Integration of Principal Component Analysis, Fuzzy C-Means and Artificial Neural Networks for Localised Environmental modelling of Tropical Climate

by

Noor Zuraidin Bin Mohd Safar

This thesis is submitted in partial fulfilment of its requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth

January 2017
“Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.”
Abstract

Meteorological processes are highly non-linear and complicated to predict at high spatial resolutions. Weather forecasting provides critical information about future weather that is important for flooding disaster prediction system and disaster management. This information is also important to businesses, industry, agricultural sector, government and local authorities for a wide range of reasons. Processes leading to rainfall are non-linear with the relationships between meteorological parameters are dynamic and disproportionate. The uncertainty of future occurrence and rain intensity can have a negative impact on many sectors which depend on the weather condition. Therefore, having an accurate rainfall prediction is important in human decisions. Innovative computer technologies such as soft computing can be used to improve the accuracy of rainfall prediction. Soft computing approaches, such as neural network and fuzzy soft clustering are computational intelligent systems that are capable of integrating humanlike knowledge within a specific domain, adapt themselves and learn in changing environments. This study evaluates the performance of a rainfall forecasting model. The data pre-processing method of Principal Component Analysis (PCA) is combined with an Artificial Neural Network (ANN) and Fuzzy C-Means (FCM) clustering algorithm and used to forecast short-term localized rainfall in tropical climate. State forecast (raining or not raining) and value forecast (rain intensity) are tested using a number of trained networks. Different types of ANN structures were trained with a combination of multilayer perceptron with a back propagation network. Levenberg-Marquardt, Bayesian Regularization and a Scaled Conjugate Gradient training algorithm are used in the network training. Each neuron uses linear, logistic sigmoid and hyperbolic tangent sigmoid as a transfer function. Preliminary analysis of input parameter data pre-processing and FCM clustering were used to prepare input data for the ANN forecast model. Meteorological data such as atmospheric pressure, temperature, dew point, humidity and wind speed
have been used as input parameters. The magnitude of errors and correlation coefficient were used to evaluate the performance of trained neural networks. The predicted rainfall forecast for one to six hour ahead are compared and analysed. One hour ahead for state and value forecast yield more than 80% accuracy. The increasing hours of rain prediction will reduce the forecast accuracy because input-output mapping of the forecast model reached termination criterion early during validation test and no improvement of convergence in the consecutive number of epochs. Result shows that, the combination of PCA-FCM-ANN forecast model produces better accuracy compared to a basic ANN forecast model.
Pursuing a postgraduate study is both a painful and an enjoyable experience. It is like climbing a high peak, step by step, accompanied with bitterness, hardships, frustration, encouragement, trust and with so many kinds of people help. When I found myself at the endpoint of writing, I realized that it was, in fact, teamwork that got me there. Though, it will not be enough to express my gratitude in words to all those people who helped me.

First of all, I would like to express my gratitude to God who has given me the strength and patience to complete this study. I would like to give my sincere thanks to my honorific supervisor, Dr. David L. Ndzi, who offered me valuable advice, patiently supervise me, and always guiding me to the right path. I have learned a lot from him, without his help I could not have finished my work successfully. I would also like to thank the following individuals for their support during the completion of this thesis, Dr. Shikun Zhou, Dr. Linda Young and Dr. Ioannis Kagalis for the help and insight they offered into my study.

My gratitude also goes to Majlis Amanah Rakyat (MARA) and Ministry of Rural and Regional Development of Malaysia for the doctoral sponsorship and other financial support throughout the course of my PhD study.

Finally yet importantly, my deepest gratitude goes to my wife Fadzlina Hani Samani, my beloved daughter and son, Nur Qistina Faqihah and Muhammad Syahmi Wafiq for their support and blessing. Deepest appreciation also goes to my parents Mohd Safar Mullorejo, Sinah Sidal, Samani Posan and Rohani Ahmad for their support and encouragement. Lastly, to those who are involved direct or indirectly towards my study. Your help is highly appreciated. Thank you.
# Table of Contents

Abstract ................................................................................................................................. iii

Acknowledgement .............................................................................................................. v

Table of Contents ................................................................................................................. vi

Glossary of Terms and Abbreviations .................................................................................. xi

List of Figures ...................................................................................................................... xiii

List of Tables ......................................................................................................................... xvii

Chapter 1  Introduction ........................................................................................................... 1
    1.1 Introduction ............................................................................................................... 1
    1.2 Background and Motivation ..................................................................................... 2
    1.3 Research Aims and Objectives ................................................................................. 3
    1.4 Thesis Overview ....................................................................................................... 4

Chapter 2  Study Area and Meteorological Data ................................................................. 6
    2.1 Introduction ............................................................................................................. 6
    2.2 Study Area ............................................................................................................. 6
    2.3 Rainfall Analysis ................................................................................................... 8
    2.4 Data Pre-processing ............................................................................................... 10
2.4.1 Missing data imputation using Principal Component Analysis ................................................................. 12
2.4.2 Results and Discussions .................................................................................................................. 15
2.5 Meteorological Parameter Analysis ................................................................................................. 20
2.5.1 Descriptive statistical of meteorological parameter ........................................ 21
2.5.2 Meteorological patterns and parameters distribution .................. 22
2.6 Summary ........................................................................................................................................ 28

Chapter 3

3.1 Introduction .................................................................................................................................. 29
3.2 Neural Network .......................................................................................................................... 29
3.3 Artificial Neural Network in Weather Forecasting ...................................................... 30
3.4 Artificial Neural Network in Rainfall Forecasting ....................................................... 33
3.5 Artificial Neural Network Architectures ............................................................................. 33
  3.5.1 Training Algorithm ............................................................................................................. 37
    3.5.1.1 Levenberg-Marquardt (LM) ......................................................................................... 37
    3.5.1.2 Bayesian Regularization (BR) ................................................................................... 37
    3.5.1.3 Scaled Conjugate Gradient (SCG) .......................................................................... 38
  3.5.2 Transfer Function ................................................................................................................ 38
  3.5.3 Input Model .......................................................................................................................... 39
3.6 Data Pre-processing for Meteorological Parameters ......................................................... 40
3.7 Fuzzy C-Means Clustering ........................................................................................................ 41
3.8 Performance Indices .................................................................................................................. 42
3.9 Summary ........................................................................................................................................ 43
Chapter 4  Artificial Neural Network in Weather Forecasting .......................... 44
  4.1  Introduction .......................................................................................... 44
  4.2  Meteorological Dataset ........................................................................ 45
  4.3  Weather Forecast Implementation Flowchart ....................................... 45
  4.4  ANN Design ......................................................................................... 46
      4.4.1  ANN Preliminary implementation .................................................. 47
      4.4.2  Data normalisation ........................................................................ 47
      4.4.3  Input layer, hidden layer and output layer ..................................... 48
      4.4.4  Training algorithm ........................................................................ 49
      4.4.5  Transfer function ........................................................................... 49
      4.4.6  Prediction setup ............................................................................. 50
      4.4.7  Input parameters and selection ...................................................... 51
      4.4.8  ANN Configuration ........................................................................ 52
  4.5  Results and Discussions ....................................................................... 52
  4.6  Summary .............................................................................................. 59

