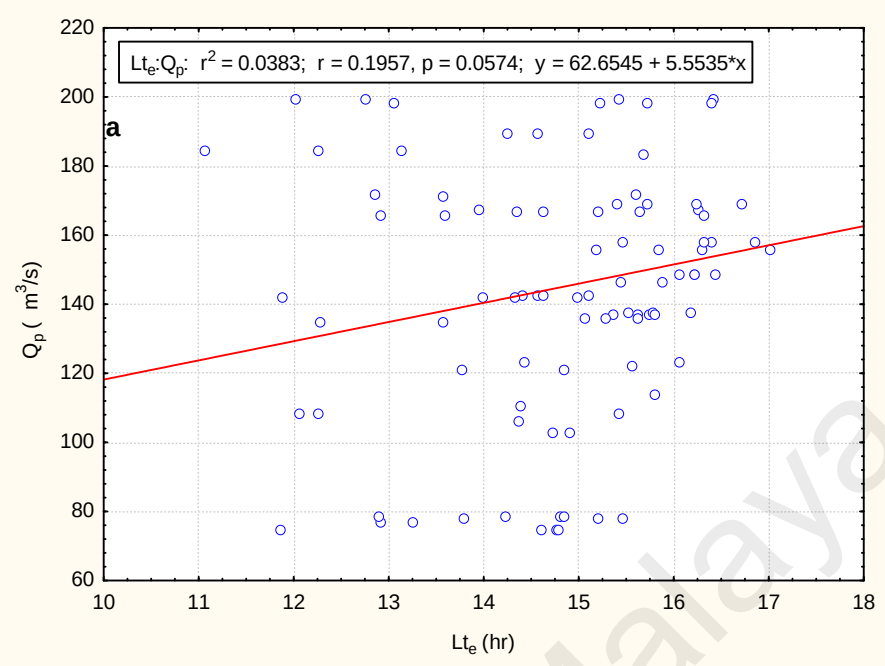
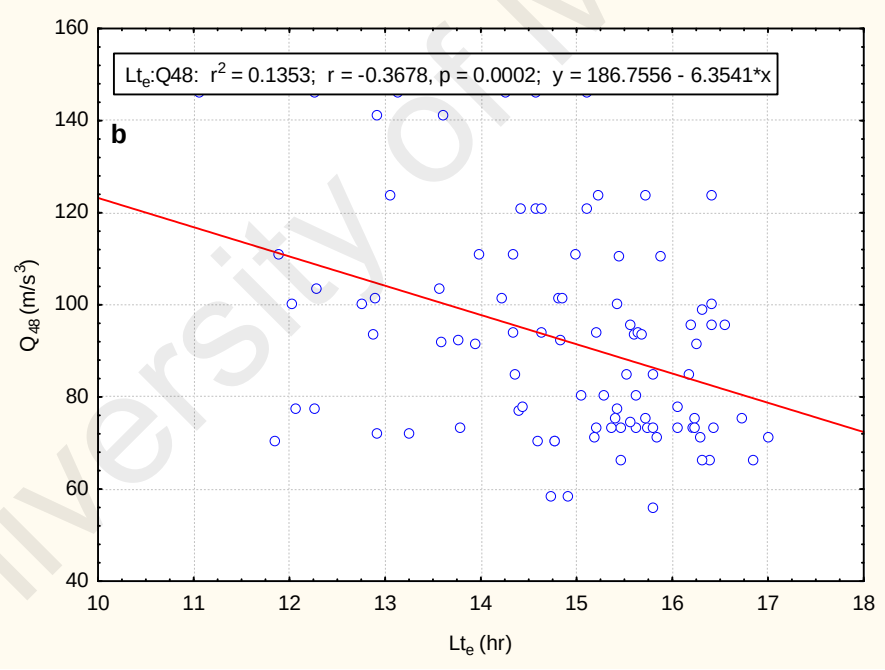
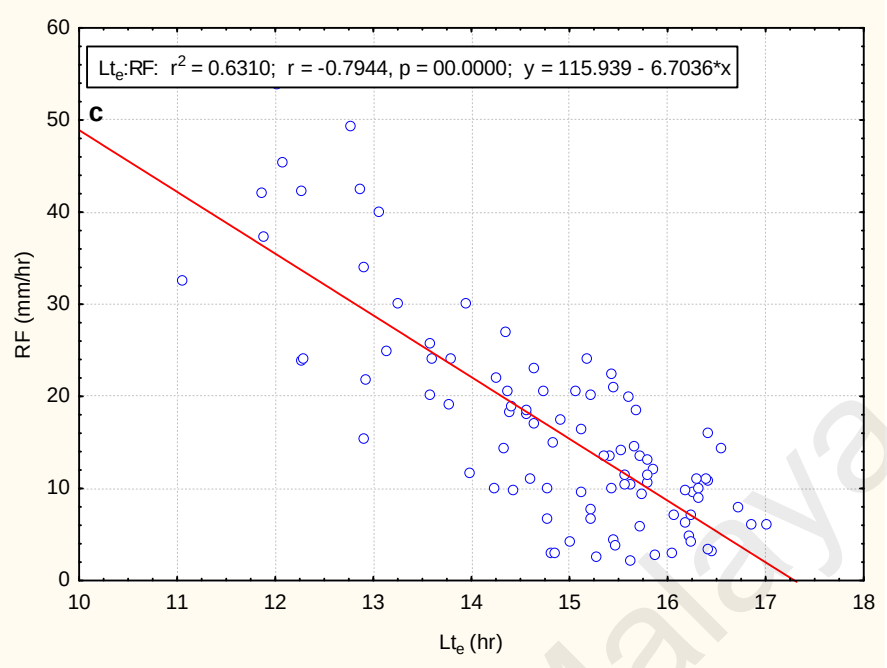


Figure 4.13: Correlation between the observed lag time and the estimated lag time by linear equation







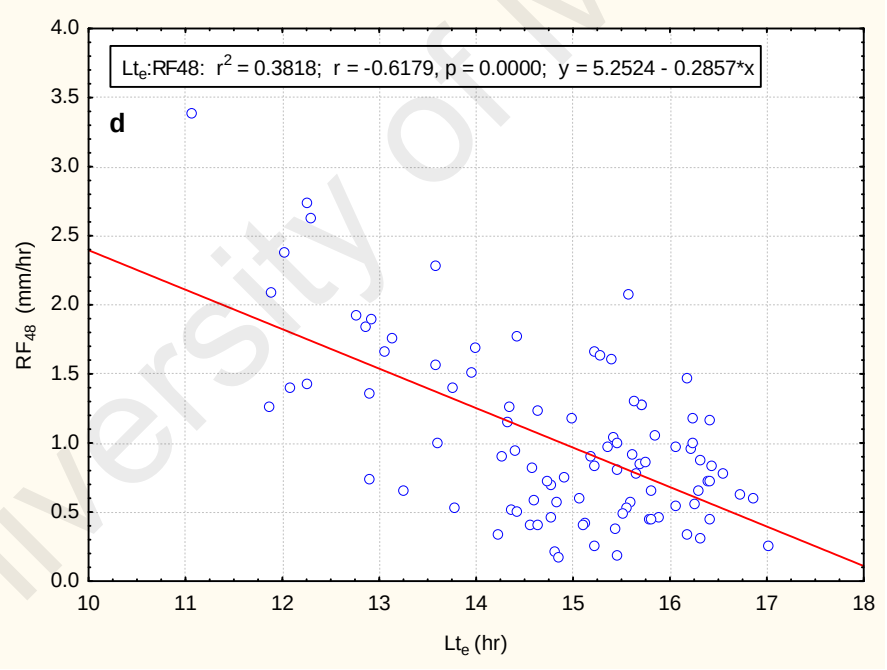


Figure 4.14: Hydrological variables versus estimated lag time by the linear equation: (a) peak stream flow, (b) previous 48 hour stream flow, (d) peak rainfall intensity and (c) previous 48 hour rainfall.

4.3.4.2 Nonlinear Empirical Formula

A new nonlinear empirical formula was derived to estimate the Lt between the upstream and downstream stations (Equation 4.2). Several nonlinear formulas were derived to estimate Lt, and a polynomial to-the-third-degree equation proved the highest estimation performance. The R value between Lt_o and Lt_e is 0.6306. The independent variables of the formula are Rf_p, Rf_{48}, Q_p and Q_{48} and the dependent variable is Lt. A very good agreement was found between the Lt_o and Lt_e as shown in Figure 4.15.

$$Lt=24.08+0.26*Q_p-.0019*Q_p^2+0.4698E^{-5}*Q_p^3-0.73*Q_{48}+0.0082*"Q_{48}^2-0.2908E^{-4}*Q_{48}^3-0.19*Rf_p+0.0045*Rf_p^2-0.4249E^{-4}*Rf_p^3+7.37*Rf_{48}-5.4997\ *Rf_{48}^2+0.9818*Rf_{48}^3 \quad (4.2)$$

Table 4.7 shows the p-value and R between the hydrological variables and the Lt_e by the polynomial equation. The R between Lt_o and Lt_e was 0.6306. The significance levels (p-values) of Q_p, Q_{48}, Rf_p and Rf_{48} are 0.1187, 0.0028, 0.00 and 0.00, respectively. It reveals the existence of relationship between all the hydrological variables and Lt, excluding Q_p. The regression analysis results between Lt_e and Q_p, Q_{48}, Rf_p and Rf_{48} are 0.1612, −0.3029, −0.6543 and −0.5091 respectively.

Figure 4.16 shows the correlation between the hydrological variables and Lt_p. The results of the regression analysis between Lt_e and the hydrological variables indicate that Lt_e is strongly inversely proportional to Rf_p and Rf_{48}, whereas it is moderately inversely proportional to Q_{48}. Based on the results, Lt_e is directly proportional to Q_p through a weak-strength relationship.

Table 4.7: p-value and correlation coefficient between the hydrological variables and the estimated lag time by the linear and polynomial formula.

Variable	Polynomial equation		Linear equation	
	p-level	r	p-level	r
Q_p	0.1187	0.1612	0.0574	0.1957
Q_{48}	0.0028	-0.3029	0.0002	-0.3678
RF_p	0.00	-0.6543	0.00	-0.7944
RF_{48}	0.00	-0.5091	0.00	-0.6179

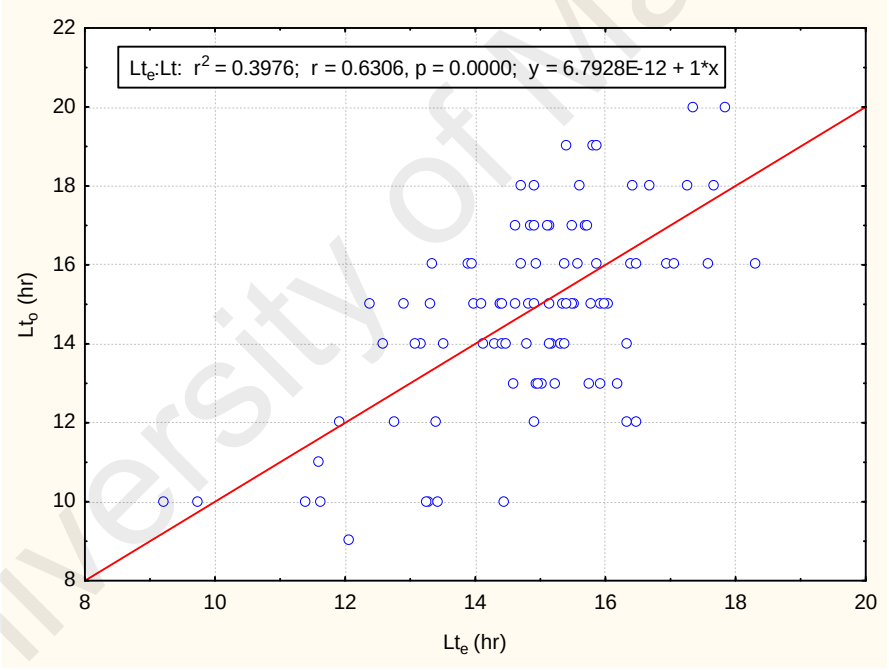
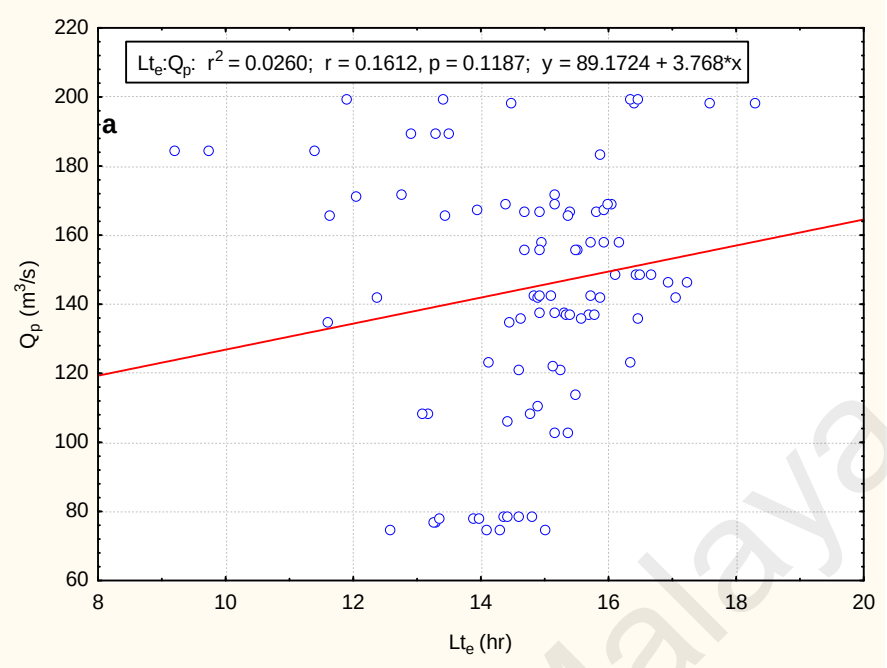
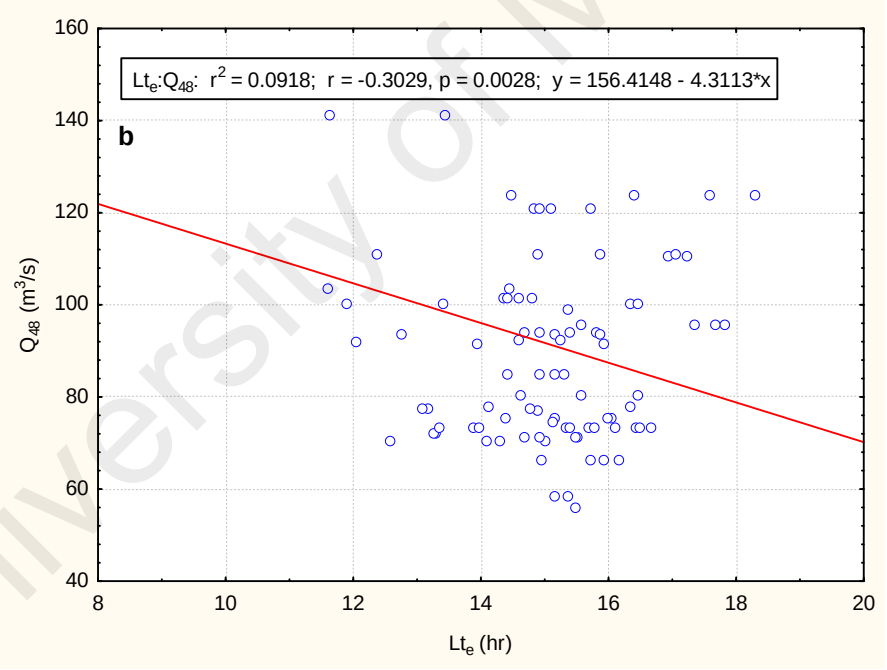


Figure 4.15: Correlation between the observed lag time and the estimated lag time by the polynomial equation





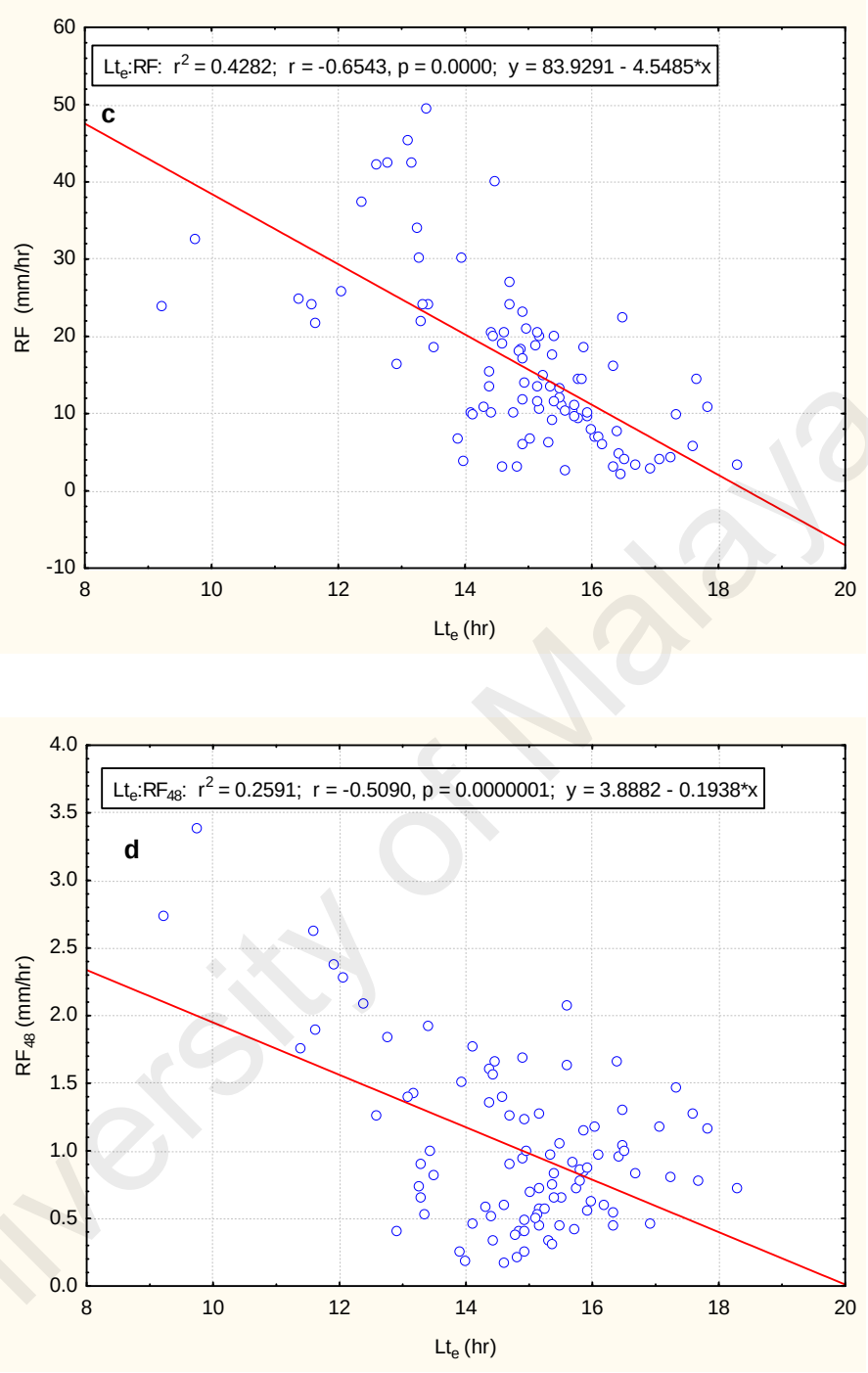


Figure 4.16: Hydrological variables versus the estimated lag time by the polynomial equation: (a) peak stream flow (b) previous 48 hour stream flow, (c) peak rainfall intensity, and (d) previous 48 hour rainfall.

4.3.5 General Discussion about the Lag Time Estimation

Three approaches were applied to estimate the Lt: the empirical formulas, CCA, and NGA. The results of both the second and third approaches were employed in the SF modelling process, particularly in the selection of the lag intervals between the input and output variables of AI-based models. The first approach was performed only to provide an initial approximation of the Lt.

The mean values of estimated Lt obtained with the four empirical formulas between the downstream station and the WL stations at Ulu Yam, Batang Kali, Kerling and Ampang Pecah stations are 12.4, 11.58, 12.6 and 13.90 hr respectively. The R between the results of the HGA and empirical formulas methods was 0.72, signifying a good accord among results. However, the R between the Lt estimated in the two approaches is not an indicator of which method provides the more accurate estimation.

The Lt estimation results of 100 RF-SF events by the HGA were applied in the derivation of two empirical formulas to estimate the Lt between upstream and downstream stations. The input (independent) variables of the empirical formulas are Rf_p, Rf_{48}, Q_p, and Q_{48}, and the output variable (dependent) is the Lt between the upstream and downstream stations. The first empirical formula is a linear equation, and the second is a polynomial to the third degree with R between the observed and estimated Lt 0.5194 and 0.6306, respectively.

The derived empirical formulas significantly simplify the Lt estimation process by a quick and easy approach directly based on the RF and SF records without the necessity of identifying the full description of all the parameters that affect the Lt. The empirical formulas are applicable only for the Selangor River basin, but they can be modified for other humid tropical river basins based on the results of the Lt estimation by the NGA, which is applicable for all humid tropical rivers.

The HGA and derived empirical formulas have the potential to be used in many future hydrological applications, especially those related to the surface water hydrology and river basin integrated management. The results of the CCA were applied in the first phase of the modelling process, and those of HGA were applied in the second phase of the modelling process.

4.4 AI-based Models to Predict Real-time Hourly Stream Flow

The modelling process was performed in two phases. First, the results of the Lt estimated by CCA were applied to select the lag intervals between the input and output variables of the AI-based models, whereas the results of the HGA were then applied to in the second phase of the modelling process. Three scenarios in selecting the input variables of the models were considered. Two input vectors were applied for these three scenarios. Given six input vectors, every one of them includes different combinations of input and output variables.

In the first phase of the modelling process, six models with different combinations of input variables were trained and developed by four AI techniques—MLP, RBF, GRNN, and SVM—resulting in the development of 24 AI-based models to predict the Q. In the second phase of the modelling process, only two models, those that achieved the highest R among the six models of the first phase, were selected for the second phase, resulting in the development of eight AI-based models to predict the Q. The total number of developed AI-based models throughout the two modelling phases is 32.

The performances of the developed models were assessed based on training, testing and overall data set performances. The best-fitting model for predicting the Q is determined based on the performance evaluation of the testing data sets. The performance evaluation criteria are mentioned in Section 3.6.8.

4.4.1 AI-based Models: First Phase of the Modelling Process

In the first phase of the modelling process, six AI-based models with different combinations of input variables were selected to predict the Q in the Rantau Panjang station. The six models were trained and developed by four AI techniques—MLP, RBF, GRNN, and SVM—resulting in the development of 24 AI-based models to predict the Q, as shown in Table 4.8.

The lag intervals between the input and output variables of the AI-based models were selected based on the results of the CCA to estimate the Lt between the upstream and downstream stations. The different combinations of the input and output variables of six AI-based models are shown in Table 4.9. Figure 4.17 shows the lag intervals between the input and output variables for the AI-based models of the first modelling phase.

Table 4.10 shows a group of 15 modelling cases of M6 as example of 8872 modelling cases. A larger group of modelling cases of M6 for three days is presented in Appendix B.

Table 4.8: AI-based models of the first modelling phase

Modelling technique	**Model No.**					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
MLP	MLP-M1	MLP-M2	MLP-M3	MLP-M4	MLP-M5	MLP-M6
RBF	RBF-M1	RBF-M2	RBF-M3	RBF-M4	RBF-M5	RBF-M6
GRNN	GRNN-M1	GRNN-M2	GRNN-M3	GRNN-M4	GRNN-M5	GRNN-M6
SVM	SVM-M1	SVM-M2	SVM-M3	SVM-M4	SVM-M5	SVM-M6

Table 4.9: Input and output variables of the AI-based models

Model	Inputs	Output	No. input Variables
M1	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+17)}$	5
M2	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Q_{(t)}$	$Q_{(t+17)}$	5
M3	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Q_{(t)}$	$Q_{(t+12)}$	5
M4	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Q_{(t)}$	$Q_{(t+12)}$	5
M5	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-5)}$, $Rf_{b(t-5)}$, $Rf_{k(t-5)}$, $Rf_{a(t-5)}$, $Q_{(t)}$	$Q_{(t+12)}$	9
M6	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-5)}$, $Rf_{b(t-5)}$, $Rf_{k(t-5)}$, $Rf_{a(t-5)}$, $Q_{(t)}$	$Q_{(t+12)}$	9

Table 4.10: Group of modelling cases of M6

date	time	$Q_{(t)}$	$Wl_{u(t)}$	$Wl_{b(t)}$	$Wl_{k(t)}$	$Wl_{a(t)}$	$Rf_{u(t-5)}$	$Rf_{b(t-5)}$	$Rf_{k(t-5)}$	$Rf_{a(t-5)}$	$Q_{(t+12)}$
10/01/2011	01:00	32.31	32.55	32.61	44.17	50.03	0.00	0.00	0.00	0.00	30.70
10/01/2011	02:00	32.43	32.55	32.60	44.17	50.03	0.00	0.00	0.00	0.00	30.15
10/01/2011	03:00	32.55	32.55	32.61	44.17	50.03	0.00	0.00	0.00	0.00	29.74
10/01/2011	04:00	32.77	32.55	32.61	44.17	50.03	0.00	0.00	0.00	0.00	29.73
10/01/2011	05:00	32.96	32.55	32.62	44.17	50.03	0.00	0.00	0.00	0.00	29.91
10/01/2011	06:00	33.08	32.55	32.60	44.17	50.03	0.00	0.07	0.00	0.20	30.11
10/01/2011	07:00	33.08	32.55	32.62	44.18	50.03	0.00	0.17	0.00	0.50	30.44
10/01/2011	08:00	32.90	32.55	32.62	44.18	50.03	0.00	0.27	0.20	0.80	31.71
10/01/2011	09:00	32.46	32.55	32.63	44.17	50.04	0.00	0.30	0.60	0.90	33.66
10/01/2011	10:00	32.11	32.55	32.62	44.16	50.04	0.00	0.30	1.00	0.90	36.02
10/01/2011	11:00	31.60	32.54	32.60	44.15	50.04	0.00	0.30	1.20	0.90	38.64
10/01/2011	12:00	31.14	32.54	32.58	44.14	50.03	0.00	0.30	1.20	0.90	40.92
10/01/2011	13:00	30.70	32.55	32.57	44.13	50.03	0.00	0.30	1.20	0.90	42.83
10/01/2011	14:00	30.15	32.55	32.58	44.12	50.03	0.00	0.30	1.20	0.90	44.65
10/01/2011	15:00	29.74	32.55	32.59	44.13	50.03	0.00	0.30	1.20	0.90	46.76

Lag time	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15	t+16	t+17
Rf_u	•																	
Rf_b	•																	
Rf_k	•																	
Rf_a	•																	
Wl_u																		
Wl_b																		
Wl_k																		
Wl_a																		
Sf	•																	▲

a

Lag time	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15	t+16	t+17
Rf_u	•	•	•															
Rf_b	•	•	•															
Rf_k	•	•	•															
Rf_a	•	•	•															
Wl_u																		
Wl_b																		
Wl_k																		
Wl_a																		
Sf	•																	▲

b

Lag time	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15	t+16	t+17
Rf_u																		
Rf_b																		
Rf_k																		
Rf_a																		
Wl_u	•																	
Wl_b	•																	
Wl_k	•																	
Wl_a	•																	
Sf	•												▲					

c

Lag time	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15	t+16	t+17
Rf_u																		
Rf_b																		
Rf_k																		
Rf_a																		
Wl_u	●	●	●															
Wl_b	●	●	●															
Wl_k	●	●	●															
Wl_a	●	●	●															
Sf	●												▲					

d

Lag time	t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
Rf_u	●																	
Rf_b	●																	
Rf_k	●																	
Rf_a	●																	
Wl_u						●												
Wl_b						●												
Wl_k						●												
Wl_a						●												
Sf						●												▲

e

Lag time	t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
Rf_u	●	●	●															
Rf_b	●	●	●															
Rf_k	●	●	●															
Rf_a	●	●	●															
Wl_u						●	●	●										
Wl_b						●	●	●										
Wl_k						●	●	●										
Wl_a						●	●	●										
Sf						●												▲

f

where ● is the input variables and ▲ is the output variable

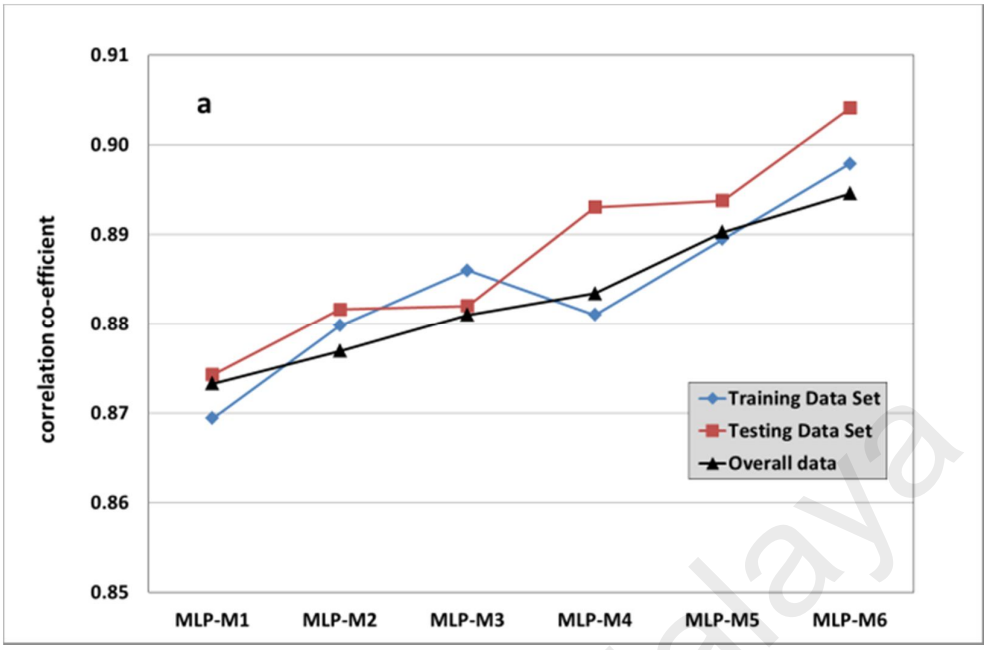
Figure 4.17: Lag intervals between the input and output variables of the AI-models: a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5 and f) Model 6

4.4.1.1 MLP-based Models

Six models with different combination of input variables were trained and developed by MLP to predict Q. The developed models' performance was assessed based on the training and testing data sets, as well as the overall performance of the data sets. The best fit model to predict Q is thus determined according to the performance of the testing data sets. Table 4.11 present the performance evaluation results as denoted by the R and MAE of the MLP-based models.

