Forecasting Model for Extreme Rainfall using Artificial Neural Network

Yasir Hilal Hadi Al-Qurayshi

Supervisors:

Prof. Dr. Ku Ruhana Ku Mahamud
Wan Hussain Wan Ishak

Abstract (English)

Successive days of rainfall are known to cause flood. The forecasting of daily rainfall helps to estimate the occurrences of rainfall and number of wet days, while with a maximum of five consecutive days of rainfall, the magnitude of rainfall within a specified period can predict what may signify rainfall extremes. In this study, data mining and back propagation neural network (BPNN) have been established in developing the extreme rainfall forecasting models. Four forecasting models were developed to forecast the maximum five consecutive days of rainfall amount (PX5D) of the next month. The models only use the extreme rainfall indices outlined by STARDEX as predictors in forecasting. The first developed model uses six extreme rainfall indices in forecasting, the second model uses the values of the PX5D index of a three-month delay, the third model uses the previous six-month PX5D values, while the fourth model was developed to forecast the PX5D using the values of the same index of a twelve-month delay. It was found that when using the six extreme rainfall core indices, the forecasting error was the lowest. A regression model has been developed using the six extreme rainfall indices to compare the performance measurements with the BPNN model that uses the same indices.

Abstrak (Bahasa Malaysia)

Hujan berterusan diketahui menyebabkan banjir. Ramalan hujan harian membantu untuk anggaran kejadian hujan dan bilangan hari basah, manakala dengan maksimum lima hari berturut-turut hujan, magnitud hujan dalam tempoh yang dinyatakan boleh diramal yang mungkin menandakan keterlaluan hujan. Dalam kajian ini, pengumpulan data dan rangkaian neural rambatan balik (BPNN) telah dikenal pasti berpotensi dalam pembangunan model peramalan hujan melampau. Empat model ramalan telah dibangunkan bagi meramal jumlah hujan maksimum lima hari berturut-turut (PX5D) bagi bulan berikutnya. Model-model berkenaan hanya menggunakan indeks hujan melampau seperti yang ditetapkan oleh STARDEX sebagai peramal. Model pertama yang dibangunkan menggunakan enam indeks hujan melampau dalam peramalan, model kedua menggunakan nilai indeks PX5D bagi tiga bulan sebelumnya, model ketiga menggunakan nilai indeks PX5D enam bulan sebelumnya, manakala model keempat dibangunkan untuk meramal PX5D menggunakan nilai indeks yang sama bagi 12 bulan sebelumnya. Kajian mendapati apabila menggunakan enam indeks teras hujan melampau, ralat bagi peramalan adalah paling rendah. Model regresi telah dibangunkan menggunakan enam indeks hujan melampau untuk dibandingkan dengan prestasi model BPNN yang menggunakan indeks yang sama.

Acknowledgement

In the name of ALLAH, Most Gracious, Most Merciful. Praise be to ALLAH, The Lord of the worlds and blessing be upon all his Prophets and upon the last Prophet and messenger, Mohammed and upon his family.

I would like to express my sincere gratitude and appreciation to my supervisors Prof. Dr. Ku Ruhana Ku Mahamud and Mr. Wan Hussain Wan Ishak for their extreme support, guidance and valuable advices throughout this work.

I am grateful to my parents for their unlimited help and the great support they offered to me.

To my family and friends for all the help and encouragement they offered.

A sincere gratitude to Reem, my classmate, my colleague, my best friend and my wife for everything.

May Allah surround you with his protection and bestow upon you all the blessing that you deserve.

Table of Contents

Abstract (English) ii
Abstrak (Bahasa Malaysia) iii
Acknowledgement iv
Chapter One: Introduction 1
1.1 Background 1
1.2 Problem Statement 4
1.3 Research Questions 5
1.4 Objective 6
1.5 Scope of the Study 6
1.6 Significance of the Study 6
Chapter Two: Literature Review 8
2.1 Extreme Events 8
2.2 Modeling of Rainfall Event 12
2.2.1 Mining Rainfall Data 13
2.2.2 Grades of Rainfall 14
2.2.3 Features for Rainfall Forecasting 16
2.3 Artificial Neural Network 17
2.3.1 ANN and Statistical Methods in Hydrological Forecasting 18
2.3.2 ANN Applications 21
2.4 Rainfall Forecasting Models 23
2.4.1 General Rainfall Forecasting Models 24
2.4.2 Extreme Rainfall Forecasting Models 27
2.6 Research Gap 30
2.7 Summary 31
Chapter Three: Methodology 32
3.1 Overview of Methodology 32
3.2 KDD Steps 36
3.2.1 Data Collection and Selection 36
3.2.2 Preprocessing 38
3.2.3 Transformation 39
3.2.4 Data Mining 45
3.2.5 Evaluation 49
3.3 Summary 49

Chapter Four: Proposed Models **51**
4.1 BPNN Model **51**
4.1.1 Experiment-A **56**
4.1.2 Experiment-B1 ..... **58**
4.1.3 Experiment-B2 ..... **59**
4.1.4 Experiment-B3 ..... **61**
4.2 Statistical Model **64**
4.3 Summary **66**
Chapter Five: Results **67**
5.1 BPNN Forecasting Model **67**
5.1.1 BPNN Experiments Results **67**
5.1.2 BPNN Experiments Comparison **73**
5.2. Multiple Regression Forecasting Model **75**
5.2.1 Multiple Regression Model Results **75**
5.3. Evaluation **77**
5.4. Summary **80**
Chapter Six: Conclusion and Future Work **81**
6.1. Conclusion **81**
6.2. Future Work **83**
References **84**

List of Tables

Table 2.1: Rainfall grades by mms……...15
Table 2.2: Summary of rainfall forecasting models...……..……..……26
Table 2.3 : Summary of extreme rainfall forecasting models...……..……30
Table 3.1: KDD, SEMMA and CRISP-DM steps...……..……35
Table 3.2 : Example of the rainfall data (in mm) by gauging station...……..…..…….38
Table 3.3 : Example of the rainfall data (in mm) with missing value...……..…..…….39
Table 3.4: STARDEX extreme rainfall indices ……..…40
Table 3.5: Daily rainfall amounts (in mm) for November 2012……..……43
Table 3.6 : Calculated values of the six core indices for November 2012…….43
Table 3.7 : Summary of calculated indices by gauging stations…….……44
Table 3.8 : Data division…….……..……...47
Table 4.1 : BPNN training parameters ……..…….…….54
Table 4.2: Example of Dataset-A…….56
Table 4.3 : Experiment-A's BPNN architecture…...……..57
Table 4.4: Example of Dataset-B1……...…58
Table 4.5 : Experiment-B1's BPNN architecture…..……..59
Table 4.6: Example of Dataset-B2…….……..60
Table 4.7 : Experiment-B2's BPNN architecture…..…….…….……61
Table 4.8: Example of Dataset-B3…….……..62
Table 4.9 : Experiment-B3's BPNN architecture…..……..63
Table 5.1: Experiment-A's final BPNN architecture……..……..……..…68
Table 5.2: Experiment-B1's final BPNN architecture…….....…..…….69
Table 5.3: Experiment-B2's final BPNN architecture…….……..…..…….71
Table 5.4: Experiment-B3's final BPNN architecture…….....…….……72
Table 5.5: RMSE of BPNN and MR models……78
Table 6.1: Calculated PQ90 index by gauging stations……..……..80

List of Figures

Figure 2.1: Rainfall measurement gauge……..……...……...…15
Figure 3.1: Overview of the KDD process steps……...…….……..……...36
Figure 3.2: Timah Tasoh reservoir and the five gauging stations ….……..….…….….. 37
Figure 4.1: Basic BPNN architecture for Experiment-A ……..….…...……57
Figure 4.2: Basic BPNN architecture for Experiment-B1 …….……..…….59
Figure 4.3: Basic BPNN architecture for Experiment-B2 ……..……..…….…61
Figure 4.4: Basic BPNN architecture for Experiment-B3.…….……..…….64
Figure 5.1: Final BPNN architecture for Experiment-A…….....….……68
Figure 5.2: Final BPNN architecture for Experiment-B1 …….……..…...……70
Figure 5.3: Final BPNN architecture for Experiment-B2 …….…..……71
Figure 5.4: Final BPNN architecture for Experiment-B3.……..….……73
Figure 5.5: MAE value of the four BPNN model experiments…….……74
Figure 5.6: MAE of BPNN model (Experiment-A) and the multiple regression model…77

Chapter One
Introduction

1.1 Background

Rainfall is one of the main sources of water for the hydrological cycle, and causes a serious impact to water source, rain related activities and the environment. In case the rainfall of one location significantly deviates from the regular condition, this can be considered that the event will be less likely to occur. Currently, it is most common to appoint a percentile value as a threshold. The values above this threshold are considered extreme values, which are the values (event) that are not likely to happen (Gu et al., 2010).

Malaysia's climate has the following characteristic features: copious rainfall, high humidity and uniform temperature. Winds are generally light. Located in the equatorial area, even with periods of severe drought, it is a high rarity to have a full day without clouds. On the other hand, it is also rare to have completely no sunshine for a stretch of a few days except during the northeast monsoon seasons (MET Malaysia, 2013). Malaysia experiences rain almost all year long, and for some regions, it is heavier. With the average yearly rainfall, Peninsular Malaysia receives around 2,440mm, while Sarawak and Sabah receive 3,830mm and 2630mm respectively (Hazenberg et al., 2011). The West Coast of the Peninsular is exposed to convective and localized storms caused by the intermonsoon seasons. Convective storms are extremely variable in space and time and can lead to very intensive rainfall rates that produce floods. The extreme flood event that happened between December 2006 and January 2007

in southern Peninsular Malaysia, for example, resulted in economic losses of more than 500 million U.S. dollars, involving more than 200,000 people and 16 deaths (Juneng et al., 2010). For a developing country such as Malaysia, which is prone to flood disasters, having a rainfall forecasting model is a very vital matter (El-Shafie et al., 2011b).

Extreme rainfall takes considerable part in evaluating flood and drought events which have catastrophic effects on the infrastructures and population. Floods are one of the most powerful forces on Earth, causing huge damage around the world. Over the past ten years, floods have affected more than 0.8 billion and killed approximately 50,000 people (OFDA/CRED, 2014). Statistics have shown that floods have a large impact on economy and human well-being (Rana, 2013). Economic damage, ecosystem damage, and loss of cultural and historical values constitute the direct outcomes of floods. Floods also lead to negative health effects on human and cause loss of life (Rana, 2013).

The extreme rainfall index is an important measure to determine the severity of rainfall throughout the world. The World Meteorological Organization (WMO) and the European Union Statistical and Regional dynamical Downscaling of Extremes for European regions project (STARDEX) have developed descriptive indices of extremes for rainfall (Sulaiman et al., 2014; Gu et al., 2010). The purpose of these indices is to observe the changes of weather and climate extremes using a uniform measurement. A detailed discussion about the six indices is presented in Chapter Three.

Temporal data mining is concerned with the data mining of large sequential datasets. Sequential data refer to data that are ordered with respect to some indices. For example, a time series constitutes a popular class of sequential data, where records are indexed by time. Other examples of sequential data could be text, gene sequences, protein sequences, lists of moves in a chess game etc. Here, although there is no notion of time as such, the ordering among the records is very important and is central to the data description/modeling. Temporal data mining concerns with the analysis of events ordered by one or more dimensions of time. The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events (Geetha et al., 2008).

Neural networks have seen an explosion of interest over the last few years and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology, physics, and hydrology (Cilimkovic, 2011). It all started way back in 1943 when McCulloch and Pitts proved that neuron can have two states and that those states could be dependent on some threshold values, in which they presented the first artificial neuron model. According to Rojas (2005), many new and more sophisticated models have been presented since McCulloch and Pitts' discovery opened the door to intelligent machines. According to the Muller et al. (1995), there are two main reasons for NN investigation: the first is to try to obtain an understanding on how a human brain functions; and the second is the desire to build machines that are capable of solving complex problems that sequentially operating computers were unable to solve.

Regression analysis is one of the widely applied methods in modeling the relationships between one or more dependent variables and independent variables. It can be defined as a conventional method for forecasting. In statistics, regression analysis is defined as the process used to estimate the parameter values of a function, in which the function predicts the value of a response variable in terms of the value of other variables (Kleinbaum et al., 2013). Many methods are developed to fit functions and these methods typically depend on the type of function being used. For example, linear regression, logistic regression, and nonlinear regression.

1.2 Problem Statement

Due to the critical impact of extreme rainfall, modeling and forecasting extreme rainfall might reduce its impact, as the authorities can be alerted when the rainfall had reached the extreme level. Previous studies such as Khalili et al. (2011) and El-Shafie et al. (2012) developed a forecasting model to forecast normal rainfall conditions. While Gu et al. (2010), Zeng et al. (2011), and Wan et al. (2012) developed a forecasting model by considering the extreme rainfall condition. However, the models are constrained to the geographical location, hydrological, and metrological variables. Therefore, the model cannot be directly applied to the Malaysian context. This study will investigate and adapt the STARDEX extreme rainfall indices into Malaysia's context.

Studies by Junaida et al. (2012), Sulaiman et al. (2013), and Sulaiman et al. (2014) have focused on the modeling and forecasting of extreme rainfall in the

Malaysian context, but these studies are limited to specific locations. In Malaysia, different states may have different characteristics, and the rainfall patterns may vary among the states.

This study aims to develop a general model that can be used to forecast extreme rainfall. Different combinations of variables are used as compared to the previous studies. In certain areas, floods which might occur due to extreme rainfall are observed from several stations. The extreme rainfalls from various stations need to be modeled in order to understand their pattern and the extreme rainfall threshold, thus, this will help to develop the forecasting model.

1.3 Research Questions

The research questions of this study are as follows:

- What is the threshold value for the extreme rainfall event?
- How many forecasting models can be built based on the threshold and the identified indices?
- Which combination of variables or model gives the best performance of forecasting?

1.4 Objective

The aim of this study is to develop a forecasting model for extreme rainfall using ANN. The specific objectives are:

1) To identify extreme rainfall threshold using percentile value.
2) To develop a forecasting model using the identified extreme rainfall threshold and descriptive indices.
3) To evaluate the developed rainfall forecasting model.

1.5 Scope of the Study

This study focuses on the modeling and forecasting of extreme rainfall from several gauging stations. The case study is the Perlis upstream rivers, specifically on the upstream of Timah Tasoh reservoir. In this study, the data from five gauging stations are collected; which are Lubuk Sireh, Tasoh, Padang Besar, Wang Kelian, and Kaki Bukit.

Data mining technique, such as temporal data mining, has been employed to extract the extreme rainfall pattern from the rainfall data. ANN, specifically back propagation neural network, is used in the development of the forecasting models.

1.6 Significance of the Study

Typically, intense rainfall occurrences in short temporal scales or persistent rainfall over a long period of time often lead to massive floods (Syafrina et al.,

2014). These floods represent one of the most important impacts of extreme climatic events resulting to hazardous situations that cause negative effects on human and infrastructures. The knowledge of extreme rainfall might be useful in planning, especially when designing new infrastructures to withstand the impact of disasters such as flood (Nigatu, 2011). The extreme rainfall events in Malaysia showed an increasing trend in recent years. It is one of the major causes of severe floods in Malaysia in the past ten years. The impacts of these floods are huge and the recovery cost reaches millions of Malaysian ringgit (Syafrina et al., 2014; Abdullah, 2013).

This study is vital as Malaysia is one of the countries prone to floods that are caused by extreme rainfall. The modeling of extreme rainfall will increase the understanding of the extreme rainfall pattern in Malaysia, thus, assist the authorities to prepare and plan for recovery actions. This might reduce the negative impact of flood on human life and infrastructure.

Chapter Two
Literature Review

This chapter discusses the review of related literature. Extreme event is defined and related studies are discussed in Section 2.1. Section 2.2 describes the modeling of the rainfall event; the classification of rainfall and features and models used for rainfall forecasting by researchers in the related area are also discussed in this section. Methods used by the reviewed researches for extreme rainfall forecasting are discussed in Section 2.3, whereas the last section is a summary of the literature review.

2.1 Extreme Events

During the evolution of the Earth's surface, state economies, and political structures, to name three examples, extreme events have obviously had significant roles to play: they shape the future courses of such systems. Indeed, the worst earthquakes in California, with a recurrence rate of about once every two centuries, account for a significant fraction of the region total tectonic deformation; landscapes are changed by the "millennium" flood, which is more effective than the concerted action of all other eroding agents; the largest volcanic eruptions lead to major topographic changes and to severe climatic disruptions; financial crashes, which in an instant can cause the loss of trillions of dollars, loom and affect the psychological state of investors, society, and the world economy.

If there are commonalities in cause, there are many more commonalities in effect. Indeed, extreme events entail casualties: deaths, heavy financial costs, environmental destruction, and undermining the fabric of society. These result from side effects or secondary events deriving from primary extreme events, such as the disruption of communication networks, the contamination of water, and the breakdown of health support, energy supplies, and so on.

The general definition of extreme events that was adopted by the Intergovernmental Panel on Climate Change (IPCC) for its Special Report on Extremes (Field et al., 2012), describes an extreme as the "occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends of the range of observed values of the variable".