Chapter 5  ANN Rainfall Forecasting Model.................................................. 61
  5.1  Introduction .......................................................................................... 61
  5.2  Study Area and Meteorological Data ..................................................... 62
  5.3  ANN Design ......................................................................................... 64
  5.4  Rainfall Forecast ANN Architecture ....................................................... 66
  5.5  Forecasting System Setup ..................................................................... 67
  5.6  Results and Discussions ....................................................................... 68
<table>
<thead>
<tr>
<th>Chapter 6</th>
<th>PCM-FCM-ANN Rainfall Forecasting Model</th>
<th>71</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>71</td>
</tr>
<tr>
<td>6.2</td>
<td>Meteorological Data</td>
<td>72</td>
</tr>
<tr>
<td>6.3</td>
<td>Implementation Framework of PCA-FCM-ANN Forecasting Model</td>
<td>72</td>
</tr>
<tr>
<td>6.4</td>
<td>PCA-ALS Imputation Method</td>
<td>74</td>
</tr>
<tr>
<td>6.5</td>
<td>FCM Clustering</td>
<td>74</td>
</tr>
<tr>
<td>6.6</td>
<td>ANN Design</td>
<td>77</td>
</tr>
<tr>
<td>6.7</td>
<td>ANN Architecture</td>
<td>78</td>
</tr>
<tr>
<td>6.8</td>
<td>Results and Discussions</td>
<td>80</td>
</tr>
<tr>
<td>6.8.1</td>
<td>PCA-ALS Imputation</td>
<td>80</td>
</tr>
<tr>
<td>6.8.2</td>
<td>FCM clustering</td>
<td>82</td>
</tr>
<tr>
<td>6.8.2.1</td>
<td>FCM Clustering for wet condition</td>
<td>83</td>
</tr>
<tr>
<td>6.8.2.2</td>
<td>FCM Clustering for dry condition</td>
<td>85</td>
</tr>
<tr>
<td>6.8.3</td>
<td>Forecast result</td>
<td>88</td>
</tr>
<tr>
<td>6.8.3.1</td>
<td>Results for Chuping weather station</td>
<td>88</td>
</tr>
<tr>
<td>6.8.3.2</td>
<td>Results for Alor Setar weather station</td>
<td>94</td>
</tr>
<tr>
<td>6.8.3.3</td>
<td>Results for MARDI Bukit Tangga weather station</td>
<td>98</td>
</tr>
<tr>
<td>6.8.4</td>
<td>Summary of forecasting result</td>
<td>99</td>
</tr>
<tr>
<td>6.9</td>
<td>Summary</td>
<td>101</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Conclusions</td>
<td>103</td>
</tr>
<tr>
<td>7.1</td>
<td>Research Contribution</td>
<td>105</td>
</tr>
</tbody>
</table>
7.2 Research Limitation........................................................................................................ 106

7.3 Future Work..................................................................................................................... 107

References................................................................................................................................ 109

Appendix A. Publications ........................................................................................................ 124
Appendix B. Monthly Rainfall Rate Average ........................................................................ 125
Appendix C. Missing Data ...................................................................................................... 126
Appendix D. Weather Station Instruments ............................................................................ 128
Appendix E. Certificate of Ethics Review .............................................................................. 130
Appendix F. Research Certificate Review Checklist ............................................................... 131
Glossary of Terms and Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS</td>
<td>alternating least square</td>
</tr>
<tr>
<td>ANN</td>
<td>artificial neural network</td>
</tr>
<tr>
<td>AWS</td>
<td>automatic weather stations</td>
</tr>
<tr>
<td>BPL</td>
<td>back propagation layer</td>
</tr>
<tr>
<td>BR</td>
<td>Bayesian regularization</td>
</tr>
<tr>
<td>CDF</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>DID</td>
<td>Department of Irrigation and Drainage</td>
</tr>
<tr>
<td>DP</td>
<td>dew point</td>
</tr>
<tr>
<td>FCM</td>
<td>fuzzy c-means clustering</td>
</tr>
<tr>
<td>HT</td>
<td>Hyperbolic Tangent Sigmoid</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>KDE</td>
<td>kernel density estimation</td>
</tr>
<tr>
<td>LM</td>
<td>Levenberg Marquardt</td>
</tr>
<tr>
<td>LT</td>
<td>linear</td>
</tr>
<tr>
<td>MAE</td>
<td>mean absolute error</td>
</tr>
<tr>
<td>MIN</td>
<td>minimum</td>
</tr>
<tr>
<td>MISO</td>
<td>multiple input single output</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MLP</td>
<td>multilayer perceptron</td>
</tr>
<tr>
<td>MSE</td>
<td>mean squared error</td>
</tr>
<tr>
<td>MSL</td>
<td>mean sea level</td>
</tr>
<tr>
<td>ND</td>
<td>new dataset matrix</td>
</tr>
<tr>
<td>NWP</td>
<td>numerical weather prediction</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>probability density function</td>
</tr>
<tr>
<td>RF</td>
<td>radio frequency</td>
</tr>
<tr>
<td>RMSE</td>
<td>root mean squared error</td>
</tr>
<tr>
<td>SC</td>
<td>soft computing</td>
</tr>
<tr>
<td>SCG</td>
<td>Scaled Conjugate Gradient</td>
</tr>
<tr>
<td>ST</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>WS</td>
<td>wind speed</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1: Study area and weather stations (A=26km, B=35km, C=32km) .................. 8
Figure 2.2: Average monthly rainfall rate for all weather stations ............................... 9
Figure 2.3: Cumulative distribution of rainfall rate for three years ........................... 10
Figure 2.4: Missing data handling flowchart ............................................................. 13
Figure 2.5: PCA Analysis comparison between Raw Data (a) and Imputed Data (b). .......................................................... 18
Figure 2.6: Kernel Density Estimation for Wind Speed dataset with raw data and imputed data .......................................................... 19
Figure 2.7: KDE for: (a) Pressure, (b) Temperature, (c) Dew point and (d) Humidity .......................................................... 20
Figure 2.8: Rainfall rate, temperature, dew point and humidity in (a)1000 and (b)120 step .......................................................... 23
Figure 2.9: Pressure and rainfall rate observation in 1000 and 120 hours .................. 25
Figure 2.10: Wind speed and rainfall rate observation in 1000 and 120 hours ......... 26
Figure 2.11: Parameters cumulative probability distribution during rain and not
raining condition, (a) pressure, (b) temperature, (c) dew point, (d) humidity and (e) wind speed..........................27

Figure 3.1: Human brain biological neuron [32][33] .......................................................30

Figure 3.2: Single neuron.........................................................................................34

Figure 3.3: ANN layer representation ........................................................................35

Figure 4.1: Weather forecast implementation flowchart ...........................................46

Figure 4.2: A four layer ANN model with three layer feed forwards ANN. ..........51

Figure 4.3: Regression plot for temperature forecast using ANN with LM algorithm.
.................................................................................................................................56

Figure 4.4: One hour forecast results for pressure using LM training algorithm .....56

Figure 4.5: One hour forecast results for temperature using LM training algorithm 57

Figure 4.6: One hour forecast results for humidity using LM training algorithm .....57

Figure 4.7: One hour forecast results for dew point using LM training algorithm ...58

Figure 4.8: One hour forecast results for wind speed using LM training algorithm.58

Figure 4.9: Three hours forecast results for temperature using LM training algorithm
.................................................................................................................................59

Figure 4.10: Six hours forecast results for temperature using LM training algorithm
.................................................................................................................................59

Figure 5.1: Map of Malaysia and Chuping .................................................................63

Figure 5.2: Monthly rainfall rate average for Chuping..............................................63

Figure 5.3: ANN step by step design .......................................................................64

Figure 5.4: ANN architecture for rainfall forecasting model ...............................66

Figure 6.1: Flowchart of forecasting for modular PCA-FCM-ANN implementation.
Figure 6.2: Schematic illustration of FCM clustering integration into ANN rainfall forecasting model .......................................................... 73

Figure 6.3: ANN flowchart implementation ........................................................................... 78