Figure 4.18 compares the performance levels of the six MLP models and determines that the best fit model is MLP-M6. Over MLP models, this model displays the highest *R* values (0.989 and 0.904) and the lowest MAE (10.83 and 10.922) in both the training and testing data sets, respectively.

Figure 4.19 shows the correlation between the observed and predicted Q in MLP-M6 model giving training and testing data set. The observed and predicted Q of the training and testing data sets, seem to be in good accord with R^2 0.806 and 0.817, respectively. In Figure 4.20, a comparison between the observed and predicted Q by MLP-M6 for the period of September 2013 can be seen. Acceptable agreement with small error between the observed and predicted Q is observed. The results verified the high performance of the model. The full records of the observed and predicted Q by MLP-M6 for September 2013 are presented in Appendix D.



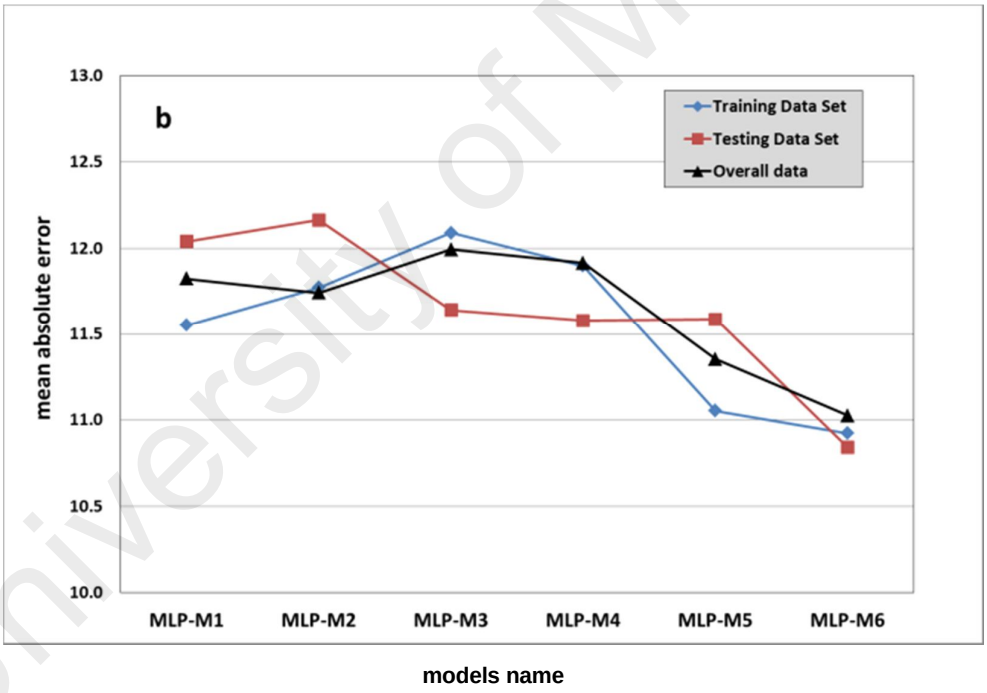


Figure 4.18: Performance values of MLP-based models: (a) correlation coefficient and (b) mean absolute error

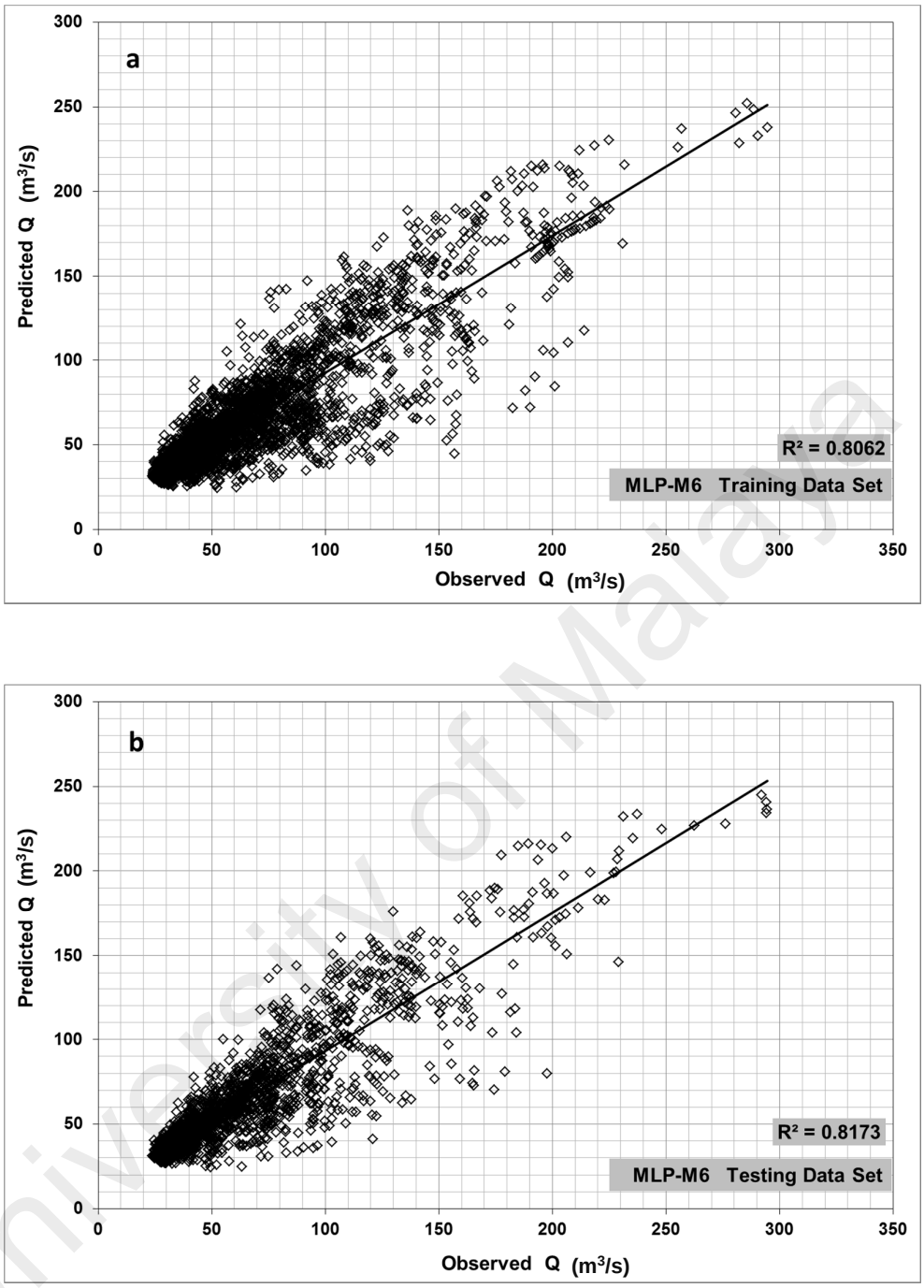


Figure 4.19: Correlation between the observed and predicted hourly stream flow by MLP-M6 model: (a) Training data set and (b) Testing data set

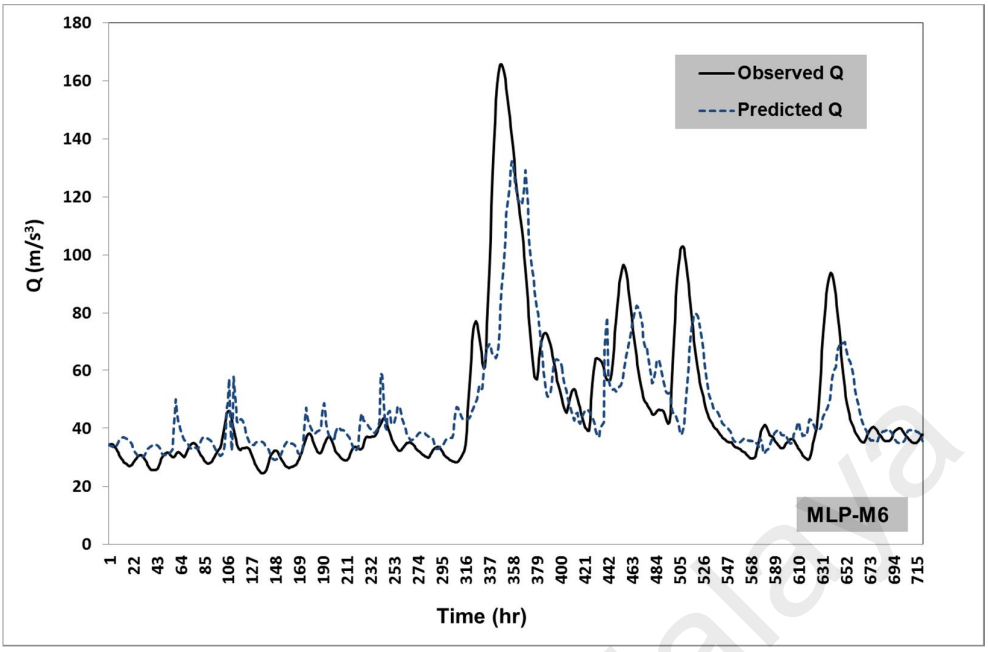


Figure 4.20: Comparison between the observed and predicted hourly stream flow by the MLP-M6 model for the period of September 2013

Table 4.11: Performance values of MLP-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
MLP-M1	0.869	11.550	0.874	12.039	0.873	11.823
MLP-M2	0.880	11.771	0.882	12.163	0.877	11.741
MLP-M3	0.886	12.088	0.882	11.640	0.881	11.991
MLP-M4	0.881	11.897	0.893	11.578	0.883	11.913
MLP-M5	0.889	11.052	0.894	11.586	0.890	11.352
MLP-M6	0.898	10.922	0.904	10.839	0.895	11.025

4.4.1.2 RBF-based Models

Six models with different input variable combinations were trained and developed using RBF to predict Q. The performance of the developed models was evaluated via the training data set, testing data set and overall data performance. The results of the performance evaluation criteria (i.e. R and MAE) of the RBF-based models are presented in Table 4.12.

A comparison of the performance evaluation for the six RBF-based models is provided in Figure 4.21. This figure indicates that the best fit RBF model is RBF-M6 with the highest values of R and lowest value of MAE for the training and testing data sets. The R between the observed and predicted Q by the RBF-M6 model is 0.987 and 0.965, while MAE is 3.37 and 6.141 for the training and testing data sets, respectively.

Figure 4.22 shows the correlation between the observed and predicted Q by RBF-M6 model, (a) training data set and (b) Testing data set. The observed and predicted Q of the training and testing data sets seem to be in good accord with R^2 0.975 and 0.930 respectively. In Figure 4.23, a comparison between the observed and predicted Q by RBF-M6 for September 2013 can be seen. Good agreement with small error between the observed and predicted Q was is evident. The results verified the high performance of the model. The full records of the observed and predicted Q by RBF-M6 for September 2013 are presented in Appendix D.

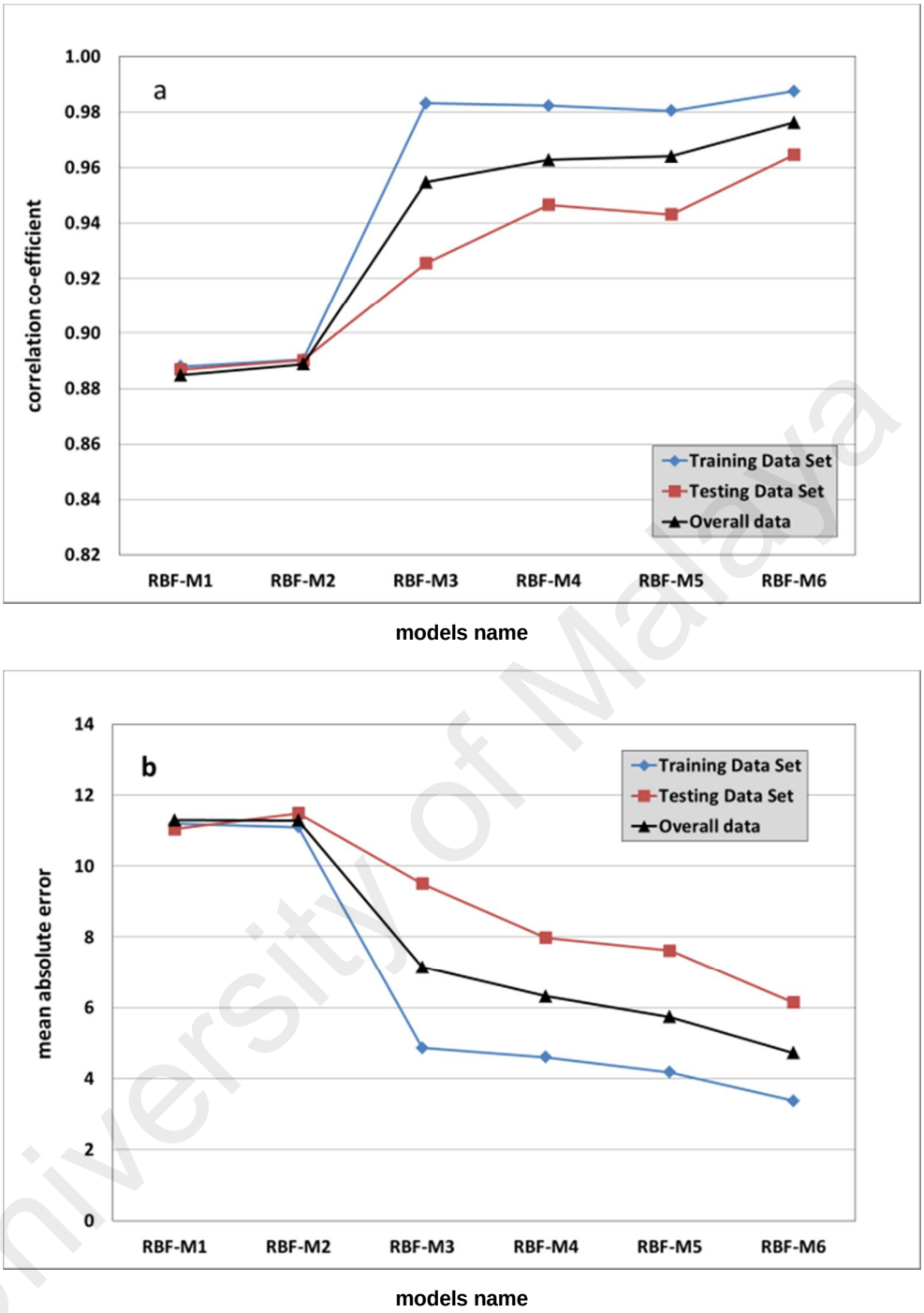
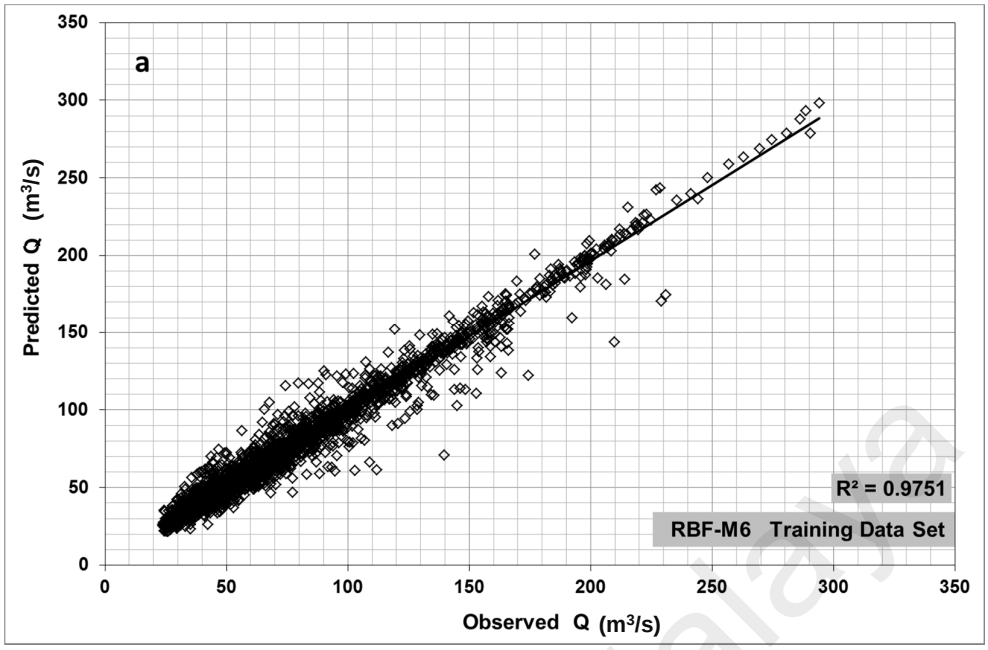


Figure 4.21: Performance values of RBF-based models: (a) correlation coefficient and (b) mean absolute error



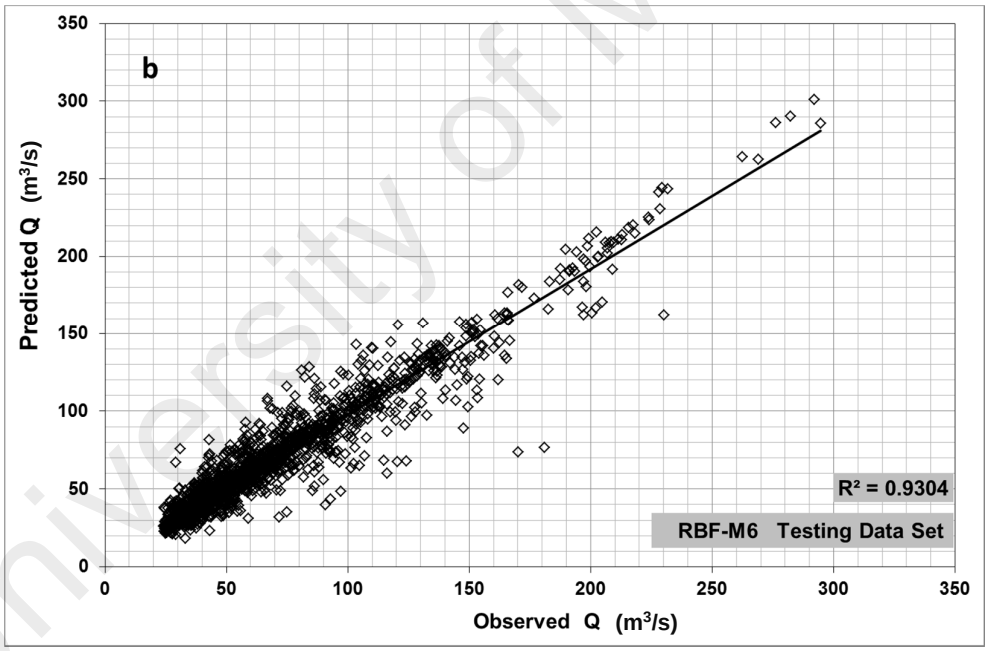


Figure 4.22: Correlation between the observed and predicted hourly stream flow by RBF-M6 model: (a) training data set and (b) Testing data set

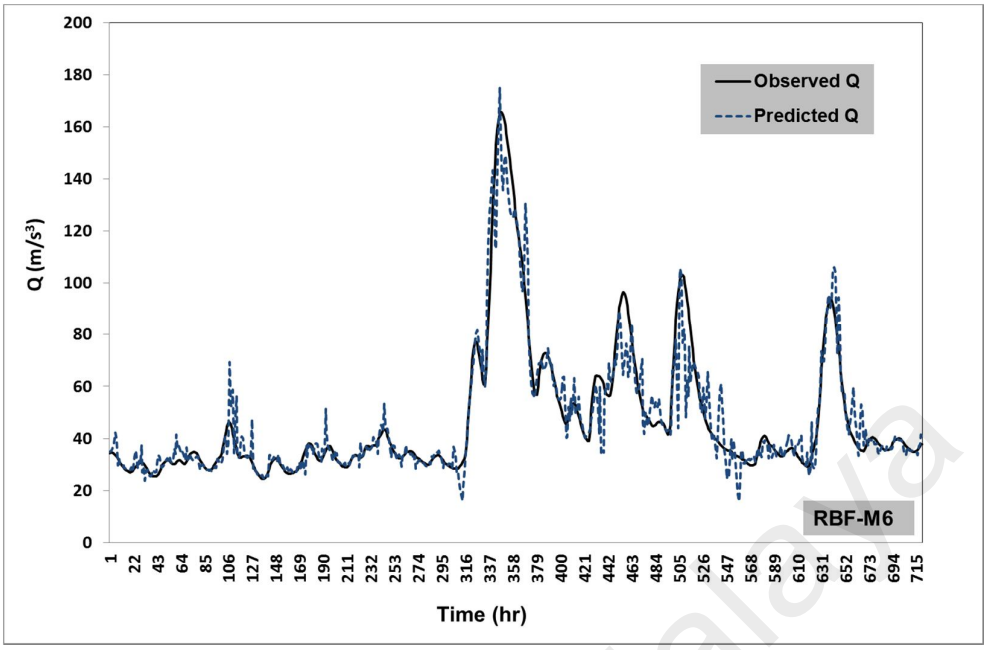


Figure 4.23: Comparison between the observed and the predicted hourly stream flow by the RBF-M6 model for the period of September 2013

Table 4.12: Performance values of RBF-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
RBF-M1	0.888	11.193	0.887	11.046	0.885	11.292
RBF-M2	0.890	11.099	0.890	11.488	0.889	11.287
RBF-M3	0.983	4.864	0.925	9.513	0.955	7.147
RBF-M4	0.982	4.604	0.947	7.984	0.963	6.316
RBF-M5	0.980	4.176	0.943	7.623	0.964	5.735
RBF-M6	0.987	3.370	0.965	6.141	0.976	4.720

4.4.1.3 GRNN-based Models

Six models with different input variable combinations were trained and developed using GRNN to predict Q. The performance of the developed models was evaluated via the training data set, testing data set and overall data performance. Table 4.13 provides the results of the performance evaluation criteria (i.e. R and MAE) of the GRNN-based models.

A comparison of the performance evaluation for the six GRNN-based models can be seen in Figure 4.24. Evidently, the best fit GRNN model is GRNN-M6 with the highest *R* values and lowest MAE value for the training and testing data sets. The *R* between the observed and predicted Q by the GRNN-M6 model is 0.99 and 0.965, while MAE is 2.634 and 5.24 for the training and testing data sets, respectively.

Figure 4.25 presents the correlation between the observed and predicted Q by the GRNN-M6 model, (a) training data set, and (b) testing data set. The observed and predicted Q of the training and testing data sets seem to be in good accord with R^2 0.9796 and 0.931, respectively. In Figure 4.26, a comparison between the observed and predicted Q by GRNN-M6 for September 2013 can be seen. There is good agreement with small error between the observed and predicted Q. The results verified the high performance of the model. The full records of the observed and predicted Q by GRNN-M6 for September 2013 are presented in Appendix D.

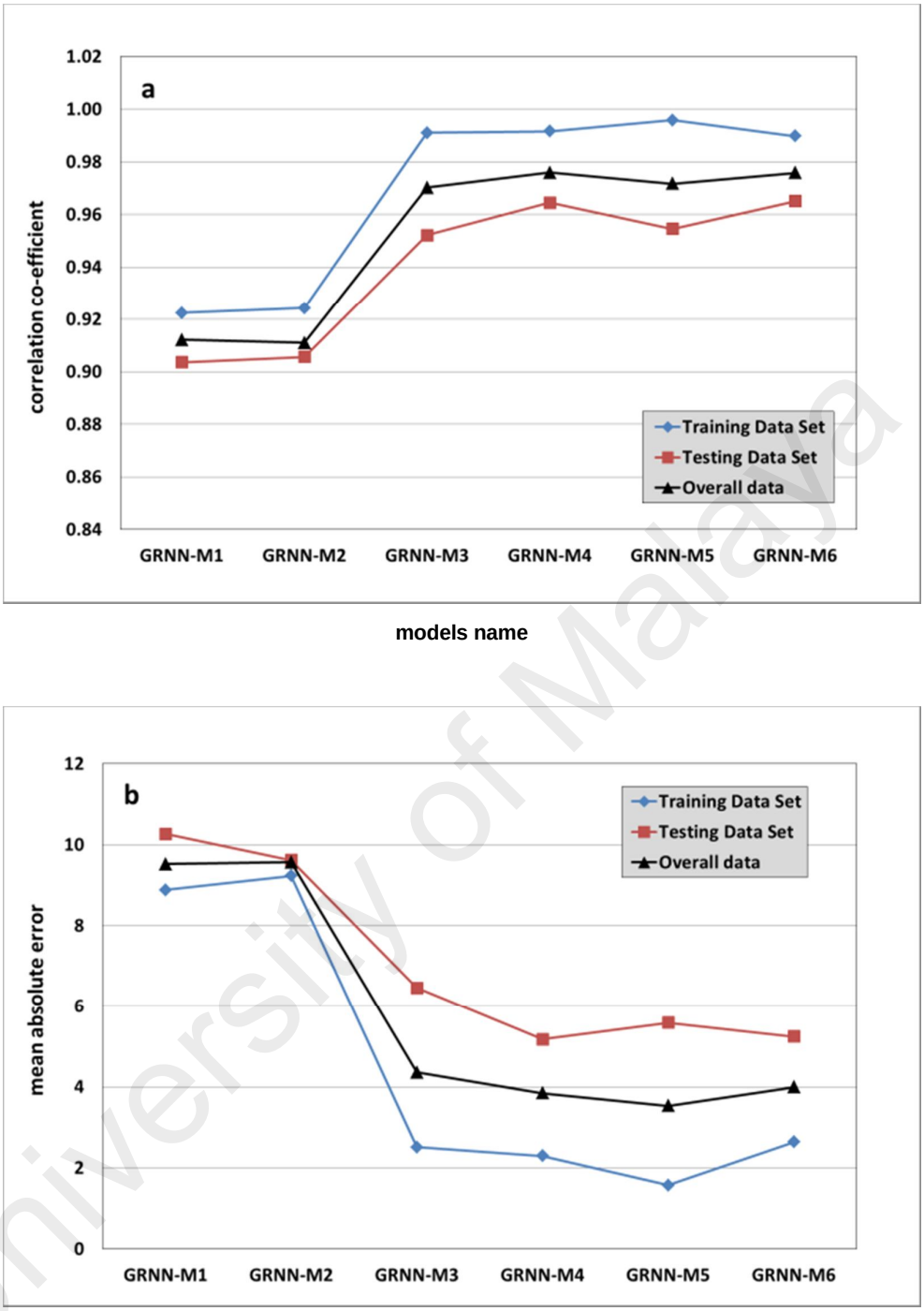
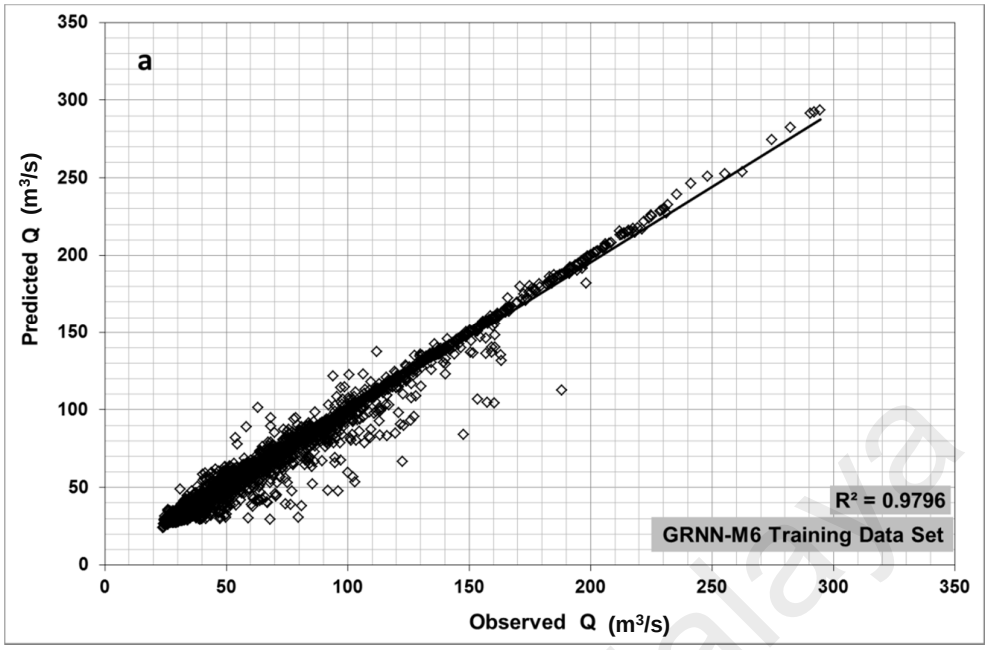


Figure 4.24: Performance values of GRNN-based models: (a) correlation coefficient and (b) mean absolute error



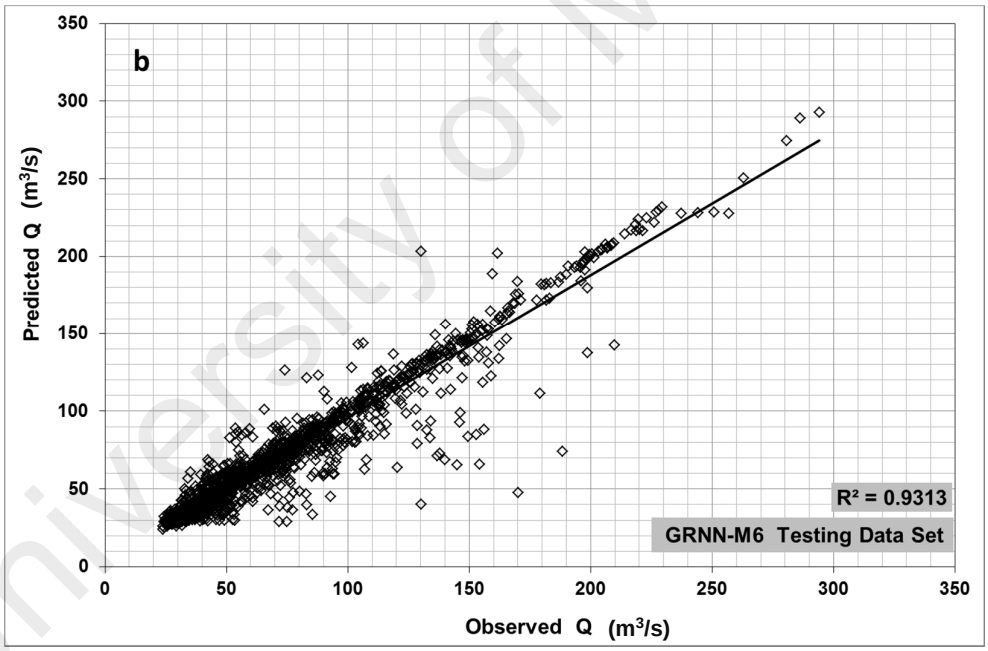


Figure 4.25: Correlation between the observed and predicted hourly stream flow by GRNN-M6 model: (a) training data set and (b) Testing data set