Extreme events, by definition, can be both rare at any given location, and common in a global sense. In any one place, the chance of a once in a 100–year heat extreme is so rare that, in principle, it only occurs, on average, once every 100 years. This also means that, on average and with a stationary climate, every year, one percent of the world would be expected to experience a once in a 100–year heat extreme, and another one percent of the world with a cold extreme. On the other hand, a few extremes such as tornadoes are rare even in a global sense, while other extremes such as loss of sea ice are limited to specific regions (Peterson et al., 2013).

Extreme climate and weather events have received an increased interest over the recent years, as a result of the unfortunate losses of human life and significantly increasing costs related to them (Easterling et al., 2000). Numerous metrological

departments study and report extreme events from around the world; in addition to that, many researchers conduct researches on climate change and how they affect the extreme events. Peterson et al. (2012) made a research on the Thailand extreme flooding in 2011, during and after the extreme wet monsoon (July - September) in north Thailand. During the floods, rivers plains within the center and the south flooded their banks and inundated wide regions of the country, including neighborhoods in the present capital Bangkok and the former capital Ayuttha.

NOAA (2013) referred to the year 2012 as the hottest in the history of the United States, simply because of the extremely high temperatures over most of the central and eastern regions of the United States during spring and summer. The high summer temperature was related to one of the most extreme droughts in history.

Peterson et al.'s (2013) study marked that Europe experienced strongly contrasting precipitation anomalies in the summer of 2012. For example, the United Kingdom experienced extreme wet summer, which led to widespread flooding. Spain, in contrast, suffered drought and wildfires associated with extreme low summer rainfall.

Peterson et al. (2013) also concluded in their research that the extreme event of the 2011–2012 winter drought over the Iberian Peninsula was extreme in its magnitude and spatial extent. And the reason was that the Iberian Peninsula is dominated by a large-scale circulation pattern from North Atlantic Oscillation and Eastern Atlantic.

The southwestern part of Japan experienced an extreme rainfall during the late Japanese "Baiu" rainy season, from 11th July to 14th July 2012. This record-breaking event occurred at many sites and caused devastating damages with 31 deaths, 3 missing persons, floods, mudslides, and damaged homes in the southwestern part of the mainland of Japan (Peterson et al., 2013).

In the end of summer 2012, a widespread flooding had resulted from heavy rain across eastern Australia (Bureau of Meteorology, 2012) and caused swollen rivers that swamped agricultural land, forced tens of thousands of people to evacuate their homes, and caused loss of life.

Standard economic theory maintains that the complex trajectory of stock market prices is the faithful reflection of the continuous flow of news that is interpreted and digested by an army of analysts and traders. Accordingly, large shocks should result from really bad surprises. It is a fact that exogenous shocks exist, as epitomized by the recent events of 11th September 2001, and there is no doubt about the existence of utterly exogenous bad news that moves stock market prices and creates strong bursts of volatility. One case that cannot be refuted is the market turmoil observed in Japan, following the Kobe earthquake of 17th January 1995, the estimated cost of which was around $200 billion dollars. Indeed, destructive earthquakes cannot be not endogenized in advance in stock market prices by rational agents ignorant of seismological processes. One may also argue that the invasion of Kuwait by Iraq on 2nd August 1990 and the coup against Gorbachev on 19th August 1991 were strong exogenous shocks. However, a few could also argue that precursory fingerprints of these events

were known to some insiders, suggesting the possibility that the action of these informed agents may have been reflected in part in stock markets prices (Albeverio et al., 2006).

Based on the above literature review, it can be summarized that there are various ways to define extreme climate events, such as extreme daily rainfall amounts, extreme daily temperatures, large areas experiencing uncommon monthly temperatures, warm, storm events like hurricanes, and financial. Moreover, extreme events may be defined by the effect that event has on the human. That effect can involve excessive economic or monetary losses, excessive loss of life, or both.

2.2 Modeling of Rainfall Event

Rainfall is considered as one of the most complicated and difficult elements of the hydrology cycle to understand and to model because the atmospheric processes are complex, which produce rainfall and the great range of variance over a number of scales both in time and space (Hung et al., 2009). A number of weather parameters must be used to obtain sufficient knowledge about the characteristics of the rainfall event. El-Shafie et al. (2011a) used a series consisted of daily readings of minimum and maximum temperatures, rainfall and wind speed. These readings were taken in Alexandria city, Egypt, from 1957 to 2009 and applied to forecast the rainfall in the mentioned city.

While Hung et al. (2009), in his study, focused on the Bangkok area only, and utilized 75 stations that collected meteorological information of six parameters that contained hourly data observed within the mast station, which are: wind speed, wet bulb temperature, relative humidity, cloudiness, dry bulb temperature, and air pressure for the same period as rainfall data. Another value used is the average intensity of the hourly rainfall of all of the rain gauges, in which this value is arithmetically averaged, computed and provided with the meteorology data.

The monthly rainfall records at National Observatory of Athens (NOA) for a 115-year period (1891–2005) were utilized by Moustris et al. (2011). The average monthly rainfall through the analyzed period for each single station has been calculated after that, the average monthly values of the four stations were calculated, and four classes were extracted based on these averages, which correspond to four distinct types of rainfall. Whereas Mebrhatu et al. (2007) modeled for prediction groups of rainfall (below, above, normal) in the highlands of Eritrea.

2.2.1 Mining Rainfall Data

Over the past years, data mining-based approaches have been widely applied in many areas where physical models were infeasible. The related applications of data mining approaches include weather forecasting, storm detection, manufacturing, engineering, and science (Marzano et al., 2010; Kusiak et al.,

2013). Rainfall prediction is a good problem to be solved by data mining techniques (Liu et al., 2001).

Five data mining algorithms: k-nearest neighbor (K-NN), neural network (NN), support vector machine (SVM), classification and regression tree (C&RT), and random forest, were used by Kusiak et al. (2013) to build a rainfall prediction model. Among the presented five data mining algorithms, NN has achieved the best performance. It has been selected to predict rainfall in Iowa City.

Hluchy et al. (2010) applied the temporal spatial data mining technique to develop a model for short-term rainfall prediction by mining a database of radar images in connection with historical rainfall data, measured by a rainfall gauging station network which contains more than 600 stations in Slovakia. The data were normalized, represented and integrated to have temporal and spatial synchronization.

2.2.2 Grades of Rainfall

The standard rain gauge is composed of a funnel attached to a graduated tube that fits into a larger cylinder, as shown in Figure 2.1. The outside container would catch the water that overflows from the graduated tube. So for measurement, the tube will be measured and then the excess water will be put into another tube and measured. Usually, the tube is marked in mm and it may be found in 100-mm (4-in) plastic and 200-mm (8-in) metal varieties.

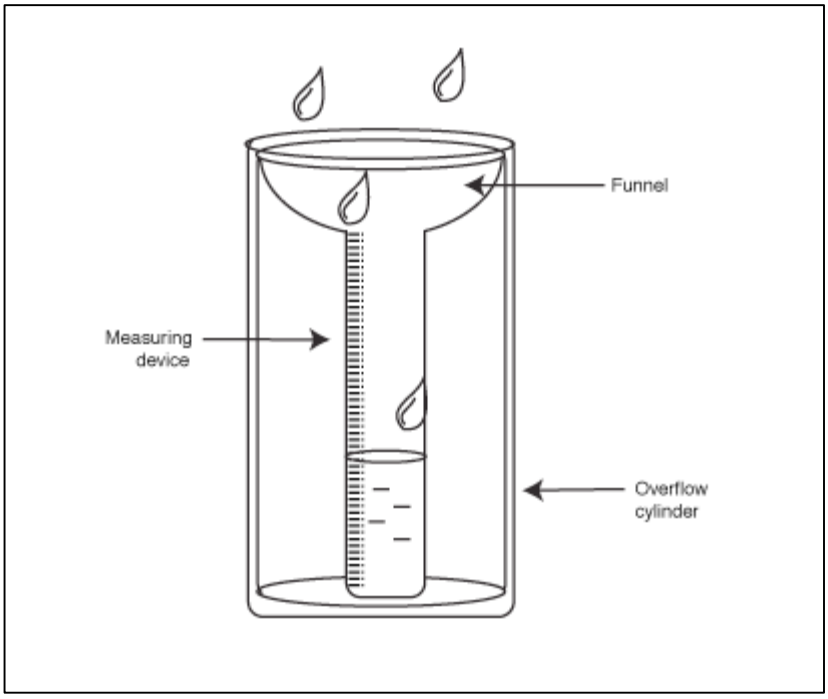

Figure 2.1: Rainfall measurement gauge

Depending on the aforementioned tool to measure the rainfall metrology, the scientist classifies the rain into grades. Table 2.1 shows the rainfall grades by mms (AMS, 2014; IMD, 2014).

Table 2.1: Rainfall grades by mms

Descriptive Term used	**Rainfall amount in mms**
No Rain	0.0
Very light Rain	0.1- 2.4
Light Rain	2.5 – 7.5
Moderate Rain	7.6 – 35.5
Rather Heavy	35.6 – 64.4
Heavy Rain	64.5 – 124.4
Very Heavy Rain	124.5 – 244.4
Extremely Heavy Rain	$\geq$ 244.

2.2.3 Features for Rainfall Forecasting

Weather and climate have intense influence on life on Earth. Weather talks about the state of the atmosphere for a given time and place, with respect to weather elements such as temperature, pressure, precipitation, humidity, direction, and speed of wind and so forth. The weather is highly variable, e.g. in respect of daily basis or weekly basis observations. While climate describes the average weather in terms of the mean and its variability over a certain time span and a certain area. The variety of climate from place to place depends on latitude, vegetation, distance to the sea, existence or nonexistence of mountains or other geographical factors. It also varies in terms of time; such as from season to season, year to year, decade to decade. The basic information of weather and climate focuses on those variables that affect daily life most directly: minimum, maximum and average temperature, precipitation in its various forms, wind near the surface of the Earth, solar radiation, cloud amount, and type and humidity. All these variables are observed on an hourly basis by a large number of satellites and weather stations around the globe (Panigrahi et al., 2014).

In forecasting, the model that is being developed is used to estimate new and unseen cases. Daily rainfall data can be presented as a time series set, as it consists of a sequence of values in time. A preliminary testing has been performed by Srikalra et al. (2006) with different sizes of time series, from which, the size of five inputs gives a better result. Each input pattern consists of five values with the interval of 15 minutes from the rainfall data in the period of 2002-2005 from 20 rainfall monitor stations around the Chao Phraya River in Thailand.

Using the rainfall data of Yangon, Myanmar (1970-1997), Mar et al. (2008) performed test cases with the changes from 1 to 10 days of delay to check how the prediction errors of the model are being affected and choose the most accurate days of delay. The test case results showed a lower forecasting error when using a three-day delay in forecasting the rainfall in Yangon region.

The daily data for 23 years from 1986 to 2010 were arranged as a time series by Khalili et al. (2011). Their work showed that using the rainfall data for the corresponding day in the previous year and the last five bi-daily rainfall moving-averages resulted in a much better forecasting performance from using only the last five bi-daily rainfall moving-averages.

2.3 Artificial Neural Network

ANN is a mimic of biological neural system. The ability of human to learn from examples had pioneered McCulloch-Pitts with the first structure of neural network in 1943. Since then, ANN has been progressively explored by researchers with new improvements on its architecture, learning algorithms and other add-ons. Nowadays, many variants of ANN exist with improved performance on different areas of applications.

2.3.1 ANN and Statistical Methods in Hydrological Forecasting

After publishing the paper of Box and Jenkins (1976), ARIMA and ARMA models or Box–Jenkins models became one of the general time series models to hydrological forecasting. An ARIMA model is a generalization of an ARMA model. Access to basic information requires integration from the series (for a continuous series) or calculating all of the differences of the series (for a continuous series). Since the constant of integration in derivation or differences is deleted, the probability of using of these amounts or middle amounts in this process is not possible. Therefore, ARIMA models are nonstatic and cannot be used to reconstruct the missing data. However, these models are very useful for forecasting changes in a process (Karamouz & Araghinejad, 2012).

One famous black box model that is used to forecast river flow in recent decades is the artificial neural network model. Artificial neural networks are free-intelligent dynamic system models that are based on the experimental data, and the knowledge and covered law beyond data changes to network structure by trends on these data (Menhaj, 2012). The use of time series models (ARMA and ARIMA) and artificial neural networks has been greatly prevalent in different fields of hydrology.

Balaguer et al. (2008) used the method of Time Delay Neural Network (TDNN) and ARMA model to forecast asking for help in support centers for crisis management. The correlation results were obtained for the TDNN and ARMA models, which were 0.88 and 0.97, respectively. This research showed that the ARMA model was better than TDNN. Toth et al. (2000) used the artificial neural

network and ARMA models to forecast rainfall. The results showed the success of both short-term rainfall forecasting models for flood forecasting in real time. Mohammadi et al. (2005) forecasted the Karaj reservoir inflow using the data of melting snow, artificial neural network and ARMA methods, and regression analysis. 60% of the inflow in the dam happened between April and June, thus, forecasting the inflow for this season is very important for the dam's performance. The highest inflows were in the spring due to snowmelt because of draining in threshold winter. The results showed that artificial neural network has lower significant faults as compared with other methods (Chegini, 2012). Valipour (2012a) determined the critical areas of Iran using the data of 50 years of rainfall and the ARIMA model. He concluded that the ARIMA model was an appropriate tool in forecasting annual rainfall. Kisi and Cigizoglu (2005) used dynamic artificial neural networks to forecast the monthly inflow, storage, and evaporation on Canak Dere basin. The results for monthly saving and monthly evaporation were satisfactory, but the forecasting of the monthly inflow as compared with monthly saving and evaporation had a lower accuracy. They used both radial and sigmoid activity functions in dynamic artificial neural network. Their research results showed that sigmoid activity function has a priority over radial activity function. Valipour (2012b), in another research using time series analysis (AR, MA, ARMA, and ARIMA), investigated a number of required observation data for rainfall forecasting according to the climate conditions. The results showed that time series models were better appropriate for rainfall forecasting in a semi-arid climate. The numbers of required observation data for forecasting one year ahead were 60 rainfall data in a semi-arid climate. Banihabib et al. (2008) also forecasted the inflow to the Dez reservoir at a daily time scale, using static artificial neural network model and simple linear

regression model, based on the discharge data from hydrometric stations located upstream of the desired station on minor and major river branches. This research showed that static artificial neural network is better than linear regression models.

Therefore, considering the aforementioned performed researches, it can be seen that artificial neural network showed a higher efficacy in forecasting and sampling in the hydrologic field as compared with other statistic models such as linear and nonlinear regression models.

ANN has been deployed in many different disciplines and also successfully applied to water resources management problems. Even though hydrologists attempt to provide rational answers to problems that arise in the design and the management of water resources, most hydrological processes exhibit a high degree of temporal and spatial variability and are further plagued by the issues of nonlinearity of physical process, conflicting spatial and temporal scales, and the uncertainty in parameter estimate. Therefore, hydrologists often apply ANN to the problems of prediction and estimation of rainfall-runoff relationship, contaminant, concentration and water level (Uneset al., 2013).

An attractive feature of ANN is its ability to extract the relationship between the inputs and outputs of a process without the physics behind these processes being explicitly provided. It has been shown by many researchers that the ANN model performs much better than the conventional models (Uneset al., 2013), since different applications in water resources have very complex and highly nonlinear relationships.

Numerous studies have employed ANN in precipitation data forecasting (Hung et al., 2009; Zeng et al., 2011). This is due to its nonlinear ability that is appropriate for seasonal data. Many research focus on forecasting the precipitation amount on different intervals of time, i.e. daily, monthly or yearly. Their purpose of studies is mainly to forecast precipitation occurrences or number of wet days. However, it is different if the focus of interest is in predicting the magnitude of precipitation that can cause severe floods which normally happen with simultaneous days of rainfall. Although the studies on the applicability of ANN in accessing the precipitation extreme are still few in numbers, yet their findings are promising for further exploration (Zeng et al., 2011; Sulaiman et al., 2013).

ANN has been deployed in various studies such as the suspended sediment modeling (Cigizoglu et al., 2004; Cigizoglu & Kisi, 2006; Kisi & Shiri, 2012), water quality modeling in reservoir (Maier & Dandy, 2000), forecasting density plunging depth in dam reservoir (Unes, 2010a), forecasting rainfall and wastewater relationship (El-Din & Smith, 2002), forecasting dam reservoir level (Unes, 2010b), and the modeling of lake and reservoir water level variation (Ondimu & Murase, 2007; Kisi et al., 2012).

2.3.2 ANN Applications

ANNs have a broad applicability to real world business problems. In fact, they have already been successfully applied in many industries (Singh et al., 2014).

ANNs are well suited for prediction or forecasting needs, including: sales forecasting (Lu et al., 2012), risk management (Jin et al., 2011), and target marketing (Marques et al., 2014). ANNs are also used in the recognition of speakers in communications (Dahl et al., 2012), the diagnosis of hepatitis (Basci & Temurtas, 2011), the interpretation of multi-meaning Chinese words, texture analysis (Karahaliou et al., 2014), three-dimensional object recognition (Piekniewski et al., 2012), handwritten word recognition (Rehman & Saba, 2014), and facial recognition (Yesu et al., 2012).