Figure 6.4: PCA-FCM-ANN architecture for rainfall forecasting model ............... 79

Figure 6.5: KDE for: (a) pressure, (b) temperature, (c) dew point and (d) humidity (e) wind speed ................................................................. 82

Figure 6.6: FCM Clustering for Pressure and Temperature during raining condition ............................................................................. 83

Figure 6.7: FCM Clustering for Dew Point and Humidity during raining condition 84

Figure 6.8: FCM Clustering for Humidity during raining condition ...................... 84

Figure 6.9: FCM Clustering for Wind Speed during raining condition .............. 85

Figure 6.10: Wet and dry comparison for (a) pressure, (b) temperature, (c) dew point, (d) humidity and (e) wind speed ........................................................................... 87

Figure 6.11: Dew Point and Temperature clustering for dry condition .............. 87

Figure 6.12: Humidity and Pressure clustering for dry condition ...................... 88

Figure 6.13: Comparison between observed and forecast for (a) one hour, (b) three hours and (c) six hours state forecast using PCA-FCM-ANN for Chuping weather station ............................................................................. 92

Figure 6.14: Comparison between observed and forecast rain intensity for Chuping, (a) one hour, (b) three hours and (c) six hours ........................................... 94

Figure 6.15: Regression plot for (a) one, (b) three and (c) six hour rain intensity forecast for PCA-FCM-ANN model for Chuping ........................................... 94

Figure 6.16: Comparison between observed and forecast for (a) one hour, (b) three
hours and (c) six hours state forecast using PCA-FCM-ANN for Alor Setar. ................................................................. 97

Figure 6.17: Scatter plot of observed and forecast for (a) one hour, (b) three hours and (c) six hours state forecast using PCA-FCM-ANN for Alor Setar. .... 97

Figure 6.18: Comparison between observed and forecast rainfall rate for one hour rain intensity forecast using PCA-FCM-ANN implementation for Alor Setar. ........................................................................................................ 98

Figure 6.19: Regression plot for (a) one, (b) three and (c) six hour rain intensity forecast using PCA-FCM-ANN model for Alor Setar................................. 98

Figure 6.20: Performance comparison of forecasting model for state prediction .... 100

Figure 6.21: Performance comparison of forecasting model for rainfall rate .......... 100

Figure 6.22: MAE performance comparison of forecasting model for state prediction ........................................................................................................... 101

Figure 6.23: RMSE performance comparison of forecasting model for state prediction........................................................................................................ 101
List of Tables

Table 2.1: General information for all weather stations................................. 7
Table 2.2: Categorisation of rainfall ........................................................... 9
Table 2.3: Rainfall rate preliminary analysis for all stations .......................... 10
Table 2.4: Parameters available and its unit measurement from weather station .... 11
Table 2.5: Missing data for all weather stations ............................................. 12
Table 2.6: Covariance matrices for missing data and non-missing data .......... 16
Table 2.7: Mean weighted parameters for NM and M matrices ...................... 16
Table 2.8: Parameter analysis for all stations in any rain condition ............... 21
Table 2.9: Parameter analysis for all stations in rain condition ...................... 22
Table 2.10: Parameter analysis for all stations in no rain condition ............... 22
Table 3.1: Transfer function ........................................................................ 39
Table 4.1: List of transfer function used in the experiment ......................... 50
Table 4.2: One hour parameter forecast ....................................................... 53
Table 4.3: Three hour parameter forecast .................................................... 55
Table 4.4: Six hour parameter forecast ......................................................... 55
Table 5.1: Hourly and monthly average rainfall rate statistic for Chuping ............... 63
Table 5.2: Transfer function................................................................................. 65
Table 5.3: Rain prediction for ANN rainfall forecasting model ......................... 68
Table 6.1: Rain prediction for basic ANN forecasting model for Chuping............... 90
Table 6.2: PCA-FCM-ANN rainfall forecasting model for Chuping....................... 90
Table 6.3: PCA-FCM-ANN rainfall forecasting model for Alor Setar ...................... 95
Table 6.4: PCA-FCM-ANN rainfall forecasting model for MARDI Bukit Tangga .... 99
Chapter 1

Introduction

1.1 Introduction

Weather forecasting requires a combination of science and technology application in order to predict the condition of the atmosphere in advance at a given location at a given time. Environmental parameters such as temperature, humidity, atmospheric pressure, dew point, solar radiation, cloud, wind direction and wind speed could potentially be used to provide weather forecasts that are more accurate. Weather forecasting is important for planning human daily activities. The effect of weather humans is very significant, affecting human comfort and planning of their activities. The structural design of houses, location of industries, airport and harbour activities are predominantly influenced by weather and climatic conditions. Farmers need weather information to help them plan crop planting, harvesting and irrigation. In the aviation industry, weather information is not only important for flight scheduling but is critical for safety. Weather forecasting is key to the prediction and prevention of catastrophic floods, droughts and hurricanes. One of the crucial weather parameter to forecast is rain prediction. Rainfall is important for
agricultural production, especially for vegetables and grains. Rain is a major component of the water cycle and source of fresh water. It is well known that atmospheric parameters such as humidity, temperature, pressure, wind speed and dew point influence precipitation factors [1]. In general, weather forecasting can be divided into two methods; empirical and dynamical approach [2]. The empirical approach is based on climatological data and is known as analogue forecasting [4]. Dynamical approach is based on forward simulations of the atmosphere and is often referred to as computer modelling [3]. This thesis investigates the use of soft computing approach based on artificial neural network in short term localised weather forecasting in tropical climates. This research is multidisciplinary and involves both computer science and meteorological study.

This chapter provides the background and motivation behind this research, introduces its objectives and expected outcomes. Later, an outline of the thesis structure is presented.

1.2 Background and Motivation

Meteorological processes are highly non-linear and complicated to predict at high spatial resolutions. Weather forecasting provides critical information about future weather that is important for disaster prediction system and disaster management. This information is also important to businesses, industry, agricultural sector [5], government and local authorities for a wide range of reasons. Soft Computing (SC) techniques such as Artificial Neural Network (ANN) can be used to predict the behaviour of such non-linear conditions [6]. Since weather processes are non-linear and follow an irregular trend, an ANN is envisaged to be a better technique for analysing and identifying the structural relationship between the meteorological parameters
Large data from satellite networks, radars and weather stations are processed continuously on a daily basis. This data is transformed into useful information that is used to forecast the weather in the next hours or days. Weather forecasting systems use complex computer algorithms that demand high performance computers and require high resolution spatial data [8]. In tropical climates, rain is localised and more difficult to predict compared to temperate climates. Weather forecasting techniques and measurement systems over wide area using satellites are therefore ineffective in tropical regions. The average rain footprint in Malaysia is about 1.5 km radius [9], radar system coverage is about 20 to 40 km radius. Instead of heavy rain in smaller area, radar weather systems predict drizzle or light rain. The actual rain condition is not forecast correctly due to spatial resolution. Therefore, a localised rain prediction system is needed.

1.3 Research Aims and Objectives

The goal of the research is the development of ANN models based on advanced computing technique for the purpose of forecasting rain events (known as state prediction in this study) and quantitative rain prediction (rainfall rate or rain intensity) for localised area in the tropics. This study also evaluates the use of ANN in environmental parameter forecasting such as temperature, atmospheric pressure, dew point, humidity and wind speed. In order to achieve the research aims the research objectives were defined as follows:

- Obtain and study weather data such as temperature, atmospheric pressure, dew point, humidity, wind speed, rain event occurrences, rainfall amount and rainfall rate from North-West Malaysia.
- Investigate ANN and Fuzzy Logic techniques for possible application.
in a forecast model.