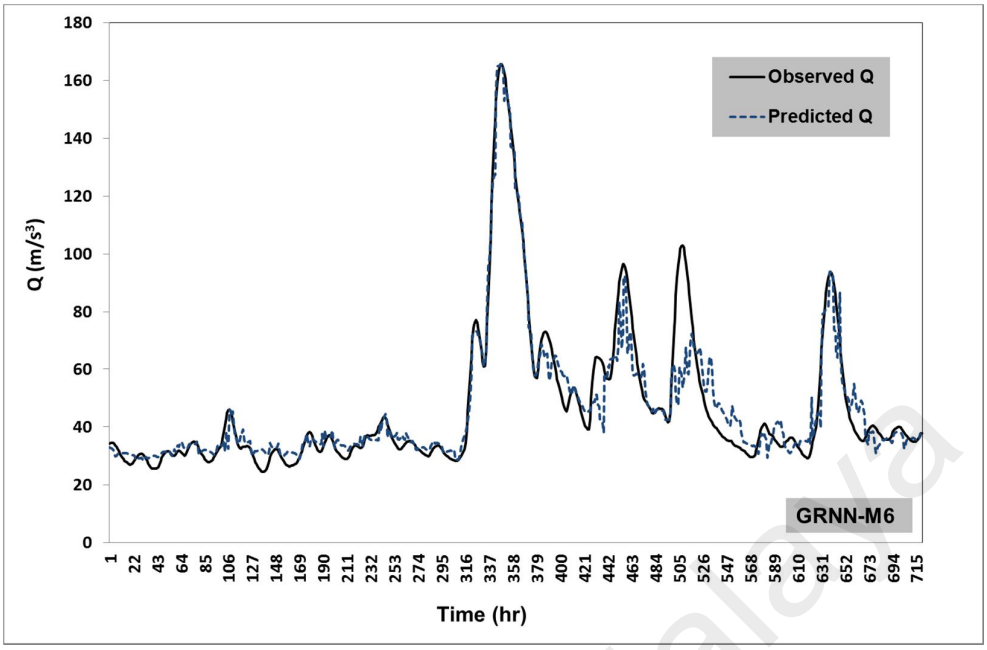


Figure 4.26: Comparison between the observed and predicted hourly stream flow by the GRNN-M6 model for the period of September 2013

Table 4.13: Performance values of GRNN-based models

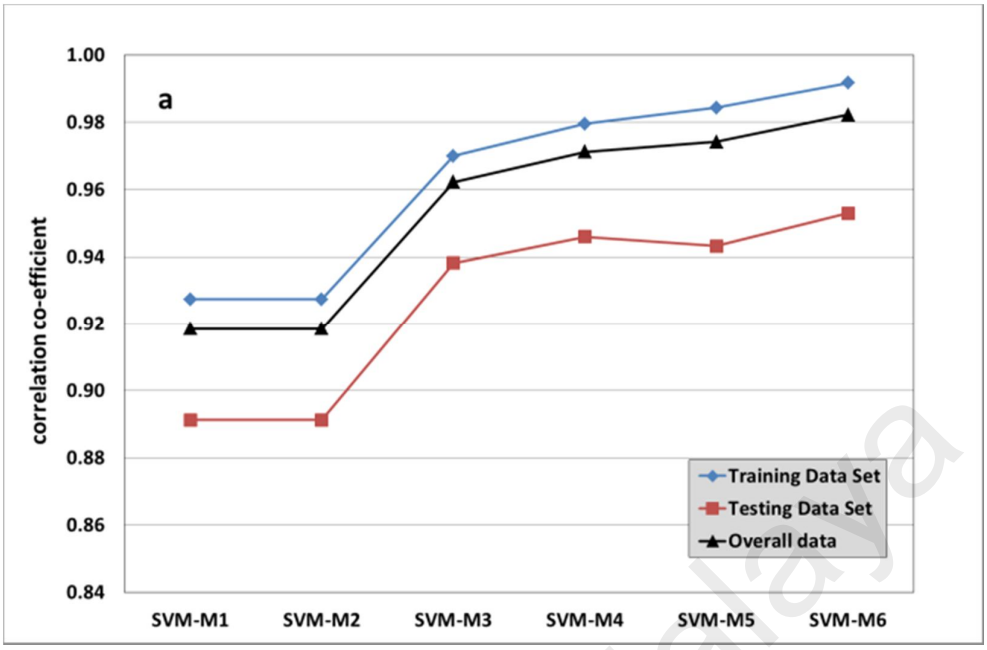
Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
GRNN-M1	0.922	8.878	0.904	10.267	0.912	9.524
GRNN-M2	0.924	9.231	0.906	9.620	0.911	9.570
GRNN-M3	0.991	2.510	0.952	6.456	0.970	4.357
GRNN-M4	0.992	2.291	0.964	5.174	0.976	3.839
GRNN-M5	0.996	1.571	0.955	5.584	0.972	3.529
GRNN-M6	0.990	2.634	0.965	5.240	0.976	3.987

4.4.1.4 SVM-based Models

Six models with different combination of input variables were trained and developed by SVM to predict Q. The performances of the models were assessed based on the training and testing data sets, as well as the overall performance of the data sets. The best fit model to predict Q is thus determined according to the performance of the testing data sets. Table 4.14 shows the performance evaluation results as denoted by the R and MAE of the SVM-based models.

Figure 4.27 compares the performance levels of the six SVM-based models and determines that the best fit model is SVM-M6. This model displays the highest *R* values (0.992 and 0.953) and the lowest MAE (0.061 and 0.253) in both the training and testing data sets, respectively.

Figure 4.28 shows the correlation between the observed and predicted Q in the SVM-M6 model given training and testing data sets. The observed and predicted Q of the training and testing data sets seem to be in good accord with R^2 0.986 and 0.909, respectively. Figure 4.29 compares the observed and predicted Q in SVM-M6 for the period of September 2013. These flows are highly consistent with very small error. The results verified the high performance of the model. The full records of the observed and predicted Q by SVM-M6 for September 2013 are presented in Appendix D.



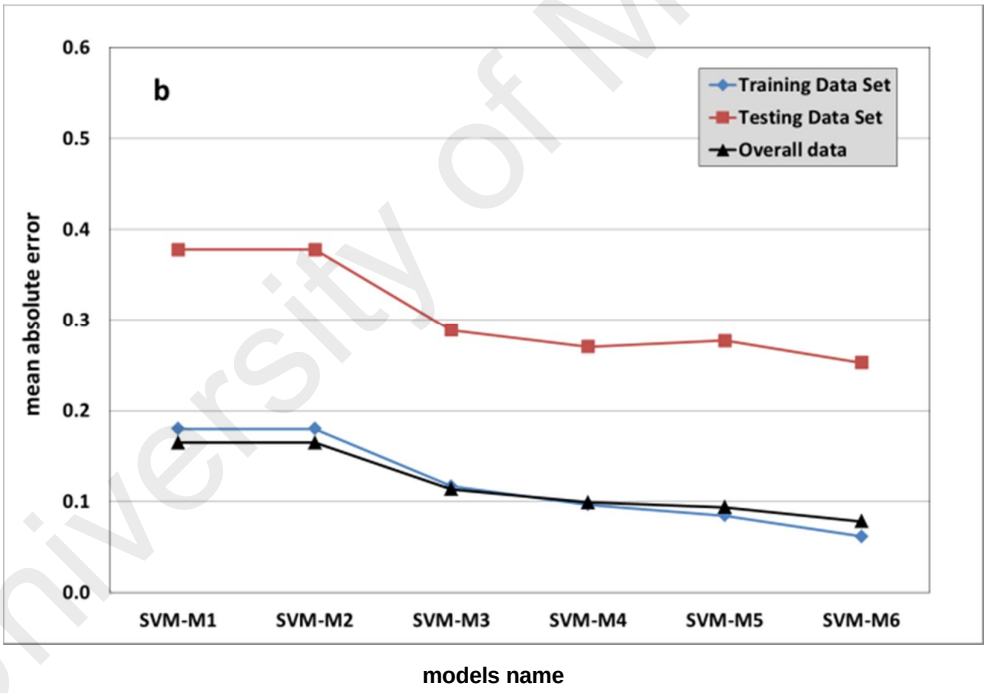
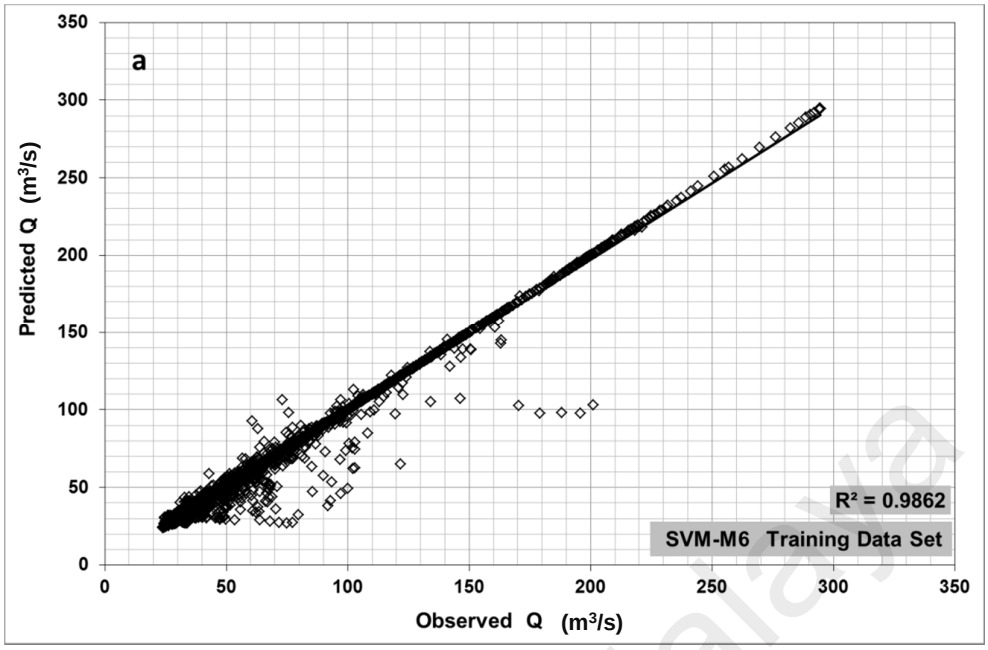


Figure 4.27: Performance values of SVM-based models: (a) correlation coefficient and (b) mean absolute error



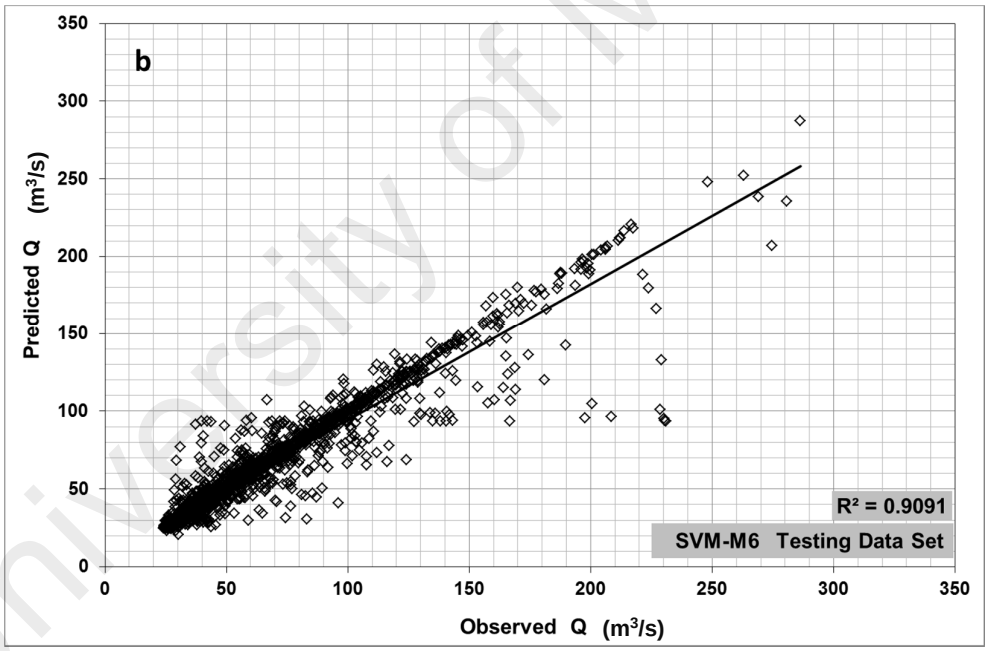


Figure 4.28: Correlation between the observed and predicted hourly stream flow by SVM-M6 model: (a) training data set and (b) Testing data set

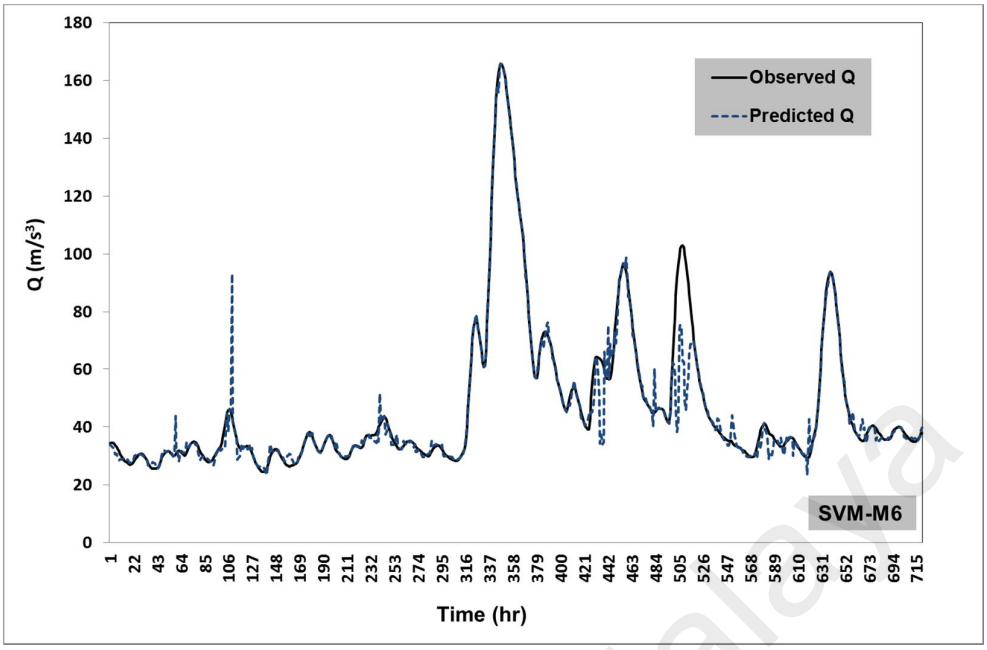


Figure 4.29: Comparison between the observed and the predicted hourly stream flow by the SVM-M6 model for the period of September 2013

Table 4.14: Performance values of SVM-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
SVM-M1	0.927	0.180	0.891	0.378	0.919	0.165
SVM-M2	0.927	0.180	0.891	0.378	0.919	0.165
SVM-M3	0.970	0.117	0.938	0.288	0.962	0.113
SVM-M4	0.979	0.097	0.946	0.270	0.971	0.099
SVM-M5	0.984	0.084	0.943	0.277	0.974	0.094
SVM-M6	0.992	0.061	0.953	0.253	0.982	0.078

4.4.1.5 Comparison between the Performances of the Four AI techniques: First Phase

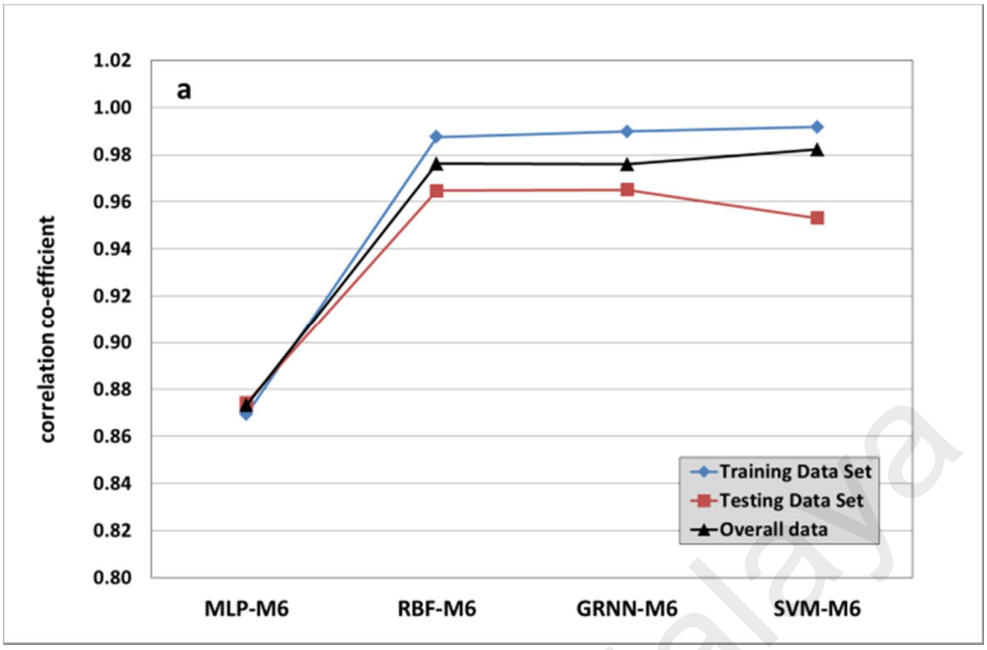
The best fit model for predicting Q in first modelling phase is determined based on evaluating the performance of the testing data sets. Through the detailed discussion in the previous sub-sections, it appears that M6 provides the best performance out of the other models while M5 could be considered the second, based on its performance evaluation results shown in Tables (4.11,4.12, 4.13 and 4.14).

M6 was trained and developed using four AI modelling techniques: MLP, RBF, GRNN and SVM. The results of performance evaluation criteria (i.e. R and MAE) of the MLP-M6, RBF-M6, GRNN-M6 and SVM-M6 models are presented in Table 4.15.

The comparison of the performance evaluation with respect to R and MAE for the four M6 models can be seen in Figure 4.30. Here, the R values for the RBF-M6, GRNN-M6 SVM-M6 models are similar, while the R value for MLP-M6 is lower than the other models.

It is noted that SVM-M6 provides the best MAE with lowest values, 0.061 and 0.253 for the training and testing data sets, respectively. GRNN could be considered second, as GRNN-M6 provides low values with 2.634 and 5.240 for the training and testing data sets, respectively. MLP provides highest values of MAE, with 11.55 and 12.039 for the training and testing data sets respectively.

Even the SVM-M6, GRNN-M6 models provide similar R values between the observed and predicted Q. However, it can be said that among all models, SVM-M6 is considered the best model for Q prediction according to its lowest MAE value between the observed and predicted Q which is significantly lower than the MAE of other three techniques.



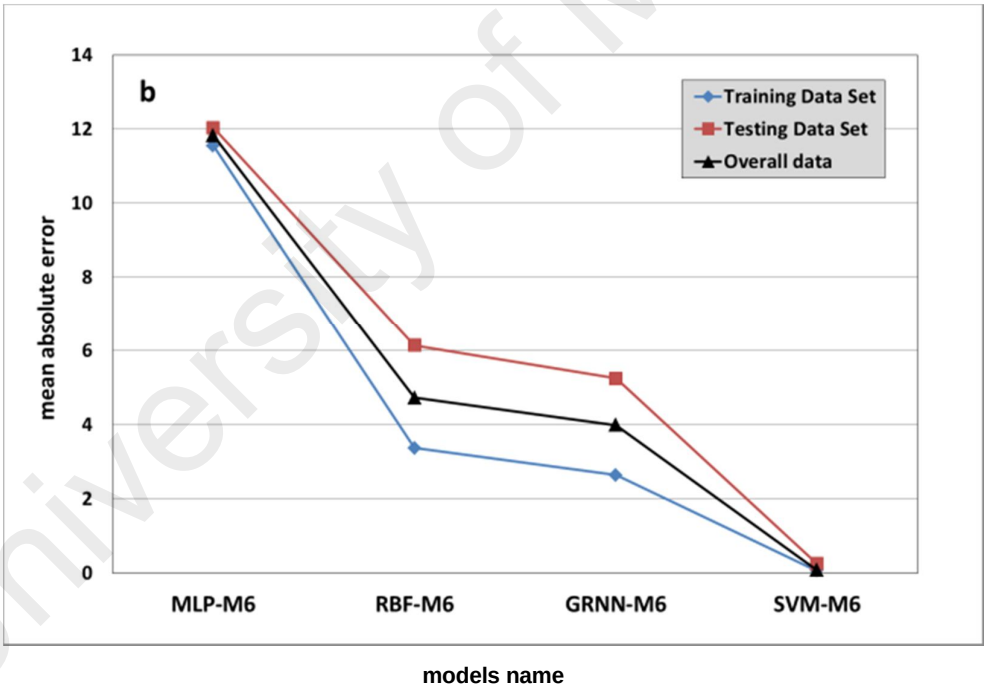


Figure 4.30: Performance values of the best fit AI-based models: (a) correlation coefficient and (b) mean absolute error

Table 4.15: Performance values of the four AI techniques applied in Model 6

Model	Training Data Set		Testing Data Set		Overall data	
	R	**MAE**	**R**	**MAE**	**R**	**MAE**
MLP-M6	0.869	11.550	0.874	12.039	0.87327	11.823
RBF-M6	0.987	3.370	0.965	6.141	0.976	4.720
GRNN-M6	0.990	2.634	0.965	5.240	0.976	3.987
SVM-M6	0.992	0.061	0.953	0.253	0.982	0.078

4.4.2 AI-based Models: Second Phase of Modelling Process

In the second phase of modelling process, only two models, those achieved the highest R among the six models of the first phase, were selected for the second phase. The new two models were named as M5a and M6a. They were trained and developed by the same four AI techniques: MLP, RBF, GRNN and SVM resulting in the development of eight AI-based models to predict the Q as shown in Table 4.16.

In this modelling phase, the lag intervals between the input and output variables of the AI-based models were selected based on the results of the HGA to estimate the Lt between the upstream and downstream stations.

New companions of variables and modelling cases of the M5a and M6a were employed giving, the changes of the lag intervals between the input and output variables which were selected based on the results of the HGA as shown in Table 4.17. Figure 4.31 presents the Lag intervals between the input variables and output variable for M5a and M6a models. Table 4.18 presents a group of 15 modelling cases of M6a as example of 8872

modelling cases. A larger group of modelling cases of M6 for three days is presented in Appendix B.

Table 4.16: AI-based models of the second modelling phase

Modelling technique	Model No.	
	M5a	**M6a**
MLP	MLP-M5a	MLP-M6a
RBF	RBF-M5a	RBF-M6a
GRNN	GRNN-M5a	GRNN-M6a
SVM	SVM-M5a	SVM-M6a

Table 4.17: Input and output variables of the AI-based models

Model	Inputs	Output	No. input Variables
M5a	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-2)}$, $Rf_{b(t-2)}$, $Rf_{k(t-2)}$, $Rf_{a(t-2)}$, $Q_{(t)}$	$Q_{(t+13)}$	9
M6a	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-2)}$, $Rf_{b(t-2)}$, $Rf_{k(t-2)}$, $Rf_{a(t-2)}$, $Q_{(t)}$	$Q_{(t+13)}$	9

Table 4.18: Group of modelling cases of M6a

date	time	$Q_{(t)}$	$Wl_{u(t)}$	$Wl_{b(t)}$	$Wl_{k(t)}$	$Wl_{a(t)}$	$Rf_{u(t-2)}$	$Rf_{b(t-2)}$	$Rf_{k(t-2)}$	$Rf_{a(t-2)}$	$Q_{(t+13)}$
10/01/2011	01:00	32.31	32.55	32.61	44.17	50.03	0.00	0.00	0.00	0.00	30.15
10/01/2011	02:00	32.43	32.55	32.60	44.17	50.03	0.00	0.00	0.00	0.00	29.74
10/01/2011	03:00	32.55	32.55	32.61	44.17	50.03	0.00	0.07	0.00	0.20	29.73
10/01/2011	04:00	32.77	32.55	32.61	44.17	50.03	0.00	0.17	0.00	0.50	29.91
10/01/2011	05:00	32.96	32.55	32.62	44.17	50.03	0.00	0.27	0.20	0.80	30.11
10/01/2011	06:00	33.08	32.55	32.60	44.17	50.03	0.00	0.30	0.60	0.90	30.44
10/01/2011	07:00	33.08	32.55	32.62	44.18	50.03	0.00	0.30	1.00	0.90	31.71
10/01/2011	08:00	32.90	32.55	32.62	44.18	50.03	0.00	0.30	1.20	0.90	33.66
10/01/2011	09:00	32.46	32.55	32.63	44.17	50.04	0.00	0.30	1.20	0.90	36.02
10/01/2011	10:00	32.11	32.55	32.62	44.16	50.04	0.00	0.30	1.20	0.90	38.64
10/01/2011	11:00	31.60	32.54	32.60	44.15	50.04	0.00	0.30	1.20	0.90	40.92
10/01/2011	12:00	31.14	32.54	32.58	44.14	50.03	0.00	0.30	1.20	0.90	42.83
10/01/2011	13:00	30.70	32.55	32.57	44.13	50.03	0.00	0.30	1.20	0.90	44.65
10/01/2011	14:00	30.15	32.55	32.58	44.12	50.03	0.00	0.30	1.20	0.90	46.76
10/01/2011	15:00	29.74	32.55	32.59	44.13	50.03	0.00	0.30	1.20	0.90	49.52

Lag time	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13
Rf_u	●															
Rf_b	●															
Rf_k	●															
Rf_a	●															
Wl_u			●													
Wl_b			●													
Wl_k			●													
Wl_a			●													
Sf			●													▲

a

Lag time	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13
Rf_u	●	●	●													
Rf_b	●	●	●													
Rf_k	●	●	●													
Rf_a	●	●	●													
Wl_u			●	●	●											
Wl_b			●	●	●											
Wl_k			●	●	●											
Wl_a			●	●	●											
Sf			●													▲

b

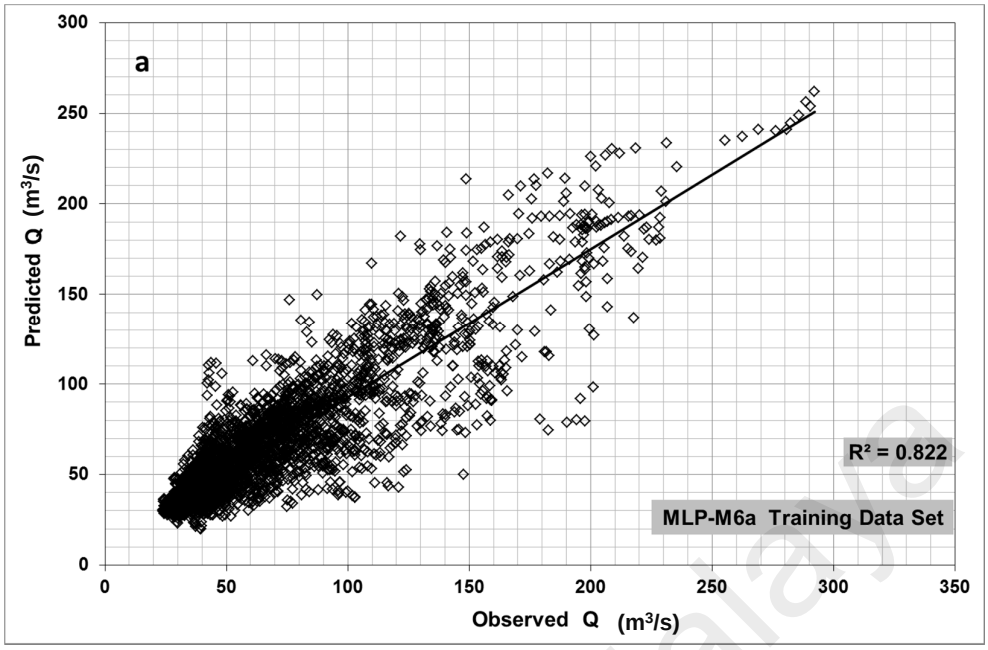
where ● is the input variables and ▲ is the output variable

Figure 4.31: Lag intervals between the input and output variables of the AI- based models: a) Model 5a, b) Model 6a

4.4.2.1 MLP-based Models

M5a and M6a were trained and developed by MLP to predict Q. The performance of the developed models was assessed based on the training and testing data sets, as well as the overall performance of the data sets. The best fit model to predict Q is thus determined according to the performance of the testing data sets. Table 4.19 presents the performance evaluation results as denoted by the R and MAE of M5a and M6a. This table shows that the best fit model is MLP-M6a. This model displays the highest *R* values (0.902 and 0.894) and the lowest MAE (10.721 and 11.076) in both the training and testing data sets, respectively.

Figure 4.32 shows the correlation between the observed and predicted Q in MLP-M6a model giving training and testing data set. The observed and predicted Q of the training and testing data sets, seem to be in good accord with R^2 0.822 and 0.80, respectively. In Figure 4.33, a comparison between the observed and predicted Q by MLP-M6a for the period of September 2013 can be seen. Acceptable agreement with small error between the observed and predicted Q is apparent. The results verified the high performance of the model. The full records of the observed and predicted Q by MLP-M6a for the period of September 2013 are presented in Appendix D.