Due to the increasing applications of ANN, many types of ANNs have been designed. When ANNs are used for data analysis, it is important to distinguish between ANN models and ANN algorithms. The ANN models are the network's arrangement, whereas ANN algorithms are computations that eventually produce the network outputs. The structure of a network depends on its application. When a network is structured according to its application, it is ready to be trained. The two approaches to training are supervised and unsupervised. The aim in supervised learning is to predict one or more target values from one or more input variables. Therefore, in this case, both inputs and outputs should be known. The supervised ANNs are useful for prediction or classification purposes. ANNs with unsupervised algorithms are known as Kohonen or self-organizing maps. These algorithms are excellent in finding relationships among complex sets of data.

The power of artificial neural networks (ANNs) as pattern classifiers, feature selectors, or paradigms for modeling complex data has been used in many fields (Dreyfus, 2005) such as character recognition, image compression, stock market

prediction, medicine, electronic noses, security, and loan applications, as well as for modeling bioactivity. Many of these applications use very large networks with hundreds or thousands of neurons.

2.4 Rainfall Forecasting Models

Rainfall is considered one of the main elements for the management of water resources for the purpose of making decisions and planning, most particularly with regards to agricultural. Providing an accurate prediction for rainfall quantity contributes hugely in many departments, such as deciding crop planting, allocating water reservoir, controlling traffic, maintaining the operations of sewer systems (Hung et al., 2009) and making the appropriate preparation against water influenced disasters like flood or draught (Chantasut et al., 2004), especially in countries where agriculture is an important factor to the wealth and financial state of the country. Hence, an accurate forecast of rainfall will help in avoiding natural disaster (Htike et al., 2010).

Rainfall forecasting can be applied to various time horizons, such as long term, medium term, and short term periods (Htike et al., 2010). Some of the systems are designed to forecast monthly data, and others forecast yearly data (Mar et al., 2008; El- Shafie et al., 2011a). And a few other try to forecast daily data (Srikalra et al., 2006; Wang et al., 2011).

2.4.1 General Rainfall Forecasting Models

Mar et al. (2008) developed a model to predict rainfall of a monthly precipitation in Yangon region, Myanmar. The model used was a 3-layer neural network with different network architectures to find the optimal parameters. Khalili et al. (2011) used the daily rainfall data of March as a month with high humidity, and May and December for medium humidity from 1986 to 2010 in the modeling of ANN for daily rainfall forecasting in Mashhad.

An hourly data collected from 75 rain gauge stations for a period of four years in Bangkok, Thailand were used by Hung et al. (2009) to develop an ANN model, which has been used for real time flood management and rainfall forecasting in Bangkok, Thailand. Monthly rainfall records in the period of 1941-1999 from 245 rainfall gauging stations in Thailand were used by Chantasut et al. (2004) to develop a back propagation neural network (BPNN) model to predict the monthly rainfall in Thailand.

Lee et al. (2012) made an experiment of forecasting heavy rainfall using weather data obtained from European Center of Medium-Range Weather Forecasts (ECMWF), which produces weather data every six hours for the area around the Korean Peninsula. Single feature and combination of weather data were used for forecasting. As a result of this, the forecast using the combination of the features which produces a stable result has a better performance compared to the forecast result using single feature. In addition, feature selection using genetic algorithms is conducted to realize Support Vector Machine, which produces a higher accuracy rate with lower dimensions.

Hou et al. (2013) studied the impact of three-dimensional variation data assimilation (3DVAR) on forecasting two heavy rainfall events in southern China in June and July. They used two heavy rainfall events: one event which affected several states in southern China with severe flooding and heavy rain; the other is characterized by nonuniformity and extremely high rainfall rates in localized areas. Their study outcome showed that the assimilation of all surface, radar, and radiosonde data has a more positive impact on the forecasting skill than the assimilation of other types of data for the two rainfall events. The forecasting system used in the study is characterized by combining the Advanced Weather Research and Forecasting (WRF-ARW) model and the Advanced Regional Prediction System (ARPS).

Afandi et al. (2013) in his study used the Weather Research and Forecasting (WRF) model to investigate the flash flood over the Sinai Peninsula that was caused by heavy rainfall events. This flood occurred on 18th January 2010, in Egypt, which has been predicted and analyzed using the WRF-ARW Model. The rainfall that has been predicted in four dimensions (time and space) has been calibrated with a recorded measurement at rain gauging stations. The result of the study showed that the WRF model can capture the heavy rainfall events over different regions of Sinai. The authors also found that the accuracy of the prediction of the WRF model was high when compared with real measurements.

Routray et al. (2012) studied a performance-based comparison of simulations carried out using the three-dimensional variation (3DVAR) data assimilation system and nudging (NUD) technique of a heavy rainfall event that occurred in

June 2005 in the west coast of India. In the experiment, after observations using the 3DVAR data assimilation technique, the model was able to simulate a better structure of the convective organization, as well as distinct overall features related to the mid-tropospheric cyclones (MTC), than the NUD experiment, and was almost similar to the observed data. Table 2.2 shows the data, methods and algorithms used by the reviewed studies in this chapter for rainfall forecasting.

Table 2.2: Summary of rainfall forecasting models

Author	Reading basis	Data (Predictors)	Method	Forecasting type (output)	Period of data	Area
Mar et al. (2008)	Daily	rainfall records	BPNN	Monthly rainfall	1970 - 1997	Yangon region, Myanmar
Hung et al. (2009)	Hourly	Wind speed, wet bulb temperature, relative humidity, cloudiness, dry bulb temperature, and air pressure and rainfall data.	BPNN	Monthly rainfall	1991 - 2003	Bangkok, Thailand
Khalili et al. (2011)	Daily	rainfall records	BPNN	Daily rainfall forecasting	1986 - 2010	Mashhad
Lee et al. (2011)	Six hours	Height, Humidity, Temperature, Uwind and Vwind	SVM	Heavy rainfall	1989 – 2009	Korea Peninsula

Routr-ay et al. (2012)	Hourly	- Upper-air profiles of humidity, wind and temperature - Wind profiles - Upper-level wind and temperature - observed cloud motion vectors, wind and total perceptible water	(NUD) technique and (3DVAR)	Heavy rainfall	25th – 28th June 2005	India
Hou et al. (2013)	Daily	radars, surface AWS, radiosonde, and rain gauge measurements	The (WRF-ARW) model and the (ARPS) 3DVAR	Heavy rainfall	-	Southern China
Afandi et al. (2013)	Hourly	rainfall records	(WRF) Model	Heavy rainfall	-	Sinai Peninsula, Egypt

2.4.2 Extreme Rainfall Forecasting Models

In recent years, an increased interest was shown in regards of the obvious rise in the frequency and/or severity of predicting extreme events for many countries around the world. The development of an accurate and timely extreme event monitoring and predicting system stands as one of the most important ways in avoiding the potential impacts that climate variations and extreme weather pose (Zeng et al., 2011). The traditional techniques for the forecasting of statistical weather include ARMA models, Multivariate Adaptive Regression Splines, and Box-Jenkins Models; and when machine learning became popular, many

attempts have been made to develop rainfall forecasting models which used feedforward neural networks, recurrent neural networks, that include input delays and BPNN. Apart from that, many attempts were also made to involve extra weather parameters in rainfall forecasting models for better forecasting (Htike et al., 2010).

Gu et al. (2010) chose the Yangtze River daily precipitation data, and used stepwise regression analysis on the six extreme precipitation indices (PX5D, PFL90, PNL90, PXCDD, PQ90 and PINT) proposed by (STARDEX) to gain the main indicants. Table 1.1 in Chapter One shows the details of the indices. The results showed that the impact of PX5D, PFL90, PNL90, and PXCDD are significant in forecasting average extreme rainfall, while PINT and PQ90's impact are insignificant. These data and indices were used to develop an average extreme rainfall prediction model that is based on back propagation neural network (BPNN).

Based on data mining, Wan et al. (2012) have developed a forecasting model to predict the next year's average extreme rainfall using BPNN, in which the input data that were used in the neural network are the six core rainfall indices proposed by the European Union Statistical and Regional dynamical Downscaling of Extremes for European regions project (STARDEX). The annual average extreme rainfall prediction model has been combined with the stepwise discriminant method and uses the Bayesian statistical method to further improve the neural network's generalization ability and model forecasting accuracy.

Junaida et al. (2012) used the stepwise regression for input variable selection (IVS) method, while the ANN method was selected for model development. Based on the IVS results, it is revealed that average temperature, minimum temperature and daily precipitation at one day and three days back are the significant subset of input variables when predicting heavy precipitation. The meteorological elements were collected from a weather station located in Kuantan, Pahang (a state in the east coast of Peninsular Malaysia). Later, these inputs were presented to ANN models. The primary benefit that was found of IVS coupled with the ANN approach is that it identifies the minimum number of ANN input variables required in the prediction of heavy precipitation, without much loss of prediction accuracy. In order to evaluate the forecasting model output, MAE and RMSE performance metrics have been calculated to find which model performs better.

Sulaiman et al. (2013) applied a method for a one-step ahead forecasting of precipitation extreme using the maximum five-day precipitation values. The data used were from 1980 to 2011 on a monthly basis. The original form of data are the daily precipitation collected from a weather station located in Kuantan, Pahang. The best inputs are determined by the stepwise regression method to be used in the ANN model which was developed. The auto regressive moving average (ARIMA) model has been developed using the same dataset to evaluate the ANN model output.

Sulaiman et al. (2014) developed a model to forecast monthly extreme precipitation using artificial neural networks (ANNs) and utilizing past PX5D data and global climate indices such as Madden Julian Oscillation (MJO),

Southern Oscillation Index (SOI), and Dipole Mode Index (DMI) in Kuantan and Kota Bharu, Malaysia. Two statistical methods, multiple liner regression and ARIMA models, were developed using the sample data used in the ANN model, the performance metrics, MAE and RMSE, were calculated and compared with all the models in order to evaluate the developed ANN model.

Table 2.3: Summary of extreme rainfall forecasting models

Authors	Area	Variables	Forecasting Basis	Target	Method	Learning algorithm
Gu et al. (2010)	Kelantan and Terengganu	Eight Metrological, two climate indices	Monthly	PX5D	BPNN	LM
Wan et al. (2012)	Kuantan	PX5D	Monthly	PX5D	BPNN	PSO
Junaida et al. (2012)	Kota Bharu, Kelantan	PX5D, three climate indices	Monthly	PX5D	BPNN, MLR	LM,PSO
Sulaiman et al. (2013)	China	Six indices	Annual	AVY of above 90^{th} %	BPNN	LM, GDX, BR
Sulaiman et al. (2014)	China	Six indices	Annual	AVY of above 90^{th} %	BPNN	-

2.6 Research Gap

In the previous researches in forecasting extreme rainfall, the six extreme rainfall indices have been used to forecast the average of the rainfall for the next year,

but have not been used to forecast the maximum consecutive five days for the next month. Furthermore, other studies used the values of the maximum five days for the last 24 months. This study focuses on forecasting extreme rainfall in Perlis, a state in the north of Peninsular Malaysia, using artificial neural network models, and by applying the six extreme rainfall indices as predictors. Furthermore, different lag length of maximum five consecutive days are experimented to find the best forecasting accuracy, and to compare these two models with a statistical model.

2.7 Summary

Based on the literature review, different computational intelligence models have been developed to forecast different weather events. ANN has been widely applied for rainfall forecasting by a number of researchers in various countries and has also been applied for extreme rainfall forecasting. Thus, in this study, ANN is explored in the forecasting of extreme rainfall event.

Chapter Three
Methodology

This chapter describes the methodology which is based on knowledge discovery in databases (KDD) that is adopted to develop the extreme rainfall forecasting models. Section 3.1 provides an overview of the methodology. The steps of the methodology are discussed in Section 3.2. The summary of this chapter is presented in Section 3.3.

3.1 Overview of Methodology

The KDD process, as presented in Fayyad et al. (1996), is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformation of the database. There are five stages: (i) Selection - this stage aims in creating a target dataset, or focuses on a subset of variables or data samples, on which discovery is to be performed; (ii) Preprocessing - this stage consists on the target data cleaning and preprocessing in order to obtain consistent data; (iii) Transformation - this stage consists of the transformation of the data using dimensionality reduction or transformation methods; (iv) Data Mining - this stage consists of the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction); (v) Evaluation - this stage comprises the evaluation of the mined patterns.

The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. The SAS Institute considers a

cycle with five stages for the process: (i) Sample - this stage consists of sampling the data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly; (ii) Explore - this stage involves the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas; (iii) Modify - this stage consists of the modification of the data by creating, selecting, and transforming the variables to focus on the model selection process; (iv) Model - this stage comprises modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome; (v) Assess - this stage involves assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

CRISP-DM stands for CRoss-Industry Standard Process for Data Mining. It consists of a cycle that comprises six stages: (i) Business Understanding - this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan is designed to achieve the objectives; (ii) Data Understanding - the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information; (iii) Data Preparation - the data preparation phase covers all activities to construct the final dataset from the initial raw data; (iv) Modeling - in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values; (v) Evaluation - at this stage, the model (or models) obtained are more thoroughly

evaluated and the steps executed to construct the model are reviewed to be certain they properly achieve the business objectives; (vi) Deployment – the creation of the model is generally not the end of the project.

By performing a comparison of the KDD and SEMMA stages, on a first approach, it could be affirmed that they are equivalent: Sample can be identified with Selection; Explore can be identified with Preprocessing; Modify can be identified with Transformation; Model can be identified with DM; Assess can be identified with Evaluation. Examining it thoroughly, it may be affirmed that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS Enterprise Miner software. Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation. Nevertheless, it can be observed that the CRISP-DM methodology incorporates the steps that must precede and follow the KDD process, that is to say: the Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user; the Deployment phase can be identified with the consolidation by incorporating this knowledge into the system. Concerning the remaining stages, it can said that: the Data Understanding phase can be identified as the combination of Selection and Preprocessing; the Data Preparation phase can be identified with Transformation; the Modeling phase can be identified with DM; the Evaluation phase can be identified with Evaluation. In Table 3.1, a summary of the three methodologies' steps are presented.

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Brachman & Anand, 1996). The KDD process is preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It must be continued by the knowledge consolidation, incorporating this knowledge into the system (Fayyad et al., 1996).

Table 3.1: KDD, SEMMA and CRISP-DM steps

KDD	**SEMMA**	**CRISP-DM**
Pre KDD	----------	Business Understanding
Selection	Sample	Data Understanding
Preprocessing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

The development of extreme rainfall forecasting models are performed based on the knowledge discovery in databases (KDD) methodology (Fayyad et al., 1996). The KDD process involves the use of the database with all the necessary steps of selection, preprocessing, transformation, applying data mining algorithms (techniques) to find patterns from the data, and the evaluation of the

data mining output (Figure 3.1). The KDD process may contain a number of iterations and there may be loops between any two of the steps.

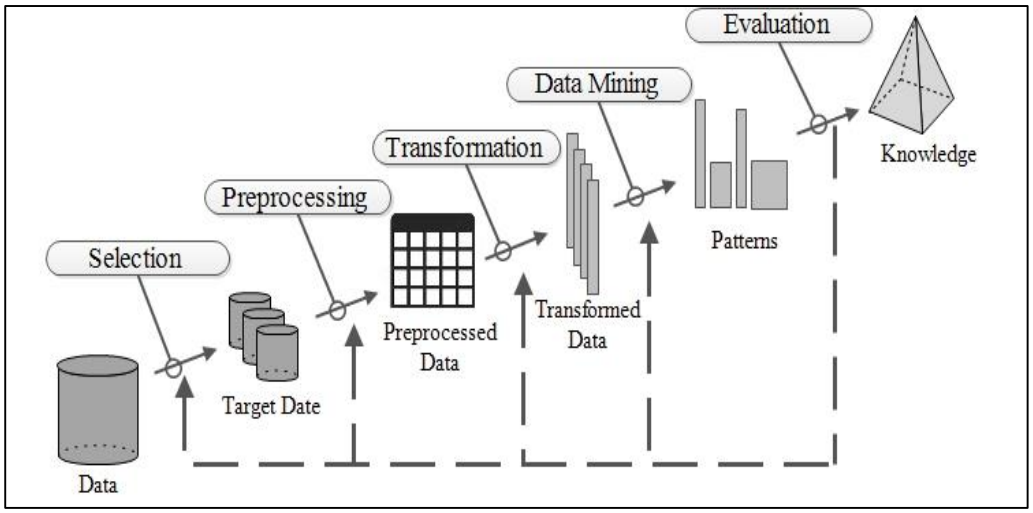

Figure 3.1: Overview of the KDD process steps

3.2 KDD Steps

The main goal of the KDD process is to extract knowledge from the data in the context of large databases. The knowledge discovery process is repetitive, interactive, and consists of a number of steps. The process is repetitive at each step, meaning one might have to move back to the previous steps.

3.2.1 Data Collection and Selection

Flood mitigation is one of the tasks under the Department of Irrigation & Drainage Malaysia (DID). DID has built and managed 16 reservoirs in Malaysia. Seven of the reservoirs were built with flood mitigation as one of their purposes.