- Develop an environmental parameter prediction model (temperature, atmospheric pressure, dew point, humidity and wind speed) using ANN and Fuzzy Logic at hourly interval.
- Develop a rainfall forecast model for state prediction and rainfall rate prediction.
- Evaluate and validate the developed model(s).

1.4 Thesis Overview

This section presents a brief outline of the structure of this thesis. There are altogether seven chapters, each of which is summarized as follows:

Chapter 1

This chapter presents a brief introduction, background and motivation of the research. It also presents research aim and objectives and thesis overview.

Chapter 2

This chapter primarily introduces data availability and discusses the study geography. It also describes the application of the Principal Component Analysis (PCA) with Alternating Least Squares (ALS) in data pre-processing technique. Rainfall and meteorological parameter analysis are presented.

Chapter 3

This chapter introduces the literature and related works in the relevant topic of the research. A description of neural networks in weather and rainfall forecasting is reviewed. Modelling method, present studies in computational forecasting and ANN architecture. PCA-ALS pre-processing technique is discussed and Soft Clustering method is also explored. These methods of pre-
processing and clustering techniques have potential capabilities that can improve the mapping process between model inputs and expected output (forecast/prediction result). Finally, this chapter reviews the performance indices used to evaluate the performance of the forecasting model.

Chapter 4

This chapter describes the process of ANN forecasting to predict environmental parameter conditions for one, three and six hours ahead for localised areas. It describes the use of atmospheric pressure, temperature, dew point, humidity and wind speed data from a single weather station. The evaluation of the performance is determined by evaluating the magnitude of the error and the correlation coefficient value between observed and forecasted values.

Chapter 5

This chapter describes the application and evaluation of basic the ANN model in rainfall event and rainfall rate prediction without any pre-processing data or clustering. A straight forward ANN implementation is used.

Chapter 6

This chapter describes the application and evaluation of the ANN model in rainfall event and rainfall rate prediction by using PCA-ALS pre-processing and Fuzzy C-Means (FCM) soft clustering methods.

Chapter 7

The final chapter is the summary of the thesis. The novelty of the research, research contributions and the future works is presented.
Chapter 2

Study Area and Meteorological Data

2.1 Introduction

This chapter describes the study area, meteorological data used, imputation technique of missing data and finally, an analysis of the meteorological parameters. The main focus in this study is to determine the best computer based rainfall forecasting technique. Before the model can be developed, a detail initial study is essential in order to gain a better understanding of the meteorological parameters and their relationships. Meteorological data from three weather stations in North-West Malaysia have been used in this study. This chapter is organised as follows; the study area is described, followed by rainfall rate analysis, data management and imputation and, finally, the meteorological parameters are analysed.

2.2 Study Area

The study focuses on meteorological data from Chuping, Alor Setar and MARDI Bukit Tangga weather stations. These stations are located in Perlis and
Kedah which have similar climatic characteristics and topology. The area has a tropical climate with long hours of sunny days, uniform temperature and high humidity. The average annual relative humidity is between 70 % to 90 % [10] and often exceed 80 % [11]. Malaysia has an annual average surface temperature of 26.7 °C [12] and for the study area, the annual average temperature is 27.5 °C. For the whole year, the average day period is from 7 a.m. to 7 p.m. Hourly meteorological data for pressure, dry bulb temperature, dew point, humidity, wind speed, rainfall amount and rainfall rate have been used in this study. Malaysia is situated near the equator and has a typically tropical climate, with abundant rainfall, high temperatures and high humidity all year round. Many of its climate variable, such as temperature and humidity do not show large monthly variations. However, many variables exhibit prominent daily variations from hour to hour, indicating highly dynamic conditions of the local climate. Table 2.1 is a summary of the weather stations information, latitude and longitude, elevation and weather station identification number. Figure 2.1 shows the study area and the distances between all of the stations.

Table 2.1: General information for all weather stations.

<table>
<thead>
<tr>
<th>Station</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation</th>
<th>Station No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alor Setar</td>
<td>6° 12' N</td>
<td>100° 24' E</td>
<td>39 m</td>
<td>48603</td>
</tr>
<tr>
<td>Chuping</td>
<td>6° 29' N</td>
<td>100° 16' E</td>
<td>21.7 m</td>
<td>48604</td>
</tr>
<tr>
<td>MARDI Bukit Tangga</td>
<td>06° 28' N</td>
<td>100° 30’ E</td>
<td>50 m</td>
<td>41625</td>
</tr>
</tbody>
</table>
2.3 Rainfall Analysis

The distributions of the average rainfall rate for all weather stations are depicted in Figure 2.2 and the distribution for specific weather stations is available in Appendix B-1, Appendix B-2 and Appendix B-3. The resolution of data is one hour and rainfall events recorded for all weather stations is 5121 hours. Most of the rainfall events are primarily as a result of localised events. Categorisation of the rainfall in Malaysia [13] is shown in Table 2.2. The Malaysian Department of Irrigation and Drainage’s flood management information shows that, whenever convective rainfall rate exceeds 60 mm within 2 to 4 hours duration it may cause flash floods [13]. The flash flood can be defined as sudden and unexpected flooding caused by local heavy rainfall or rainfall in another area [13]. From this categorisation only 27 incidences fall
under the Very Heavy rainfall category that may cause flash flood. Descriptive statistics such as minimum, maximum, mean, median, variance and standard deviation were used to determine the initial characteristics of the data before further statistical analyse were carried out. Table 2.3 tabulates the descriptive statistic preliminary analysis of rainfall rate for each weather station.

![Monthly Average Rainfall Rate: ALL Stations](image)

Figure 2.2: Average monthly rainfall rate for all weather stations

<table>
<thead>
<tr>
<th>Rainfall rate at any given time</th>
<th>Number of rain incidences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>1 ( \text{mmh}^{-1} ) to 10 ( \text{mmh}^{-1} )</td>
</tr>
<tr>
<td>Moderate</td>
<td>11 ( \text{mmh}^{-1} ) to 30 ( \text{mmh}^{-1} )</td>
</tr>
<tr>
<td>Heavy</td>
<td>31 ( \text{mmh}^{-1} ) to 60 ( \text{mmh}^{-1} )</td>
</tr>
<tr>
<td>Very Heavy</td>
<td>Greater than 60 ( \text{mmh}^{-1} )</td>
</tr>
</tbody>
</table>

According to the Malaysia Meteorological Department, three types of monsoon seasons occur in West Malaysia; Northeast Monsoon typically from November to March, Southeast Monsoon between May and September and Inter-Monsoon from April to October [11]. Coastal areas are highly affected by monsoon rains compared to central locations.

Figure 2.3 shows the cumulative probability distribution for rain events over three year. Most of the rainfall intensity recorded is less than 30 mmh\(^{-1}\),
which is 95% of all rain events. The variability of the rain condition can be used to study the relationship between the meteorological parameters. Before further analysis is performed, estimating the reliability of the data is necessary. In the following section, data pre-processing methodology is described.