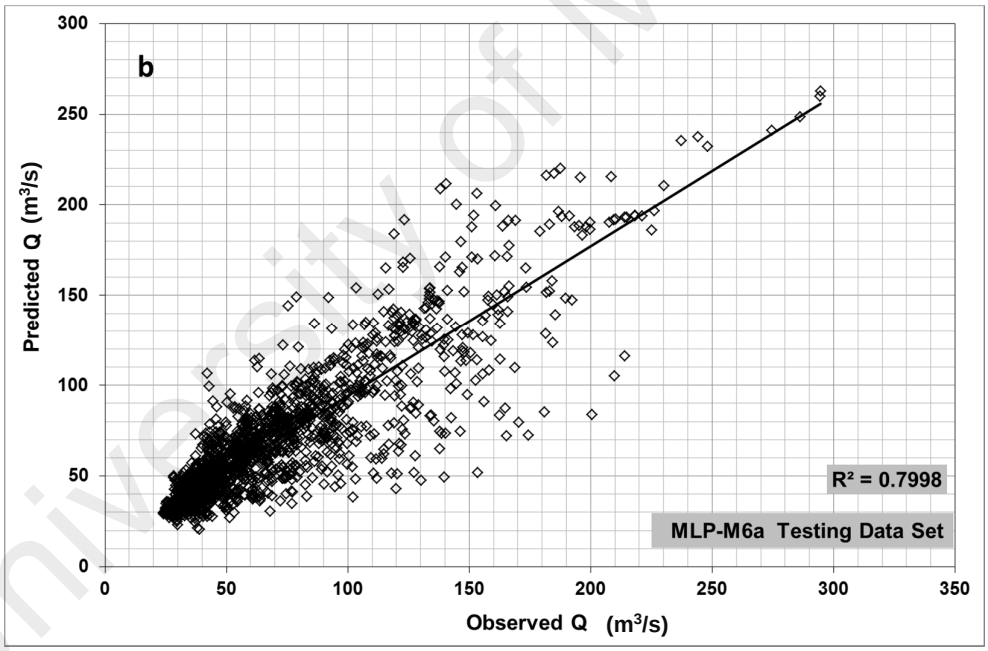


Figure 4.32: Correlation between the observed and predicted hourly stream flow by M6a-MLP model: (a) training data set and (b) testing data set

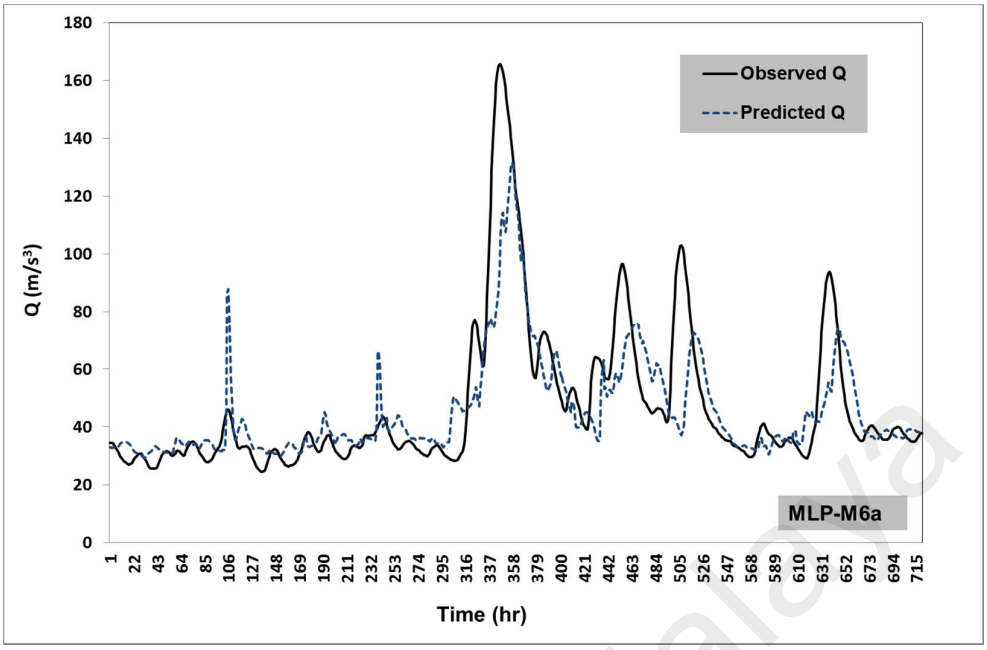


Figure 4.33: Comparison between the observed and predicted hourly stream flow via the M6a-MLP model for the period of September 2013

Table 4.19: Performance values of MLP-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
MLP-M5a	0.878	12.062	0.885	12.376	0.879	11.957
MLP-M6a	0.902	10.721	0.894	11.076	0.903	10.681

4.4.2.2 RBF-based Models

M5a and M6a were trained and developed by RBF to predict Q. Table 4.20 presents the performance evaluation results as denoted by the R and MAE of M5a and M6a. This table shows that the best fit model is RBF-M6a. This model displays the highest *R* values (0.984 and 0.962) and the lowest MAE (3.965 and 6.690) in both the training and testing data sets, respectively.

Figure 4.34 presents the correlation between the observed and predicted Q in RBF-M6a model giving training and testing data set. The observed and predicted Q of the training and testing data sets, seem to be in good accord with R^2 0.969 and 0.926, respectively. In Figure 4.35, a comparison between the observed and predicted Q by RBF-M6a for the period of September 2013 can be seen. Good agreement with small error between the observed and predicted Q is apparent. The results verified the high performance of the model. The full records of the observed and predicted Q by RBF-M6a for the period of September 2013 are presented in Appendix D.

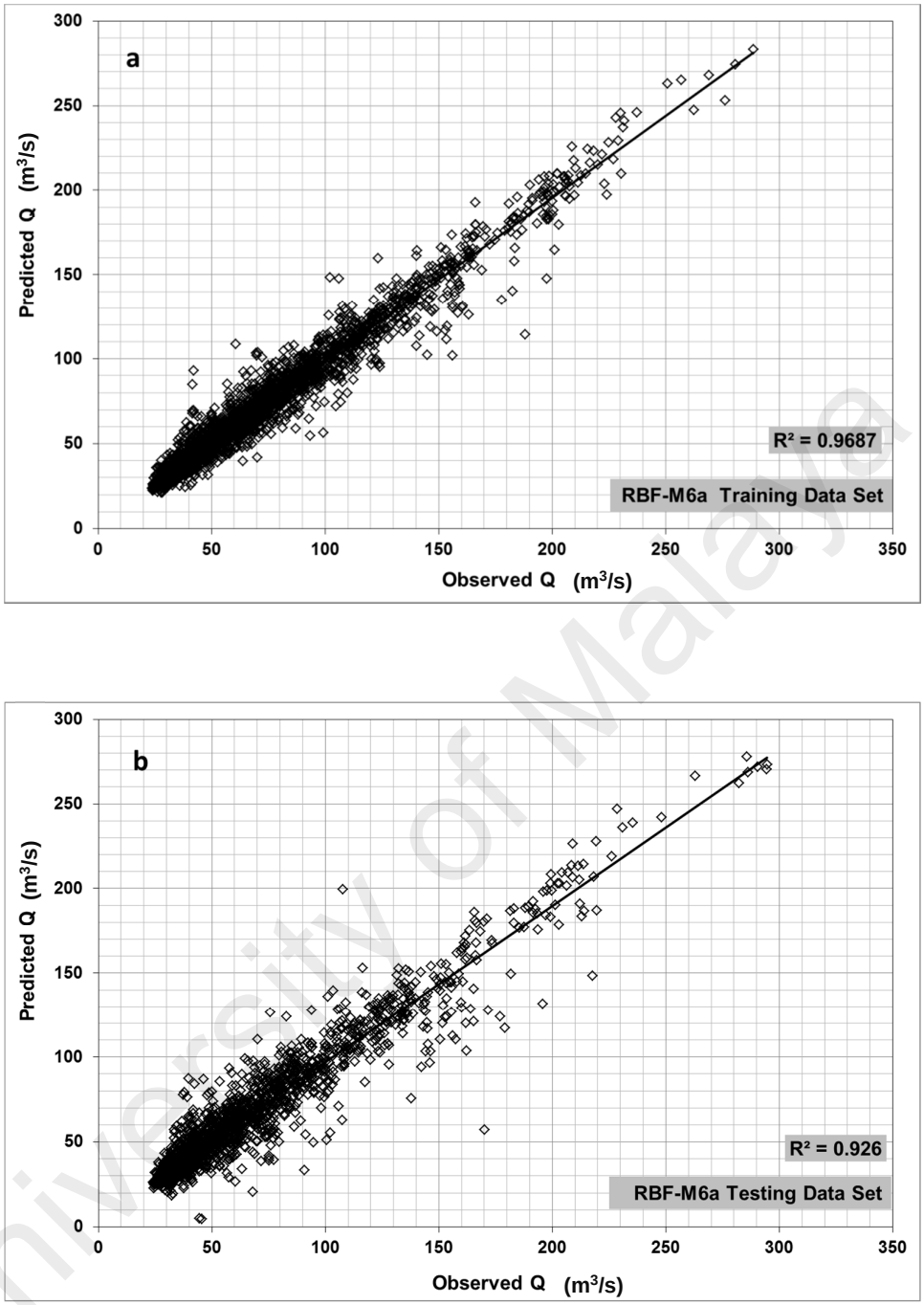


Figure 4.34: Correlation between the observed and predicted Q by RBF-M6a model:

(a) training data set and (b) Testing data set

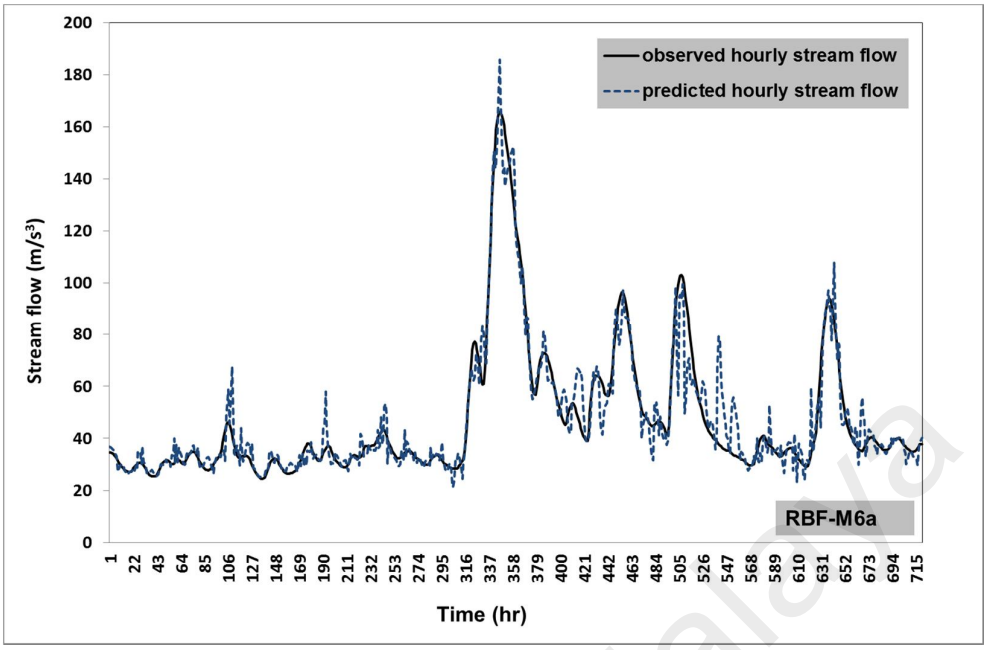


Figure 4.35: Comparison between observed and the predicted Q via the RBF-M6a model for the period of September 2013

Table 4.20: Performance values of RBF-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
RBF-M5a	0.985	3.90	0.945	7.686	0.965	5.850
RBF-M6a	0.984	3.965	0.962	6.690	0.971	5.318

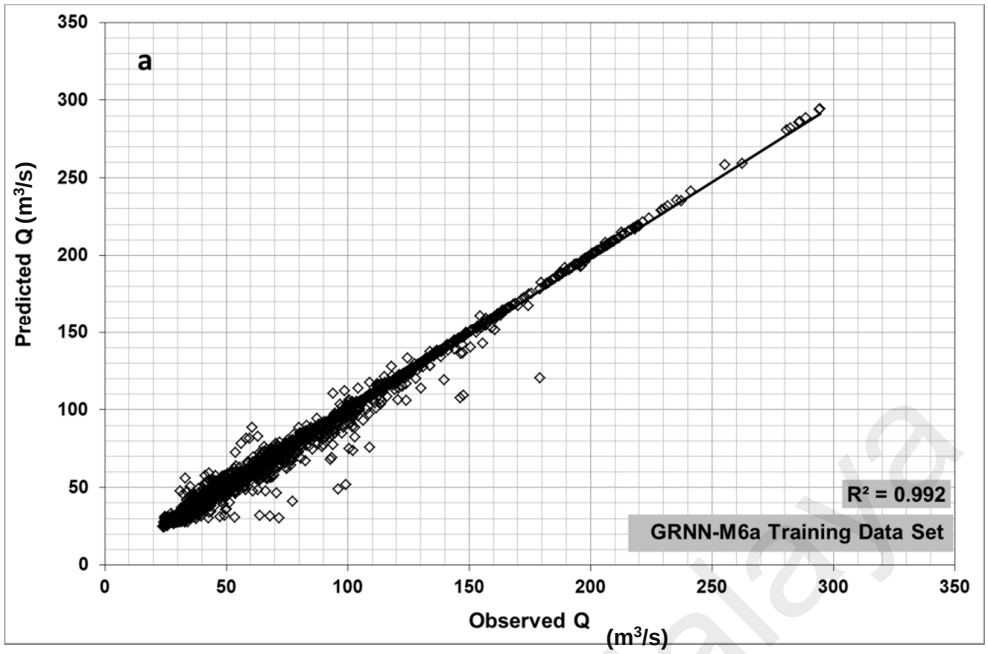
4.4.2.3 GRNN-based Models

M5a and M6a models were trained and developed using GRNN to predict Q. The results of performance evaluation criteria i.e. R and MAE of the GRNN models are presented in Table 4.21. The best fit model is GRNN-M6a with the highest *R* values and lowest MAE value for the training and testing data sets. The *R* between the observed and predicted Q by the GRNN-M6a model is 0.996 and 0.964 while the MAE is 1.421 and 4.271 for the training and testing data sets respectively.

Figure 4.36 presents the correlation between the observed and predicted Q in GRNN-M6a model giving training and testing data set. The observed and predicted Q of the training and testing data sets, seem to be in very good accord with R^2 0.992 and 0.93, respectively. In Figure 4.37, a comparison between the observed and predicted Q by GRNN-M6a for the period of September 2013 can be seen. High agreement with small error between the observed and predicted Q is apparent. The results verified the high performance of the model. The full records of the observed and predicted Q by GRNN-M6a for the period of September 2013 are presented in Appendix D.

Table 4.21: Performance values of GRNN-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
GRNN-M5a	0.997	1.247	0.951	5.690	0.974	3.397
GRNN-M6a	0.996	1.421	0.964	4.271	0.978	3.013



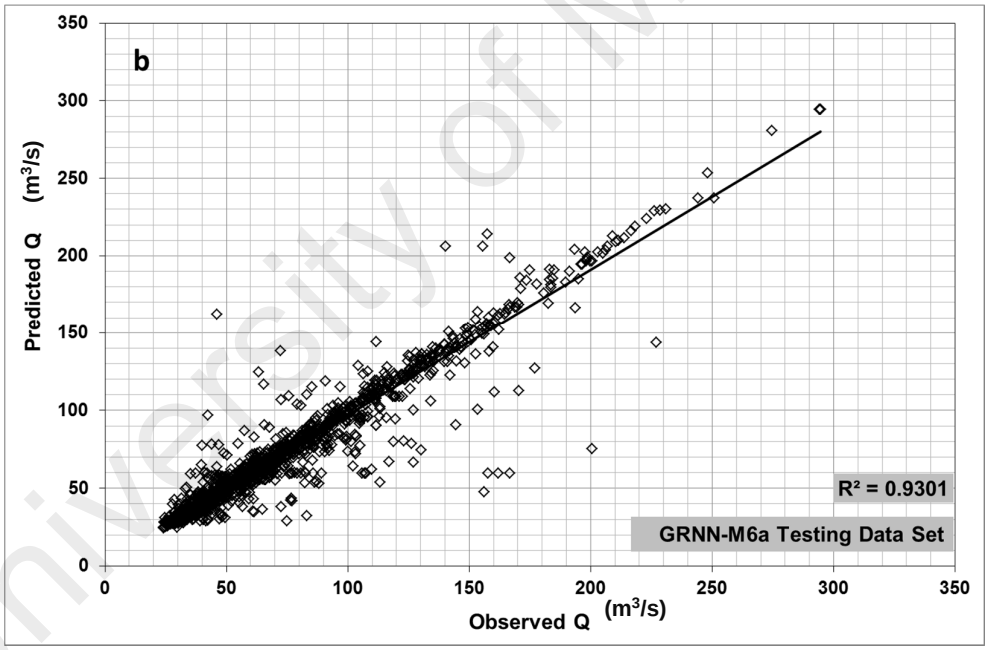


Figure 4.36: Correlation between the observed and predicted Q by GRNN-M6a model: (a) training data set and (b) Testing data set

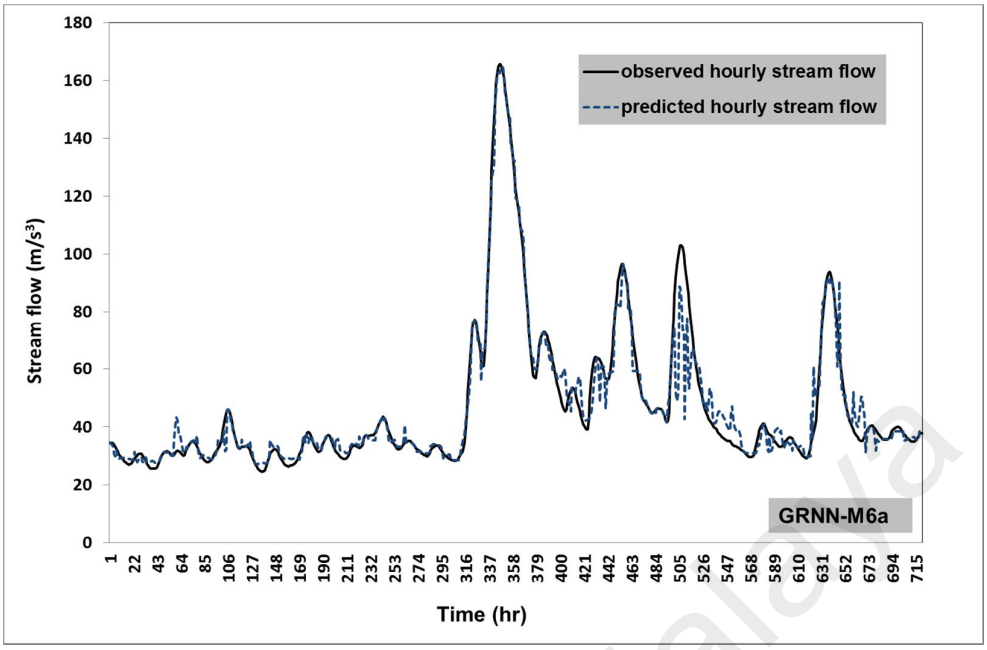


Figure 4.37: Comparison between the observed and predicted Q via the GRNN-M6a model for the period of September 2011

4.4.2.4 SVM-based Models

M5a and M6a models were trained and developed using SVM to predict Q. The results of performance evaluation criteria i.e. R and MAE of the SVM models are presented in Table 4.22. The best fit SVM model is SVM-M6a with the highest *R* values and lowest MAE value for the training and testing data sets. The *R* between the observed and predicted Q by the SVM-M6a model is 0.985 and 0.952 while the MAE is 0.083 and 0.254 for the training and testing data sets, respectively.

Figure 4.38 presents the correlation between the observed and predicted Q in SVM-M6a model giving training and testing data set. The observed and predicted Q of the training and testing data sets, seem to be in very high accord with R^2 0.980 and 0.907, respectively. In Figure 4.39, a comparison between the observed and predicted Q by SVM-M6a for the period of September 2013 can be seen. A very high agreement with very small error

between the observed and predicted Q is apparent. The results verified the high performance of the model. The full records of the observed and predicted Q by SVM-M6a for the period of September 2013 are presented in Appendix D.

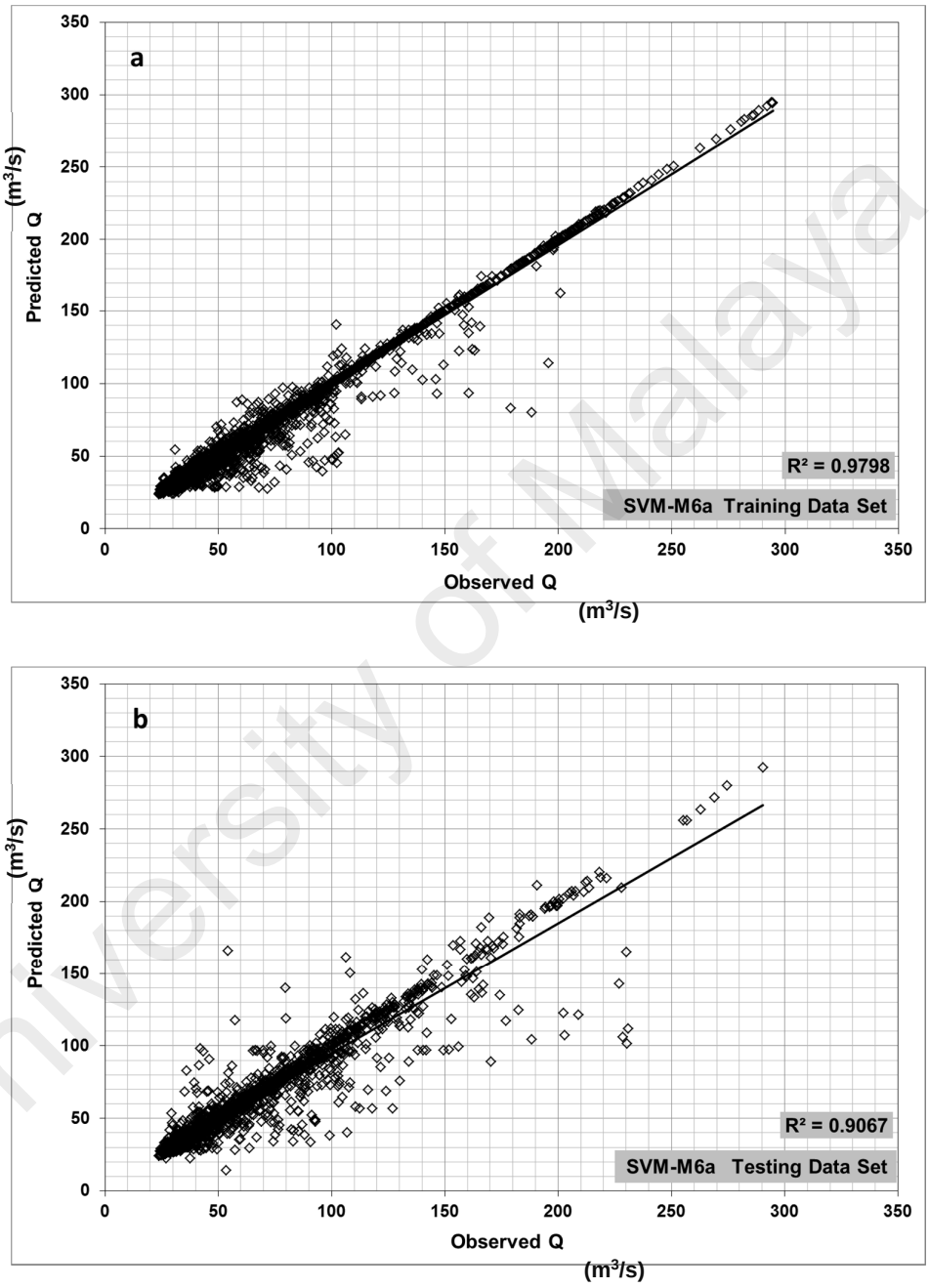


Figure 4.38: Correlation between the observed and the predicted hourly stream flow by M6a-SVM model: (a) training data set and (b) Testing data set

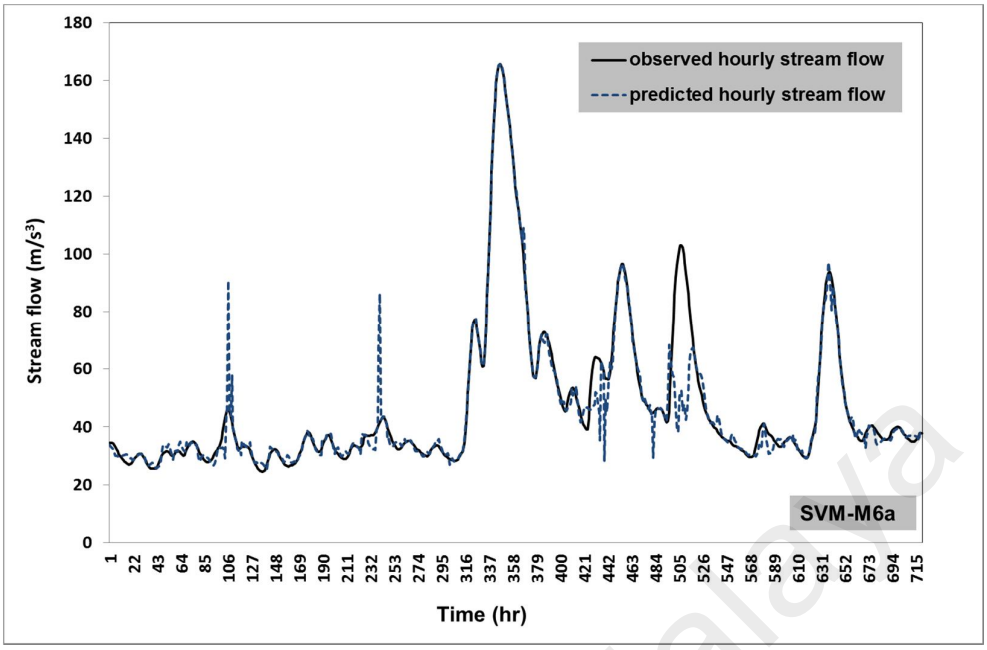


Figure 4.39: Comparison between the observed and the predicted hourly stream flow via the M6a-SVM model for the period of September 2013

Table 4.22: Performance values of SVM-based models

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
SVM-M5a	0.978	0.098	0.942	0.278	0.969	0.102
SVM-M6a	0.985	0.083	0.952	0.254	0.977	0.089

4.4.2.5 Comparison between the Performances of the Four AI techniques: Second Phase

The comparison between the four AI techniques was performed by analysis the performance of the best fit model for predicting Q which is determined based on the performance evaluation of testing data sets of the developed models over the second modelling phase. Through the detailed discussion in the previous sections, it appears that M6a provides better performance than M5a, based on its performance evaluation results shown in Tables (4.19, 4.20, 4.21 and 4.22).

M6a was trained and developed using the four AI modelling techniques: MLP, RBF, GRNN and SVM. The results of performance evaluation criteria, i.e. R and MAE of the MLP-M6a, RBF-M6a, GRNN-M6a and SVM-M6a models are presented in Table 4.23.

The performance evaluation comparison with respect to R and MAE for the four M6 models can be seen in Figure 4.40. As presented in this figure, the results of comparison is similar to the results of the first phase of modelling as the value of R of RBF-M6a, GRNN-M6a and SVM-M6a models are very similar, while the value of R of MLP-M6 is lower than other models. For the MAE values, it is noted that SVM-M6a provides the minimum values with 0.083 and 0.254 for the training and testing data sets respectively, while others models provide higher value of MAE such as MLP-M6a with 11.824 and 11.411 for the training and testing data sets respectively.

Even the RBF, GRNN and SVM provide similar values R between the observed and predicted hourly Q, but is can be said that SVM are considered the best technique to the predict the hourly Q depending on its lowest value of MAE between the observed and predicted Q which are significantly lower than other techniques.

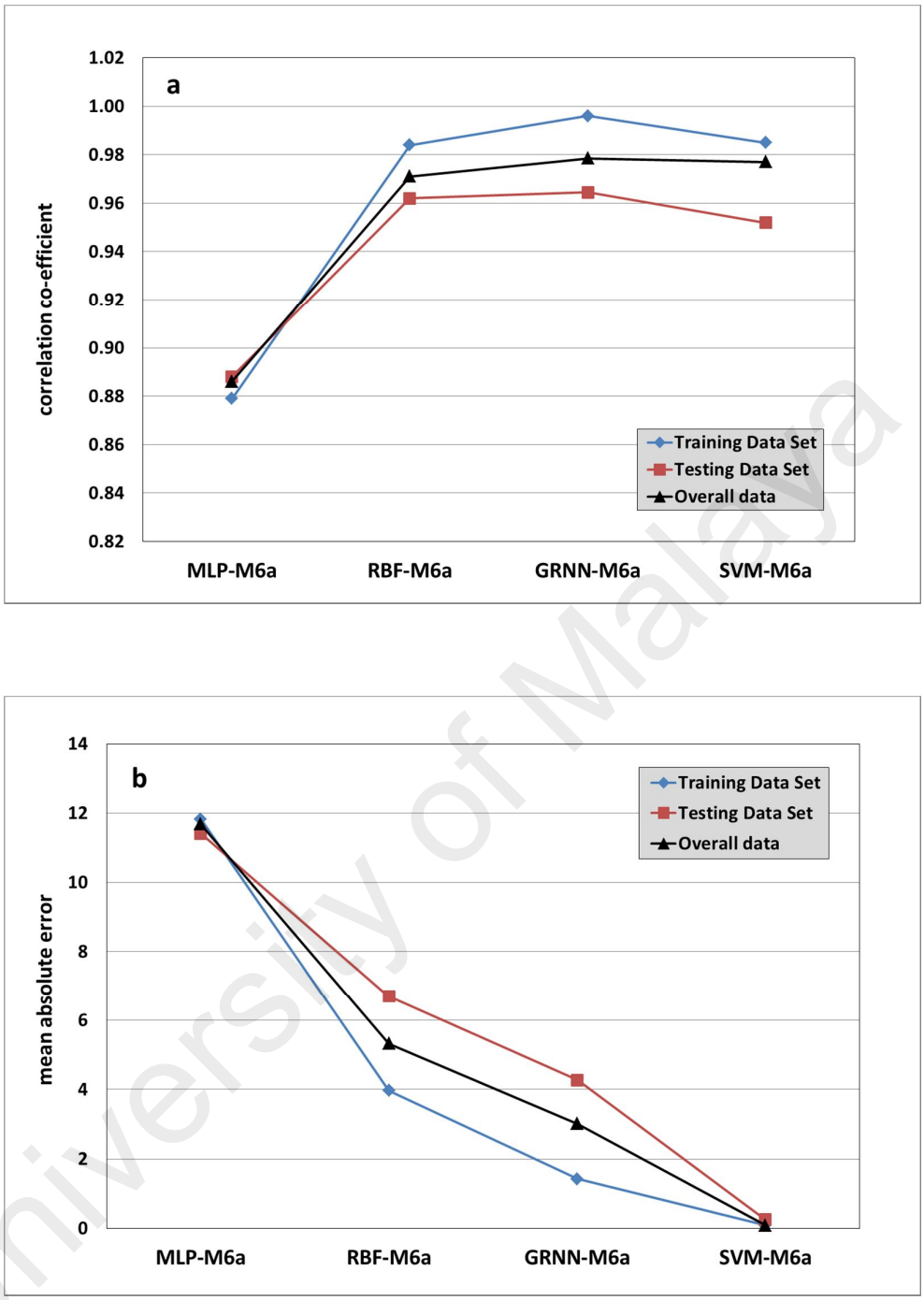


Figure 4.40: Performance values of the best fit AI-based models: (a) correlation coefficient and (b) mean absolute error

Table 4.23: Performance values of the four AI techniques applied in Model 6a

Model	Training data set		Testing data set		Overall data	
	R	MAE	R	MAE	R	MAE
MLP-M6a	0.879	11.824	0.888	11.411	0.886	11.683
RBF-M6a	0.984	3.965	0.962	6.69	0.971	5.318
GRNN-M6a	0.996	1.421	0.964	4.271	0.978	3.013
SVM-M6a	0.985	0.083	0.952	0.254	0.977	0.089

4.5 Applications of AI-based Models

The developed AI-based models can be utilized in several hydrological applications. In this study, they were employed as prediction tools, as shown in the previous section, and as analytical tools to investigate the influence of WL and RF on Q. They were also applied to estimate the missing Q records. Finally, they were employed in flood early warning throughout the advance detection of hydrological conditions that could lead to formations of floods.