Out of these seven, only Timah Tasoh reservoir is fully operated based on human decision making. Other reservoirs combine both human and mechanical operations. Therefore, this study focuses on Timah Tasoh reservoir which is located in the state of Perlis.

Five gauging stations that are located at Timah Tasoh upstream have been identified for the purpose of this study. The gauging stations are Padang Besar, Tasoh, Lubuk Sireh, Kaki Bukit, and Wang Kelian. Figure 3.2 shows the location of Timah Tasoh reservoir and the five gauging stations. These gauging stations record the volume of rainfall that fall and disperse into the river flows into Timah Tasoh reservoir. The data are recorded on a daily basis, and consist of 187 months of daily rainfall records for the period of April 1998 until October 2013. Table 3.2 shows an example of the daily records of rainfall for the period between 16th October 2006 and 23rd October 2006 for each of the five gauging stations measured in millimeter.

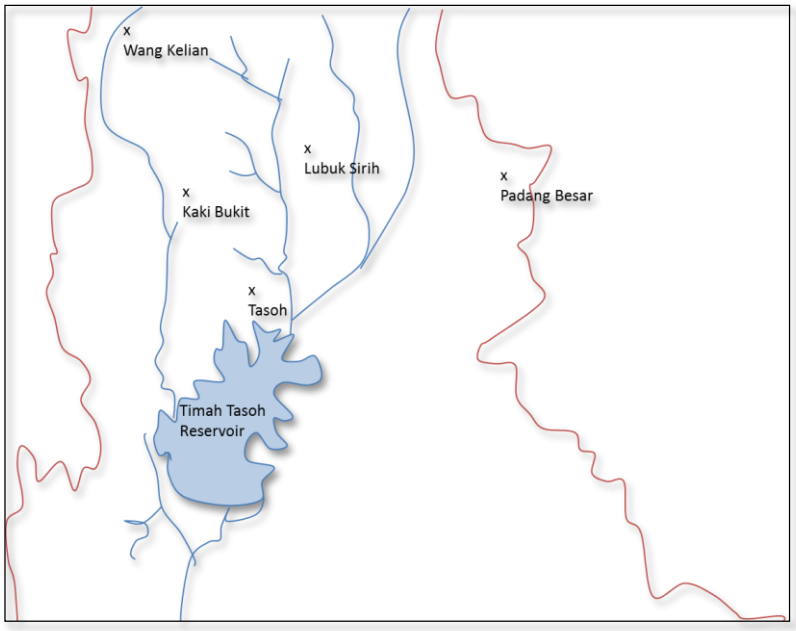

Figure 3.2: Timah Tasoh reservoir and five gauging stations

Table 3.2: Example of the rainfall data (in mm) by gauging station

Date	Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
16-Oct-06	35.00	55.00	50.00	0.00	49.50
17-Oct-06	0.00	0.00	0.00	10.00	39.50
18-Oct-06	0.00	0.00	0.00	0.00	29.50
19-Oct-06	0.00	50.00	91.00	20.00	47.50
20-Oct-06	0.00	0.00	2.00	11.00	2.50
21-Oct-06	5.50	0.00	25.00	100.00	7.50
22-Oct-06	25.00	43.00	9.00	20.00	4.50
23-Oct-06	14.50	0.00	6.00	10.00	5.50

3.2.2 Preprocessing

In this stage, the data are preprocessed. The aim of this stage is to clean and prepare the data for the next stage. The data are examined to detect any mistakes during data entry and missing values. The missing value is replaced with the average of previous and after value such as in Equation 3.1. This equation has been deployed by Papalexiou and Koutsoyiannis (2013) to replace missing daily climate records.

$$new\_value = \frac{v_{t-1}+v_{t+1}}{2} \qquad \text{(Equation 3.1)}$$

For example, it is found that there are missing rainfall data of Wang Kelian gauging station, as shown in Table 3.3. The value is replaced with the average value from 12th February 2006 and 14th February 2006. Therefore, the value on 13th February 2006 is 33.

Table 3.3: Example of the rainfall data (in mm) with missing value

Date	Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
12-Feb-06	85.00	32.00	55.00	0.00	58.50
13-Feb-06	2.00	9.50	6.00	0.00	?
14-Feb-06	32.50	22.00	34.00	0.00	7.50

3.2.3 Transformation

Once the data have been preprocessed, the six core indices (Table 3.4) are calculated on a monthly basis in order to find the characteristics of extreme rainfall; the six extreme precipitation indices are PQ90, PX5D, PINT, PFL90, PNL90, and PXCDD.

Most of the indices are based on thresholds defined using percentile values rather than fixed values. The six STARDEX core indices for rainfall is shown in Table 3.4. The indices encompass frequency (e.g. days of heavy rainfall) and persistence (e.g. longest dry period) of extremes. The rainfall indices provide a good mix of measures of intensity (PX5D, PINT, and PQ90), frequency

(PXCDD and PNL90), and proportion of total (PFL90). All thresholds are percentile-based and so can be used for a wide variety of climates (Haylock, 2005). Some of the indices consider properties of merely the rain day distribution (PQ90, SDII), while the others use the entire distribution.

Table 3.4: STARDEX extreme rainfall indices (Gu et al., 2010)

Index Name	Details	Description
PQ90	90^{th} percentile of rain day amount (mm/day)	Heavy rainfall threshold
PX5D	Greatest five-day total rainfall (mm)	Greatest five-day rainfall (amount)
PINT	Simple daily rainfall intensity (rain per rain day)	Average wet-day rainfall (amount)
PFL90	% of total rainfall from events>long-term 90^{th} percentile	Heavy rainfall proportion
PNL90	Number of events>long-term 90^{th} percentile of rain days	Heavy rainfall days
PXCDD	Maximum number of consecutive dry days	Longest dry period

The maximum five-day accumulated precipitation (PX5D) as an index was proposed by both WMO and STARDEX to identify extreme events that could affect human life and the natural environment. Previous studies have indicated the importance of evaluating extreme precipitation events based on successive days of precipitation amounts (Zeng et al., 2011; Foresti et al., 2010). This is significant because the risk of flood increases after several days of precipitation.

The PQ90 index is the 90th percentile value of the rainy day, this index is calculated for each month to be used as one of the inputs in the forecasting models. To calculate the percentile, all the rainfall records that are smaller than 1 mm are removed; the reason is to calculate the percentile of the rainy days only, and then the 90th percentile of the remaining records is calculated.

The second index is the greatest five consecutive days of rainfall amount (mm); this index represents the maximum total amount of five consecutive days of rainfall amount. To calculate the PX5D index, Equation 3.2 (Sulaiman et al., 2014) is used.

$$PX5D = max\{(\sum_{n=1}^{5} RR_n)\ in\ a\ month\} \qquad \text{(Equation 3.2)}$$

Where RRn is the daily rainfall amount.

The PINT index represents the simple daily rainfall intensity, in which the days with the rainfall amount of below 1.0 mm are removed in order to find the average rainfall amount of wet days.

The fourth index that is PFL90 represents the percentage of the rainfall amount that is greater than the 90th percentile to the number of rainfall events with a rainfall amount greater than 1.0 mm (wet days) in a month.

The PNL90 index is the number of rainfall events that is greater than the 90th percentile of rain days (days with a rainfall amount greater than 1.0 mm). This index is calculated by removing the days that have rainfall amounts lower than

the 90th percentile of the specified month, wherein the remaining days are counted to find the PNL90 index value.

The last calculated index is the PXCDD index; this index represents the maximum number of consecutive dry days (rainfall amount less than 1.0 mm) within a month.

Table 3.5 shows an example of the daily rainfall records of November 2012 for all the five gauging stations, while Table 3.6 shows the calculated values of the six indices of the same month. A summary of the six extreme rainfall indices is shown in Table 3.7.

Table 3.5: Daily rainfall (mm) amounts for November 2012

Date	Gauging Station				
	Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
1-Nov-12	10.00	0.00	0.00	0.00	20.50
2-Nov-12	0.00	0.00	0.00	0.00	0.00
3-Nov-12	0.00	0.00	0.00	0.00	0.00
4-Nov-12	13.00	5.50	10.00	5.00	11.00
5-Nov-12	14.00	0.00	4.00	6.00	2.00
6-Nov-12	0.00	3.50	0.00	0.00	8.00
7-Nov-12	0.00	0.00	0.00	0.00	0.00
8-Nov-12	5.00	0.00	0.00	0.00	1.00
9-Nov-12	0.00	0.00	0.00	0.00	0.50
10-Nov-12	4.00	0.00	0.00	0.00	0.00
11-Nov-12	0.00	3.50	7.00	10.00	0.00
12-Nov-12	0.00	69.50	10.00	25.00	4.50
13-Nov-12	0.00	15.50	0.00	0.00	6.50
14-Nov-12	0.00	0.00	0.00	0.00	0.00
15-Nov-12	0.00	0.00	0.00	0.00	0.00
16-Nov-12	0.00	0.00	0.00	0.00	0.00
17-Nov-12	0.00	0.00	20.00	8.00	0.00
18-Nov-12	0.00	5.50	19.00	11.00	12.00
19-Nov-12	45.00	4.50	75.00	55.00	16.00
20-Nov-12	0.00	0.00	8.00	6.00	29.00
21-Nov-12	0.00	0.00	0.00	0.00	7.00
22-Nov-12	0.00	25.30	1.00	2.00	0.00
23-Nov-12	26.00	60.50	34.00	16.00	0.00
24-Nov-12	0.00	100.00	51.00	60.00	21.00
25-Nov-12	0.00	25.60	13.00	8.00	45.00
26-Nov-12	65.00	0.00	8.00	8.00	6.50
27-Nov-12	98.00	0.00	6.00	18.00	7.00
28-Nov-12	4.00	4.50	0.00	0.00	46.50
29-Nov-12	15.00	7.50	25.00	25.00	0.00
30-Nov-12	61.00	0.00	0.00	0.00	41.00

Table 3.6: Calculated values of the six core indices for November 2012

Index	Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
PQ90	64.60	67.70	44.20	43.00	42.60
PX5D	243.00	211.40	122.00	110.00	126.00
PINT	18.56	14.59	12.37	13.07	17.93
PFL90	0.17	0.15	0.13	0.13	0.12
PNL90	2	2	2	2	2
PXCDD	8	4	5	5	4

Table 3.7: Summary of calculated indices by gauging stations

Index		Gauging Station				
		Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
PQ90	Minimum	0	0	0	0	0
	Average	31.05	34.04	30.45	25.46	31.66
	Maximum	97.80	115.00	101.40	128.00	116.80
PX5D	Minimum	0	0	0	0	0
	Average	77.42	79.35	81.46	70.29	96.69
	Maximum	281.00	269.00	248.00	331.00	387.00
PINT	Minimum	0	0	0	0	0
	Average	15.79	18.99	14.82	13.64	14.74
	Maximum	48.82	68.00	51.32	58.25	43.60
PFL90	Minimum	0	0	0	0	0
	Average	0.37	0.37	0.37	0.30	0.38
	Maximum	0.91	0.93	0.94	0.96	0.89
PNL90	Minimum	0	0	0	0	0
	Average	1.22	1.19	1.39	1.24	1.59
	Maximum	3.00	3.00	3.00	3.00	3.00
PXCDD	Minimum	2	2	2	2	1
	Average	10.86	11.24	9.49	8.84	9.05
	Maximum	31	31	31	31	31.00

WMO and STARDEX define the extreme rainfall threshold as the 90^{th} percentile of the rainfall records for a specified period (e.g. one month). Table 3.5 shows the average 90^{th} percentile for each gauging station, calculated on a monthly basis for 187 months. The average of extreme rainfall threshold of Pedang Besar is 31.05 mm, Tasoh 34.04 mm, Lubuk Sireh 30.45 mm, Kaki Bukit 25.46 mm, and Wang Kelian 31.66 mm.

The value of the extreme rainfall threshold can reach a maximum value of 97.80, 115.00, 101.40, 128.00, and 116.80 for Pedang Besar, Tasoh, Lubuk Sireh, Kaki Bukit, and Wang Kelian respectively.

3.2.4 Data Mining

Data mining development has naturally led to the exploration of application domains that employ data mining. Knowledge discovery in databases (KDD) is the process of extracting meaningful information from the databases. Temporal data mining task is so crucial when compared to the ordinary mining, because it handles the time related features of the dataset (Shahnawaz et al., 2011). Temporal data mining is an important extension of data mining and it is a nontrivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content from a large database.

Three main steps are followed for discovering the relations between sequences of events: (i) the modeling and representing of the data sequence in a suitable form; (ii) defining the similarity measures between sequences; and (iii) the

application of models and representations to the actual mining problems. A sequence consists of a series of nominal symbols from a particular alphabet which is usually called a temporal sequence, and a sequence of continuous, real-valued elements, is known as a time series (Geetha et al., 2008).

The temporal data mining procedure is executed to frame the temporal pattern of the rainfall. Temporal data mining is applied to extract patterns that represent sequences of extreme rainfall events. Temporal data usually represent sequences of events which are usually the impact of certain causes due to time delays. The sliding window technique is used to extract temporal patterns in the data mining step. The sliding window technique was proven to be able to detect patterns from the temporal data (Keogh et al., 2001; Wan Ishak et al., 2011). This technique is used to capture the time delay between the cause of the event and the actual event.

In this stage, four datasets have been formed, namely Dataset-A, Dataset-B1, Dataset-B2 and Dataset-B3. Dataset-A has the six core indices as predictors and the target value is the PX5D value for the next month. Previous studies (Gu et al., 2010; Wan et al., 2012) showed the importance of forecasting extreme rainfall by using the six extreme rainfall indices as predictors.

Dataset-B1, Dataset-B2, and Dataset-B3 contain only the PX5D index value that is calculated based on a monthly basis. Previous studies (Zeng et al., 2011; Junaida et al., 2012; Sulaiman et al., 2013; Sulaiman et al., 2014) indicated that the PX5D index value for the next month can be forecasted by using the values of the same index for the previous months, different lag lengths have been used

to forecast the PX5D value for the next month. Thus, in this study, three different lag lengths are experimented.

The sliding window technique is applied to segment the temporal pattern for Dataset-B1, Dataset-B2, and Dataset-B3. In this study, three different sliding window sizes are used to find which number of months' delay have the most effect on the accuracy of forecasting the PX5D index value, one for each dataset. Each window size represents the time duration of the delays. Dataset-B1 is created using a sliding window size of 3 that represents a three-month delay. Dataset-B2's sliding window size is 6, while Dataset-B3's sliding window size is 12 which represents 12 months of delay. All of these three datasets have the PX5D value of the next month as a target.

Each dataset is divided into three continues series of data (Bigus, 1996): training set (80%), validation set (10%), and testing set (10%). The training set is used in the training phase of ANN, while validation set is used to validate ANN's performance during the training. Testing set is used to test the performance of ANN after the training has completed. Table 3.8 shows the division of collected data by period.

Table 3.8: Data division

Series	**Series Period**	
	From	**To**
Training	April 1998	September 2010
Validation	October 2010	March 2012
Testing	April 2012	September 2013

Back propagation neural network is a multilayer feedforward neural network that consists of three or more layers of neurons. It includes input layer, hidden layer (middle layer) and output layer. BPNN in rainfall forecasting has been utilized by numerous studies. Mar et al. (2008) applied BPNN to forecast rainfall in Myanmar, Hung et al. (2009) used the data collected from 75 metrological gauging stations to train a BPNN model that could forecast rainfall in Bangkok, Thailand, Gu et al. (2010) established an average extreme rainfall forecasting model based on BPNN, Khalili et al. (2011) used BPNN to design a rainfall forecasting model in Mashhad city.

In this study, four experiments using back propagation neural network models are developed, namely Experiment-A, Experiment-B1, Experiment-B2 and Experiment-B3. In all the experiments, NN consists of three layers: one input layer, one hidden layer, and one output layer.

For each experiment, different combinations of inputs are used. The number of hidden units is determined by the trial and error method, that is by training and testing NN with different numbers of hidden units to find which number performs better. The unit in the output layer is the same for all networks that is the PX5D index value of the next month.

Dataset-A is used to train and test NN in Experiment-A, the inputs of NN is the six indices as predictors. Therefore, the number of input units are six, while Experiment-B1, Experiment-B2 and Experiment-B3 are developed using Dataset-B1, Dataset-B2 and Dataset-B3, respectively. This means the number of

input units in Experiment-B1 is three, Experiment-B2 has six input units, and Experiment-B3 has twelve input units in the input layer.

The design and architecture of the proposed models and experiments are discussed in detail in Chapter Four.

3.2.5 Evaluation

The result of the four BPNN experiments are compared to find which experiment has the lowest error rate. The mean absolute error formula (Equation 3.3) is applied to calculate the error between the network output and the target of each experiment.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|t_k - y_k| \quad \text{(Equation 3.3)}$$

Where *n* is the number of the output.

The BPNN model that has the least error measurement is evaluated by comparing the MAE value of the model with an equivalent regression model.

3.3 Summary

The extreme rainfall forecasting model is developed based on the KDD methodology. According to KDD, the first step is data collection, followed by data selection, in which data from several gauging stations are selected and the others are ignored. Then the data are transformed into the extreme rainfall descriptive indices based on a monthly basis; after applying the aforementioned

steps, temporal data mining is conducted to frame the temporal pattern of the rainfall. Finally, four experiments utilizing the BPNN model are developed for each one of the created dataset.