Table 2.3: Rainfall rate preliminary analysis for all stations

<table>
<thead>
<tr>
<th>Weather Stations</th>
<th>MIN (mmh^{-1})</th>
<th>MAX (mmh^{-1})</th>
<th>MEAN (mmh^{-1})</th>
<th>(\sigma)</th>
<th>MED (mmh^{-1})</th>
<th>(\sigma^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL Stations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>1.520</td>
<td>8.471</td>
<td>5.087</td>
<td>1.570</td>
<td>5.050</td>
<td>2.464</td>
</tr>
<tr>
<td>Hour</td>
<td>0.044</td>
<td>95.280</td>
<td>5.021</td>
<td>8.985</td>
<td>1.500</td>
<td>80.736</td>
</tr>
<tr>
<td>Chuping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>0.105</td>
<td>4.856</td>
<td>2.844</td>
<td>1.147</td>
<td>2.710</td>
<td>1.317</td>
</tr>
<tr>
<td>Hour</td>
<td>0.090</td>
<td>95.280</td>
<td>2.937</td>
<td>7.200</td>
<td>0.540</td>
<td>51.836</td>
</tr>
<tr>
<td>Alor Setar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>0.122</td>
<td>8.289</td>
<td>4.381</td>
<td>1.852</td>
<td>4.486</td>
<td>3.431</td>
</tr>
<tr>
<td>Hour</td>
<td>0.200</td>
<td>70.320</td>
<td>4.484</td>
<td>8.249</td>
<td>1.400</td>
<td>68.040</td>
</tr>
<tr>
<td>MARDI Bukit Tangga</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td>0.122</td>
<td>18.423</td>
<td>7.817</td>
<td>3.650</td>
<td>7.782</td>
<td>13.324</td>
</tr>
<tr>
<td>Hour</td>
<td>0.044</td>
<td>72.300</td>
<td>7.458</td>
<td>10.916</td>
<td>2.586</td>
<td>119.161</td>
</tr>
</tbody>
</table>

MIN=minimum, MAX=maximum, MEAN=average, \(\sigma\)=Standard deviation, MED=median, \(\sigma^2\)=variance

Figure 2.3: Cumulative distribution of rainfall rate for three years

2.4 Data Pre-processing

A common problem in meteorological study is missing data due to insufficient sampling, faulty data acquisition or instrument measurement error. Data pre-processing methods have been used in this study to eliminate faulty
data, reduce noise and data discrepancies. Due to high dimensionality of the parameters, Principal Component Analysis (PCA) algorithm was used to interpolate some missing data [14]. Hourly meteorological data used in this study was provided by the Malaysian Meteorological Department. One of the aims of this study was to investigate the weather forecast for each weather station independently and localised. Table 2.4 summarises the parameters that are available:

Table 2.4: Parameters available and its unit measurement from weather station.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmospheric Pressure</td>
<td>hPa</td>
</tr>
<tr>
<td>Dry bulb temperature</td>
<td>°C</td>
</tr>
<tr>
<td>Dew point</td>
<td>°C</td>
</tr>
<tr>
<td>Humidity</td>
<td>%</td>
</tr>
<tr>
<td>Wind speed</td>
<td>ms⁻¹</td>
</tr>
<tr>
<td>Rainfall amount</td>
<td>mm</td>
</tr>
<tr>
<td>Rainfall duration</td>
<td>minute</td>
</tr>
<tr>
<td>Rainfall rate</td>
<td>mmh⁻¹</td>
</tr>
</tbody>
</table>

The rain-gauges used were self-emptying gauges or tipping buckets. The gauge rotates and empties when 0.1 mm of rain has fallen. The system will count the number of times the meter has emptied every minute and sends the appropriate rainfall value to the base station which is accumulated into hourly rainfall data. When the rainfall is less than 0.1 mm the system will record no rain event. All of the weather stations used anemometers to measure wind speed and wind direction [15]. Instruments used to measure and send the data to the storage system is available at Appendix D. The percentage of the missing value for each weather station is presented in Table 2.5. Details parameters availability is tabulated in Appendix C.
### Table 2.5: Missing data for all weather stations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>% Missing data</th>
<th>No. of missing data</th>
<th>Total number of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmospheric Pressure</td>
<td>2.6</td>
<td>2069</td>
<td>78912</td>
</tr>
<tr>
<td>Dry bulb temperature</td>
<td>1.3</td>
<td>993</td>
<td>78912</td>
</tr>
<tr>
<td>Dew point</td>
<td>1.2</td>
<td>976</td>
<td>78912</td>
</tr>
<tr>
<td>Humidity</td>
<td>2.9</td>
<td>2295</td>
<td>78912</td>
</tr>
<tr>
<td>Wind speed</td>
<td>9.1</td>
<td>7187</td>
<td>78912</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17.1</strong></td>
<td><strong>13520</strong></td>
<td><strong>394560</strong></td>
</tr>
</tbody>
</table>

#### 2.4.1 Missing data imputation using Principal Component Analysis

Interpolating to handle missing data is referred to as imputation [16]. In one dimensional data, basic statistical analysis such as minimum, maximum, median, standard deviation and variance are important statistical parameters to interpret data trend and changes [17]. This study makes use of multi-dimensional data to build the forecasting model. However, before it can be implemented, denoising data and missing data management is essential. In this study, a non-linear PCA-ALS algorithm has been used to handle missing data for all weather stations. PCA is a well-established technique in reducing data dimensionality [18]. Data compression, image processing, visualisation, exploratory data analysis, pattern recognition and time series prediction are among the applications that engaging PCA technique. The entire process flow diagram in missing data handling is presented in Figure 2.4.
The relationship between weather parameters can be generalized using covariance. However, the covariance analysis typically works for two dimensional data. Since the available data consists of six parameters, the covariance matrix is a good alternative to determine the relationship between all parameters [19]. The covariance matrix of five weather parameters (excluding rain data) was calculated in determination of missing data. The covariance can be simplified in the following form:

\[ C_{n\times n} = (C_{ij}, C_{ji} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \]  

(2.1) where \( C_{n\times n} \) is matrix with \( n \) rows and \( n \) column, \( \text{Dim}_x \) is the \( x^{th} \) matrix
The following matrix is the covariance matrix data representation in this study:

$$C_{n,n} = \begin{pmatrix}
\text{cov}(P, P) & \text{cov}(P, T) & \text{cov}(P, DP) & \text{cov}(P, H) & \text{cov}(P, WS) \\
\text{cov}(T, P) & \text{cov}(T, T) & \text{cov}(T, DP) & \text{cov}(T, H) & \text{cov}(T, WS) \\
\text{cov}(DP, P) & \text{cov}(DP, T) & \text{cov}(DP, DP) & \text{cov}(DP, H) & \text{cov}(DP, WS) \\
\text{cov}(H, P) & \text{cov}(H, T) & \text{cov}(H, DP) & \text{cov}(H, H) & \text{cov}(H, WS) \\
\text{cov}(WS, P) & \text{cov}(WS, T) & \text{cov}(WS, DP) & \text{cov}(WS, H) & \text{cov}(WS, WS)
\end{pmatrix}$$ (2.2)

where P is pressure, T is temperature, DP is dew point, H is humidity, and WS is wind speed. Equation (2.2) shows that the matrix is symmetrical at the main diagonal and indicates that the covariance is one of the dimension of itself [20].

The covariance matrix together with its eigenvalues and eigenvectors will be used in implementing the non-linear PCA-ALS technique. In statistics, PCA is utilized to extract core relationships in high dimensionality data. Basically, PCA is a technique used to identify patterns in data by expressing the similarity and differences of the data [20]. Principal components can be defined by calculating the eigenvectors of the correlated data matrix. These vectors provide the directions in which the group of data is stretched most. The projections of the data on the eigenvectors are the principal components. The corresponding eigenvalues from eigenvectors provide an indication of the amount of information to the respective principal components. Principal components corresponding to large eigenvalues represent considerable information in the dataset and consequently tell the relations between the data points. PCA assumes that all data were measured in ratio and interval scale. In implementing PCA-ALS, the raw data must be quantified using optimal scaling. The ALS algorithm estimates the least squares values by updating alternately between a set of data in matrix. The regularization stops when it converge [21]. Step by step implementation of the missing data handling using non-linear PCA-ALS can be summarized in the following steps:
Step 1. Load data: five parameters from observed meteorological data into matrix NM.