However, not all AI-based models have the ability to understand and investigate the physical behavior of hydrological systems. AI-based models must be able to explore the physical behavior of hydrological systems so that they can be applied in hydrological applications.

All the developed models were checked so that the best-fitting model in investigation of the physical behavior of hydrological systems is selected for the applications. The result shows that MLP–M6a provides the most hydrological sounds in the investigation of the influence of the input variables on Q. This result means that this model achieves the

highest ability in exploring the hydrological description of the applied process. The hydrological applications in this research were performed using the MLP-6a model.

4.5.1 Utilizing AI-based Models as Analytical Tool

Hypothetical modelling cases of input variables were assumed and modeled to study the influence of the input variables, WL and RF of the four upstream stations, on Q. The hypothetical cases were prepared by gradually changing the values of the input variable from the minimum to the maximum value within the range of the input variables. The ranges of the input variables were divided into ten steps, and the values were gradually increased from the minimum to the maximum value of the range. Table 4.24 shows the hypothetical values of the gradual change of the input variables.

The hypothetical values were used to determine the input cases and develop eight input matrices to study the influence of both WL and RF of the four upstream stations on Q. The input combinations of the modelling cases of these matrices were predicted using MLP-6a to obtain the output of these hypothetical cases. The investigation of the relationship between the hydrological variables and the results of the hypothetical cases can help to determine the influence of the WL and RF on Q.

Table 4.24: Hypothetical cases of input variables

	Q	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a
Min.	**23.9**	**30.6**	**27.0**	**43.9**	**49.6**	**0.0**	**0.0**	**0.0**	**0.0**
Max.	**294.6**	**35.5**	**34.7**	**45.6**	**50.9**	**19.3**	**22.7**	**25.3**	**28.0**
Mean	**32.24**	**32.42**	**44.18**	**50.16**	**0.16**	**0.24**	**0.25**	**0.24**	
1	23.9	30.6	27.0	43.9	49.6	0.0	0.0	0.0	0.0
2	51.0	31.1	27.8	44.1	49.7	1.9	2.3	2.5	2.8
3	78.1	31.5	28.6	44.3	49.9	3.9	4.5	5.1	5.6
4	105.2	32.0	29.3	44.4	50.0	5.8	6.8	7.6	8.4
5	132.2	32.5	30.1	44.6	50.1	7.7	9.1	10.1	11.2
6	159.3	33.0	30.9	44.8	50.3	9.7	11.3	12.7	14.0
7	186.4	33.5	31.6	44.9	50.4	11.6	13.6	15.2	16.8
8	213.4	34.0	32.4	45.1	50.5	13.5	15.9	17.7	19.6
9	240.5	34.5	33.2	45.3	50.6	15.5	18.1	20.3	22.4
10	267.6	35.0	33.9	45.4	50.8	17.4	20.4	22.8	25.2
11	294.6	35.5	34.7	45.6	50.9	19.3	22.7	25.3	28.0

4.5.1.1 Influence of the Water Level in Upstream Stations on the Stream Flow

Four input matrices were employed to investigate the influence of WL of the upstream stations on SF. The input values of these matrices were derived from the hypothetical values mentioned in Table 4.24. Each matrix was employed to study the influence of the WL of one upstream station on the SF. The values of WL and RF of the other stations were fixed at the mean value, as shown in Table 4.25, to provide the model with the ability to investigate the influence of changes on one specific station. Table 4.25 shows the hypothetical cases to study the influence of WL on SF and the predicted results of the investigation process using MLP-6a.

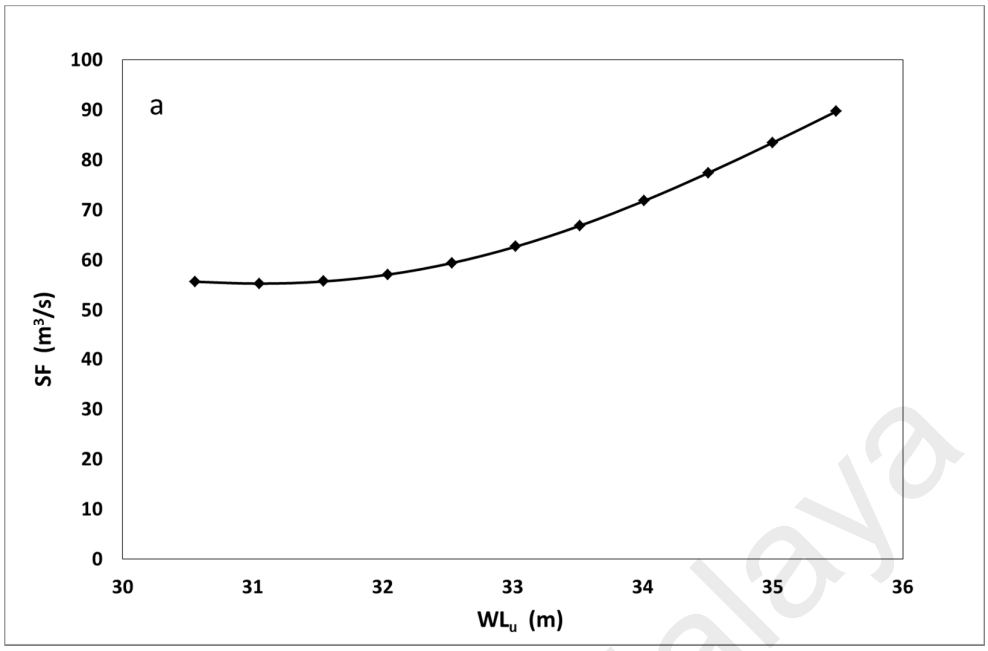
The axiomatic theory indicates that the Q in the downstream station is directly proportional to the WL in the upstream stations, which was proved through Figure 4.41. This figure shows the results of the investigation of the influence of all the WL upstream stations on the Q. The WL in the upstream stations directly affects the SF in the Rantau Panjang station (downstream station). Figure 4.41 shows that the influence of the WL is valid over the full ranges of the WL records of the downstream station, meaning that the MLP-6a model can be used as a hydrological sound tool in investigating the influence of WL on SF.

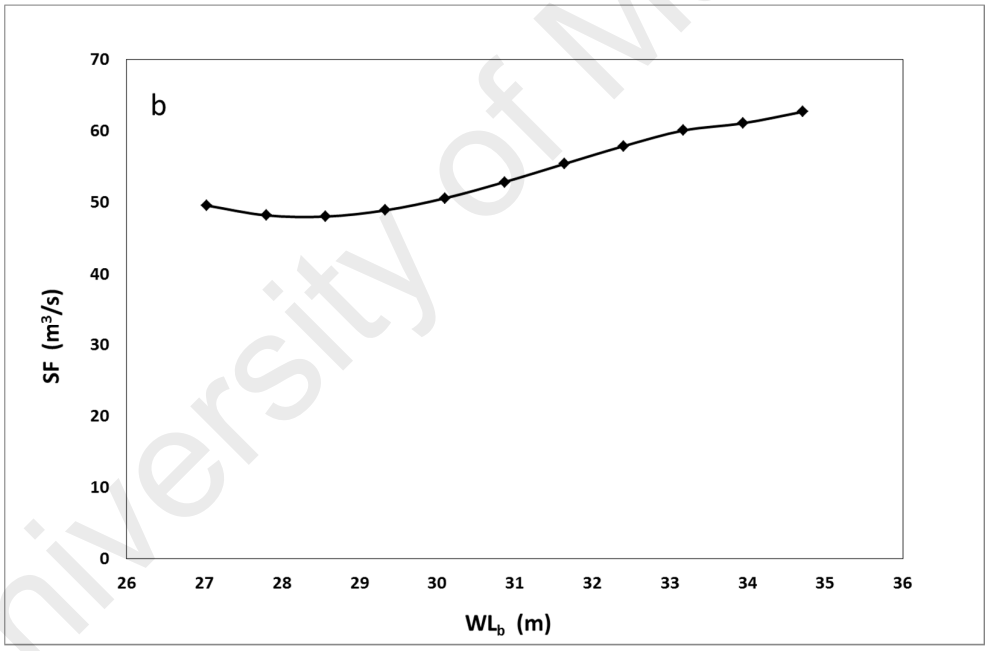
Table 4.25: Results of the hypothetical cases to study the influence of Water Level in upstream stations on Stream Flow

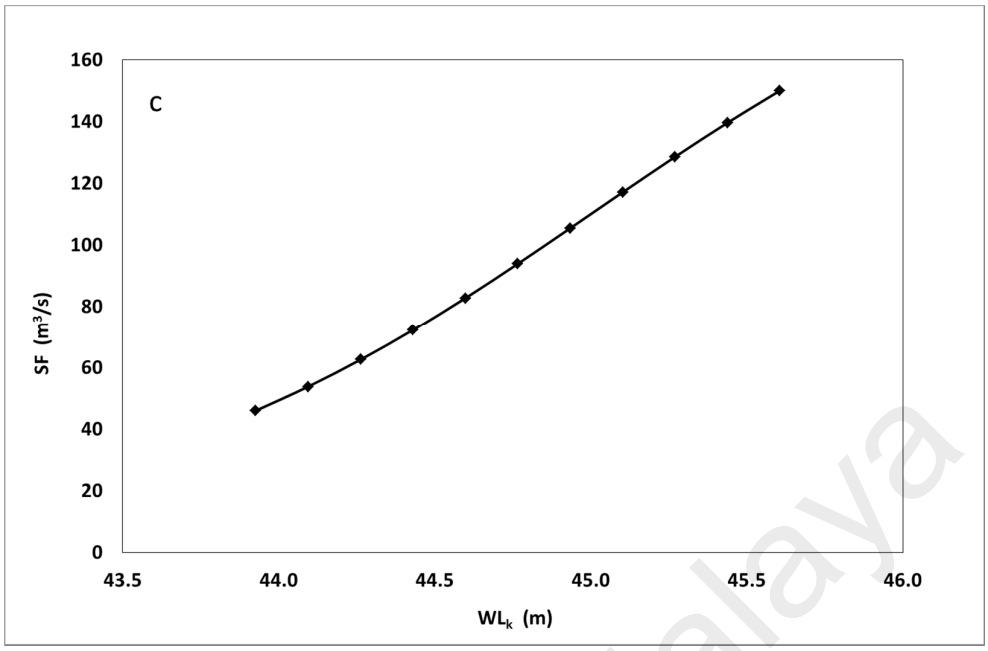
$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{(t+13)}$
Ulu Yam station									
60.34	30.56	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**55.67**
60.34	31.05	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**55.29**
60.34	31.55	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**55.72**
60.34	32.04	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**57.08**
60.34	32.53	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**59.40**
60.34	33.03	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**62.69**
60.34	33.52	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**66.86**
60.34	34.01	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**71.81**
60.34	34.50	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**77.39**
60.34	35.00	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**83.41**
60.34	35.49	32.42	44.18	50.16	0.16	0.24	0.25	0.24	**89.70**
Batang Kali station									
60.34	32.24	27.03	44.18	50.16	0.16	0.24	0.25	0.24	**49.58**
60.34	32.24	27.80	44.18	50.16	0.16	0.24	0.25	0.24	**48.20**
60.34	32.24	28.57	44.18	50.16	0.16	0.24	0.25	0.24	**48.03**
60.34	32.24	29.33	44.18	50.16	0.16	0.24	0.25	0.24	**48.90**
60.34	32.24	30.10	44.18	50.16	0.16	0.24	0.25	0.24	**50.59**
60.34	32.24	30.87	44.18	50.16	0.16	0.24	0.25	0.24	**52.85**

Table 4.25 Continued

$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{(t+13)}$
60.34	32.24	31.64	44.18	50.16	0.16	0.24	0.25	0.24	**55.39**
60.34	32.24	32.41	44.18	50.16	0.16	0.24	0.25	0.24	**57.88**
60.34	32.24	33.17	44.18	50.16	0.16	0.24	0.25	0.24	**60.06**
60.34	32.24	33.94	44.18	50.16	0.16	0.24	0.25	0.24	**61.07**
60.34	32.24	34.71	44.18	50.16	0.16	0.24	0.25	0.24	**62.67**
Kerling station									
60.34	32.24	32.42	43.93	50.16	0.16	0.24	0.25	0.24	**45.97**
60.34	32.24	32.42	44.09	50.16	0.16	0.24	0.25	0.24	**53.79**
60.34	32.24	32.42	44.26	50.16	0.16	0.24	0.25	0.24	**62.56**
60.34	32.24	32.42	44.43	50.16	0.16	0.24	0.25	0.24	**72.25**
60.34	32.24	32.42	44.60	50.16	0.16	0.24	0.25	0.24	**82.77**
60.34	32.24	32.42	44.77	50.16	0.16	0.24	0.25	0.24	**93.91**
60.34	32.24	32.42	44.93	50.16	0.16	0.24	0.25	0.24	**105.45**
60.34	32.24	32.42	45.10	50.16	0.16	0.24	0.25	0.24	**117.11**
60.34	32.24	32.42	45.27	50.16	0.16	0.24	0.25	0.24	**128.58**
60.34	32.24	32.42	45.44	50.16	0.16	0.24	0.25	0.24	**139.62**
60.34	32.24	32.42	45.61	50.16	0.16	0.24	0.25	0.24	**150.02**
Ampang Pecah station									
60.34	32.24	32.42	44.18	49.61	0.16	0.24	0.25	0.24	**49.62**
60.34	32.24	32.42	44.18	49.74	0.16	0.24	0.25	0.24	**51.65**
60.34	32.24	32.42	44.18	49.87	0.16	0.24	0.25	0.24	**53.69**
60.34	32.24	32.42	44.18	50.00	0.16	0.24	0.25	0.24	**55.66**
60.34	32.24	32.42	44.18	50.12	0.16	0.24	0.25	0.24	**57.44**
60.34	32.24	32.42	44.18	50.25	0.16	0.24	0.25	0.24	**58.93**
60.34	32.24	32.42	44.18	50.38	0.16	0.24	0.25	0.24	**60.05**
60.34	32.24	32.42	44.18	50.50	0.16	0.24	0.25	0.24	**60.75**
60.34	32.24	32.42	44.18	50.63	0.16	0.24	0.25	0.24	**61.01**
60.34	32.24	32.42	44.18	50.76	0.16	0.24	0.25	0.24	**60.84**
60.34	32.24	32.42	44.18	50.89	0.16	0.24	0.25	0.24	**60.30**







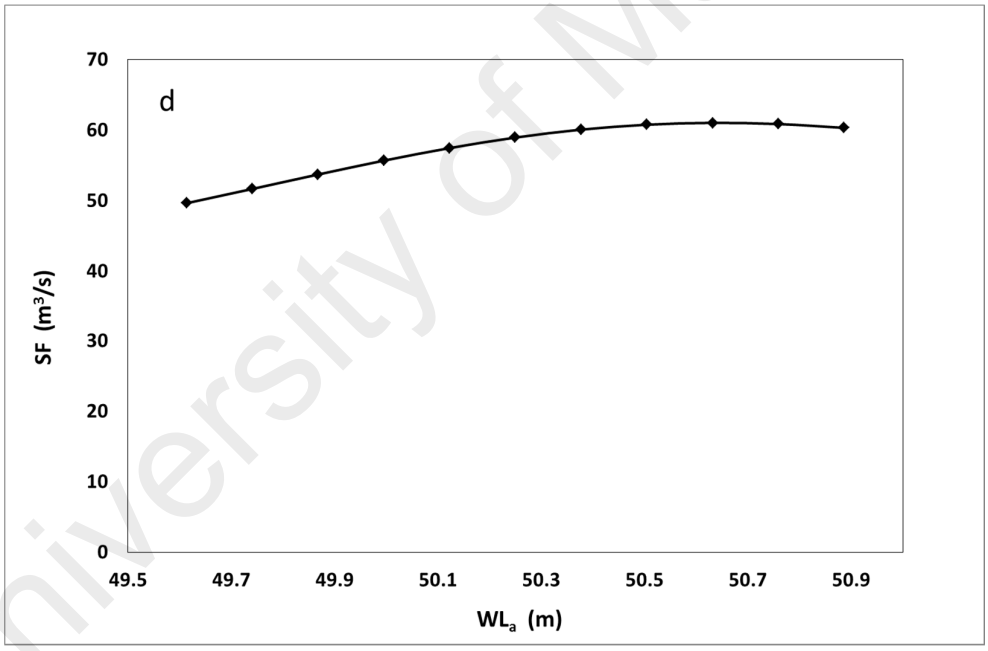


Figure 4.41: Influence of the water level variables in the stream flow: (a) Ulu Yam station (b) Batang Kali station (c) Kerling station (d) Ampang Pecah station

4.5.1.2 Influence of the Rainfall in Upstream Stations on the Stream Flow

Four input matrices were employed to investigate the influence of RF of the upstream stations on SF. The input values of these matrices were derived from the hypothetical values mentioned in Table 4.24. Each matrix was employed to study the influence of the RF of one upstream station on the SF. The values of WL and RF of the other stations were fixed at the mean value, as shown in Table 4.26, to provide the model with the ability to investigate the influence of changes on one specific station. Table 4.26 shows the hypothetical cases to study the influence of RF on SF and the predicted results of the investigation process using MLP-6a.

The axiomatic theory indicates that the Q in downstream station is directly proportional to the RF in the upstream stations, which was verified through Figure 4.42. This figure presents the results of the investigation of the influence of four RF stations on the SF. The RF in the upstream stations directly affects the SF in the Rantau Panjang station. Through the Figure 4.42, it can be reached that the influence of the WL records is valid over the almost ranges of RF records of upstream excluding Batang Kali and Kerling stations, meaning that the MLP-6a model can be used as a hydrological sound tool in investigating the influence of RF on SF.

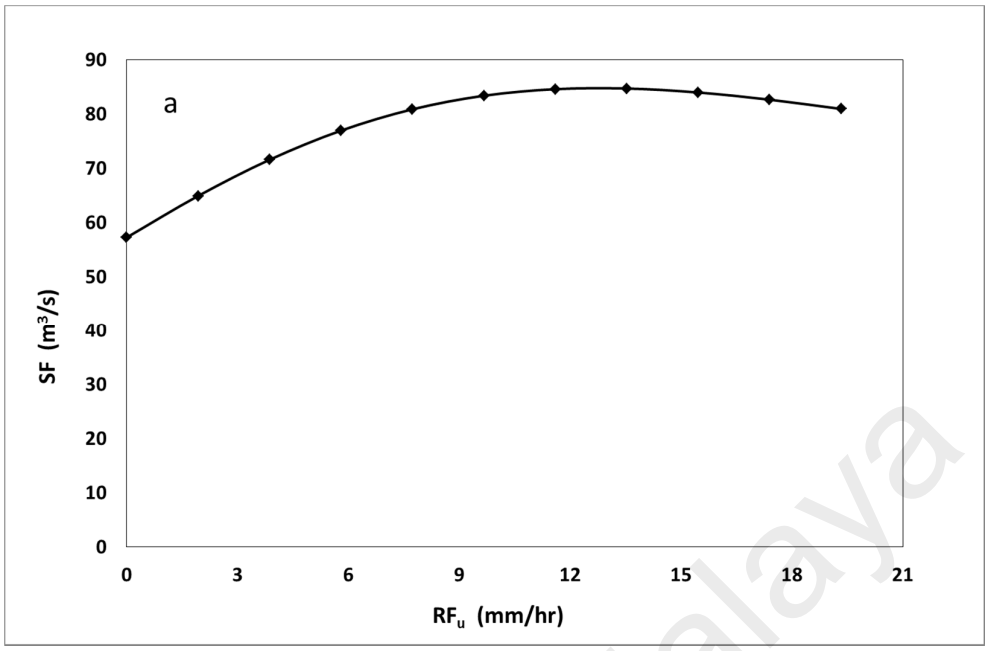
In high RF events of Batang Kali and Kerling stations, the MLP-6a model failed to investigate the effect of RF on the SF. There are many potential reasons to justify this behavior of the model, such as non-enough modelling cases of the high RF events and the potential errors that may be found in the modelling data.

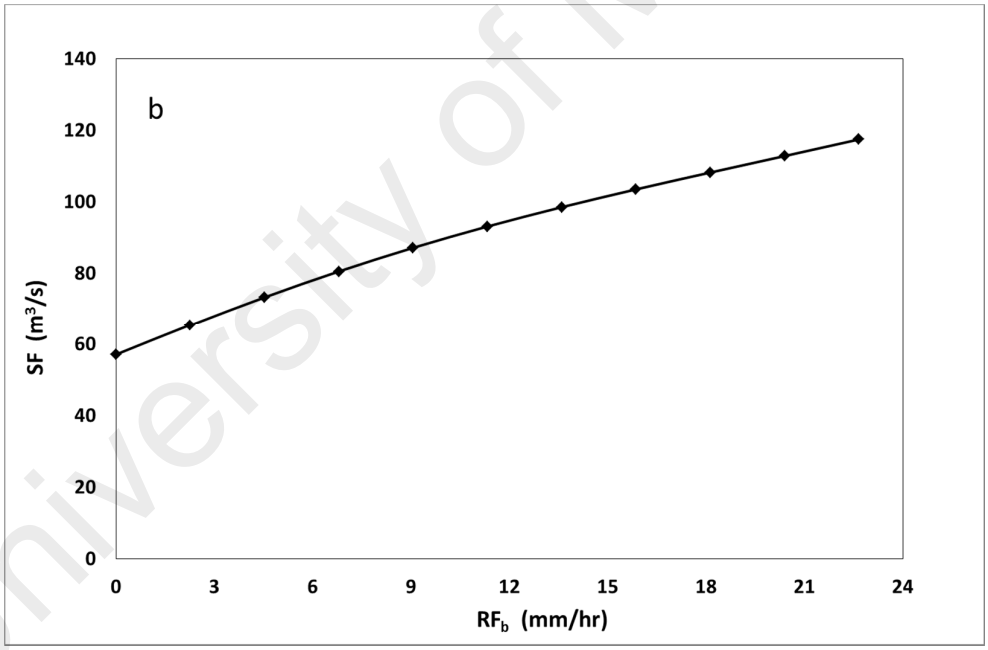
Table 4.26: Results of the hypothetical cases to investigate the influence of Rainfall on Stream Flow

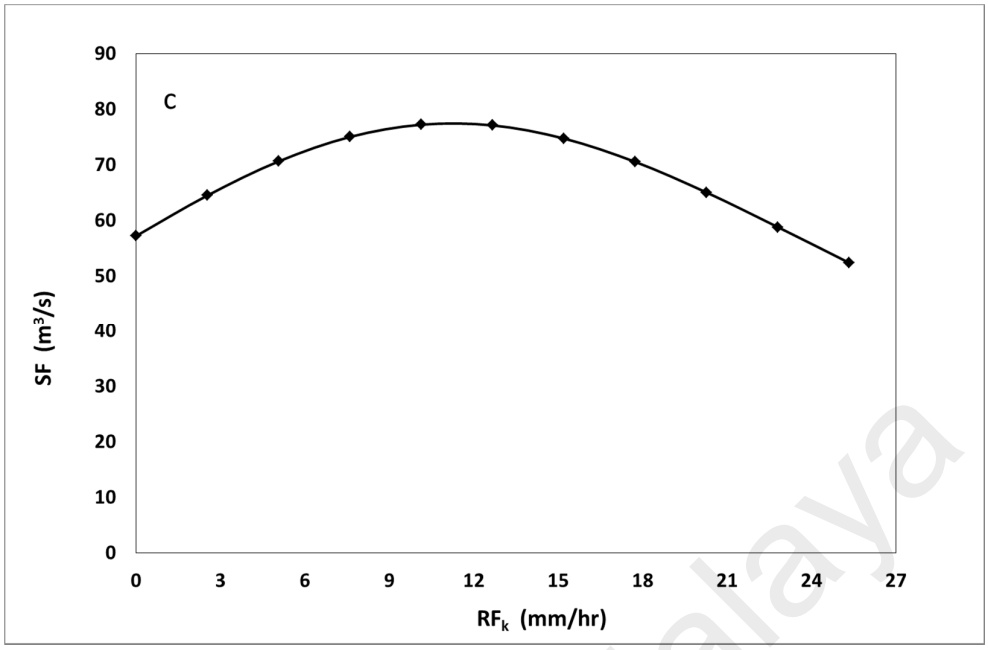
Ulu Yam station									
$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{(t+13)}$
60.34	32.24	32.42	44.18	50.16	0.00	0.24	0.25	0.24	57.27
60.34	32.24	32.42	44.18	50.16	1.93	0.24	0.25	0.24	64.92
60.34	32.24	32.42	44.18	50.16	3.87	0.24	0.25	0.24	71.59
60.34	32.24	32.42	44.18	50.16	5.80	0.24	0.25	0.24	76.96
60.34	32.24	32.42	44.18	50.16	7.73	0.24	0.25	0.24	80.88
60.34	32.24	32.42	44.18	50.16	9.67	0.24	0.25	0.24	83.38
60.34	32.24	32.42	44.18	50.16	11.60	0.24	0.25	0.24	84.59
60.34	32.24	32.42	44.18	50.16	13.53	0.24	0.25	0.24	84.71
60.34	32.24	32.42	44.18	50.16	15.47	0.24	0.25	0.24	83.98
60.34	32.24	32.42	44.18	50.16	17.40	0.24	0.25	0.24	82.65
60.34	32.24	32.42	44.18	50.16	19.33	0.24	0.25	0.24	80.95
Batang Kali station									
60.34	32.24	32.42	44.18	50.16	0.16	0.00	0.25	0.24	57.04
60.34	32.24	32.42	44.18	50.16	0.16	2.27	0.25	0.24	65.36
60.34	32.24	32.42	44.18	50.16	0.16	4.53	0.25	0.24	73.25
60.34	32.24	32.42	44.18	50.16	0.16	6.80	0.25	0.24	80.53
60.34	32.24	32.42	44.18	50.16	0.16	9.07	0.25	0.24	87.13
60.34	32.24	32.42	44.18	50.16	0.16	11.33	0.25	0.24	93.09
60.34	32.24	32.42	44.18	50.16	0.16	13.60	0.25	0.24	98.49
60.34	32.24	32.42	44.18	50.16	0.16	15.87	0.25	0.24	103.47
60.34	32.24	32.42	44.18	50.16	0.16	18.13	0.25	0.24	108.19
60.34	32.24	32.42	44.18	50.16	0.16	20.40	0.25	0.24	112.79
60.34	32.24	32.42	44.18	50.16	0.16	22.67	0.25	0.24	117.45
Kerling station									
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.00	0.24	57.17
60.34	32.24	32.42	44.18	50.16	0.16	0.24	2.53	0.24	64.46
60.34	32.24	32.42	44.18	50.16	0.16	0.24	5.07	0.24	70.60

Table 4.26 Continued,

$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{(t+13)}$
60.34	32.24	32.42	44.18	50.16	0.16	0.24	7.60	0.24	74.99
60.34	32.24	32.42	44.18	50.16	0.16	0.24	10.13	0.24	77.20
60.34	32.24	32.42	44.18	50.16	0.16	0.24	12.67	0.24	77.06
60.34	32.24	32.42	44.18	50.16	0.16	0.24	15.20	0.24	74.70
60.34	32.24	32.42	44.18	50.16	0.16	0.24	17.73	0.24	70.50
60.34	32.24	32.42	44.18	50.16	0.16	0.24	20.27	0.24	65.00
60.34	32.24	32.42	44.18	50.16	0.16	0.24	22.80	0.24	58.78
60.34	32.24	32.42	44.18	50.16	0.16	0.24	25.33	0.24	52.36
Ampang Pecah station									
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	0.00	56.75
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	2.80	69.78
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	5.60	81.34
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	8.40	90.44
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	11.20	96.60
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	14.00	99.81
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	16.80	100.33
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	19.60	98.63
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	22.40	95.21
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	25.20	90.59
60.34	32.24	32.42	44.18	50.16	0.16	0.24	0.25	28.00	85.24







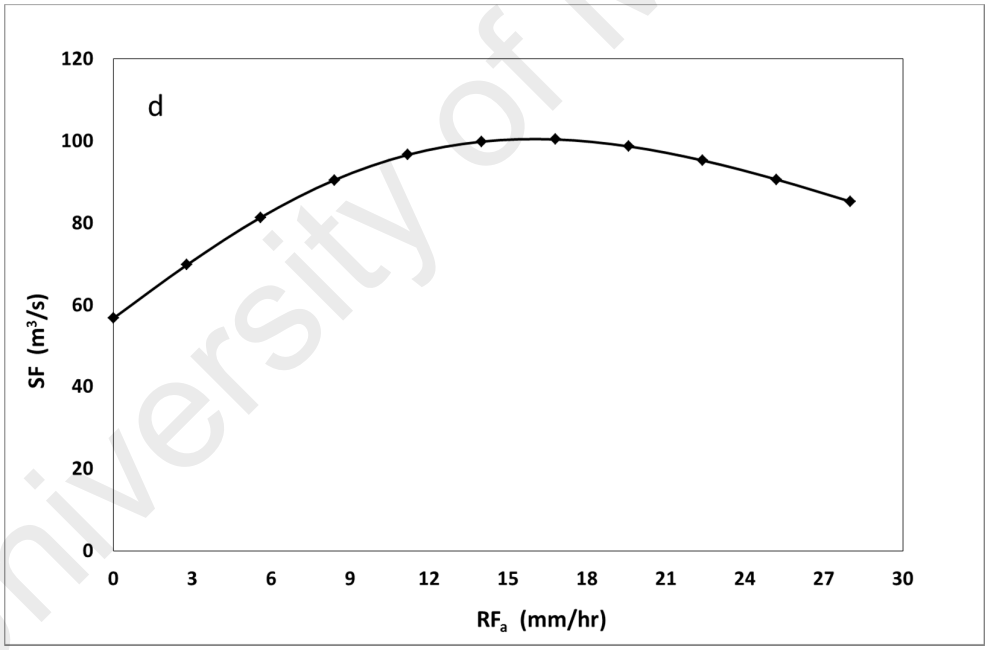


Figure 4.42: Influence of rainfall variables in the stream flow: (a) Ulu Yam station (b) Batang Kali station (c) Kerling station (d) Ampang Pecah station

4.5.2 Utilizing AI-based Models to Estimate the Missing Stream Flow Records

The AI-based models were utilized to estimate the missing records of Q, which could be lost because of some errors or was unread from the beginning. Suppose the 24 hourly records of Q on August 28, 2010, have been lost or have not been read at that time. These records have not entered the developed model before, and the results are considered a real indication of the performance of the MLP-6a model.