Chapter Four
Proposed Models

This chapter focuses on the models that have been developed to forecast extreme rainfall. Section 4.1 describes the experiments that utilize BPNN in forecasting. Section 4.2 illustrates the regression model, and in Section 4.3, a summary of the models are stated.

4.1 BPNN Model

Four experiments, namely Experiment-A, Experiment-B1, Experiment-B2 and Experiment-B3, have been conducted using the developed BPNN Model. Each experiment uses different combinations of inputs: Experiment-A applies the six extreme rainfall indices as predictors; in Experiment-B1, the value of PX5D for the previous three months is applied as predictors; for Experiment-B2, the previous six months of PX5D index values are applied as predictors; whereas in Experiment-B3, the values of PX5D for the last twelve months are used as predictors. The experiments' target is the monthly maximum five consecutive days of rainfall amount (PX5D) index for one month ahead. MATLAB software was used in developing the BPNN model of the experiments using neural network tool.

For each experiment, a dataset is created; these datasets are Dataset-A, Dataset-B1, Dataset-B2 and Dataset-B3. The datasets are used to train, validate and test the BPNN of the model; each dataset has a predictor part and a target part. The predictor part varies from one set to another, while the target is the same.

BPNN has been utilized in all the four experiments. One of the most popular NN algorithms is back propagation algorithm (Sulaiman et al., 2013). Each neural network consists of three layers: input, hidden and output layers. The reason why two or more hidden layers have not been adopted is that the intermediate units which are not directly connected to the output units will have small weight changes and will learn very slowly (Gallant, 1993).

Choosing the number of nodes for each layer will depend on the problem NN is trying to solve, the types of data network it is dealing with, the quality of data and some other parameters. The number of input and output nodes depends on the training set in hand.

The number of hidden units is determined by empirical approach, in which NN is retrained with varying numbers of hidden neurons and the output error is observed as a function of the number of hidden units (Rojas, 1996; Priddy & Keller, 2005).

BPNN has been trained using Levenberg-Marquardt back propagation training function (TrainLM). In time series forecasting, particularly in the climate domain, LM is seen as a common choice as a BPNN learning algorithm (El-Shafie et al., 2012; Sulaiman et al., 2014). The Levenberg-Marquardt (LM) algorithm is an iterative technique that locates the minimum of a multivariate function that is expressed as the sum of squares of nonlinear real-valued functions. It has become a standard technique for nonlinear least-squares problems, and widely adopted in a broad spectrum of disciplines. LM can be

thought of as a combination of the steepest descent and the Gauss-Newton methods. When the current solution is far from the correct one, the algorithm behaves like a steepest descent method: slow, but guaranteed to converge. When the current solution is close to the correct solution, it becomes a Gauss-Newton method (Lourakis, 2005; Sapna et al., 2012).

The algorithm used for the initialization of BPNN weights is the Nguyen-Widrow weight initialization algorithm. The Nguyen-Widrow method generates initial weights and bias values for a layer, so that the active regions of the layers of neurons will be distributed approximately evenly over the input space (Nguyen & Widrow, 1990; Castillo et al., 2006).

According to Cilimkovic (2011), activation functions are needed for the hidden layer of NN to introduce nonlinearity. Without them, NN would be same as plain perceptions. If linear function were used, NNs would not be as powerful as they are. Activation functions can be linear, threshold or sigmoid functions. The sigmoid activation function is usually used for the hidden layer because it combines nearly linear behavior, curvilinear behavior and nearly constant behavior, depending on the input value (Larose, 2005). The activation function that is used to generate the output of the BPNN model in the hidden layer units is the tangent-sigmoid function, while the linear activation function is used in the output layer unit (Gu et al., 2010).

Mean Squared Error (MSE) has been used to evaluate the network performance during the training phase. MSE measures the neural network performance according to the mean of squared errors (Gu et al., 2010; Wan et al., 2012).

Two main training parameters were set to terminate the training. First, a goal of the performance function, MSE. The goal is the minimum error needs to be achieved. Second, the maximum epoch/iteration for the training, that has been specified prior to the training. When the validation error continues to arise for several epochs, early stopping is executed. The goal of this procedure is to develop a BPNN model that achieves the best result.

Since the Levenberg-Marquardt (LM) algorithm has been used to train the network, four additional parameters are also needed to be specified, namely, an initial value for the Marquardt parameter (mu), a decrease factor (mu_dec), an increase factor (mu_inc), and the maximum step size (mu_max). Table 4.1 shows the training parameters.

Table 4.1: BPNN training parameters

General ANN Training Parameters	
Goal (Minimum error to be achieved)	0
Performance Function	MSE
Maximum number of epochs	1000
Levenberg- Marquardt (LM) Parameter	
Initial Marquardt Parameter(mu)	0.001
decrease factor (mu_dec)	0.1
Increase factor (mu_inc)	10
Maximum step size (mu_max)	$1*10^{10}$

Rojas (2005) claimed that the BP algorithm could be broken down into four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections. The algorithm can be decomposed in the following four steps:

i. Feedforward computation
ii. Back propagation of error to the output layer
iii. Back propagation of error to the hidden layer
iv. Weight updates

The algorithm is stopped when the value of the error function has become sufficiently small. In the last step, weights are updated throughout the algorithm.

Feedforward computation or forward pass is a two-step process. The first part is getting the values of the hidden layer nodes, and the second part is using those values from the hidden layer to compute the value of the output layer. Then, the error is calculated; once the error is known, it will be used for backward propagation and weights adjustment. Error is propagated from the output layer to the hidden layer first. This is where the learning rate and momentum are brought to equation. Before weights can be updated, the rate of change needs to be found. This is done by the multiplication of the learning rate, error value and node value. In the third step (back propagation to the hidden layer), the error is propagated from the hidden layer down to the input layer. After computing all partial derivatives, the network weights are updated.

4.1.1 Experiment-A

Experiment-A is developed using the six extreme rainfall core indices (outlined by STARDEX) of the previous month as inputs, while the output is the PX5D index value of the next month.

The dataset used to train and test BPNN in this experiment is Dataset-A. Dataset-A has been created to have the values of the six core indices in the predictor part. Previous extreme rainfall forecasting studies (Gu et al., 2010; Wan et al., 2012) indicated the importance of forecasting extreme rainfall based on the six extreme rainfall indices as predictors. While the target part has only one variable, which is the PX5D value of the next month. Table 4.2 shows an example of the created Dataset-A.

Table 4.2: Example of Dataset-A

Predictors						**Target**
PQ90(t)	PX5D(t)	PINT(t)	PFL90(t)	PNL90(t)	PXCDD(t)	PX5D(t+1)
28.00	77.00	15.73	6	0.21	1	45.50
29.50	81.00	20.07	8	0.25	1	56.50
24.60	68.00	11.94	4	0.30	2	34.00
30.40	76.00	13.06	4	0.36	2	39.50
30.70	75.00	12.55	4	0.41	2	85.25

The number of neural network input units is six, the number of hidden neurons is determined using the heuristic approach by training the neural network with different numbers of hidden neurons, the selection criterion is the network output's MAE, the number of units in the output layer is one. Table 4.3 shows

Experiment-A's BPNN architecture. Figure 4.1 shows the basic BPNN architecture of Experiment-A.

Table 4.3: Experiment-A's BPNN architecture

Number of Layers	3
Number of Input Units	6
Number of Hidden Units	To be determined
Number of Output Units	1
BPNN-Learning Algorithm	TrainLM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear

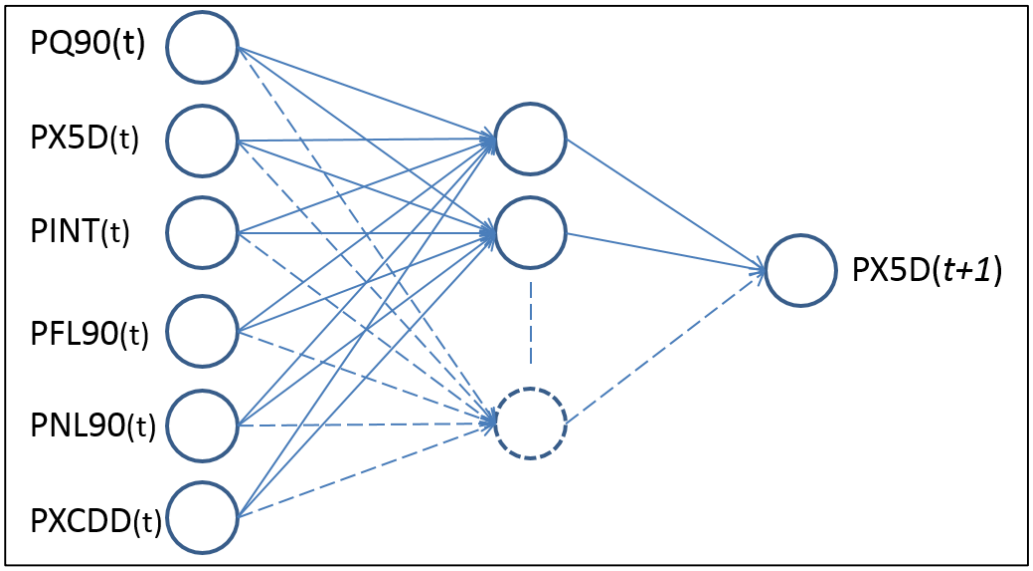

Figure 4.1: Basic BPNN architecture for Experiment-A

4.1.2 Experiment-B1

Experiment-B1 has been developed to forecast the maximum five consecutive days for a month ahead using the values of the PX5D of three months before. Dataset-B1 was used to train and test BPNN of this experiment. Dataet-B1 was created using a sliding window size of 3. The predictors of this model are the three-month delay of the PX5D value, as previous studies (Junaida et al., 2012; Sulaiman et al., 2013; Sulaiman et al., 2014) showed that the PX5D index can be forecasted based on the previous values of the same index; the previous three months' values were selected to examine the effect of these values on the forecasting accuracy. The output is the PX5D index value of the next month. Table 4.4 shows an example of the created Dataset-B1.

Table 4.4: Example of Dataset-B1

Predictors			**Target**
$PX5D_{(t-2)}$	$PX5D_{(t-1)}$	$PX5D_{(t)}$	$PX5D_{(t+1)}$
56.00	90.00	140.00	64.00
90.00	140.00	64.00	110.00
140.00	64.00	110.00	76.00
64.00	110.00	76.00	18.00
110.00	76.00	18.00	123.00

The BBNN input layer consists of three input units, one input unit for each month. The number of hidden units is determined by trail. Only one unit is found in the output layer. Table 4.5 shows Experiment-B1's BPNN architecture. Figure 4.2 shows the basic BPNN architecture of Experiment-B1.

Table 4.5: Experiment-B1's BPNN architecture

Number of Layers	3
Number of Input Units	3
Number of Hidden Units	To be determined
Number of Output Units	1
BPNN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear

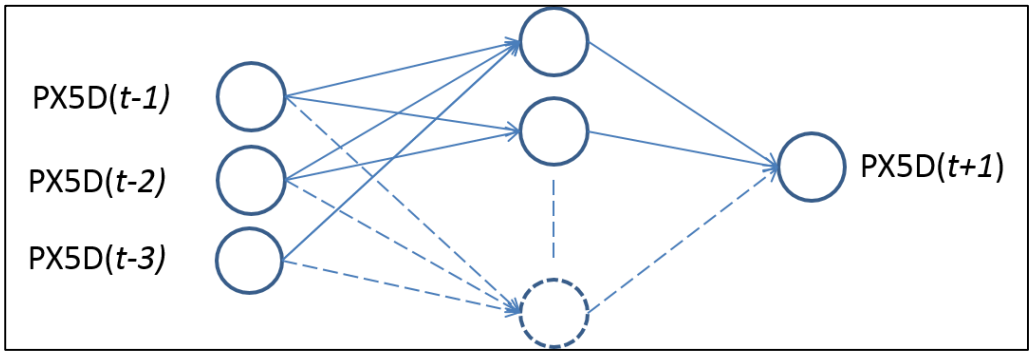

Figure 4.2: Basic BPNN architecture for Experiment-B1

4.1.3 Experiment-B2

Experiment-B2 has been developed to forecast the maximum five consecutive days of rainfall amount (PX5D index) of the next month by using the lagged values of the same index for a six-month delay.

Dataset-B2 is developed to fit the BPNN architecture of Experiment-B2, This dataset has been created to have the values of the previous six months in the predictor part, as previous studies (Junaida et al., 2012; Sulaiman et al., 2013; Sulaiman et al., 2014) indicated that the PX5D index for the next month can be forecasted based on the previous months' values of the same index; previous six months' values were selected to examine the effect of these values on the forecasting accuracy. The value of the same index for the next month is applied in the target part. Dataset-B2 was created using the sliding window technique with a window size of 6. The values of the PX5D index of a six-month delay is set to be introduced to the BPNN of this model as inputs, and the target is the PX5D value of the next month. Table 4.6 shows an example of the created Dataset-B2.

Table 4.6: Example of Dataset-B2

Predictors						**Target**
$PX5D_{(t-5)}$	$PX5D_{(t-4)}$	$PX5D_{(t-3)}$	$PX5D_{(t-2)}$	$PX5D_{(t-1)}$	$PX5D_{(t)}$	$PX5D_{(t+1)}$
76.00	39.50	87.00	56.00	90.00	140.00	64.00
39.50	87.00	56.00	90.00	140.00	64.00	110.00
87.00	56.00	90.00	140.00	64.00	110.00	76.00
56.00	90.00	140.00	64.00	110.00	76.00	18.00
90.00	140.00	64.00	110.00	76.00	18.00	123.00

Experiment-B2's BBNN's input layer consists of six units, one input neuron for each month. The number of hidden neurons has been determined by training several BPNNs with different numbers of hidden neurons. However, there is

only one neuron in the output layer. Table 4.7 shows Experiment-B2's BPNN architecture. Figure 4.3 shows the basic BPNN architecture of Experiment-B2.

Table 4.7: Experiment-B2's BPNN architecture

Number of Layers	3
Number of Input Units	6
Number of Hidden Units	To be determined
Number of Output Units	1
BPNN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear

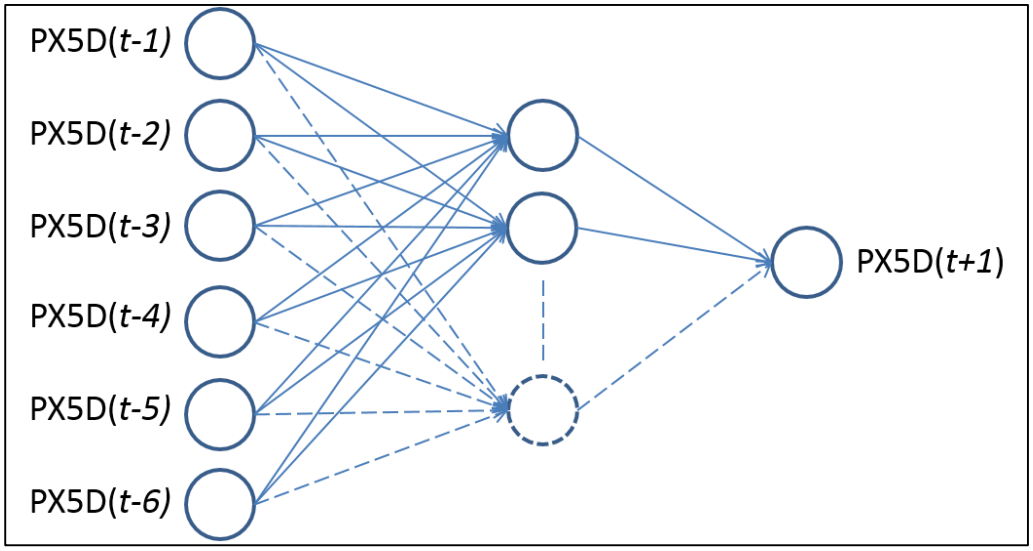

Figure 4.3: Basic BPNN architecture for Experiment-B2

4.1.4 Experiment-B3

Model-B3 has been developed using the values of the PX5D index of the past year (12 months) to forecast the PX5D value of the next month.

Dataset-B3 was created to train and test the BPNN of Experiment-B3. Dataset-B3 was created using the sliding window technique with a window size of 12. In order to obtain the values of the PX5D index for the last twelve months in the predictor part of the dataset, previous studies (Junaida et al., 2012; Sulaiman et al., 2013; Sulaiman, et al., 2014) applied the previous month values of the PX5D index to forecast the next month value of the PX5D, previous one year values were selected to examine the effect of these values on the forecasting accuracy. Whilst the target part of this dataset is the same like other datasets, which is the PX5D index value for the next month. An example of Dataset-B3 can be seen in Table 4.8.