Step 2. Find missing value and replace -1, -1.1, -1.0 or N/A to NaN and store data into M matrix.

Step 3. Determine principal component for each parameter.

Step 4. Calculate the covariance matrix of the whole dataset.

Step 5. Calculate eigenvectors and corresponding eigenvalues.

Step 6. Calculate matrix of vectors to determine principal component.

Step 7. Perform principal component coefficient for matrices NM and M using PCA.

Step 8. Calculate mean for each weighted parameters in M matrix using PCA-ALS.

Step 9. Construct new dataset with estimates value.

Step 10. Evaluate estimated data:

   Step 10.1. Comparing coefficient of M and NM matrices angle subspace; if the angle is small, the two spaces are nearly linear dependent.

   Step 10.2. Comparing probability distribution between matrices M and NM.

where M is matrix with missing data and NM is matrix for non-missing data.

2.4.2 Results and Discussions

PCA-ALS has been implemented on the data from all weather stations. This is very important as preparation for providing the input data to a model that will be developed later [22]. The amount of missing data in the MARDI Bukit Tangga station was considerably high compared to other stations. Table 2.5 shows that the largest number of missing data is wind speed, the PCA-ALS imputation technique started with that data. After the observed data is store into NM matrix, the next step is to replace default and missing data to NaN. NaN is one of the data types available in Matlab, NaN returns the IEEE® arithmetic representation for Not-a-Number (NaN) [23]. The method of imputation in this study will make use of existing PCA pre-defined subroutine available in Matlab [23]. Table 2.6 shows the covariance of NM (cov_{NM}) and M (cov_{M}) matrices derived from the covariance matrix in Equation (2.2). There are positive and negative covariance coefficients in both matrices; these imply positive and negative relationship between parameters.
Once the covariance coefficients were determined, the next step was to calculate the estimated mean of weighted parameters. The average of weighted parameters will be utilised in constructing a new dataset with estimated filling value for missing data. Table 2.7 represents the mean weighted value for NM and M matrices. Mean weighted parameter values for both matrices are equal except for wind speed.

\[
\text{Table 2.6: Covariance matrices for missing data and non-missing data}
\]

\[
\text{Covariance matrix for non-missing data}
\]

\[
\text{cov}_{\text{NM}} = 
\begin{pmatrix}
0.9328 & -0.3574 & -0.0415 & -0.0136 & -0.0145 \\
-0.0368 & -0.2016 & 0.6661 & -0.0252 & 0.7167 \\
0.0296 & 0.0169 & 0.7348 & 0.0632 & -0.6744 \\
0.3572 & 0.9113 & 0.1164 & 0.0187 & 0.1671 \\
0.0032 & -0.0281 & -0.0325 & 0.9974 & 0.0576 \\
\end{pmatrix}
\]

\[
\text{Covariance matrix for missing data}
\]

\[
\text{cov}_{\text{M}} = 
\begin{pmatrix}
0.9314 & -0.3612 & -0.0415 & 0.0114 & 0.0128 \\
-0.0374 & -0.2013 & 0.6649 & -0.1046 & 0.7107 \\
0.0296 & 0.0167 & 0.7349 & 0.0321 & -0.6765 \\
0.3599 & 0.9092 & 0.1194 & 0.0309 & 0.1693 \\
0.0268 & -0.0459 & 0.0430 & 0.9935 & 0.0916 \\
\end{pmatrix}
\]

\[
\text{Table 2.7: Mean weighted parameters for NM and M matrices}
\]

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Meteorological Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pressure</td>
</tr>
<tr>
<td>NM</td>
<td>1003.63</td>
</tr>
<tr>
<td>M</td>
<td>1003.63</td>
</tr>
</tbody>
</table>

The next step is to construct a dataset with new filling values using the following equation:

\[
\text{ND} = (\text{score}_{\text{M}} \cdot \text{cov}_{\text{M}}) + \mu_{\text{M}}
\]

(2.3)

where \text{ND} is a new dataset matrix, \text{score}_{\text{M}} is the representation of matrix M in
the principal component space, \( \text{cov}_M \) is the covariance coefficient of matrix \( \mathbf{M} \) and \( \mu_M \) is the mean weighted value for all parameters in matrix \( \mathbf{M} \).

The PCA-ALS algorithm estimates the missing values. The appropriateness of the substituted values can be determined by comparing the angle between the two spaces spanned by the coefficient vectors. The missing data is close and nearly linearly dependent if the angle yield is small [24]. The results produced by the function for the angle of the coefficient vectors is 5.778e-16. Figure 2.5 shows the reliability of the imputed data compared to the raw data. The linear relationship between the dataset's principal component can be visualized in Figure 2.5(a) and Figure 2.5(b). From the figures, it is evident that differences between raw data and imputed data are very small.

![PCA Analysis for Raw Data](image)

(a)
Another approach to verify the validity of the estimated data is by using Gaussian Kernel Density Estimation (KDE). KDE is a non-parametric method used to estimate the probability density function (PDF) of random variables [25]. Comparison between KDE for the wind speed parameter is presented in Figure 2.6. The PDF between the raw data and imputed data for wind speed datasets is identically distributed and it can be concluded that the imputation method using PCA-ALS is an acceptable technique. For other meteorological parameters, the differences between the measured and the data with missing data estimates using KDE plot are shown in Figure 2.7. This shows the reliability of the PCA-ALS algorithm in imputation.
Figure 2.6: Kernel Density Estimation for Wind Speed dataset with raw data and imputed data

(a) 

(b)
Figure 2.7: KDE for: (a) Pressure, (b) Temperature, (c) Dew point and (d) Humidity

2.5 Meteorological Parameter Analysis

In the previous section, imputation has been applied to estimate missing data. The dataset created is used to further analyse weather parameters. Three years of data from 01/01/2012 to 31/12/2014 has been used to develop a forecast model. The resolution of the data is one hour. In general, annual reports presented by the Malaysia Meteorological Department covers all areas across Malaysia [26]. Because weather fronts are dynamic and change over short periods of time and space [27], the report does not often reflect the actual
localised weather conditions. There is no specific statistical report available for the study area. In this section, an overview of meteorological parameters analyses is divided into two sub-section; preliminary analysis and parameters analyses. In the preliminary analysis, basic statistical analysis is conducted by calculating the simple descriptive statistics from three years of data. Meanwhile, parameters analyses are used to present the parameters’ variations during rain or non-raining conditions. Temperature, dew point and humidity have a homogenous relationship with the amount of moisture in the air and rain condition [28]. For pressure and wind speed, separate representations of their patterns and relationships with rain condition will be presented. All of these analyses are crucial and a major contribution to the forecasting model development using ANN. ANN can forecast output by mapping input and output parameters using appropriate algorithms [29]. Therefore, the parameter analysis is considered as an important initial step in the development of a forecasting model.