The MLP-6a model was used to model these cases and estimate their Q. Table 4.27 presents the records of Q on August 28, 2010, and the predicted results using the MLP-6a model. Figure 4.43 shows the comparison between the observed and predicted Q via the MLP-6a model. The observed Q were compared with the predicted to check the capability of the MLP-6a model in estimating lost Q. A good agreement between the observed and estimated SF was observed, meaning that the MLP-6a model can be used in estimating the missing records of Q.

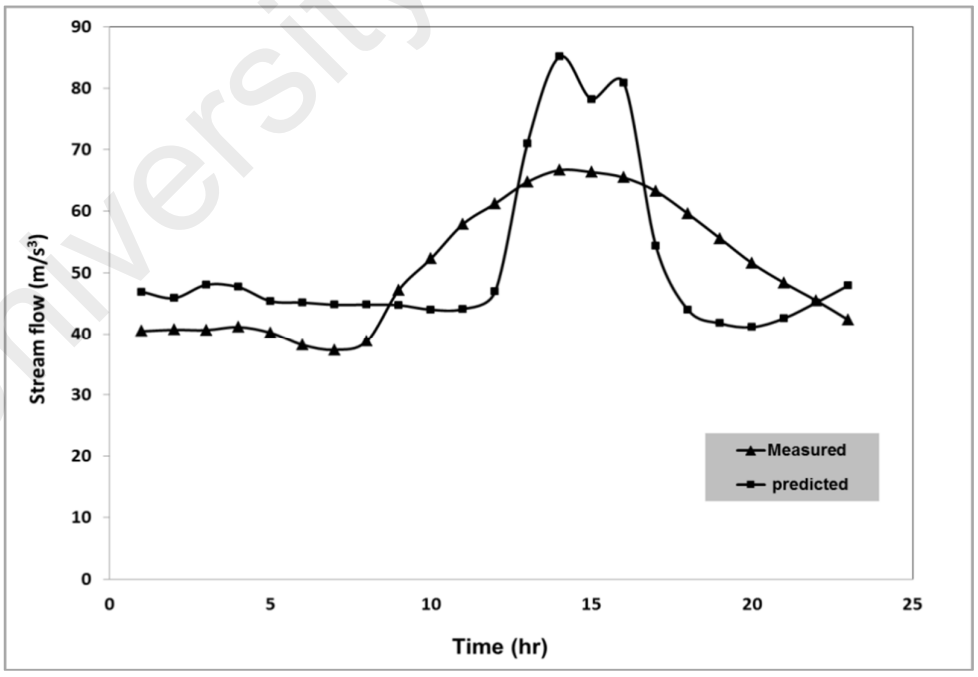


Figure 4.43: Comparison between the observed and preidected Q by MLP-6a model on 28th August 2010

Table 4.27: Observed and predicted Q by MLP-6a model on 28th August 2010

$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{o\ (t+13)}$	$Q_{p\ (t+13)}$
32.33	32.62	32.84	44.23	50.21	0.60	0.00	0.80	0.00	40.46	46.88
32.56	32.62	32.84	44.22	50.15	0.60	0.00	0.80	0.00	40.69	45.86
31.88	32.60	32.75	44.26	50.29	0.60	0.00	0.80	0.00	40.63	48.03
32.11	32.61	32.81	44.24	50.27	0.60	0.00	0.80	0.00	41.13	47.66
32.79	32.63	32.84	44.21	50.10	0.60	0.00	0.80	0.00	40.20	45.38
33.02	32.64	32.83	44.21	50.09	0.60	0.00	0.80	0.00	38.14	45.14
33.25	32.64	32.81	44.20	50.08	0.60	0.00	0.80	0.00	37.33	44.78
33.47	32.65	32.77	44.20	50.08	0.60	0.00	0.80	0.00	38.76	44.79
34.13	32.69	32.73	44.18	50.08	0.60	0.00	0.80	0.00	47.21	44.70
36.03	32.68	32.72	44.17	50.08	0.43	0.00	0.53	0.00	52.29	43.97
38.17	32.66	32.72	44.18	50.08	0.23	0.00	0.27	0.00	57.91	44.09
39.67	32.65	32.78	44.23	50.11	0.03	0.00	0.00	0.20	61.24	46.98
40.63	32.74	32.94	44.26	50.16	0.00	0.00	3.00	3.03	64.74	70.95
41.13	32.94	33.08	44.28	50.18	4.97	0.00	9.00	6.20	66.66	85.09
40.46	33.00	33.12	44.24	50.17	6.67	0.00	9.00	6.13	66.34	78.19
40.69	33.02	33.02	44.23	50.12	6.67	0.00	6.00	3.30	65.47	80.90
40.20	32.94	32.93	44.20	50.10	1.70	0.00	0.00	0.13	63.26	54.37
38.14	32.91	32.86	44.17	50.09	0.00	0.00	0.00	0.00	59.58	43.95
37.33	32.79	32.82	44.14	50.09	0.00	0.00	0.00	0.00	55.57	41.74
38.76	32.68	32.78	44.12	50.09	0.00	0.00	0.00	0.00	51.53	41.13
42.11	32.62	32.75	44.12	50.09	0.00	0.00	0.00	0.00	48.32	42.58
47.21	32.63	32.73	44.11	50.09	0.00	0.00	0.00	0.00	45.44	45.15
52.29	32.61	32.70	44.11	50.09	0.00	0.00	0.00	0.00	42.29	47.94

4.5.3 Utilizing AI based Model in the Early Warning of High Stream Flow Events

Based on the results of previous sections, it has been verified that the SF is affected in direct relationship with WR and RF. MLP-6a model was employed to predict the potential high SF events and determine the hydrological conditions lead to form these events. The high SF events were investigated through three levels: danger level when the SF is higher than 250 m^3/s; warning level when the SF is above 180 m^3/s; and alert level when the SF is more than 160 m^3/s. The three levels are just indictors of high SF events, they have been determined by the Department of Irrigation and Drainage Malaysia (DID).

Six scenarios have been arranged through hypothetical cases of input variables to investigate how the RF and WL in upstream stations can produce high SF events in downstream area and to investigate how the changes of RF intensity lead to formation of three levels of high SF events (i.e. Alert, Warning, and Danger levels). The initial SF was changed for each scenario in order to include wide range of hydrological situations in this study.

The hypothetical cases of the six scenarios were prepared through gradual changing of input RF variables values' from minimum value to maximum value within the range of MLP-6a validity which is start from 0 mm/hr to about 20 mm/hr of Batang Kali station, from 0 mm/hr to about 15 mm/hr of Ulu Yam station and Ampang Pecah station while Kerling station is start from 0 mm/hr to about 12 mm/hr.

In all the scenarios, the investigation process was performed via three situations (i.e., A, B and C) related to the existing saturation of the river basin, which was represented by the WL of the upstream stations. The WL values in situation (A) were assumed to be stable around the average value. The values of the Ulu Yam, Batang Kali, Kerling, and Ampang Pecah stations were 32.2, 32.4, 44.2, and 50.2 m, respectively. Meanwhile, the WL values in situation (B) were assumed to be stable around the average value plus SD,

and the values of the Ulu Yam, Batang Kali, Kerling, and Ampang Pecah stations were 32.73, 33.20, 44.30, and 50.31 m, respectively. The WL values in situation (C) were assumed to be stable around the average plus double SD, and the values of the Ulu Yam, Batang Kali, Kerling and Ampang Pecah stations were 33.21, 33.98, 44.42, and 50.47 m, respectively. Every situation is represented by a set of 11 hypothetical cases to provide a sufficient number of results in the investigation process.

4.5.3.1 Scenario No. 1

The initial Q in the Rantau Panjang downstream station was assumed to be stable around the average value of Q, which is 60 m^3/s. The RF in the Batang Kali station was selected to investigate the role of the changes of RF intensity in the formation of the high SF events in the downstream area. The RF intensity in the Batang Kali station gradually increased from the minimum to the maximum value over the validity range of the model from 0 mm/h to 20 mm/h. The RF of the other upstream stations was assumed to be stable around their average plus the double SD of each RF station.

The hypothetical cases of scenario No. 1 and the predicted Q are shown in Table 4.28. The predicted Q of the three situations of scenario No. 1 is shown in Figure 4.44. The figure shows that increasing the RF of the Batang Kali station from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 1 is 60 m^3/s for all situations. The predicted Q in situation A increases from 78.80 m^3/s to 138.33 m^3/s by increasing the RF intensity in the Batang Kali station from 0 mm/h to 20 mm/h, whereas in situation B, it increases from 93.12 m^3/s to 140.51 m^3/s. In situation C, it increases from 110.86 m^3/s to 149.50 m^3/s.

The predicted Q is still in the normal level zone for all situations of scenario No. 1. The Q in the high RF intensity is near the alert level but does not reach it within the range of

the MLP-6a validity, which starts from 0 mm/hr to approximately 20 mm/hr of the Batang Kali station.

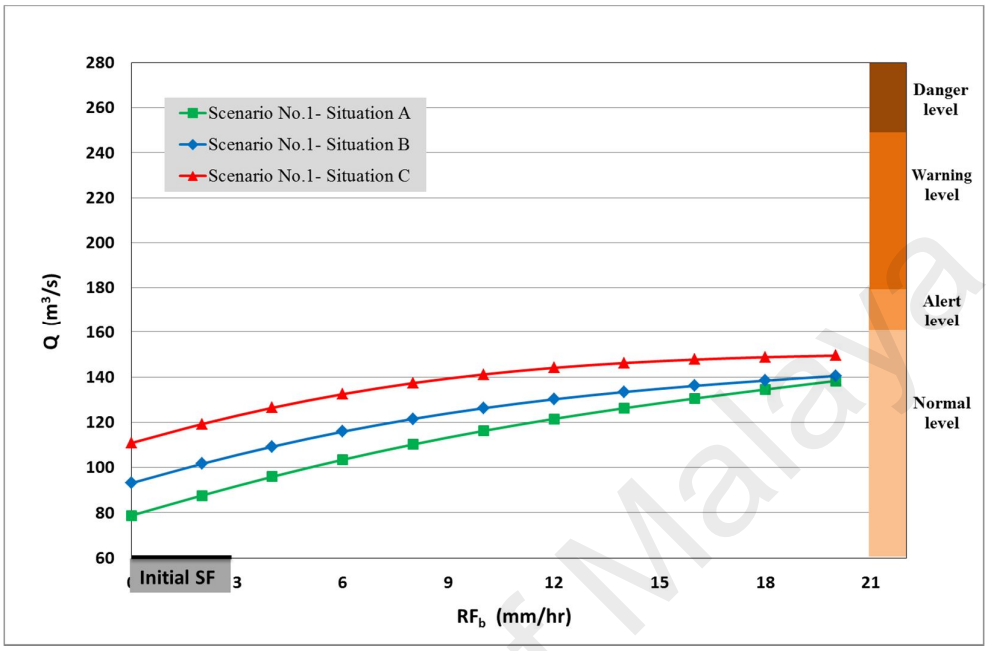


Figure 4.44: Predicted hourly stream flow of scenario No. 1

Table 4.28: Results of scenario No. 1

Scenario No.1 Situation A									
$Q_{(t)}$	Wl_u	Wl_b	Wl_k	Wl_a	Rf_u	Rf_b	Rf_k	Rf_a	$Q_{p\,(t+13)}$
60.00	32.2	32.4	44.2	50.2	1.62	**0.00**	2.36	2.40	78.80
60.00	32.2	32.4	44.2	50.2	1.62	**2.00**	2.36	2.40	87.66
60.00	32.2	32.4	44.2	50.2	1.62	**4.00**	2.36	2.40	95.93
60.00	32.2	32.4	44.2	50.2	1.62	**6.00**	2.36	2.40	103.47
60.00	32.2	32.4	44.2	50.2	1.62	**8.00**	2.36	2.40	110.23
60.00	32.2	32.4	44.2	50.2	1.62	**10.00**	2.36	2.40	116.22
60.00	32.2	32.4	44.2	50.2	1.62	**12.00**	2.36	2.40	121.54
60.00	32.2	32.4	44.2	50.2	1.62	**14.00**	2.36	2.40	126.27
60.00	32.2	32.4	44.2	50.2	1.62	**16.00**	2.36	2.40	130.56
60.00	32.2	32.4	44.2	50.2	1.62	**18.00**	2.36	2.40	134.54
60.00	32.2	32.4	44.2	50.2	1.62	**20.00**	2.36	2.40	138.33
Scenario No.1 Situation B									
60.00	32.73	33.20	44.30	50.31	1.62	**0.00**	2.36	2.40	93.12
60.00	32.73	33.20	44.30	50.31	1.62	**2.00**	2.36	2.40	101.63
60.00	32.73	33.20	44.30	50.31	1.62	**4.00**	2.36	2.40	109.25
60.00	32.73	33.20	44.30	50.31	1.62	**6.00**	2.36	2.40	115.88
60.00	32.73	33.20	44.30	50.31	1.62	**8.00**	2.36	2.40	121.53
60.00	32.73	33.20	44.30	50.31	1.62	**10.00**	2.36	2.40	126.26
60.00	32.73	33.20	44.30	50.31	1.62	**12.00**	2.36	2.40	130.18
60.00	32.73	33.20	44.30	50.31	1.62	**14.00**	2.36	2.40	133.42
60.00	32.73	33.20	44.30	50.31	1.62	**16.00**	2.36	2.40	136.13
60.00	32.73	33.20	44.30	50.31	1.62	**18.00**	2.36	2.40	138.44
60.00	32.73	33.20	44.30	50.31	1.62	**20.00**	2.36	2.40	140.51
Scenario No.1 Situation C									
60.00	33.21	33.98	44.42	50.47	1.62	**0.00**	2.36	2.40	110.86
60.00	33.21	33.98	44.42	50.47	1.62	**2.00**	2.36	2.40	119.21
60.00	33.21	33.98	44.42	50.47	1.62	**4.00**	2.36	2.40	126.43
60.00	33.21	33.98	44.42	50.47	1.62	**6.00**	2.36	2.40	132.46
60.00	33.21	33.98	44.42	50.47	1.62	**8.00**	2.36	2.40	137.35
60.00	33.21	33.98	44.42	50.47	1.62	**10.00**	2.36	2.40	141.19
60.00	33.21	33.98	44.42	50.47	1.62	**12.00**	2.36	2.40	144.10
60.00	33.21	33.98	44.42	50.47	1.62	**14.00**	2.36	2.40	146.24
60.00	33.21	33.98	44.42	50.47	1.62	**16.00**	2.36	2.40	147.76
60.00	33.21	33.98	44.42	50.47	1.62	**18.00**	2.36	2.40	148.80
60.00	33.21	33.98	44.42	50.47	1.62	**20.00**	2.36	2.40	149.50

4.5.3.2 Scenario No. 2

The same hydrological conditions of scenario No. 1 were employed in the investigation process of scenario No. 2, except for the initial Q in the Rantau Panjang downstream station, which was assumed to be 120 m^3/s.

The hypothetical cases and the predicted Q of scenario No. 2 are shown in Appendix E. The predicted Q of the three situations of scenario No. 2 is shown in Figure 4.45. The figure shows that increasing the RF intensity of the Batang Kali station from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 2 is 120 m^3/s for the three situations. The predicted Q in situation A increases from 127.04 m^3/s to 186.91 m^3/s by increasing the RF intensity in the Batang Kali station from 0 mm/hr to 20 mm/hr whereas, in situation B, it increases from 146.62 m^3/s to 187.46 m^3/s. In situation C, it increases from 160.27 m^3/s to 193.78 m^3/s.

In situation B, the predicted Q approaches the alert level zone at the RF intensity of 8 mm/hr while approaches the warning level at RF 16 mm/hr in the Batang Kali station. In situation B, the Q approaches the alert level at the RF intensity of 4 mm/hr while approaches the warning level at the RF intensity of 13 mm/hr. In situation C, the Q approaches the alert level at the RF intensity of 0 mm/hr while approaches the warning level at the RF intensity of 6 mm/hr.

The predicted Q is still in warning level zone in the high RF intensity, for all situations of scenario No. 2 and doesn't approach the danger level within the range of MLP-6a validity.

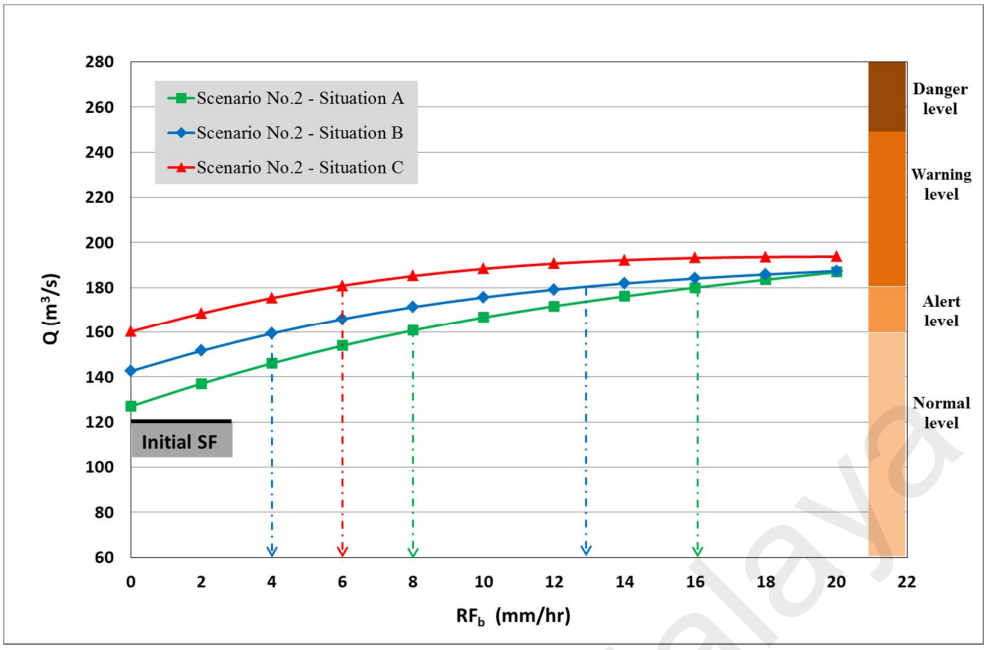


Figure 4.45: Predicted hourly stream flow of scenario No. 2

4.5.3.3 Scenario No. 3

The same hydrological conditions of scenario No. 1 were employed in the investigation process of scenario No. 3, except for the initial Q in the Rantau Panjang downstream station, which was assumed to be 160 m^3/s.

The hypothetical cases of scenario No. 3 and the predicted Q are shown in Appendix E. The predicted Q of the three situations of scenario No. 3 is shown in Figure 4.46. This figure shows that increasing the RF intensity of the Batang Kali station from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 3 is 160 m^3/s for the three situations. The predicted Q in situation A increases from 171.46 m^3/s to 212.32 m^3/s by increasing the RF intensity in the Batang Kali station from 0 mm/hr to 20 mm/hr whereas, in situation B, it increases

from 171.46 m^3/s to 212.32 m^3/s. In situation C, it increases from 187.27 m^3/s to 216.78 m^3/s.

In situation A, the Q approaches the warning level at the RF intensity of 5 mm/hr, whereas in situation B, the Q approaches the warning level at the RF intensity of 2 mm/hr. In situation C, the Q approaches in the warning zone at the RF intensity of 0 mm/hr.

The predicted Q is still in warning level zone for the most range of RF intensity, for all situations of scenario No. 3 and doesn't approach the danger level within the range of MLP-6a validity.

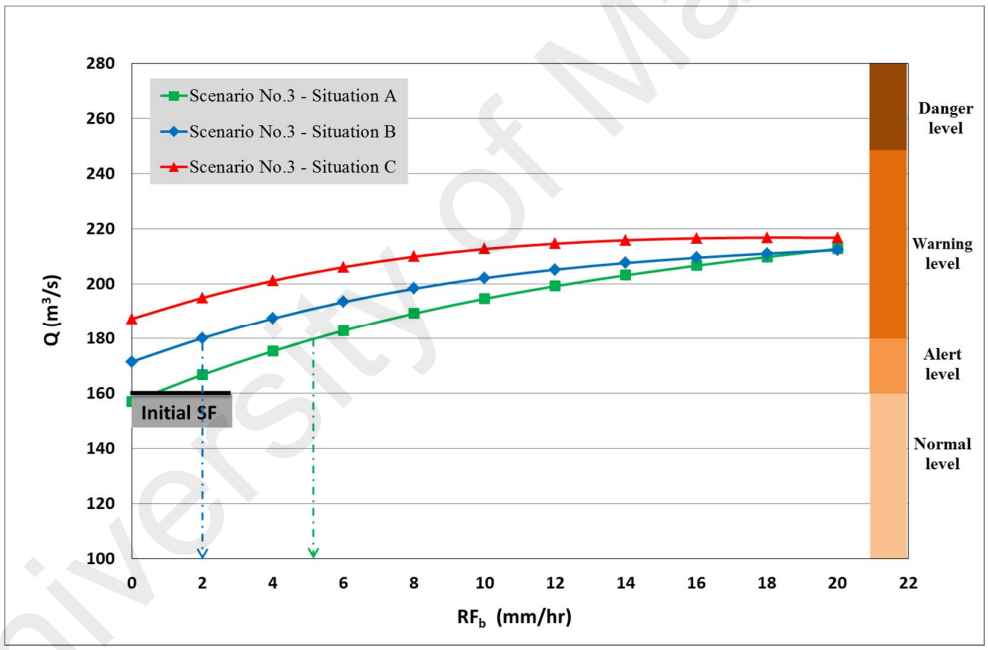


Figure 4.46: Predicted hourly stream flow of scenario No. 3

4.5.3.4 Scenario No. 4

The same hydrological conditions of scenario No. 1 were employed in the investigation process of scenario No. 4, except for the initial Q in the Rantau Panjang downstream station, which was assumed to be 180 m^3/s.

The hypothetical cases and the predicted Q of scenario No. 4 are shown in Appendix E. The predicted Q of the three situations of scenario No. 4 is shown in Figure 4.47. This figure shows that increasing RF intensity of Batang Kali station from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 4 is 180 m^3/s for the three situations. The predicted Q in situation A increases from 170.16 m^3/s to 223.55 m^3/s by increasing the RF intensity in the Batang Kali station from 0 m/hr to 20 mm/hr, whereas in situation B, it increases from 183.99 m^3/s to 222.73 m^3/s. In situation C, it increases from 198.64 m^3/s to 226.36 m^3/s.

In situation A, the Q approaches the warning level at the RF intensity of 2 mm/hr, whereas in situation B and situation C, the Q approaches the warning zone from the RF intensity of 0 mm/hr. The predicted Q is still in warning level zone for the most range of RF intensity, for all situations of scenario No. 4 and does not approach the danger level within the range of MLP-6a validity.

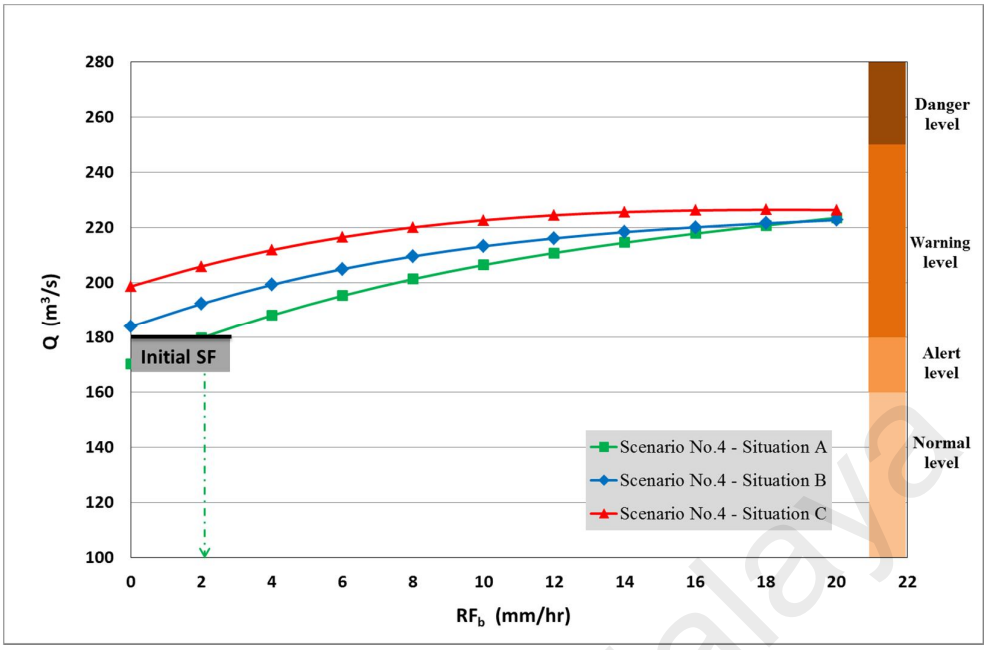


Figure 4.47: Predicted hourly stream flow of scenario No. 4

4.5.3.5 Scenario No. 5

The same hydrological conditions of scenario No. 1 were employed in the investigation process of scenario No. 5, except for the initial Q in the Rantau Panjang downstream station, which was assumed to be 200 m^3/s.

The hypothetical cases of scenario No. 5 and the predicted Q are shown in Appendix E. The predicted Q of the three situations of scenario No. 4 is shown in Figure 4.48. The figure shows that increasing RF intensity of Batang Kali station from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 5 is 200 m^3/s for the three situations. The predicted Q in situation A, increases from 208.67 m^3/s to 231.90 m^3/s by increasing the RF intensity in the Batang Kali station from 0 mm/hr to 20 mm/hr whereas, in situation B, it increases

from 195.20 m^3/s to 233.07 m^3/s. In situation C, it increases from 182.29 m^3/s to 234.79 m^3/s.

The predicted Q approaches the warning zone in the 3 situations of scenario No. 5 at the RF intensity of 0 mm/hr. The predicted Q is still in the warning level zone and does not approach the danger level for all situations. The Q in the high RF intensity is near the danger level but doesn't reach it within the range of MLP-6a validity which starts from 0 mm/hr to about 20 mm/hr of Batang Kali station.

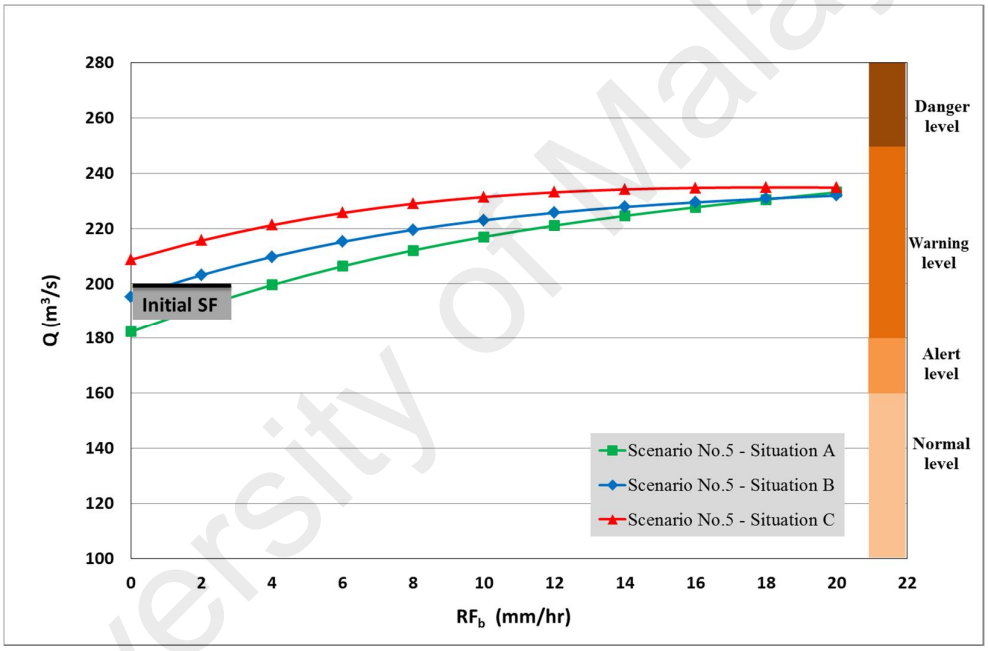


Figure 4.48: Predicted hourly stream flow of scenario No. 5

4.5.3.6 Scenario No. 6

Given that the Q did not approach the danger level in all previous scenarios, which mainly focus on changing the RF in only one station, new hydrological conditions have been assumed to detect the conditions that may lead Q to approach the danger level.

The same hydrological conditions in scenario No. 5 were used in the investigation process of scenario No. 6, except for the RF intensity in the Kerling station, which gradually increased from 0 mm/hr to 20 mm/hr, such as the RF intensity in the Batang Kali station. Only situation C was applied in this scenario because it provided the highest Q through the scenario No. 5.

The hypothetical cases of scenario No. 6 and the predicted Q are shown in Appendix E. The predicted Q of scenario No. 6 is shown in Figure 4.49. The figure shows that increasing the RF intensity of both the Batang Kali and Kerling stations from 0 mm/hr to 20 mm/hr results in a significant influence on the predicted Q.

The initial Q in scenario No. 6 is 200 m^3/s. The predicted Q increases from 205.18 m^3/s to 250.85 m^3/s by increasing the RF intensity of the Batang Kali and Kerling stations from 0 mm/h to 20 mm/h. The predicted Q approaches the danger level at around the RF intensity of 20 mm/hr of both the Batang Kali and Kerling stations.

Based on the results of scenario No. 6 and the previous scenarios, reaching the danger level in the downstream area required high RF intensity in at least two upstream stations. The danger levels cannot be reached if the high RF intensity is present in only one upstream station.

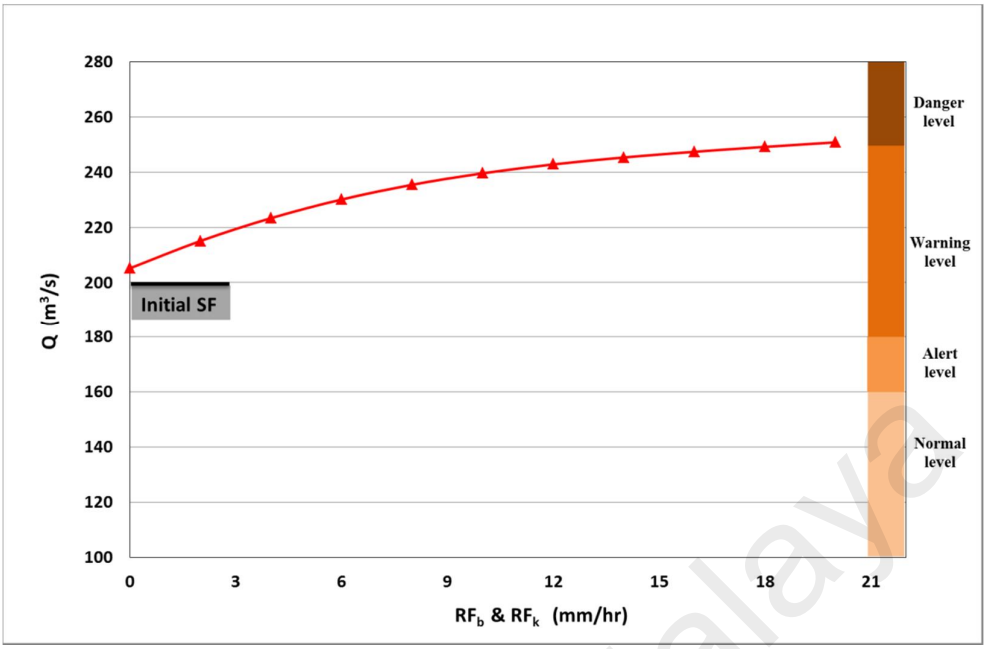


Figure 4.49: Predicted hourly stream flow of scenario No. 6

4.6 General Discussion about the AI-based Models and its' Applications

This section provides a general discussion about the AI-based models and it applications which were covered in sections 4.4 and 4.5. The ability of the AI techniques for Q prediction in the downstream area from the upstream WL and RF records in the humid tropical area was successfully explored. A total of six AI-based models with different combinations of input variables were developed using four AI techniques: MLP, RBF, GRNN and SVM through two modelling phases.