Table 4.8: Example of Dataset-B3

Predictors												**Target**
$PX5D_{(t-11)}$	$PX5D_{(t-10)}$	$PX5D_{(t-9)}$	$PX5D_{(t-8)}$	$PX5D_{(t-7)}$	$PX5D_{(t-6)}$	$PX5D_{(t-5)}$	$PX5D_{(t-4)}$	$PX5D_{(t-3)}$	$PX5D_{(t-2)}$	$PX5D_{(t-1)}$	$PX5D_{(t)}$	$PX5D_{(t+1)}$
38.5	52	73	78	223	82.5	91	172	62	32	131	77	45.5
52	73	78	223	82.5	91	172	62	32	131	77	45	44
73	78	223	82.5	91	172	62	32	131	77	45.5	44	61
78	223	82.5	91	172	62	32	131	77	45.5	44	61	38.2
223	82.5	91	172	62	32	131	77	45.5	44	61	38.2	71

Experiment-A's BPNN has twelve units in the input layer, and one unit in the output layer. Concerning the number of units in the hidden layer, it is determined by developing several neural networks having different numbers of units in the hidden layer. Table 4.9 shows Experiment-B3's BPNN model architecture. Figure 4.4 shows the basic BPNN architecture of Experiment-B3.

Table 4.9: Experiment-B3's BPNN architecture

Number of Layers	3
Number of Input Units	12
Number of Hidden Units	To be determined
Number of Output Units	1
BPNN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear

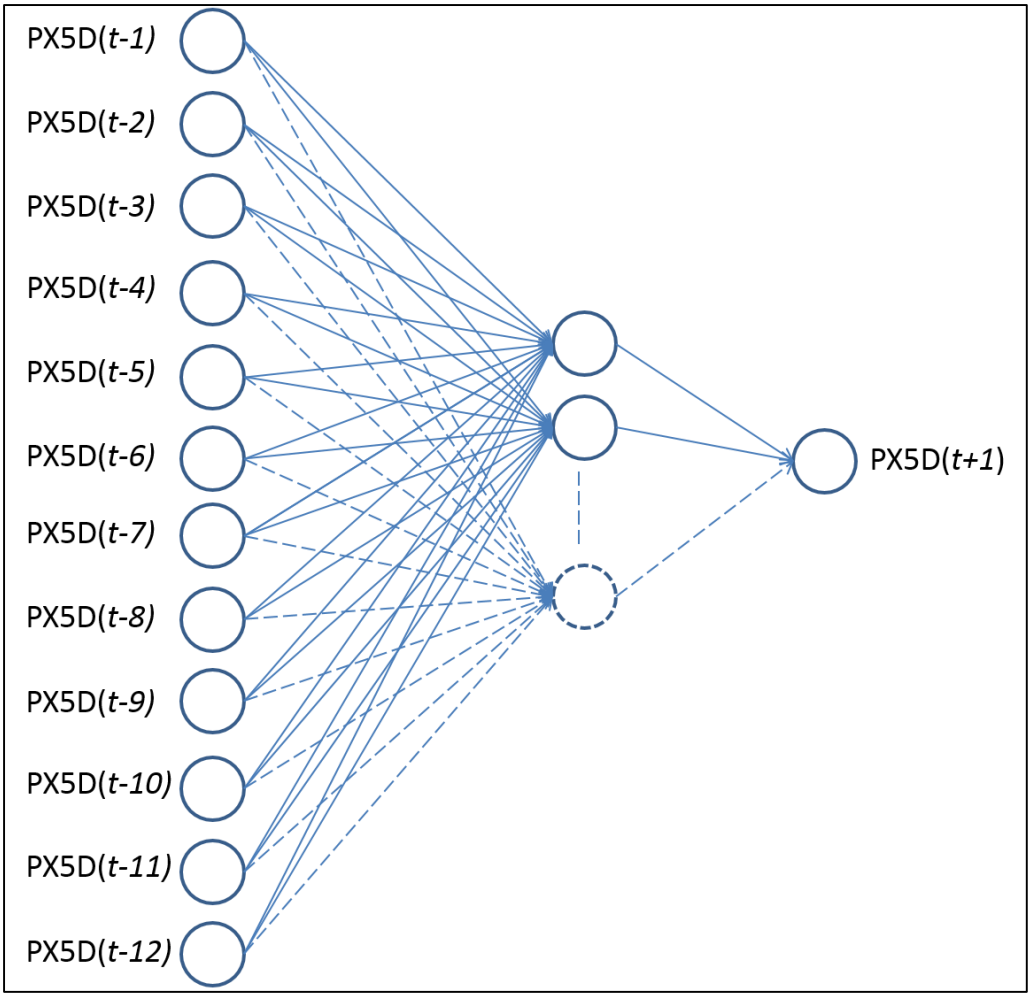

Figure 4.4: Basic BPNN architecture for Experiment-B3

4.2 Statistical Model

Multiple regression is used to find the relationship between the response variable and the observed variables, just like simple linear regression. From the name, the difference is clear that more than one observed variables are used in the multiple regression. The multiple regression takes the form as in Equation 4.1.

$$y = \varepsilon + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \beta_n x_n + e \quad \text{(Equation 4.1)}$$

Where

y is still the response variable.

ε is the constant.

β_1, β_2, β_3…βn are the coefficients.

X_1, X_2, X_3…X_n are the observed variables.

In this model, the coefficients indicate the information about how the response variable shifts when the observed variables change. The magnitudes of the coefficients represent the strength corresponding to the contribution of certain observed variables and the signs of the number point out the direction of the changes. In other words, the coefficients represent the amount of the response variable that changes, when the observed variables change one unit. Furthermore, the positive coefficients denote the response variable will increase when the observed variables increase, vice versa; the negative coefficients mean the response variable will decrease when the observed variables increase, and vice versa.

Multiple regression forecasting model is developed to forecast the maximum rainfall amount of five consecutive days. The six extreme rainfall indices were applied to the model as predictors to forecast the PX5D index value of the next month.

Since both Experiment-A in the BPNN Model and the multiple regression model use the six extreme rainfall indices, the same dataset (Dataset-A) was used to find the coefficients of the multiple regression model.

Regression analysis was performed to formulate the multiple regression model. The multiple regression model is shown in Equation 4.2.

$$PX5D_{(t+1)} = \varepsilon + \beta_1 * PQ90_{(t)} + \beta_2 * PX5D_{(t)} + \beta_3 * PINT_{(t)} + \beta_4 * PFL90_{(t)} + \beta_5 * PNL90_{(t)} + \beta_6 * PXCCD_{(t)} \quad \text{(Equation 4.2)}$$

In the multiple regression equation, $PX5D_{(t+1)}$ on the left side is the maximum accumulated five days of rainfall amount of the next month (t+1). On the right side, the indices $PQ90_{(t)}$, $PX5D_{(t)}$, $PINT_{(t)}$, $PFL90_{(t)}$, $PNL90_{(t)}$ and $PXCCD_{(t)}$ are the values of current time (t).

4.3 Summary

A total of four experiments utilizing BPNN model and a statistical model have been developed. Different types of predictors were used for each experiment.

Experiment-A and the statistical model used the values of the six extreme rainfall indices as predictors. Experiment-B1 used the lagged values of the PX5D index with a lag length of three. Experiment-B2 forecasted the PX5D index value depending on the values of the same index for the last six months. In Experiment-B3, the previous twelve-month values of PX5D index were used to forecast the value of the same index for the next month.

Chapter Five
Results

In this chapter, the results of the forecasting models are presented. Section 5.1 describes the BPNN model's results. In Section 5.2, the output of the multiple regression model is presented, while Section 5.3 discusses the evaluation of the forecasting models. The summary of the forecasting models are presented in Section 5.4.

5.1 BPNN Forecasting Model

Four experiments using the BPNN model, namely Experiment-A, Experiment-B1, Experiment-B2 and Experiment-B3, have been developed using different combinations of inputs. The experiments target to forecast the PX5D index of the next month.

5.1.1 BPNN Experiments Results

Experiment-A has been developed using Dataset-A. The number of input units is six. The number of hidden units has been determined using the heuristic approach by training BPNN with different numbers of hidden units. The selection criterion was the model output's MAE. It was found that when the number of hidden units is 3, and the error is the lowest. The MAE measurement value for this experiment is 31.62, the BPNN architecture for this experiment is 6-3-1. Table 5.1 shows Experiment-A's BPNN architecture. Figure 5.1 shows the developed BPNN architecture of Experiment-A.

Table 5.1: Experiment-A's final BPNN architecture

Number of Layers	3
Number of Input Units	6
Number of Hidden Units	3
Number of Output Units	1
ANN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear
Lowest MAE For Test Series	31.6250

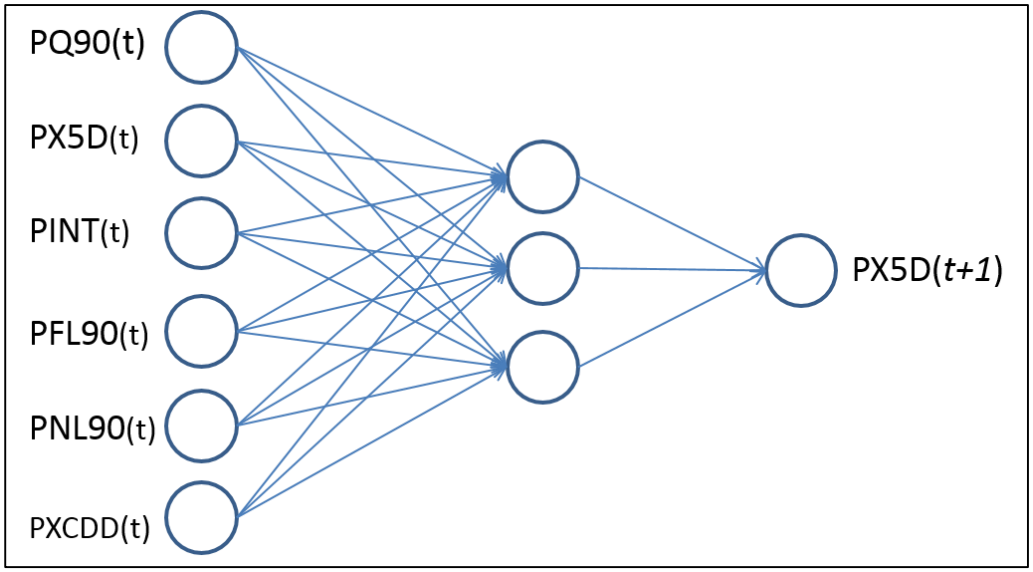

Figure 5.1: Final BPNN architecture for Experiment-A

Experiment-B1 was developed using Dataset-B1. This means Experiment-B1's BPNN has three units in the input layer. The number of hidden units has been determined by training BPNN with different numbers of hidden units and then selecting the network with the lowest output MAE. It was found that when the

network was trained with five units in the hidden layer, the network output error was the lowest, equal to 33.44. Table 5.2 shows Experiment-B1's BPNN architecture. Figure 5.2 shows the developed BPNN architecture of Experiment-B1.

Table 5.2: Experiment-B1's final BPNN architecture

Number of Layers	3
Number of Input Units	3
Number of Hidden Units	11
Number of Output Units	1
ANN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear
Lowest MAE For Test Series	33.4439

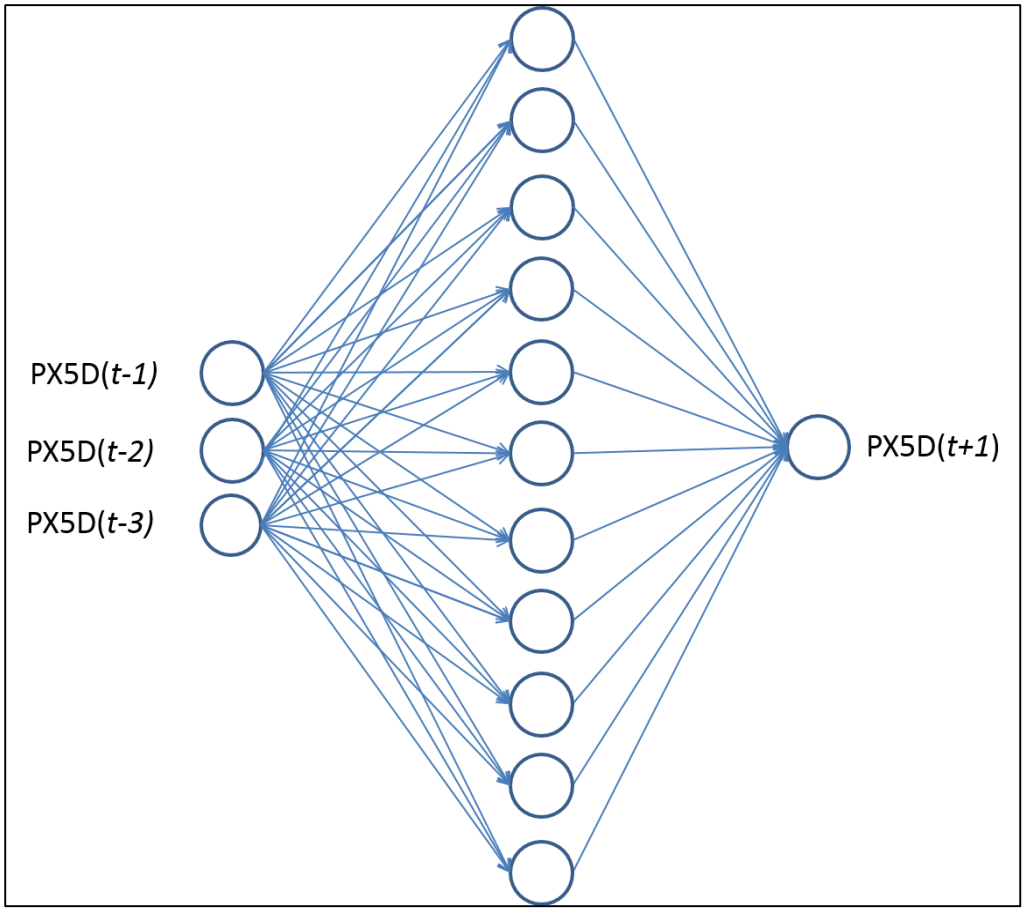

Figure 5.2: Final BPNN architecture for Experiment-B1

Experiment-B2's BPNN has been developed to forecast the maximum five consecutive days of rainfall amount of one month ahead by using Dataset-B2. The best MAE of this model output was equal to 32.95, when the network was trained with seven units in the hidden layer. Table 5.3 shows Experiment-B2's BPNN model architecture. Figure 5.3 shows the developed BPNN architecture of Experiment-B2.

Table 5.3: Experiment-B2's final BPNN architecture

Number of Layers	3
Number of Input Units	6
Number of Hidden Units	7
Number of Output Units	1
ANN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear
Lowest MAE For Test Series	32.9558

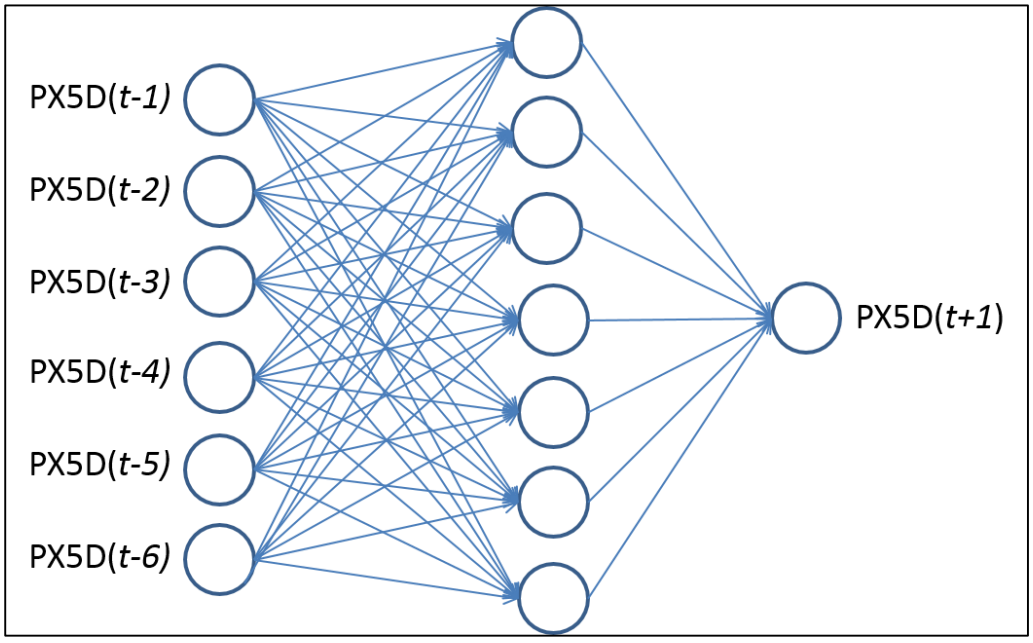

Figure 5.3: Final BPNN architecture for Experiment-B2

Experiment-B3 has been developed using the values of the PX5D index of the past year (Dataset-B3) to forecast the PX5D value of the next month. This model has twelve units in the input layer, seven units in the hidden layer and one unit in the output layer. The number of units in the hidden layer has been determined by training BPNN with different numbers of hidden units. The best MAE

obtained from the Model-B3 output was 33.82. Table 5.4 shows Model-B3's BPNN architecture. Figure 5.4 shows the developed BPNN architecture of Model-B3.