2.5.1 Descriptive statistical of meteorological parameter

In order to gain a better understanding of the parameters’ behaviour, initial basic analyses were carried out. Table 2.8, Table 2.9 and Table 2.10 show the descriptive statistics of the parameters for any rain condition, rain and no rain.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Max</th>
<th>Mean</th>
<th>Min</th>
<th>σ</th>
<th>σ²</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure</td>
<td>1017.0</td>
<td>1009.6</td>
<td>1001.2</td>
<td>2.0</td>
<td>4.0</td>
<td>1009.6</td>
</tr>
<tr>
<td>Temperature</td>
<td>38.1</td>
<td>27.5</td>
<td>16.1</td>
<td>3.2</td>
<td>10.2</td>
<td>26.4</td>
</tr>
<tr>
<td>Dew point</td>
<td>29.4</td>
<td>24.5</td>
<td>15.3</td>
<td>1.5</td>
<td>2.2</td>
<td>24.5</td>
</tr>
<tr>
<td>Humidity</td>
<td>100</td>
<td>83.7</td>
<td>29.0</td>
<td>12.2</td>
<td>148.3</td>
<td>88.0</td>
</tr>
<tr>
<td>Wind speed</td>
<td>10.7</td>
<td>1.1</td>
<td>0.0</td>
<td>0.9</td>
<td>0.7</td>
<td>1.1</td>
</tr>
<tr>
<td>Rainfall rate</td>
<td>64.8</td>
<td>0.3</td>
<td>0.0</td>
<td>2.3</td>
<td>5.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

σ = Standard Deviation, σ² = Variance

21
The climate in the study area for three years presented an average of 27.5 °C and an average humidity of 83.7 %. The maximum temperature can reach 40.0 °C and the minimum temperature recorded is 16.1 °C. The average pressure, dew point and wind speed do not show any behavioural differences between rain and not rain condition. On the contrary, temperature and humidity show slight differences in their average values. In general, it has been shown that temperature and humidity are the parameters that have the capabilities to determine the rain condition at given times [30].

### 2.5.2 Meteorological patterns and parameters distribution

Further analysis of meteorological parameters is to investigate the attributes of all parameters in all environmental conditions. Because the main focus in this study is to develop rain forecasting model, all plotted parameter patterns are presented together with the rain condition at hourly interval. In this section, two approaches will be used to determine the meteorological patterns and probability distribution. The patterns of variations are presented...
in Figure 2.8, Figure 2.9 and Figure 2.10, while the cumulative probability distribution for each parameter is depicted in Figure 2.11.

![Figure 2.8: Rainfall rate, temperature, dew point and humidity in (a)1000 and (b)120 step.](image)

On a typical day there are variations in the temperature and humidity. As expected, the relative humidity rises at night (as the temperature falls toward the dew point) and the maximum temperatures are typically recorded
in the afternoon. Figure 2.8 (a) and (b) show that dew point is over 23 °C which is typical in the tropics. In Figure 2.8 (b), it can be seen that when it is raining or slightly before it starts to rain, humidity rises but the temperature drops. This occurs when cold air mixes with warm air, the air temperature will drop.

In pressure observation, low and high pressure areas are associated with vertical movement of the air from a high pressure area to a low pressure area. Rising air cools, thus reducing its ability to retain moisture which condenses as rain. Figure 2.9 shows the measured pressure. When plotted alongside rainfall, it indicates a drop in pressure just before the rainfall but sometime the opposite occurs. These ambiguities are infrequent and therefore a generalised relationship between pressure and rainfall can be deduced.

(a)
Figure 2.9: Pressure and rainfall rate observation in 1000 and 120 hours.

Figure 2.10 shows the plot of the wind speed and rainfall rate observation within the same time interval. The average wind speed tends to be high when it is raining. It can be concluded, in general, that wind condition affects rainfall [31]. In this study, it can be argued that the measurement of the rainfall rate is acceptable and no significant errors were introduced due to strong winds because the average of the wind speed is 1.1 ms\(^{-1}\) (Table 2.8).
In the following parameters analysis, distributions of weather parameters are used to determine the characteristic. Cumulative Distribution Function (CDF) curves are used to determine the significant parameters during raining or not raining events. The CDF of each parameter is compared between rain and no rain events.

Figure 2.11 (a),(b),(c),(d) and (e) show the curves of all parameters. In raining event, humidity varied from 65 % to 100 %. Thus the following characteristic is implied; if the humidity is less than 65 % it is most probably that it is not raining. For temperature greater than 33 °C and dew point is greater than 28 °C, it is not raining. Unfortunately, for atmospheric pressure and wind speed there is no clear distinction between rain and not raining event.
Figure 2.11: Parameters cumulative probability distribution during rain and not raining condition, (a) pressure, (b) temperature, (c) dew point, (d) humidity and (e) wind speed
2.6 Summary

This chapter has described the study area, rainfall rate analysis, data availability, imputation methodology and parameter analysis. Basic information of the study area, including their locations, elevations and fundamental climate characteristic have been presented. For rainfall rate analysis, monthly average for three years, rainfall rate analysis for each station and rainfall rate distribution using CDF have been presented to gain an insight into rain characteristic. The implementation of PCA-ALS imputation method to handle gaps in the data has been presented. This is important to ensure reliable datasets is available to develop a forecasting model. The evaluation of the reliability of imputation has been determined by evaluating the subspace point of PCA. KDE has also been used to evaluate the distribution of the data. Descriptive statistics, parameter pattern and its probability distribution have been presented. Rain is most likely going to occur when the humidity and wind speed is high and when the temperature, dew point and pressure are relatively low.

The next chapter describes the setup and methodology for weather forecast. Datasets available after the imputation method will be used to develop the weather forecasting model.
Chapter 3

Literature Review

3.1 Introduction

This chapter presents the literature in the application of ANN in weather and rainfall forecasting. It presents the Neural Network concept and the analytical methodologies used. The soft computing approach of artificial neural networks coupled with PCA data pre-processing and FCM soft clustering is presented. Earlier related works in computational weather and rainfall forecasting is also described.

3.2 Neural Network

ANN comprises a set of artificial neurons interconnected in an architecture and functionality similar to biological neurons of the brain. A biological neuron illustrated in Figure 3.1, is made up of three basic elements: the body, the dendrites and the axon [32][33].
ANN gains knowledge through a learning process that involves finding an optimal set of weights for the connections. ANNs can be characterised into single layer, bilayer and multilayer with feed forward, recurrent or self-organizing depending on the direction of information flow and processing. The ANN architecture is the multilayer feed forward network with back propagation layer (BPL), representative examples of which are the multilayer perceptron (MLP) and the radial basis functions [34][35]. A neuron is an information processing unit that obtains several signals from its input links, each of which has an assigned weight [34]. These weights correspond to synaptic efficiency in biological neurons [36][37].