The modelling process was performed in two phases. First, the results of the Lt estimated by CCA were applied to select the lag intervals between the input and output variables of AI-based models. The results of HGA were then applied in the second phase of the modelling process. The two phases of the modelling process were performed to explore the ability of improving the performance of AI-based models by the accurate selection of the lag intervals between the model variables based on the accurate estimation of the Lt.

The hourly records of Q, RF, and WL for a one-year period (2011) were applied to train and test the AI-based models. The hourly records of WL and RF in the upstream stations were used as input variables (independent variables), whereas the Q data in the downstream station were used as output variable (independent variables) of the AI-based models. The model performances were assessed based on three performance evaluation criteria: R, R^2, and MAE.

Based on the performance evaluation of the developed models, M6 from the first modelling phase and M6a from the second modelling phase achieved the best performance among all the models. The correlation between the observed and predicted Q of all the developed models, particularly by the M6 and M6a models, appears to be consistent for both training and testing data sets.

The performance of the four AI techniques is compared by analyzing the performance of the best-fitting model for predicting the Q, which is determined based on the performance evaluation of the testing data sets. The results suggest that the SVM is superior to the ANNs in predicting Q.

The developed AI-based models were then successfully utilized as prediction and analytical tools to investigate the influence of the input variables on Q. The results of the investigation of the influence of all WL and RF on Q proved that the RF and WL of the upstream stations directly affected the Q in the Rantau Panjang station (downstream station).

The AI-based models were also utilized to estimate the missing Q records. The records of Q on August 28, 2010, which were not entered into the developed models before, were predicted by the AI-based models to check the capability of the developed models to estimate Q. A good agreement between the observed (hidden records) and estimated Q was observed.

Finally, the AI-based models were applied in early warning of upcoming high SF events. Six scenarios were arranged through the hypothetical cases of the input variables to investigate how the RF and WL in the upstream stations can affect the SF in the downstream area and how the changes in RF intensity lead to the formation of high SF events, as represented by the three levels (i.e., alert, warning, and danger). Based on the results of the studied scenarios, reaching the danger level in the downstream area required high RF intensity in at least two upstream stations. The danger levels cannot be reached if the high RF intensity is present in only one upstream station.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Many conclusions, recommendations and suggestions for future work related to the research topic were obtained based on the results of this study.

5.2 Conclusions

The following conclusions were obtained based on the results of this study:

1. A better understanding of the river basin hydrology is the key to improving the performance of the AI-based models in predicting SF. Investigating the long-term changes in SF regimes and the Lt estimation was implemented in this study to enhance the understanding of the river basin hydrology, and their results were included in the SF modelling process to improve the prediction performance of Q through the accurate timing of the input and output variables of the AI-based models.

2. The analysis of the long-term variations in the SF regime in the Selangor River basin includes an investigation of the variations in nine hydrological variables that describe the yearly SF and the variations in the monthly SF, as well as the variations in the yearly duration of high and low SF over a 50-year period from 1961 to 2010. Apparent changes were observed through the analysis of the long-term variations. The results verified the existence of long-term variations in the SF regime that may result in the formation of appropriate hydrological conditions to increase the occurrence probability of flood and drought events in the future.

3. The Lt between the upstream and downstream stations was estimated using three approaches namely, empirical formulas, CCA and NGA. The estimated Lt by HGA was applied in deriving the two empirical formulas to estimate the Lt between the RF upstream and downstream stations. The derived empirical formulas significantly simplify the Lt estimation process by a quick and easy approach that is directly based on the RF and SF records without the necessity of identifying the full description of all the parameters that affect the Lt. The HGA is applicable for all humid tropical rivers, whereas the derived empirical formulas are applicable only for the Selangor River basin, but they can be modified for other humid tropical rivers.

4. The ability of the three techniques of ANNs (i.e. MLP, RBF, and GRNN) along with SVM for real-time Q prediction in the downstream area from the upstream WL and RF records in the Selangor River basin was explored and successfully achieved with high performance throughout two modelling phases. In the first phase, the estimated Lt by CCA was employed to select the lag intervals between the input and output variables of the AI-based models, whereas the estimated Lt by HGA was applied in the second phase. The two phases of the modelling process were used to explore the ability of improving the performance of the AI-based models through the accurate timing of variables of the AI-based models depending on the Lt estimation. In the first phase, six models with different combinations of input variables were trained and developed by four AI techniques resulting in the development of 24 AI-based models to predict the Q. In the second phase, only two models, those achieved the highest R among the six models of the first phase were selected, resulting in the development of eight. The total number of developed AI-based models in the two modelling phases is 32.

5. The performance evaluation of the developed AI-based models was assessed based on the three performance evaluation criteria: R, R^2, and MAE. It shows that high R

was reached for most of the developed models. The results show the ability of the accurate timing of the variables of the AI-based in enhancing the performance of the AI-based models, thereby improving the performance of the Q prediction. High agreement between the observed and predicted Q was observed also for most of the developed models. The results suggested that SVM is superior to ANNs in predicting Q given the R values between the observed and the predicted Q by the SVM–M6 model are 0.992 and 0.953, whereas the MAE values are 0.061 and 0.253 for the training and testing data sets, respectively. The achieved values of the R and MAE of SVM are generally better than those of the three ANNs techniques.

6. The developed AI-based models were successfully employed in many hydrological applications, such as prediction tools to predict the future Q and as analytical tools to investigate the influence of the RF and WL on Q. They were also employed in estimation of the missing records of Q. Furthermore, they were employed in flood early warning through the advance detection of hydrological conditions that may lead to the formation of floods via six hydrological scenarios which were arranged to select the hydrological conditions that may lead to the formations of floods. According to results of applications, it can be concluded that AI-based models are beneficial tool to the local authorities for flood control and awareness.

7. To the best of the researcher's knowledge, this research can be considered a unique contribution to the field of real-time Q prediction. The significance of this research lies in the uniqueness of the considered process and the novelty of the applied methodology in the modelling process. The integration of the hydrological description of SF in the modelling process and high performance and applicability of the developed AI-based models also have an immense role in enhancing the significance of the research.

5.3 Recommendations

The following recommendations were made based on the results of this study:

1. The analysis of the long-term changes of the SF regime in the Selangor River basin is a key toward achieving an extensive knowledge of the changes in the SF regime of this river. The outcome of exploring the long-term variations in the SF regime promotes the awareness of the demand for a better understanding of the Selangor River hydrology and draws attention to the necessity of the development of water resource management systems, considering the increasing probability of future droughts. Awareness is also drawn to improving the flood protection plans in response to the increasing possibility of future flood events to prevent the negative impacts that may result from the probable variations in the SF regime.

2. Further studies should be directed to develop and improve the performance of HGA and the empirical formulas to estimate the Lt between the upstream and downstream stations in the Selangor River basin and tropical humid rivers. The new studies mainly start with improving the availability and quality of the required hydrological data, which is one of the main challenges in this research. HGA and the derived empirical formulas have the potential to be employed in many future hydrological applications, especially those related to surface water hydrology and river systems.

3. Given the high R achieved by ANNs and SVM in real-time Q prediction in the Selangor River basin, which is a paradigm of humid tropical rivers, these techniques can be applied in the Q prediction in other river basins in humid tropical areas, especially in Southeast Asia. The results also offer a starting point to explore new possible hydrological processes in the Selangor River basin, such as RF, water quality, and sedimentation to be modelled and predicted using ANNs and SVM.

4. Although the ANNs and SVM performed well in real-time Q prediction, higher R in the Selangor River basin can be investigated by employing other AI techniques, such as FRBSs and GAs in Q prediction. The role of Lt estimation in achieving high R between the observed and predicted Q by ANNs and SVM offers enough motivation to explore possible improvements by employing the Lt estimation in real-time Q modelling process using other AI techniques.

5. The role of Lt estimation in achieving high R between the observed and predicted Q offers a starting point to explore the effect of integration Lt estimation in modelling and prediction of new possible hydrological processes, such as RF, water quality, and sedimentation in Selangor River basin or other river basins in humid tropical areas.

REFERENCES

Abraham, A. (2005). Artificial Neural Networks. In P. H. S. a. R. Thorn (Ed.), *Handbook of measuring system design.* (pp. 901-908). London: John Wiley and Sons.

Abrahart, R., Kneale, P. E., & See, L. M. (2004). *Neural networks for hydrological modeling*: Taylor & Francis.

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., . . . Wilby, R. L. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography, 36*(4), 480-513. doi: 10.1177/0309133312444943

Abrahart, R. J., Heppenstall, A. J., & See, L. M. (2007). Timing error correction procedure applied to neural network rainfall—runoff modelling. *Hydrological Sciences Journal, 52*(3), 414-431. doi: 10.1623/hysj.52.3.414

Adamowski, J., & Sun, K. (2010). Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *Journal of Hydrology, 390*(1–2), 85-91. doi: 10.1016/j.jhydrol.2010.06.033

Ahmed, J., & Sarma, A. (2007). Artificial neural network model for synthetic streamflow generation. *Water Resources Management, 21*(6), 1015-1029. doi: 10.1007/s11269-006-9070-y

Akhtar, M. K., Corzo, G. A., van Andel, S. J., & Jonoski, A. (2009). River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin. *Hydrol. Earth Syst. Sci., 13*(9), 1607-1618. doi: 10.5194/hess-13-1607-2009

Albrecher, H., Ladoucette, S. A., & Teugels, J. L. (2010). Asymptotics of the sample coefficient of variation and the sample dispersion. *Journal of Statistical Planning and Inference, 140*(2), 358-368. doi: 10.1016/j.jspi.2009.03.026

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology, 517*(0), 913-922. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.035

Allen, J. R. L. (1976). Computational models for dune time-lag: General ideas, difficulties, and early results. *Sedimentary Geology, 15*(1), 1-53. doi: http://dx.doi.org/10.1016/0037-0738(76)90020-8

Ammar, K., McKee, M., & Kaluarachchi, J. (2009). Bayesian method for groundwater quality monitoring network analysis. *Journal of Water Resources Planning and Management, 137*(1), 51-61.

Anctil, F., Perrin, C., & Andréassian, V. (2004). Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. *Environmental Modelling & Software, 19*(4), 357-368. doi: 10.1016/s1364-8152(03)00135-x

Aqil, M., Kita, I., Yano, A., & Nishiyama, S. (2007). Neural Networks for Real Time Catchment Flow Modeling and Prediction. *Water Resources Management, 21*(10), 1781-1796. doi: 10.1007/s11269-006-9127-y

Armaghani, D., Momeni, E., Abad, S., & Khandelwal, M. (2015). Feasibility of ANFIS model for prediction of ground vibrations resulting from quarry blasting. *Environmental Earth Sciences*, 1-16. doi: 10.1007/s12665-015-4305-y

ASCE. (2000). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123.

ASCE. (2000a). Artificial Neural Networks in Hydrology. I: Preliminary Concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123. doi: doi:10.1061/(ASCE)1084-0699(2000)5:2(115)

ASCE. (2000b). Artificial neural networks in hydrology II : preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 124–137.

Asefa, T., Kemblowski, M., McKee, M., & Khalil, A. (2006). Multi-time scale stream flow predictions: The support vector machines approach. *Journal of Hydrology, 318*(1–4), 7-16. doi: http://dx.doi.org/10.1016/j.jhydrol.2005.06.001

Asefa, T., Kemblowski, M., McKee, M., & Khalil, A. (2006). Multi-time scale stream flow predictions: The support vector machines approach. *Journal of Hydrology, 318*(1-4), 7-16.

Askew, A. J. (1970). Derivation of formulae for variable lag time. *Journal of Hydrology, 10*(3), 225-242. doi: http://dx.doi.org/10.1016/0022-1694(70)90251-9

Athira, P., & Sudheer, K. P. (2015). A method to reduce the computational requirement while assessing uncertainty of complex hydrological models. *Stochastic Environmental Research and Risk Assessment, 29*(3), 847-859. doi: 10.1007/s00477-014-0958-4

Bai, J., Lu, R., Xue, A., She, Q., & Shi, Z. (2015). Finite-time stability analysis of discrete-time fuzzy Hopfield neural network. *Neurocomputing, 159*(0), 263-267. doi: http://dx.doi.org/10.1016/j.neucom.2015.01.051

Banasik, K., Madeyski, M., Mitchell, J. K., & Mori, K. (2005). An investigation of lag times for rainfall–runoff–sediment yield events in small river basins/Une analyse des temps de réponse d'événements pluie–débit–transport solide au sein de petits bassins versants. *Hydrological Sciences Journal, 50*(5).

Basheer, I., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43*(1), 3-31.

Baxter, C. W., Smith, D. W., & Stanley, S. J. (2004). A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. *Journal of Environmental Engineering and Science, 3*(S1), S45-S58. doi: 10.1139/s03-081

Beaulieu, C., Seidou, O., Ouarda, T. B. M. J., & Zhang, X. (2009). Intercomparison of homogenization techniques for precipitation data continued: Comparison of two recent Bayesian change point models. *Water Resources Research, 45*(8).

Behzad, M., Asghari, K., Eazi, M., & Palhang, M. (2009). Generalization performance of support vector machines and neural networks in runoff modeling. *Expert Systems with Applications, 36*(4), 7624-7629. doi: http://dx.doi.org/10.1016/j.eswa.2008.09.053

Besaw, L. E., Rizzo, D. M., Bierman, P. R., & Hackett, W. R. (2010). Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology, 386*(1–4), 27-37. doi: http://dx.doi.org/10.1016/j.jhydrol.2010.02.037

Bhadra, A., Bandyopadhyay, A., Singh, R., & Raghuwanshi, N. S. (2010). Rainfall-Runoff Modeling: Comparison of Two Approaches with Different Data Requirements. *Water Resources Management, 24*(1), 37-62. doi: 10.1007/s11269-009-9436-z

Bierkens, M. F. P. (2006). Designing a monitoring network for detecting groundwater pollution with stochastic simulation and a cost model. *Stochastic Environmental Research and Risk Assessment, 20*(5), 335-351. doi: 10.1007/s00477-005-0025-2

Birsan, M.-V., Molnar, P., Burlando, P., & Pfaundler, M. (2005). Streamflow trends in Switzerland. *Journal of Hydrology, 314*(1–4), 312-329. doi: 10.1016/j.jhydrol.2005.06.008

Boik, R. J., & Shirvani, A. (2009). Principal components on coefficient of variation matrices. *Statistical Methodology, 6*(1), 21-46. doi: 10.1016/j.stamet.2008.02.006

Bowden, G. J., Dandy, G. C., & Maier, H. R. (2005). Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology, 301*(1–4), 75-92. doi: http://dx.doi.org/10.1016/j.jhydrol.2004.06.021

Breemen, M. T. J. v. (2008). Salt intrusion in the Selangor Estuary in Malaysia. The Netherlands: University of Twente.

Bronstert, A., de Araújo, J.-C., Batalla, R., Costa, A., Delgado, J., Francke, T., . . . Vericat, D. (2014). Process-based modelling of erosion, sediment transport and reservoir siltation in mesoscale semi-arid catchments. *Journal of Soils and Sediments, 14*(12), 2001-2018. doi: 10.1007/s11368-014-0994-1

Buishand, T. A. (1982). Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology, 58*(1–2), 11-27. doi: 10.1016/0022-1694(82)90066-x

Burn, D. H., Cunderlik, J. M., & Pietroniro, A. (2004). Hydrological trends and variability in the Liard River basin / Tendances hydrologiques et variabilité dans le basin de la rivière Liard. *Hydrological Sciences Journal, 49*(1), 53-67. doi: 10.1623/hysj.49.1.53.53994

Burn, D. H., Sharif, M., & Zhang, K. (2010). Detection of trends in hydrological extremes for Canadian watersheds. *Hydrological Processes, 24*(13), 1781-1790.

Campoli, G., Bolsterlee, B., van der Helm, F., Weinans, H., & Zadpoor, A. A. (2014). Effects of densitometry, material mapping and load estimation uncertainties on the accuracy of patient-specific finite-element models of the scapula. *Journal of the Royal Society Interface, 11*(93), 20131146. doi: 10.1098/rsif.2013.1146

Ch, S., Anand, N., Panigrahi, B. K., & Mathur, S. (2013). Streamflow forecasting by SVM with quantum behaved particle swarm optimization. *Neurocomputing, 101*(0), 18-23. doi: http://dx.doi.org/10.1016/j.neucom.2012.07.017

Chang, F.-J., Chen, P.-A., Lu, Y.-R., Huang, E., & Chang, K.-Y. (2014). Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. *Journal of Hydrology, 517*(0), 836-846. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.013

Charron, C., & Ouarda, T. B. M. J. (2015). Regional low-flow frequency analysis with a recession parameter from a non-linear reservoir model. *Journal of Hydrology, 524*(0), 468-475. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.03.005

Chen, S.-T., & Yu, P.-S. (2007). Pruning of support vector networks on flood forecasting. *Journal of Hydrology, 347*(1–2), 67-78. doi: http://dx.doi.org/10.1016/j.jhydrol.2007.08.029

Chen, W., & Chau, K. (2006). Intelligent manipulation and calibration of parameters for hydrological models. *International journal of environment and pollution, 28*(3), 432-447.

Chen, X., Peng, D., & Gao, S. (2013). SVM-Based Topological Optimization of Tetrahedral Meshes. In X. Jiao & J.-C. Weill (Eds.), *Proceedings of the 21st International Meshing Roundtable* (pp. 211-224): Springer Berlin Heidelberg.

Chen, Z., Chen, Y., & Li, B. (2013). Quantifying the effects of climate variability and human activities on runoff for Kaidu River Basin in arid region of northwest China. *Theoretical and Applied Climatology, 111*(3-4), 537-545. doi: 10.1007/s00704-012-0680-4

Chua, L. O., & Yang, L. (1988). Cellular neural networks: Applications. *IEEE transactions on circuits and systems, 35*(10), 1273-1290. doi: 10.1109/31.7601

Cigizoglu, H. K. (2005). Generalized regression neural network in monthly flow forecasting. *Civil Engineering and Environmental Systems, 22*(2), 71-81. doi: 10.1080/10286600500126256

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., . . . Marks, D. G. (2015). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, n/a-n/a. doi: 10.1002/2015WR017200

Coin, D. (2008). A goodness-of-fit test for normality based on polynomial regression. *Computational Statistics & Data Analysis, 52*(4), 2185-2198. doi: 10.1016/j.csda.2007.07.012

Crout, N., Kokkonen, T., Jakeman, A. J., Norton, J. P., Newham, L. T. H., Anderson, R., . . . Whitfield, P. (2008). Chapter Two Good Modelling Practice. In A. A. V. A. E. R. A.J. Jakeman & S. H. Chen (Eds.), *Developments in Integrated Environmental Assessment* (Vol. Volume 3, pp. 15-31): Elsevier.

Cui, H., & Singh, V. P. (2015). Configurational entropy theory for streamflow forecasting. *Journal of Hydrology*, *521*(0), 1-17. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.11.065

Dai, Q., Han, D., Zhuo, L., Huang, J., Islam, T., & Srivastava, P. K. (2015). Impact of complexity of radar rainfall uncertainty model on flow simulation. *Atmospheric Research*, *161–162*(0), 93-101. doi: http://dx.doi.org/10.1016/j.atmosres.2015.04.002

Damangir, H. (2001). *Dynamic Training of ANN for Its Application in Real-Time Flood Forecasting.* Shiraz University, Shiraz, Iran.

Daniel, E. B., Camp, J. V., LeBoeuf, E. J., Penrod, J. R., Dobbins, J. P., & Abkowitz, M. D. (2011). Watershed Modeling and its Applications: A State-of-the-Art Review. *The Open Hydrology Journal*, *5*, 26-50.

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In T.-M. Choi, C.-L. Hui & Y. Yu (Eds.), *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70): Springer Berlin Heidelberg.

Dawson, C. W., See, L. M., Abrahart, R. J., & Heppenstall, A. J. (2006). Symbiotic adaptive neuro-evolution applied to rainfall-runoff modelling in northern England. *Neural Networks*, *19*(2), 236-247.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, *25*(1), 80-108. doi: 10.1191/030913301674775671

de Vos, N. J., & Rientjes, T. H. M. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrol. Earth Syst. Sci.*, *9*(1/2), 111-126. doi: 10.5194/hess-9-111-2005

Descroix, L., Genthon, P., Amogu, O., Rajot, J.-L., Sighomnou, D., & Vauclin, M. (2012). Change in Sahelian Rivers hydrograph: The case of recent red floods of the Niger River in the Niamey region. *Global and Planetary Change*, *98–99*(0), 18-30. doi: 10.1016/j.gloplacha.2012.07.009

Dibike, Y. B., & Solomatine, D. P. (2001). River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *26*(1), 1-7. doi: http://dx.doi.org/10.1016/S1464-1909(01)85005-X

Ding, S., Huang, H., Xu, X., & Wang, J. (2014). Polynomial Smooth Twin Support Vector Machines. *Applied mathematics & information sciences, 8*(4), 2063-2071.

Elsafi, S. H. (2014). Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. *Alexandria Engineering Journal, 53*(3), 655-662. doi: http://dx.doi.org/10.1016/j.aej.2014.06.010

Fang, X., Cleveland, T., Garcia, C., Thompson, D., & Malla, R. (2005). Literature review on timing parameters for hydrographs (pp. 83): Department of Civil Engineering, Lamar University, Beaumont, Texas.

Fang, X., Thompson, D., Cleveland, T., Pradhan, P., & Malla, R. (2008). Time of Concentration Estimated Using Watershed Parameters Determined by Automated and Manual Methods. *Journal of Irrigation and Drainage Engineering, 134*(2), 202-211. doi: 10.1061/(ASCE)0733-9437(2008)134:2(202)

Figueroa-García, J. C., Ochoa-Rey, C. M., & Avellaneda-González, J. A. (2015). Rule generation of fuzzy logic systems using a self-organized fuzzy neural network. *Neurocomputing, 151, Part 3*(0), 955-962. doi: http://dx.doi.org/10.1016/j.neucom.2014.09.079

Firat, M. (2007). Artificial Intelligence Techniques for river flow forecasting in the Seyhan River Catchment, Turkey. *Hydrol. Earth Syst. Sci. Discuss., 4*(3), 1369-1406. doi: 10.5194/hessd-4-1369-2007

Firat, M. (2008). Comparison of Artificial Intelligence Techniques for river flow forecasting. *Hydrol. Earth Syst. Sci., 12*(1), 123-139. doi: 10.5194/hess-12-123-2008

Firat, M., & Turan, M. E. (2010). Monthly river flow forecasting by an adaptive neuro-fuzzy inference system. *Water and Environment Journal, 24*(2), 116-125. doi: 10.1111/j.1747-6593.2008.00162.x

Francés, F., Vélez, J. I., & Vélez, J. J. (2007). Split-parameter structure for the automatic calibration of distributed hydrological models. *Journal of Hydrology, 332*(1–2), 226-240. doi: http://dx.doi.org/10.1016/j.jhydrol.2006.06.032

Gautam, M. R., & Acharya, K. (2012). Streamflow trends in Nepal. *Hydrological Sciences Journal, 57*(2), 344-357. doi: 10.1080/02626667.2011.637042

Gopakumar, R., Takara, K., & James, E. J. (2007). Hydrologic Data Exploration and River Flow Forecasting of a Humid Tropical River Basin Using Artificial Neural

Networks. *Water Resources Management, 21*(11), 1915-1940. doi: 10.1007/s11269-006-9137-9

Green, J. I., & Nelson, E. J. (2002). Calculation of time of concentration for hydrologic design and analysis using geographic information system vector objects. *Journal of Hydroinformatics, 4*(2), 75-81.

Grimaldi, S., Petroselli, A., Tauro, F., & Porfiri, M. (2012). Time of concentration: a paradox in modern hydrology. *Hydrological Sciences Journal, 57*(2), 217-228. doi: 10.1080/02626667.2011.644244

Guo, J., Zhou, J., Qin, H., Zou, Q., & Li, Q. (2011). Monthly streamflow forecasting based on improved support vector machine model. *Expert Systems with Applications, 38*(10), 13073-13081. doi: http://dx.doi.org/10.1016/j.eswa.2011.04.114

Haddadnia, J., Faez, K., & Ahmadi, M. (2003). A fuzzy hybrid learning algorithm for radial basis function neural network with application in human face recognition. *Pattern Recognition, 36*(5), 1187-1202. doi: http://dx.doi.org/10.1016/S0031-3203(02)00231-5

Han, D., Chan, L., & Zhu, N. (2007). Flood forecasting using support vector machines. *Journal of Hydroinformatics, 9*(4), 267-276.

Hannaford, J., & Buys, G. (2012). Trends in seasonal river flow regimes in the UK. *Journal of Hydrology*(0). doi: 10.1016/j.jhydrol.2012.09.044

Harou, J. J., Pulido-Velazquez, M., Rosenberg, D. E., Medellín-Azuara, J., Lund, J. R., & Howitt, R. E. (2009). Hydro-economic models: Concepts, design, applications, and future prospects. *Journal of Hydrology, 375*(3–4), 627-643. doi: http://dx.doi.org/10.1016/j.jhydrol.2009.06.037

Hassan, A. J., Ghani, A. A., & Abdullah, R. (2004). Development Of Flood Risk Map Using GIS For Sg. Selangor Basin. Malaysia: National Hydraulic Research Institute of Malaysia.

Hassan, M., Shamim, M., Hashmi, H., Ashiq, S., Ahmed, I., Pasha, G., . . . Han, D. (2014). Predicting streamflows to a multipurpose reservoir using artificial neural networks and regression techniques. *Earth Science Informatics*, 1-16. doi: 10.1007/s12145-014-0161-7

Hatmoko, W., Radhika, Raharja, B., Tollenaar, D., & Vernimmen, R. (2015). Monitoring and Prediction of Hydrological Drought Using a Drought Early Warning System

in Pemali-Comal River Basin, Indonesia. *Procedia Environmental Sciences, 24*(0), 56-64. doi: http://dx.doi.org/10.1016/j.proenv.2015.03.009

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*: Wiley (New York).

Hoła, J., & Schabowicz, K. (2005). Application of artificial neural networks to determine concrete compressive strength based on non-destructive tests. *Journal of Civil Engineering and Management, 11*(1), 23-32. doi: 10.1080/13923730.2005.9636329

Honarbakhsh, A. H., Sadatinejad, S. J. S., Heydari, M., & Mozdianfard, M. (2012). Lag Time Forecasting in a River Basin *ENVIRONMENTAL SCIENCES, 9*(1), 39-50.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554-2558.

Iliyas, S. A., Elshafei, M., Habib, M. A., & Adeniran, A. A. (2013). RBF neural network inferential sensor for process emission monitoring. *Control Engineering Practice, 21*(7), 962-970. doi: http://dx.doi.org/10.1016/j.conengprac.2013.01.007

Jain, A., & Kumar, A. M. (2007). Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing, 7*(2), 585-592. doi: http://dx.doi.org/10.1016/j.asoc.2006.03.002

Jain, S. K., Singh, V. P., & van Genuchten, M. T. (2004). Analysis of Soil Water Retention Data Using Artificial Neural Networks. *Journal of Hydrologic Engineering, 9*(5), 415-420.

Jena, P. P., Chatterjee, C., Pradhan, G., & Mishra, A. (2014). Are recent frequent high floods in Mahanadi basin in eastern India due to increase in extreme rainfalls? *Journal of Hydrology, 517*(0), 847-862. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.021

Kalteh, A. M. (2013). Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. *Computers & Geosciences, 54*(0), 1-8. doi: http://dx.doi.org/10.1016/j.cageo.2012.11.015

Kalteh, A. M., Hjorth, P., & Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software, 23*(7), 835-845. doi: 10.1016/j.envsoft.2007.10.001

Kang, H. M., & Yusof, F. (2012). Homogeneity tests on daily rainfall series in peninsular Malaysia *International Journal of Contemporary Mathematical Sciences, 7*(1-4), 9-22.

Kentel, E. (2009). Estimation of river flow by artificial neural networks and identification of input vectors susceptible to producing unreliable flow estimates. *Journal of Hydrology, 375*(3–4), 481-488. doi: http://dx.doi.org/10.1016/j.jhydrol.2009.06.051

Kerh, T., & Lee, C. S. (2006). Neural networks forecasting of flood discharge at an unmeasured station using river upstream information. *Advances in Engineering Software, 37*(8), 533-543. doi: http://dx.doi.org/10.1016/j.advengsoft.2005.11.002

Khader, A. I., & McKee, M. (2014). Use of a relevance vector machine for groundwater quality monitoring network design under uncertainty. *Environmental Modelling & Software, 57*(0), 115-126. doi: http://dx.doi.org/10.1016/j.envsoft.2014.02.015

Kim, B., Lee, D. W., Park, K. Y., Choi, S. R., & Choi, S. (2004). Prediction of plasma etching using a randomized generalized regression neural network. *Vacuum, 76*(1), 37-43. doi: http://dx.doi.org/10.1016/j.vacuum.2004.05.018

Kirpich, Z. P. (1940). Time of concentration of small agricultural watersheds. *Civil Eng., 10*(6), 362-368.