Table 5.4: Experiment-B3's final BPNN architecture

Number of Layers	3
Number of Input Units	12
Number of Hidden Units	9
Number of Output Units	1
ANN-Learning Algorithm	LM
Activation Function of Hidden Layer	Tan-sigmoid
Activation Function of Output Layer	linear
Lowest MAE For Test Series	33.8292

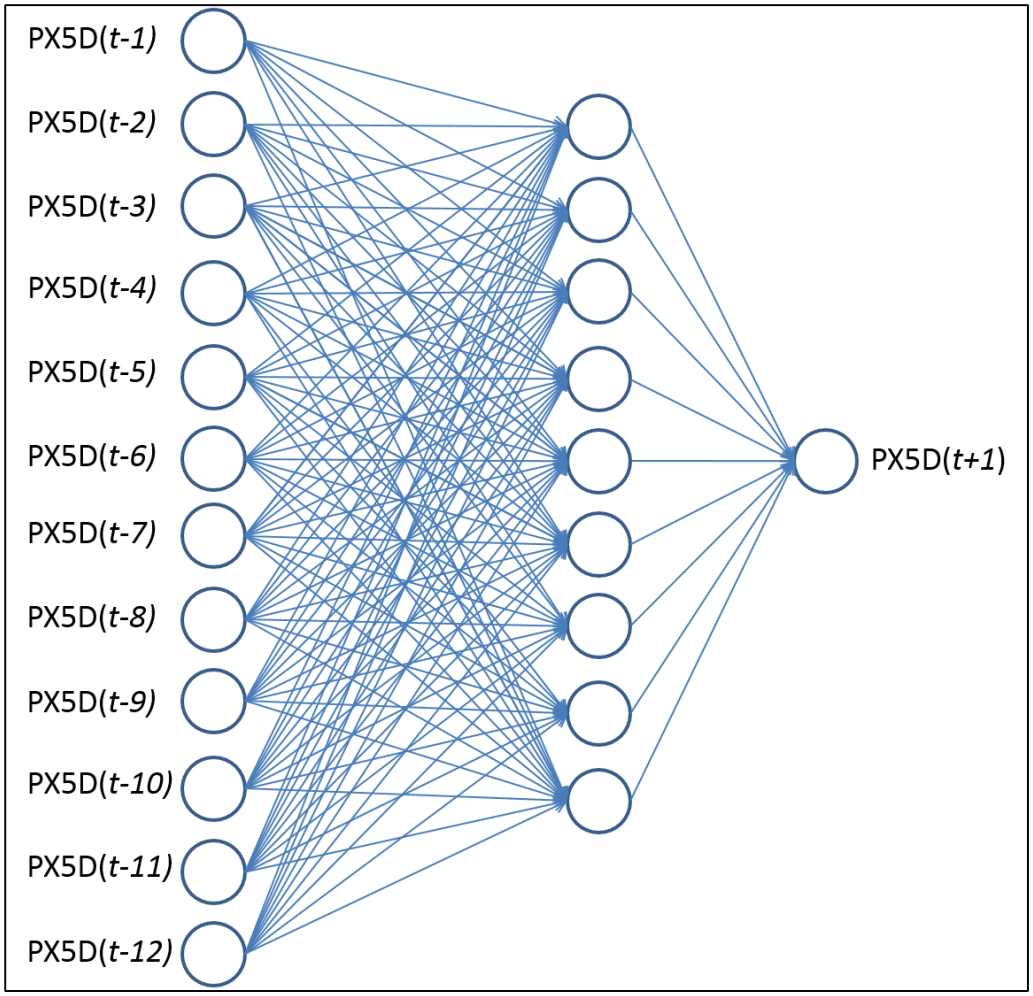

Figure 5.4: Final BPNN architecture for Experiment-B3

5.1.2 BPNN Experiments Comparison

The main aim of developing four experiments is to find which input combination would give a better forecasting accuracy. Mean Absolute Error (MAE) evaluation metric was calculated for the test period of each experiment separately. Figure 5.5 shows the obtained MAE value for each experiment.

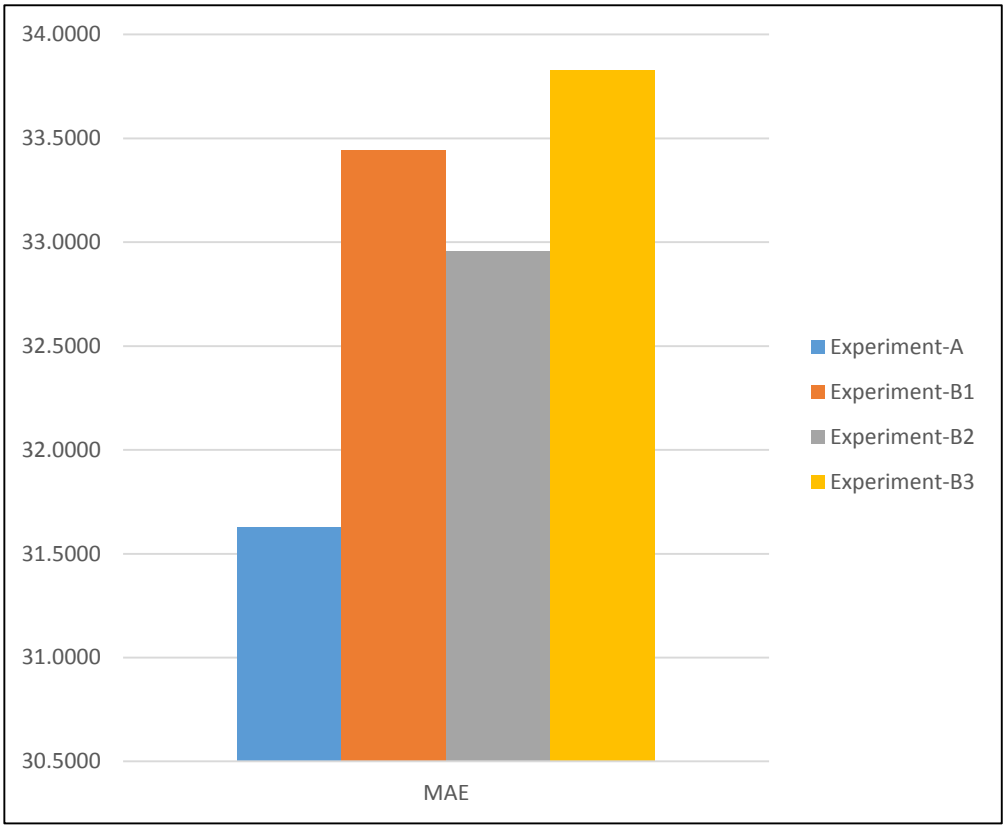

Figure 5.5: MAE value of the four BPNN model experiments

As can be seen in Figure 5.5, using the descriptive indices of extreme rainfall contribute in making the forecasting error lower than when using only the lagged value of the PX5D index. Experiment-A has outperformed Experiment-B1, Experiment-B2 and Experiment-B3 in extreme rainfall forecasting. Experiment-A, which was developed using BPNN and the six extreme rainfall indices as predictors. Using the six extreme rainfall indices as predictors led to obtaining a lower forecasting error rate than using the value of only one index. This shows that when more descriptive indices are added the lower error measurements will get lower.

5.2 Multiple Regression Forecasting Model

A multiple regression forecasting model has been developed using the six extreme rainfall indices as predictors to forecast the PX5D index value of the next month. The same dataset (Dataset-A) was used to develop Experiment-A, and the BPNN model has been used to find the multiple regression coefficients of Model-C.

5.2.1 Multiple Regression Model Results

Regression analysis was performed to formulate the multiple regression model. The least square method has been employed to obtain the best fitting line. The following multiple regression model had been generated (Equation 5.1).

$$PX5D_{(t+1)} = 79.342 + 0.236*PQ90_{(t)} - 0.061*PX5D_{(t)} + 0.005*PINT_{(t)} - 6.266*PFL90_{(t)} + 8.753*PNL90_{(t)} - 1.145*PXCCD_{(t)}$$ (Equation 5.1)

In this equation, PX5D(t+1) (target) on the left side is the maximum five consecutive days of rainfall amount of the next month (t+1). On the right side, (predictors) the indices PQ90(t), PX5D(t), PINT(t), PFL90(t), PNL90(t) and PXCCD(t) are the values of current time (t). The positive coefficients denote the response variable will increase when the index value increases, and vice versa; the negative coefficients means the response variable will decrease when the index value increases, vice versa.

In order to evaluate the model output, the same test data series used to test BPNN of Experiment-A had been applied to the generated multiple regression model. Then the MAE of the output has been calculated; the obtained MAE equals to 34.15 of the test series when using the generated multiple regression model.

5.3 Evaluation

This study presented two different models to forecast extreme rainfall. For the BPNN model, four experiments have been experimented, namely Experiment-A, Experiment-B1, Experiment-B2 and Experiment-B3.

As can be seen in Figure 5.1, Experiment-A which uses the six extreme rainfall indices in forecasting has outperformed the other three BPNN experiments that used the lagged values of the PX5D index in forecasting extreme rainfall.

As Experiment-A produced the lowest MAE among the four BPNN experiments, an equivalent statistical model to Experiment-A's BPNN model has been developed. The statistical model is a multiple regression model, which uses the six extreme rainfall indices as predictors and the target is the same like all the BPNN experiments. The same data series used to train and test Experiment-A's BPNN has been used to calculate the multiple regression model coefficients and to test the model output as well. Figure 5.6 shows the calculated MAE for Experiment-A's BPNN model and the multiple regression model.

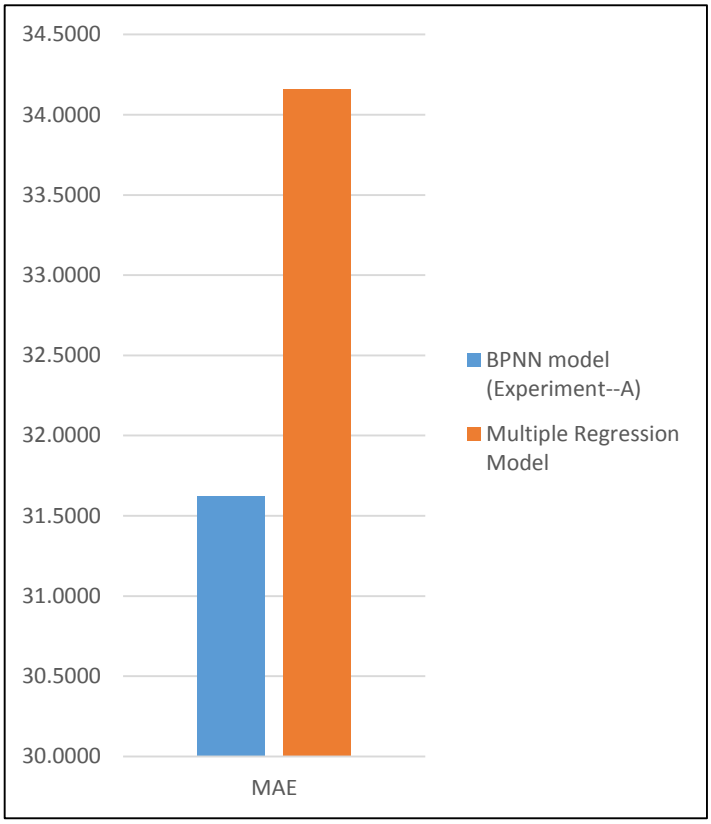

Figure 5.6: MAE of BPNN model (Experiment-A) and the multiple regression model

Besides calculating the MAE of the BPNN model (Experiment-A) and the multiple regression model output for the test data series, the Root Mean Squared Error (RMSE) has been calculated for each one of the five gauging stations separately using Equation 5.2 (Junaida et al., 2012).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - O_i)^2} \quad \text{(Equation 5.2)}$$

Where

T is the target

O is the Model output

Table 5.5 shows the calculated RMSE for both the BPNN model output of Experiment-A and the multiple regression model. From the table, it can be seen that the BPNN model has less RMSE than the multiple regression model for the five gauging stations.

Table 5.5: RMSE of BPNN and MR models

Gauging Station	BPNN Model	MR Model
Padang Besar	53.52	53.86
Tasoh	45.82	47.57
Lubuk Sireh	34.73	37.01
Kaki Bukit	36.31	39.62
Wang Kelian	36.67	36.68

Form the two error measurements used in this study to evaluate the models' output, the BPNN model has outperformed the multiple regression model by producing a lower MAE value and lower RMSE values for each one of the five gauging stations separately. Adding more descriptive indices as predictors and using the artificial neural network in forecasting the maximum five consecutive days of rainfall amount give a lower forecasting error than using the statistical models or using the previous values of one descriptive index as predictors.

5.4 Summary

A total of four experiments have been presented utilizing the developed BPNN model and one multiple regression model has been developed. Experiment-A's BPNN model and the multiple regression model used the six extreme rainfall indices as predictors, while Experiment-B1, Experiment-B2 and Experiment-B3's BPNN model used the lagged values of the PX5D index.

Experiment-A, which was developed using BPNN and the six extreme rainfall indices as predictors, has outperformed the other developed models. Using the six extreme rainfall indices as predictors contribute in obtaining a lower forecasting error rate than using the value of only one index. Utilizing BPNN in forecasting extreme rainfall based on the six extreme rainfall indices as predictors led to having a lower error rate. This proves that adding more descriptive indices helps in getting lower error measurements.

Chapter Six
Conclusion and Future Work

In this chapter, the conclusion and the future work are presented. Section 6.1 illustrates the conclusion of this study. Section 6.2 discusses the recommendations for future work.

6.1 Conclusion

In this study, the descriptive extreme rainfall indices have been calculated for the area of Perlis; the extreme rainfall indices have been calculated using the daily rainfall records of five gauging stations that represent the upstream of the Timah Tasoh reservoir. The PQ90 index has been calculated as one of the extreme rainfall indices. The PQ90 index represents the 90^{th} percentile value of the rainfall amount for a month. It has been calculated for each gauging station on a monthly basis and for each gauging station separately. According to STAREDEX and WMO, they define the extreme rainfall threshold based on the percentile value of rainfall records within a specified period. In the study, the average of the calculated 90^{th} Percentile value is as shown in Table 6.1.

Table 6.1: Calculated PQ90 index by gauging stations

Index		**Gauging Station**				
		Padang Besar	Tasoh	Lubuk Sireh	Kaki Bukit	Wang Kelian
PQ90	Average	31.05	34.04	30.45	25.46	31.66
	Maximum	97.80	115.00	101.40	128.00	116.80

This study presented five different methods for extreme rainfall forecasting. Different forecasting models have been employed using different types of variables. All the models share the same goal: forecasting the maximum five consecutive days of rainfall amount of a month ahead.

Using BPNN to forecast extreme rainfall event with six extreme rainfall descriptive indices as predictors produces a lower error measurement as compared to using the multiple regression model or applying one extreme rainfall index as a predictor.

When a comparison was conducted between the BPNN model (Experiment-A), which uses the six core extreme rainfall indices, and the other BPNN experiments that used lagged values of the maximum five consecutive days of rainfall amount, it was found that Experiment-A produced the lowest error measurement. Experiment-A also has the lowest error measurement when it was compared with the developed multiple regression model.

6.2 Future Work

The extreme rainfall forecasting error can be reduced in order to develop a forecasting model with higher accuracy. For future work, more variables can be combined with extreme rainfall indices to decrease the forecasting models' output error. Different lag lengths can be experimented to find the most significant period of previous months. Input variable section method can be applied to the six core indices or any added variables to specify the most significant forecasting variables.

References

Abdullah, J. (2013). *Distributed runoff simulation of extreme monsoon rainstorms in Malaysia using TREX*. (Thesis). Colorado State University.

Albeverio, S., Jentsch, V., & Kantz, H. (2006). *Extreme events in nature and society*. Springer Science & Business Media.

Ali, A. N. (2011). *Performance of artificial neural network and regression techniques for rainfall-runoff prediction*. International Journal of Physical Sciences, 6(8), 1997-2003.

AMS American Meteorological Society. (2014) *Rain*. Glossary of Meteorology. [Available online at http://glossary.ametsoc.org/wiki/"Rain"]

Balaguer, E., Palomares, A., Sorie, E., & Martin-Guerrero, J.D. (2008). *Predicting service request in support centers based on nonlinear dynamics, ARMA modeling and neural networks*. Expert Syst. Appl. 34, 665–672.

Banihabib, M.E., Mousavi, S.M., & Jamali, F.S. (2008). *Artificial neural network model to study the spatial and temporal correlation between stations in reservoir inflow forecasting*. In 3rd Iran Water Resources Management Conference, Tabriz, Iran.

Bascil, M. S., & Temurtas, F. (2011). *A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm*. Journal of Medical Systems, 35(3), 433-436.

Bigus, J.P. (1996). *Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*. McGraw-Hill: New York.

Box, G.E.P. & Jenkins, G.M. (1976). *Series Analysis Forecasting and Control*, (first ed.). Holden-Day: San Francisco (ISBN-10: 0816211043, p. 575).

Brachman, R. J. & Anand, T. (1996). The process of knowledge discovery in databases. In Fayyad, U. M. et al. (Eds.), *Advances in knowledge discovery and data mining*. AAAI Press / The MIT Press.

Bureau of Meteorology (2012) *Exceptionally heavy rainfall across southeast Australia*. Special Climate Statement.

Castillo, E., Guijarro-Berdiñas, B., Fontenla-Romero, O., & Alonso-Betanzos, A. (2006). *A very fast learning method for neural networks based on sensitivity analysis*. The Journal of Machine Learning Research, 7, 1159-1182.

Chantasut, N., Charoenjit, C., & Tanprasert, C. (2004, August). *Predictive mining of rainfall predictions using artificial neural networks for Chao Phraya River*. In 4th International Conference of the Asian Federation of Information Technology in Agriculture and the 2nd World Congress on Computers in Agriculture and Natural Resources, Bangkok, Thailand, August (pp. 9-12).

Chapman, P. et al. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*. Accessed from http://www.crisp-dm.org/CRISPWP-0800.pdf on May 2008.

Chegini, A.G. (2012). *MATLAB Tools*. Naghous Press. <http://www.naghoospress.ir/bookview.aspx?bookid=1485875>.

Cigizoglu, H. K. (2004). *Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons*. Advances in Water Resources, 27(2), 185-195.

Cigizoglu, H. K., & Kisi, Ö. (2006). *Methods to improve the neural network performance in suspended sediment estimation*. Journal of Hydrology, 317(3), 221-238.

Cilimkovic, M. (2011). *Neural Networks and Back Propagation Algorithm.* Institute of Technology Blanchardstown, Blanchardstown Road North Dublin, 15.

Climatic Research Unit, School of Environmental Sciences, University of East Anglia. (n.a) *Statistical and regional dynamical downscaling of extremes for European regions*. http://www.cru.uea.ac.uk/projects/stardex

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). *Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition*. Audio, Speech, and Language Processing, IEEE Transactions on, 20(1), 30-42.

Dreyfus, G. (2003). *Neural Networks: Methodology and Applications*. Springer.

Easterling, D. R., Evans, J. L., Groisman, P. Y., Karl, T. R., Kunkel, K. E., & Ambenje, P. (2000). *Observed Variability and Trends in Extreme Climate Events: A Brief Review*. Bulletin of the American Meteorological Society, 81(3), 417-425.

El Afandi, G., Morsy, M., & El Hussieny, F. (2013). *Heavy rainfall simulation over Sinai Peninsula using the weather research and forecasting model*. International Journal of Atmospheric Sciences, 2013.

El-Din, A. G., & Smith, D. W. (2002). *A neural network model to predict the wastewater inflow incorporating rainfall events*. Water research, 36(5), 1115-1126.

El-Shafie, A. H., El-Shafie, A., El Mazoghi, H. G., Shehata, A., & Taha, M. R. (2011a). *Artificial neural network technique for rainfall forecasting applied to Alexandria, Egypt*. International Journal of Physical Sciences, 6(6), 1306-1316.

El-Shafie, A., Jaafer, O., & Seyed, A. (2011b). *Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia.* Int J Phys Sci, 6(12), 2875-2888.

El-Shafie, A., Noureldin, A., Taha, M., Hussain, A., & Mukhlisin, M. (2012). *Dynamic versus static neural network model for rainfall forecasting at Klang River Basin, Malaysia*. Hydrology and Earth System Sciences, 16(4), 1151-1169.

Fausett, L. V. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*. (Vol. 40). Englewood Cliffs: Prentice-Hall.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*.

Field, C. B., and Coauthors. (Eds.) (2012) *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Summary for Policymakers*.

Foresti, L., Pozdnoukhov, A., Tuia, D., & Kanevski, M. (2010). *Extreme precipitation modelling using geostatistics and machine learning algorithms*. In geoENV VII–Geostatistics for Environmental Applications (pp. 41-52). Springer: Netherlands.

Gallant, S. I. (1993). *Neural network learning and expert systems*. MIT press.

Geetha, R., Sumathi, N., & Sathiabama, D. S. (2008). *A survey of spatial, temporal and spatio-temporal data mining*. Journal of Computer Applications, 1(4), 31-33.

Goodess, C. (2005). *STARDEX–Downscaling climate extremes*. UEA, Norwich.

Gu, N., & Wan, D. (2010, August). *Trend analysis of extreme rainfall based on BP neural network*. In Natural Computation (ICNC), 2010 Sixth International Conference on (Vol. 4, pp. 1925-1928). IEEE.

Haylock, M. (2005). *STARDEX Core Indices*. STARDEX project. http://www.cru.uea.ac.uk/projects/stardex/

Hazenberg, P., N. Yu, et al. (2011). *Scaling of raindrop size distributions and classification of radar reflectivity-rain rate relations in intense Mediterranean precipitation.* Journal of Hydrology 402(3-4): 179-192.

Hluchy, L., Seleng, M., Habala, O., & Krammer, P. (2010, August). *Mining environmental data in hydrological scenarios*. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on (Vol. 6, pp. 2988-2992). IEEE.

Hou, T., Kong, F., Chen, X., & Lei, H. (2013). *Impact of 3DVAR data assimilation on the prediction of heavy rainfall over southern China*. Advances in Meteorology, 2013.

Htike, K. K., & Khalifa, O. O. (2010, May). *Rainfall forecasting models using focused time-delay neural networks*. In Computer and Communication Engineering (ICCCE), 2010 International Conference on (pp. 1-6). IEEE.

Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009). *An artificial neural network model for rainfall forecasting in Bangkok, Thailand*. Hydrology and Earth System Sciences, 13(8), 1413-1425.

IMD Indian metrological Department (2014). *Terminologies and Glossary*. [Available online at http://www.imd.gov.in/section/nhac/dynamic/Met_Glossary.htm]

Jin, X. H., & Zhang, G. (2011). *Modelling optimal risk allocation in PPP projects using artificial neural networks*. International journal of project management, 29(5), 591-603.

Junaida, S., & Hirose, H. (2012, December). *A method to predict heavy precipitation using the artificial neural networks with an application*. In

Computing and Convergence Technology (ICCCT), 2012 7th International Conference on (pp. 663-667). IEEE.

Juneng, L., Tangang, F. T., Kang, H., Lee, W. J., & Seng, Y. K. (2010). *Statistical downscaling forecasts for winter monsoon precipitation in Malaysia using multimodel output variables.* Journal of climate, 23(1), 17-27.

Karahaliou, A., Skiadopoulos, S., Boniatis, I., Sakellaropoulos, P., Likaki, E., Panayiotakis, G., & Costaridou, L. (2014). *Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis*.

Karamouz, M., Araghinejad, Sh. (2012). *Advance Hydrology*. Amirkabir University of Technology Press.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). *An online algorithm for segmenting time series*. In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on (pp. 289-296). IEEE.

Khalili, N., Khodashenas, S. R., Davary, K., & Karimaldini, F. (2011). *Daily rainfall forecasting for Mashhad synoptic station using artificial neural networks*. In 2011 International Conference on Environmental and Computer Science IPCBEE (Vol. 19).

Kisi, I., Cigizoglu, K. (2005). *Reservoir management using artificial neural networks*. In 14th. Reg. Directorate of DSI (State Hydraulic Works), Istanbul, Turkey.

Kisi, O., & Shiri, J. (2012). *River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques*. Computers & Geosciences, 43, 73-82.

Kisi, O., Shiri, J., & Nikoofar, B. (2012). *Forecasting daily lake levels using artificial intelligence approaches*. Computers & Geosciences, 41, 169-180.

Kleinbaum, D., Kupper, L., Nizam, A., & Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Kusiak, A., Wei, X., Verma, A. P., & Roz, E. (2013). *Modeling and prediction of rainfall using radar reflectivity data: a data-mining approach*. Geoscience and Remote Sensing, IEEE Transactions on, 51(4), 2337-2342.

Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Laxman, S., & Sastry, P. S. (2006). *A survey of temporal data mining.* Sadhana, 31(2), 173-198.

Lee, J., Kim, J., Lee, J. H., Cho, I. H., Lee, J. W., Park, K. H., & Park, J. (2012, November). *Feature selection for heavy rain prediction using genetic algorithms*. In Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on (pp. 830-833). IEEE.

Liu, J. N., Li, B. N., & Dillon, T. S. (2001). *An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]*. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 31(2), 249-256.

Lourakis, M. (2005). *A brief description of the levenberg-marquardt algorithm implemented by levmar*. ICS Journal.

Lu, C. J., Lee, T. S., & Lian, C. M. (2012). *Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks*. Decision Support Systems, 54(1), 584-596.

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.

Mar, K. W., & Naing, T. T. (2008). *Optimum neural network architecture for precipitation prediction of Myanmar*. World Academy of Science, Engineering and Technology, 48, 130-134.

Margaret, H. D. (2003). *Data mining introductory and advanced topics*. Pearsons Education Inc.

Marques, A., Lacerda, D. P., Camargo, L. F. R., & Teixeira, R. (2014). *Exploring the relationship between marketing and operations: Neural network analysis of marketing decision impacts on delivery performance*. International Journal of Production Economics, 153, 178-190.

Marzano, F. S., Marchiotto, S., Textor, C., & Schneider, D. J. (2010). *Model-based weather radar remote sensing of explosive volcanic ash eruption*.

Geoscience and Remote Sensing, IEEE Transactions on, 48(10), 3591-3607.

Mebrhatu, M. T., Tsubo, M., & Walker, S. (2007). *A statistical model for seasonal rainfall forecasting over the highlands of Eritrea*. Discovery and Innovation, 19(1), 37.

Menhaj, M.B. (2012). *Artificial Neural Networks*. Amirkabir University of Technology Press.

MET Malaysia. (2013). Ministry of science, technology and innovation/Malaysian metrological department.

Mohammadi, K., Eslami, H.R., Dayyani Dardashti, Sh. (2005). *Comparison of regression ARIMA and ANN models for reservoir inflow forecasting using snowmelt equivalent (a case study of Karaj)*. J. Agric. Sci. Technol. 7, 17–30.

Moisão, R. L. M., Pires, F. M., & Ramalho, R. R. (2001). *Prediction model, based on neural networks, for time series with origin in chaotic systems.* In: Proc. of Worshop Artificial Intelligence Techniques for Financial Time Series Analysis, EPIA.

Moustris, K. P., Larissi, I. K., Nastos, P. T., & Paliatsos, A. G. (2011). *Precipitation forecast using artificial neural networks in specific regions of Greece.* Water resources management, 25(8), 1979-1993.

Nguyen, D., & Widrow, B. (1990). *Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights*. In International Joint Conference on Neural Networks, (pp. 21-26).

Nigatu, M. K. (2011). *Rainfall Intensity Duration Frequency (RIDF) Relationships under the Changing climate (Case study on Upper Blue Nile River Basin, Ethiopia).* (Thesis). Addis Ababa University.

NOAA, National Climate Data Center state of the climate. (2012). *National overview – Annual 2012*. [Available online at http://www.ncdc.noaa.gov/sotc/national/2012/13.]

OFDA/CRED (2013). *OFDA/CRED International Disaster Database*.

Ondimu, S., & Murase, H. (2007). *Reservoir level forecasting using neural networks: Lake Naivasha*. Biosystems engineering, 96(1), 135-138.

Panigrahi, S., Verma, K., Tripathi, P., & Sharma, R. (2014, April). *Knowledge Discovery from Earth Science Data.* In Communication Systems and

Network Technologies (CSNT), 2014 Fourth International Conference on (pp. 398 - 403). IEEE.

Papalexiou, S. M., & Koutsoyiannis, D. (2013). *Battle of extreme value distributions: A global survey on extreme daily rainfall*. Water Resources Research, 49(1), 187-201.

Peterson, T. C., Alexander, L. V., Allen, M. R., Anel, J. A., Barriopedro, D., Black, M. T., ... & Otto, F. E. L. (2013). *Explaining extreme events of 2012 from a climate perspective.* Bulletin of the American Meteorological Society, 94(9), S1-S74.

Peterson, T. C., Stott, P. A., &Herring, S. (2012) *Explaining extreme events of 2011 from a climate perspective.* Bulletin of the American Meteorological Society 93.7 (2012): 1041-1067.

Piekniewski, F., Izhikevich, E., Szatmary, B., & Petre, C. (2012). *Spiking neural network object recognition apparatus and methods*. U.S. Patent Application 13/465,918.

Povinelli, R. J. (2001). *Identifying temporal patterns for characterization and prediction of financial time series events.* In Temporal, Spatial, and Spatio-Temporal Data Mining (pp. 46-61). Springer: Berlin Heidelberg.

Priddy, K. L., & Keller, P. E. (2005). *Artificial neural networks: an introduction* (Vol. 68). SPIE Press.

Rana, A. (2013). *Climate Change Effects on Rainfall and Management of Urban Flooding*. (Thesis). University of Lund.

Rehman, A., & Saba, T. (2014). *Neural networks for document image preprocessing: state of the art*. Artificial Intelligence Review, 42(2), 253-273.

Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Springer.

Routray, A., Osuri, K. K., & Kulkarni, M. A. (2012). *A comparative study on performance of analysis nudging and 3DVAR in simulation of a heavy rainfall event using WRF modeling system*. International Scholarly Research Notices, 2012.

Samia, M. (2004, June). *A Representation of Time Series for Temporal Web Mining*. In Grundlagen von Datenbanken (pp. 103-107).

Sapna, S., Tamilarasi, A., & Kumar, M. P. (2012). *Backpropagation learning algorithm based on Levenberg Marquardt Algorithm*. Computer Science & Information Technology (CS & IT), 2, 393-398.

SAS Enterprise Miner – SEMMA. SAS Institute.

Shahnawaz, M., A. Ranjan, et al. (2011). *Temporal Data Mining: An Overview.* International Journal of Engineering and Advanced Technology 1(1): 20-24.

Shrivastava, G., Karmakar, S., Kowar, M. K., & Guhathakurta, P. (2012). *Application of Artificial Neural Networks in Weather Forecasting: A Comprehensive Literature Review*. International Journal of Computer Applications, 51(18), 0975-8887.

Singh, S., Vashishtha, V., & Singla, T. (2014). *Artificial Neural Network*. International Journal of Research, 1(9), 934-942.

Srikalra, N., & Tanprasert, C. (2006). *Rainfall prediction for Chao Phraya River using neural networks with online data collection.* In Proc. of the 2nd IMTGT Regional Conference on Mathematics.

Sulaiman, J. B., Darwis, H., & Hirose, H. (2014). *Monthly Maximum Accumulated Precipitation Forecasting Using Local Precipitation Data and Global Climate Modes*. Journal ref: Journal of Advanced Computational Intelligence and Intelligent Informatics, 18(6), 999-1006.

Sulaiman, J., Darwis, H., & Hirose, H. (2013, October). *Forecasting Monthly Maximum 5-Day Precipitation Using Artificial Neural Networks with Initial Lags.* In Computational Intelligence and Design (ISCID), 2013 Sixth International Symposium on (Vol. 2, pp. 3-7). IEEE

Syafrina, A. H., Zalina, M. D., & Juneng, L. (2014). *Historical trend of hourly extreme rainfall in Peninsular Malaysia.* Theoretical and Applied Climatology, 1-27.

Toth, E., Brath, A., & Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. Journal of Hydrology, 239(1), 132-147.

Unes, F. (2010). *Dam reservoir level modeling by neural network approach: A case study*. Neural Network World, 20(4), 461.

Unes, F. (2010a). *Prediction of density flow plunging depth in dam reservoirs: an artificial neural network approach*. Clean–Soil, Air, Water, 38(3), 296-308.

Unes, F., Yildirim, S., Cigizoglu, H. K., & Coskun, H. (2013). *Estimation of dam reservoir volume fluctuations using artificial neural network and support vector regression*. Journal of Engineering Research, 1(3), 53-74.

Valipour, M. (2012a). *Critical areas of Iran for agriculture water management according to the annual rainfall*. Eur. J. Sci. Res. 84 (4), 600–08.

Valipour, M. (2012b). *Number of required observation data for rainfall forecasting according to the climate conditions*. Am. J. Sci. Res. 74, 79–86.

Wan Ishak, W. H., Ku Mahamud, K. R., & Md Norwawi, N. (2011). *Mining Temporal Reservoir Data Using Sliding Window Techniques*. CiiT International Journal of Data Mining Knowledge Engineering, 3(8), pp. 473-478, 2011

Wan, D., Wang, Y., Gu, N., & Yu, Y. (2012, December). *A novel approach to extreme rainfall prediction based on data mining*. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on (pp. 1873-1877). IEEE.

Wang, L., Simões, N., Ochoa, S., Leitão, J. P., Pina, R., Onof, C., ... & David, L. (2011). *An enhanced blend of SVM and Cascade methods for short-term rainfall forecasting*. In Proceedings from the12th International Conference on Urban Drainage.

Yesu, K., Chakravorty, H. J., Bhuyan, P., Hussain, R., & Bhattacharyya, K. (2012). *Hybrid features based face recognition method using Artificial Neural Network*. In Computational Intelligence and Signal Processing (CISP), 2012 2nd National Conference on (pp. 40-46). IEEE.

Zehraoui, F., & Bennani, Y. (2005). *New self-organizing maps for multivariate sequences processing*. International Journal of Computational Intelligence and Applications, 5(04), 439-456.

Zeng, Z., Hsieh, W. W., Shabbar, A., & Burrows, W. R. (2011). *Seasonal prediction of winter extreme precipitation over Canada by support* vector regression. Hydrology and Earth System Sciences, 15(1), 65-74.