KiŞI, Ö. (2006). Generalized regression neural networks for evapotranspiration modelling. *Hydrological Sciences Journal, 51*(6), 1092-1105. doi: 10.1623/hysj.51.6.1092

Kisi, O., Nia, A., Gosheh, M., Tajabadi, M., & Ahmadi, A. (2012). Intermittent Streamflow Forecasting by Using Several Data Driven Techniques. *Water Resources Management, 26*(2), 457-474. doi: 10.1007/s11269-011-9926-7

Kiumarsi, B., Lewis, F. L., & Levine, D. S. (2015). Optimal control of nonlinear discrete time-varying systems using a new neural network approximation structure. *Neurocomputing, 156*(0), 157-165. doi: http://dx.doi.org/10.1016/j.neucom.2014.12.067

Kokkinos, Y., & Margaritis, K. G. (2015). Topology and simulations of a Hierarchical Markovian Radial Basis Function Neural Network classifier. *Information Sciences, 294*(0), 612-627. doi: http://dx.doi.org/10.1016/j.ins.2014.08.025

Kokkonen, T. S., & Jakeman, A. J. (2001). A comparison of metric and conceptual approaches in rainfall-runoff modeling and its implications. *Water Resources Research, 37*(9), 2345-2352. doi: 10.1029/2001WR000299

Korhonen, J., & Kuusisto, E. (2010). Long-term changes in the discharge regime in Finland. *Hydrology Research, 41*(3-4), 253–268

Krasovskaia, I., & Gottschalk, L. (2002). River flow regimes in a changing climate. *Hydrological Sciences Journal, 47*(4), 597-609. doi: 10.1080/02626660209492962

Kretzschmar, A., Tych, W., & Chappell, N. A. (2014). Reversing hydrology: Estimation of sub-hourly rainfall time-series from streamflow. *Environmental Modelling & Software, 60*(0), 290-301. doi: http://dx.doi.org/10.1016/j.envsoft.2014.06.017

Krose, B., & Smagt, P. v. d. (1996). *Introduction to Artificial Neural Networks* (Eighth edition

ed.). Faculty of Mathematics & Computer Science Institute of Robotics and System Dynamics.University of Amsterdam. Amsterdam.

Kumar, S., Merwade, V., Kam, J., & Thurner, K. (2009). Streamflow trends in Indiana: Effects of long term persistence, precipitation and subsurface drains. *Journal of Hydrology, 374*(1–2), 171-183. doi: 10.1016/j.jhydrol.2009.06.012

Kundzewicz, Z. W., & Robson, A. J. (2004). Change detection in hydrological records—a review of the methodology / Revue méthodologique de la détection de changements dans les chroniques hydrologiques. *Hydrological Sciences Journal, 49*(1), 7-19. doi: 10.1623/hysj.49.1.7.53993

Laraque, A., Ronchail, J., Cochonneau, G., Pombosa, R., & Guyot, J. L. (2007). Heterogeneous Distribution of Rainfall and Discharge Regimes in the Ecuadorian Amazon Basin. *Journal of Hydrometeorology, 8*(6), 1364-1381. doi: 10.1175/2007jhm784.1

Lee, C. M. (2002). Master Plan Study on Flood Mitigation and River Management for Sg. Selangor River Basin. (Vol. 2): Drainage and Irrigation Department (DID) Malaysia.

Li, K., Xu, H., Huang, W., & Huang, Z. (2013). The Upper Preferred Multiple Directed Acyclic Graph Support Vector Machines for Classification. *Applied mathematics & information sciences, 7*(2), 733-739.

Li, M.-H., & Chibber, P. (2008). Overland Flow Time of Concentration on Very Flat Terrains. *Journal of the Transportation Research Board, 2060*(15), 133-140. doi: 10.3141/2060-15

Lin, J. Y., Cheng, C. T., & Chau, K. W. (2006). Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal, 51*(4), 599-612.

Liu, X., Wang, B., & Hu, C. (2015). Some new results on dynamics of delayed Cohen-Grossberg neural networks without intr-neuron delay. *Neurocomputing*(0). doi: http://dx.doi.org/10.1016/j.neucom.2015.05.019

Lou, Y., Ke, C., & Li, L. (2013). Accurately Predicting High Temperature Flow Stress of AZ80 Magnesium Alloy with Particle Swarm Optimization-based Support Vector Regression. *Applied mathematics & information sciences, 7*(3), 1093-1102.

Ma, Z., Kang, S., Zhang, L., Tong, L., & Su, X. (2008). Analysis of impacts of climate variability and human activity on streamflow for a river basin in arid region of northwest China. *Journal of Hydrology, 352*(3–4), 239-249. doi: 10.1016/j.jhydrol.2007.12.022

Machado, F., Mine, M., Kaviski, E., & Fill, H. (2011). Monthly rainfall–runoff modelling using artificial neural networks. *Hydrological Sciences Journal, 56*(3), 349-361. doi: 10.1080/02626667.2011.559949

Maier, H. R., Ascough Ii, J. C., Wattenbach, M., Renschler, C. S., Labiosa, W. B., & Ravalico, J. K. (2008). Chapter Five Uncertainty in Environmental Decision Making: Issues, Challenges and Future Directions. In A. A. V. A. E. R. A.J. Jakeman & S. H. Chen (Eds.), *Developments in Integrated Environmental Assessment* (Vol. Volume 3, pp. 69-85): Elsevier.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software, 15*(1), 101-124. doi: http://dx.doi.org/10.1016/S1364-8152(99)00007-9

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software, 25*(8), 891-909. doi: http://dx.doi.org/10.1016/j.envsoft.2010.02.003

Mäkelä, A., Landsberg, J., Ek, A. R., Burk, T. E., Ter-Mikaelian, M., Ågren, G. I., . . . Puttonen, P. (2000). Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation. *Tree Physiology, 20*(5-6), 289-298. doi: 10.1093/treephys/20.5-6.289

May, D. B., & Sivakumar, M. (2009). Prediction of urban stormwater quality using artificial neural networks. *Environmental Modelling & Software, 24*(2), 296-302. doi: 10.1016/j.envsoft.2008.07.004

McCuen, R. (2009). Uncertainty Analyses of Watershed Time Parameters. *Journal of Hydrologic Engineering, 14*(5), 490-498. doi: doi:10.1061/(ASCE)HE.1943-5584.0000011

McCuen, R., Wong, S., & Rawls, W. (1984). Estimating Urban Time of Concentration. *Journal of Hydraulic Engineering, 110*(7), 887-904. doi: doi:10.1061/(ASCE)0733-9429(1984)110:7(887)

Mecklin, C. J., & Mundfrom, D. J. (2004). An Appraisal and Bibliography of Tests for Multivariate Normality. *International Statistical Review, 72*(1), 123-138. doi: 10.1111/j.1751-5823.2004.tb00228.x

Mediero, L., Santillán, D., Garrote, L., & Granados, A. (2014). Detection and attribution of trends in magnitude, frequency and timing of floods in Spain. *Journal of Hydrology, 517*(0), 1072-1088. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.040

Meshgi, A., Schmitter, P., Chui, T. F. M., & Babovic, V. (2015). Development of a Modular Streamflow Model to Quantify Runoff Contributions from Different Land Uses in Tropical Urban Environments Using Genetic Programming. *Journal of Hydrology*(0). doi: http://dx.doi.org/10.1016/j.jhydrol.2015.04.032

Miao, C. Y., Shi, W., Chen, X. H., & Yang, L. (2012). Spatio-temporal variability of streamflow in the Yellow River: possible causes and implications. *Hydrological Sciences Journal, 57*(7), 1355-1367. doi: 10.1080/02626667.2012.718077

Minsky, M., & Seymour, P. (1969). *Perceptrons*: Oxford, England: M.I.T. Press.

Mirabbasi, R., Fakheri-Fard, A., & Dinpashoh, Y. (2012). Bivariate drought frequency analysis using the copula method. *Theoretical and Applied Climatology, 108*(1-2), 191-206. doi: 10.1007/s00704-011-0524-7

Misra, D., Oommen, T., Agarwal, A., Mishra, S. K., & Thompson, A. M. (2009). Application and analysis of support vector machine based simulation for runoff and sediment yield. *Biosystems Engineering, 103*(4), 527-535. doi: http://dx.doi.org/10.1016/j.biosystemseng.2009.04.017

Moliere, D. R., Lowry, J. B. C., & Humphrey, C. L. (2009). Classifying the flow regime of data-limited streams in the wet-dry tropical region of Australia. *Journal of Hydrology, 367*(1–2), 1-13. doi: 10.1016/j.jhydrol.2008.12.015

Morán-Tejeda, E., López-Moreno, J. I., Ceballos-Barbancho, A., & Vicente-Serrano, S. M. (2011). River regimes and recent hydrological changes in the Duero basin (Spain). *Journal of Hydrology*, *404*(3–4), 241-258. doi: 10.1016/j.jhydrol.2011.04.034

Mustafa, M., Rezaur, R., Rahardjo, H., Isa, M., & Arif, A. (2015). Artificial Neural Network Modeling for Spatial and Temporal Variations of Pore-Water Pressure Responses to Rainfall. *Advances in Meteorology, 2015*, 12.

Nayebi, M., Khalili, D., Amin, S., & Zand-Parsa, S. (2006). Daily Stream Flow Prediction Capability of Artificial Neural Networks as influenced by Minimum Air Temperature Data. *Biosystems Engineering*, *95*(4), 557-567. doi: http://dx.doi.org/10.1016/j.biosystemseng.2006.08.012

Nelson, W. B. (2002). An unusual turbidity maximum. In C. W. Johan & K. Cees (Eds.), *Proceedings in Marine Science* (Vol. Volume 5, pp. 483-497): Elsevier.

Nilsson, P., Uvo, C. B., & Berndtsson, R. (2006). Monthly runoff simulation: Comparing and combining conceptual and neural network models. *Journal of Hydrology, 321*(1–4), 344-363. doi: http://dx.doi.org/10.1016/j.jhydrol.2005.08.007

Nolan, R. H., Lane, P. N. J., Benyon, R. G., Bradstock, R. A., & Mitchell, P. J. (2015). Trends in evapotranspiration and streamflow following wildfire in resprouting eucalypt forests. *Journal of Hydrology*, *524*(0), 614-624. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.02.045

Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A., & Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology, 401*(3–4), 177-189. doi: http://dx.doi.org/10.1016/j.jhydrol.2011.02.021

Nourani, V. (2012). Conjugation of Artificial Neural Network and Geostatistics Approaches for Groundwater Modeling.

Nourani, V., Hosseini Baghanam, A., Adamowski, J., & Kisi, O. (2014). Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. *Journal of Hydrology, 514*(0), 358-377. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.03.057

Opitz-Stapleton, S., & Gangopadhyay, S. (2011). A non-parametric, statistical downscaling algorithm applied to the Rohini River Basin, Nepal. *Theoretical and Applied Climatology, 103*(3-4), 375-386. doi: 10.1007/s00704-010-0301-z

Palani, S., Liong, S. Y., & Tkalich, P. (2008). An ANN application for water quality forecasting. *Marine pollution bulletin, 56*(9), 1586-1597.

Pavlovic, S., & Moglen, G. (2008). Discretization Issues in Travel Time Calculation. *Journal of Hydrologic Engineering, 13*(2), 71-79. doi: doi:10.1061/(ASCE)1084-0699(2008)13:2(71)

Perugu, M., Singam, A., & Kamasani, C. (2013). Multiple Linear Correlation Analysis of Daily Reference Evapotranspiration. *Water Resources Management, 27*(5), 1489-1500. doi: 10.1007/s11269-012-0250-7

Pettitt, A. N. (1979). A non parametric approach to the change point problem. *Applied Statistic, 28*(2), 126-135.

Piotrowski, A. P. (2014). Differential Evolution algorithms applied to Neural Network training suffer from stagnation. *Applied Soft Computing, 21*(0), 382-406. doi: http://dx.doi.org/10.1016/j.asoc.2014.03.039

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., . . . Stromberg, J. C. (1997). The natural flow regime: a new paradigm for riverine conservation and restoration. *BioScience, 47*(11), 769-784.

Pulido-Calvo, I., & Portela, M. M. (2007). Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. *Journal of Hydrology, 332*(1–2), 1-15. doi: http://dx.doi.org/10.1016/j.jhydrol.2006.06.015

Qi, J., Li, C., & Huang, T. (2015). Stability of inertial BAM neural network with time-varying delay via impulsive control. *Neurocomputing, 161*(0), 162-167. doi: http://dx.doi.org/10.1016/j.neucom.2015.02.052

Rakhshanehroo, G. R., Vaghefi, M., & Shafiee, M. M. (2010). Flood forecasting in similar catchments using neural networks. *Turkish Journal of Engineering and Environmental Sciences, 34*(1), 57-66.

Rather, A. M., Agarwal, A., & Sastry, V. N. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications, 42*(6), 3234-3241. doi: http://dx.doi.org/10.1016/j.eswa.2014.12.003

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Reed, D. W., Johnson, P., & Firth, J. M. (1975). A non-linear rainfall-runoff model, providing for variable lag time. *Journal of Hydrology, 25*(3–4), 295-305. doi: http://dx.doi.org/10.1016/0022-1694(75)90027-X

Remesan, R., & Mathew, J. (2015). *Hydrological Data Driven Modelling. A Case Study Approach* (1 ed.): Springer International Publishing.

Richter, B. D. (1996). A method for assessing hydrologic alteration within ecosystems. *Un metro para evaluar alteraciones hidrologicas dentro de ecosistemas, 10*(4), 1163-1174.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Rougé, C., Ge, Y., & Cai, X. (2012). Detecting Gradual and Abrupt Changes in Hydrological Records. *Advances in Water Resources*(0). doi: 10.1016/j.advwatres.2012.09.008

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing, 2*(3), 117-119. doi: 10.1007/bf01891203

Rumelhart, D. E., Hintont, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.

Saber, M., Hamaguchi, T., Kojiri, T., Tanaka, K., & Sumi, T. (2015). A physically based distributed hydrological model of wadi system to simulate flash floods in arid regions. *Arabian Journal of Geosciences, 8*(1), 143-160. doi: 10.1007/s12517-013-1190-0

Sabzevari, T., Talebi, A., Ardakanian, R., & Shamsai, A. (2010). A steady-state saturation model to determine the subsurface travel time (STT) in complex hillslopes. *Hydrol. Earth Syst. Sci., 14*(6), 891-900. doi: 10.5194/hess-14-891-2010

Saghafian, B., & Julien, P. Y. (1995). Time to equilibrium for spatially variable watersheds. *Journal of Hydrology, 172*(1–4), 231-245. doi: http://dx.doi.org/10.1016/0022-1694(95)02692-I

Sahay, R., & Srivastava, A. (2014). Predicting Monsoon Floods in Rivers Embedding Wavelet Transform, Genetic Algorithm and Neural Network. *Water Resources Management, 28*(2), 301-317. doi: 10.1007/s11269-013-0446-5

Sahoo, G. B., & Ray, C. (2006). Flow forecasting for a Hawaii stream using rating curves and neural networks. *Journal of Hydrology, 317*(1–2), 63-80. doi: http://dx.doi.org/10.1016/j.jhydrol.2005.05.008

Sahoo, G. B., Ray, C., & De Carlo, E. H. (2006). Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *Journal of Hydrology, 327*(3–4), 525-538. doi: http://dx.doi.org/10.1016/j.jhydrol.2005.11.059

Sahoo, S., & Jha, M. (2015). On the statistical forecasting of groundwater levels in unconfined aquifer systems. *Environmental Earth Sciences, 73*(7), 3119-3136. doi: 10.1007/s12665-014-3608-8

Sahu, M., Khatua, K. K., & Mahapatra, S. S. (2011). A neural network approach for prediction of discharge in straight compound open channel flow. *Flow Measurement and Instrumentation, 22*(5), 438-446. doi: http://dx.doi.org/10.1016/j.flowmeasinst.2011.06.009

Samsudin, R., Saad, P., & Shabri, A. (2011). River flow time series using least squares support vector machines. *Hydrol. Earth Syst. Sci., 15*(6), 1835-1852. doi: 10.5194/hess-15-1835-2011

Sang, Y.-F., Wang, Z., Liu, C., & Yu, J. (2013). The impact of changing environments on the runoff regimes of the arid Heihe River basin, China. *Theoretical and Applied Climatology*, 1-9. doi: 10.1007/s00704-013-0888-y

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*(0), 85-117. doi: http://dx.doi.org/10.1016/j.neunet.2014.09.003

Schulmerich, M., Leporcher, Y.-M., & Eu, C.-H. (2015). Risk Measures in Asset Management *Applied Asset and Risk Management* (pp. 1-99): Springer Berlin Heidelberg.

Seth, I. (2015). Use of Artificial Neural Networks and Genetic Algorithms in Urban Water Management: A Brief Overview. *American Water Works Association, 107*(5), 93-97.

Seyam, M., & Othman, F. (2014a). The Influence of Accurate Lag Time Estimation on the Performance of Stream Flow Data-driven Based Models. *Water Resources Management, 28*(9), 2583-2597. doi: 10.1007/s11269-014-0628-9

Seyam, M., & Othman, F. (2014b). Long-term variation analysis of a tropical river's annual streamflow regime over a 50-year period. *Theoretical and Applied Climatology*, 1-15. doi: 10.1007/s00704-014-1225-9

Shabri, A., & Suhartono. (2012). Streamflow forecasting using least-squares support vector machines. *Hydrological Sciences Journal, 57*(7), 1275-1293. doi: 10.1080/02626667.2012.714468

Shafie, A. (2009). Extreme Flood Event: A Case Study on Floods of 2006 and 2007 in Johor, Malaysia. Fort Collins, Colorado, USA: Colorado State University.

Shahla Ramzan, Faisal Maqbool Zahid, & Shumila Ramzan. (2013). Evaluating Multivariate Normality: A Graphical Approach. *Middle-East Journal of Scientific Research, 13*(2), 254-263. doi: 10.5829/idosi.mejsr.2013.13.2.1746

Sharifi, S., & Hosseini, S. (2011). Methodology for Identifying the Best Equations for Estimating the Time of Concentration of Watersheds in a Particular Region. *Journal of Irrigation and Drainage Engineering, 137*(11), 712-719. doi: doi:10.1061/(ASCE)IR.1943-4774.0000373

Shi, F., & Xu, J. (2012). Emotional cellular-based multi-class fuzzy support vector machines on product's KANSEI extraction. *Applied mathematics & information sciences, 6*(1), 41-49.

Simas, M. J. (1996). *Lag time characteristics for small watersheds in the US.* (PhD), University of Arizona, Tuscon, AZ.

Singh, K. K., & Kumar, S. (2007). Extension of stream flow series using artificial neural networks. *ISH Journal of Hydraulic Engineering, 13*(3), 55-65. doi: 10.1080/09715010.2007.10514883

Singh, R. M., & Datta, B. (2007). Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resources Management, 21*(3), 557-572.

Singh, V. P. (1976). Derivation of time of concentration. *Journal of Hydrology, 30*(1–2), 147-165. doi: http://dx.doi.org/10.1016/0022-1694(76)90095-0

Singh, V. P. (1988). *Hydrologic Systems: Rainfall-Runoff Modeling* (Vol. I). New Jersey: Prentice Hal.

Sivakumar, B., & Berndtsson, R. (2010). SUMMARY AND FUTURE *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting* (pp. 463-477).

Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Data-Driven Modelling: Concepts, Approaches and Experiences. In R. Abrahart, L. See & D. Solomatine (Eds.), *Practical Hydroinformatics* (Vol. 68, pp. 17-30): Springer Berlin Heidelberg.

Solomatine, D. P. (2006). Data-Driven Modeling and Computational Intelligence Methods in Hydrology *Encyclopedia of Hydrological Sciences*: John Wiley & Sons, Ltd.

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics, 10*(1), 3-22. doi: 10.2166/hydro.2008.015

Specht, D. F. (1991). A general regression neural network. *Neural Networks, IEEE Transactions on, 2*(6), 568-576. doi: 10.1109/72.97934

Subramaniam, V. (2004). Managing Water Supply In Selangor And Kuala Lumpur. *BULETIN INGENIEUR, 22*, 12-20.

Sudheer, K. P., Gosain, A. K., & Ramasastri, K. S. (2002). A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes, 16*(6), 1325-1330. doi: 10.1002/hyp.554

Sun, S., Chen, H., Ju, W., Song, J., Zhang, H., Sun, J., & Fang, Y. (2013). Effects of climate change on annual streamflow using climate elasticity in Poyang Lake Basin, China. *Theoretical and Applied Climatology, 112*(1-2), 169-183. doi: 10.1007/s00704-012-0714-y

Talei, A., & Chua, L. H. C. (2012). Influence of lag time on event-based rainfall–runoff modeling using the data driven approach. *Journal of Hydrology, 438–439*(0), 223-233. doi: http://dx.doi.org/10.1016/j.jhydrol.2012.03.027

Tehrany, M., Pradhan, B., & Jebur, M. (2015). Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research and Risk Assessment, 29*(4), 1149-1165. doi: 10.1007/s00477-015-1021-9

Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality. *Computational Statistics & Data Analysis*, *55*(5), 1980-1992. doi: 10.1016/j.csda.2010.12.004

Thomas, B., Lischeid, G., Steidl, J., & Dietrich, O. (2015). Long term shift of low flows predictors in small lowland catchments of Northeast Germany. *Journal of Hydrology*, *521*(0), 508-519. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.12.022

Tiwari, M., Song, K.-Y., Chatterjee, C., & Gupta, M. (2013). Improving reliability of river flow forecasting using neural networks, wavelets and self-organising maps. *Journal of Hydroinformatics*, *15*(2), 486-502.

Tiwari, M. K., & Chatterjee, C. (2010). Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. *Journal of Hydrology*, *394*(3–4), 458-470. doi: http://dx.doi.org/10.1016/j.jhydrol.2010.10.001

Tiwari, M. K., Song, K.-Y., Chatterjee, C., & Gupta, M. M. (2012). River-Flow Forecasting Using Higher-Order Neural Networks. *Journal of Hydrologic Engineering*, *17*(5), 655-666.

Toro, C. H. F., Gómez Meire, S., Gálvez, J. F., & Fdez-Riverola, F. (2013). A hybrid artificial intelligence model for river flow forecasting. *Applied Soft Computing*, *13*(8), 3449-3458. doi: http://dx.doi.org/10.1016/j.asoc.2013.04.014

Toro, C. H. F., Peña, D. G., González, B. S., & Fdez-Riverola, F. (2008). Water flows modelling and forecasting using a RBF neural network. *Sistemas & Telemática*, *6*(12).

Toth, E. (2008). Data-Driven Streamflow Simulation: The Influence of Exogenous Variables and Temporal Resolution. In R. Abrahart, L. See & D. Solomatine (Eds.), *Practical Hydroinformatics* (Vol. 68, pp. 113-125): Springer Berlin Heidelberg.

Turan, M. E., & Yurdusev, M. A. (2009). River flow estimation from upstream flow records by artificial intelligence methods. *Journal of Hydrology*, *369*(1–2), 71-77. doi: http://dx.doi.org/10.1016/j.jhydrol.2009.02.004

Viessman, W., & Lewis, G. L. (2003). *Introduction to Hydrology* (5th ed.). United States of America: Pearson Education.

Villarini, G., Smith, J. A., Baeck, M. L., Vitolo, R., Stephenson, D. B., & Krajewski, W. F. (2011). On the frequency of heavy rainfall for the Midwest of the United States. *Journal of Hydrology, 400*(1–2), 103-120. doi: 10.1016/j.jhydrol.2011.01.027

Walling, D. E., & Fang, D. (2003). Recent trends in the suspended sediment loads of the world's rivers. *Global and Planetary Change, 39*(1–2), 111-126. doi: 10.1016/s0921-8181(03)00020-1

Wang, W.-C., Chau, K.-W., Cheng, C.-T., & Qiu, L. (2009). A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology, 374*(3–4), 294-306. doi: http://dx.doi.org/10.1016/j.jhydrol.2009.06.019

Wang, X., Yu, J., Li, C., Wang, H., Huang, T., & Huang, J. (2015). Robust stability of stochastic fuzzy delayed neural networks with impulsive time window. *Neural Networks, 67*(0), 84-91. doi: http://dx.doi.org/10.1016/j.neunet.2015.03.010

Wei, C.-C. (2015). Comparing lazy and eager learning models for water level forecasting in river-reservoir basins of inundation regions. *Environmental Modelling & Software, 63*(0), 137-155. doi: http://dx.doi.org/10.1016/j.envsoft.2014.09.026

Wei, S., Yang, H., Song, J., Abbaspour, K., & Xu, Z. (2013). A wavelet-neural network hybrid modelling approach for estimating and predicting river monthly flows. *Hydrological Sciences Journal, 58*(2), 374-389. doi: 10.1080/02626667.2012.754102

Wieland, R., Mirschel, W., Zbell, B., Groth, K., Pechenick, A., & Fukuda, K. (2010). A new library to combine artificial neural networks and support vector machines with statistics and a database engine for application in environmental modeling. *Environmental Modelling & Software, 25*(4), 412-420. doi: http://dx.doi.org/10.1016/j.envsoft.2009.11.006

Woodward, D. E. (2010). National Engineering Handbook, Part 630 Hydrology, Chapter 15: Time of Concentration *National Engineering Handbook* (Vol. Part 630 Hydrology). Natural Resources Conservation Service: United States Department of Agriculture.

Wu, C. L., Chau, K. W., & Li, Y. S. (2008). River stage prediction based on a distributed support vector regression. *Journal of Hydrology, 358*(1–2), 96-111. doi: http://dx.doi.org/10.1016/j.jhydrol.2008.05.028

Xiong, L., & Guo, S. (2004). Trend test and change-point detection for the annual discharge series of the Yangtze River at the Yichang hydrological station / Test de tendance et détection de rupture appliqués aux séries de débit annuel du fleuve

Yangtze à la station hydrologique de Yichang. *Hydrological Sciences Journal, 49*(1), 99-112. doi: 10.1623/hysj.49.1.99.53998

Xu, Y.-P., Yu, C., Zhang, X., Zhang, Q., & Xu, X. (2012). Design rainfall depth estimation through two regional frequency analysis methods in Hanjiang River Basin, China. *Theoretical and Applied Climatology, 107*(3-4), 563-578. doi: 10.1007/s00704-011-0497-6

Yang, S. L., Gao, A., Hotz, H. M., Zhu, J., Dai, S. B., & Li, M. (2005). Trends in annual discharge from the Yangtze River to the sea (1865–2004) / Tendances et épisodes extrêmes dans les débits annuels du Fleuve Yangtze débouchant dans la mer (1865–2004). *Hydrological Sciences Journal, 50*(5), 836. doi: 10.1623/hysj.2005.50.5.825

Yao, C., Zhang, K., Yu, Z., Li, Z., & Li, Q. (2014). Improving the flood prediction capability of the Xinanjiang model in ungauged nested catchments by coupling it with the geomorphologic instantaneous unit hydrograph. *Journal of Hydrology, 517*(0), 1035-1048. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.06.037

Yu, P.-S., Chen, S.-T., & Chang, I. F. (2006). Support vector regression for real-time flood stage forecasting. *Journal of Hydrology, 328*(3–4), 704-716. doi: http://dx.doi.org/10.1016/j.jhydrol.2006.01.021

Yucel, I., Onen, A., Yilmaz, K. K., & Gochis, D. J. (2015). Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *Journal of Hydrology, 523*(0), 49-66. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.01.042

Yue, S., Pilon, P., & Phinney, B. O. B. (2003). Canadian streamflow trend detection: impacts of serial and cross-correlation. *Hydrological Sciences Journal, 48*(1), 51-63. doi: 10.1623/hysj.48.1.51.43478

Zakaria, Z. A., & Shabri, A. (2012). Streamflow Forecasting at Ungaged Sites Using Support Vector Machines. *Applied Mathematical Sciences, 6*(60), 3003-3014.

Zazo, S., Molina, J.-L., & Rodríguez-Gonzálvez, P. (2015). Analysis of flood modeling through innovative geomatic methods. *Journal of Hydrology, 524*(0), 522-537. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.03.011

Zhang, X.-S. (2000). Self-Organized Neural Networks *Neural Networks in Optimization* (Vol. 46, pp. 177-195): Springer US.

Zhang, X., Vincent, L. A., Hogg, W. D., & Niitsoo, A. (2000). Temperature and precipitation trends in Canada during the 20th century. *Atmosphere-Ocean, 38*(3), 395-429. doi: 10.1080/07055900.2000.9649654

Zhang, Y., Vaze, J., Chiew, F. H. S., & Li, M. (2015). Comparing flow duration curve and rainfall–runoff modelling for predicting daily runoff in ungauged catchments. *Journal of Hydrology, 525*(0), 72-86. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.03.043

Zheng, H., Zhang, L., Liu, C., Shao, Q., & Fukushima, Y. (2007). Changes in stream flow regime in headwater catchments of the Yellow River basin since the 1950s. *Hydrological Processes, 21*(7), 886-893. doi: 10.1002/hyp.6280

Zhou, Z., Chen, X., & Xiao, X. (2013). On Evaluation Model of Circular Economy for Iron and Steel Enterprise Based on Support Vector Machines with Heuristic Algorithm for Tuning Hyper-parameters. *Applied mathematics & information sciences, 7*(6), 2215-2223.

Ziaee, S., Sadrossadat, E., Alavi, A., & Mohammadzadeh Shadmehri, D. (2015). Explicit formulation of bearing capacity of shallow foundations on rock masses using artificial neural networks: application and supplementary studies. *Environmental Earth Sciences, 73*(7), 3417-3431. doi: 10.1007/s12665-014-3630-x

Zin, W., Jemain, A., & Ibrahim, K. (2013). Analysis of drought condition and risk in Peninsular Malaysia using Standardised Precipitation Index. *Theoretical and Applied Climatology, 111*(3-4), 559-568. doi: 10.1007/s00704-012-0682-2

List of Publications and Papers Presented

Peer Reviewed Journal Papers

- Mohammed Seyam, and Faridah Othman (2014). The Influence of Accurate Lag Time Estimation on the Performance of Stream Flow Data-driven Based Models. Water Resources Management, 28(9), 2583-2597. doi: 10.1007/s11269-014-0628-9
- Mohammed Seyam, and Faridah Othman (2014). Long-term variation analysis of a tropical river's annual stream flow regime over a 50-year period. Theoretical and Applied Climatology, 1-15. doi: 10.1007/s00704-014-1225-9

Conference Papers

- Mohammed Seyam, and Faridah Othman (2012). Long-term changes in the stream flow regime in Selangor River basin. Asia-Oceania Top University League on Engineering (AOTULE). University of Malaya, Kuala Lumpur, Malaysia.
- Mohammed Seyam, Faridah Othman, Alaa-Eldin M. E. (2013). Influence of the rainfall intensity on the lag time between the upstream and downstream stations, International Conference on Water and Wastewater Management (ICWWM 2013), Kuala Lumpur, Malaysia.
- Mohammed Seyam, and Faridah Othman (2015). Hourly Stream Flow Prediction in Tropical Rivers by Multi-Layer Perceptron network, The 2nd IWA Malaysia Young Water Professionals Conference (YWP15), Kuala Lumpur, Malaysia.

Submitted papers

- Prediction of Hourly Stream Flow in Humid Tropical Rivers by Support Vector Machines.
- Derivation of New Empirical Formulas to Estimate Lag Time between Upstream and Downstream Stations in Tropical Humid Rivers.

Ongoing papers

- Integrating lag time estimation with artificial neural networks for hourly stream flow prediction.
- Influence of lag time estimation on the performance of stream flow prediction by support vector machine.
- Application of artificial neural networks as early warning tool of floods.