3.3 Artificial Neural Network in Weather Forecasting

Weather forecasting is one of the challenging problems especially for tropical climate where the meteorological conditions are dynamic and constantly changing. Many studies have been carried out in weather forecasting using ANN [3][38][39][40]. [3] presents the use of an ANN to provide daily forecasts for temperature, wind speed, and humidity for southern Saskatchewan, Canada. The results show that a proposed ANN with radial basis function outperforms empirical statistical modelling. [39] obtained
temperature and relative humidity weekly forecasts using ANN time series analysis. The network model used is a MLP feed forward ANN model with BPL. The error is less than 3% for 15 weeks temperature and humidity forecast. [39] suggested that statistical parameters could be used as inputs in ANN models for weeks ahead weather forecasting. Daily temperature forecasting model is presented in [40] using ANN with an additional input. Ten years of meteorological data from Kermanshah in Iran have been used in the study. The results show that the best performance of MLP-BPL ANN architecture combines sigmoid transfer function in the hidden layer, linear function in the output layer and Scaled Conjugate Gradient (SCG) algorithm yields maximum error between 0.69 to 1.29 [40]. Rain intensity forecasting in Athens, Greece using the ANN approach was implemented in [41]. The forecast results focuses on the projected mean, maximum and minimum monthly rain intensity for the next four successive months in Athens. The meteorological data used to estimate the rain intensity were the monthly rain totals and the corresponding raining days from the National Observatory of Athens. The training datasets used was 111 years of data from 1899 to 2009. The results of the developed and applied ANN models showed a fairly reliable forecast of the rain intensity for the next four months with statistical significance level of p < 0.01. This work used historical rainfall data and concentrated on longer period of forecasting that serve the needs of monthly rain intensity prediction. However, the study in this thesis focuses on localised short period of prediction at hourly intervals based on meteorological data. Another monthly rainfall prediction using ANN was presented in [42] using data of Darjeeling, India. The proposed rainfall prediction model is the combination of wavelet analysis and ANNs. The results of monthly rainfall series modelling show that the performances of wavelet analysis and ANN models produce 0.9 correlation value. Recently, [43] used deep learning ANN method to forecast air temperature for short-term
prediction in North-Western Nevada, United States. One year of data from 2012 to 2013 was used in the study. Hourly meteorological data such as atmospheric pressure, temperature, humidity, precipitation and wind speed were also used. The results show that in deep learning ANNs, reconstructing the input parameters and combining related meteorological parameters, such as barometric pressure, humidity and wind speed data achieved a 97% accuracy whilst basic ANN implementations yielded 94% accuracy. [44] used temperature time series analysis to forecast monthly rainfall in Australia yields statistical significance value of $p < 0.05$. A neural network based algorithm in forecasting atmosphere pressure for a future time and a given location is presented in [45]. BPL networks have been used to initiate the ANN forecasting model. The proposed ANN based forecasting method integrated BPL and Hopfield Networks and were tested using three years of weather data consisting of 15000 records containing environmental parameters such as temperature, humidity and wind speed. The prediction error was found to be very small and the neural network learning showed better convergence. Their study emphasis was on predictive data mining by extracting interesting non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from enormous amount of meteorological data. The proposed models in [43], [44] and [45] were developed for temperate climate. However the study reported in this thesis is in tropical climate where the behaviour and pattern of the meteorological parameters is different. Most of the weather forecasting methods in the literature are liable on the area where the study has been made specifically. Different climatological characteristic, tend to give a different environmental relationship towards weather condition and rainfall occurrences. The current advances in ANN methodology for modelling non-linear and dynamical phenomena are the motivation to investigate the application of different ANN algorithms for hourly weather forecasting.
3.4 Artificial Neural Network in Rainfall Forecasting

Some studies of short term rain and weather forecasts using ANN have also been reported in [46],[47] and [48]. [46] implemented ANNs to forecast whether it is raining or not regardless of rainfall amount or its intensity. The proposed ANN model results were compared to U.S Weather Bureau forecasts. The ANN model was trained using 24 hour sea level pressure pattern to forecast rain for San Francisco Bay for 12 hours and 24 hours. The forecasted results show that an adaptive system using ANN pattern recognition has the ability to produce 90 % for 12 hours and 81 % accuracy for 24 hours forecasts without understand thoroughly the dynamics of the parameter pattern. A case study of a tropical climate was conducted in Thailand to improve the current rainfall forecast using ANNs [47]. Forecast results for near real time forecasting for 1 to 3 hours achieve more than 0.7 correlation [47]. A combination of spatial meteorological parameters such as humidity, air pressure, temperature and cloud condition from 75 weather stations were used to train the network. Generalized feed-forward networks and simple MLP network architecture were applied in the proposed model [47]. A similar study was conducted in the Philippines using daily meteorological parameters to train the network [48]. Results showed that the daily rainfall forecast achieved high accuracy when ANNs and Bayesian network were implemented [48]. All of these methods took advantage of high spatial and temporal resolution data. For example, data availability for the study in [47] came from 104 weather stations measured over a period of 15 years.

3.5 Artificial Neural Network Architectures

ANN is a mathematical data processing system. The key function of an ANN is to predict or classify appropriately the particular input instances.
Similarly to biological neurons, learning in ANN is achieved by adapting the synaptic weights to the input data [49]. ANN structure has a number of interconnected artificial neurons connected together. It is a mathematical model which functions like the neuron patterns of the human brain. A supervised learning ANN must be trained through a training data set and it creates the patterns and the rules governing the network [50]. In an early model of ANN, feed forward ANN was introduced as perceptron [51]. This model uses a single input perceptron layer for a single output. The disadvantage of this approach is that the perceptron is not able to train and recognize many types of patterns. Each neuron is composed of two parts: a weight coefficient and a transfer function. Figure 3.2 shows a single neuron that consists of a weighted input, summing function and transfer function.

![Single neuron diagram]

Figure 3.2: Single neuron

For a neuron receiving $n$ inputs, each input $x_i$ (where $i = 1...n$) is weighted by multiplying by weight $w_i$. The sum of the $w_ix_i$ products gives the net activation of the neuron. This activation value is subjected to a transfer function ($f$) to produce the neuron’s output, $y$. A single neuron can be described by Equation (3.1):
Adding more perceptron layers in a network topology will increase the ability of the ANN to recognize various classes of patterns. This additional layer is known as MLP. MLP works by adding more layers of nodes between input and output nodes. The neurons are arranged into an input layer, an output layer and one or more hidden layers. The learning process for MLP is known as BPL. The processes repetitively calculate an error function and adjust the weight with smaller error. The transformations of the weights depend on the following steps:

**Step 1.** Each learning step start with forcing input signal from training data set.
**Step 2.** Determine the output values for each neuron in each network layer.
**Step 3.** Finally, best weight is selected to map the input with the output.

Figure 3.3 illustrates the propagation of signals traversing through each neuron from input layer, hidden layer and finally, the target.

Suppose that there are $L$ number of layers, (including input, hidden and
output layer). \( i \) representing a layer that has \( N \) number of nodes in the form of \( N(l) \), where \( l = (0,1,...,L; l = 0 \) is the input parameter) and \( i = 1,...,N(l) \) is the node that has output from a previous layer and is the input for the next layer, \( y_{ij} \) is the output that depends on the incoming signals \( x_{ij} \) and parameters \( \alpha, \beta \) and \( \gamma \). Thus the following equation is the generalization of the output from of node:

\[
y_{ij} = f_{ij}(x_{i-1j}w_{i-1j},...,x_{1,N(l-1)}N(l-1),\alpha,\beta,\gamma,\ldots)
\] (3.2)

After the propagation of signals, the next step is to compare the output signal from the node and the desired output \((z)\) from the training data. In general form, the difference is known as the error signal \( \delta_{ij} \).

\[
\delta_{ij} = z - y_{ij}
\] (3.3)

Assuming that the training data set has \( P \) entries, the error measure for the \( P^{th} \) entry of the training data set is the sum of the squared error:

\[
E_p = \sum_{k=1}^{N(l)} (z - y_{ij})^2
\] (3.4)

Where \( k \) is the number of component \( z \) (desired output). The weight for a particular node is adjusted in direct proportion to the error of the neurons to which it is connected using back propagation Gradient descent algorithm. The Gradient descent algorithm is responsible for finding the weight that will deliver the minimised error. The algorithm can be simplified into two steps below:

**Step 1.** Obtain gradient vector.
**Step 2.** Calculate the error signal \( e_{ij} \) as the derivative of the error measure \( E_p \) with respect to the output node \( i \) in layer \( l \) in both direct and indirect paths. The ordered derivative can be express in the following equation:
\[
\varepsilon_{ij} = \frac{\partial^2 E_p}{\partial y_{ij}}
\]  

(3.5)

Every input parameter that is used to train the network is associated with the output pattern. The ANN model use standard supervised learning MLP trained with BPL algorithm [52]. The whole process of training in MLP-BPL implementations is summarized in the following steps:

**Step 1.** Initialize the weights from training data by mapping the input data to the desired output data.

**Step 2.** Initialize bias by randomly selecting data from the training dataset.

**Step 3.** Compute the output of neurons, error and update weights.

**Step 4.** Update all weights and bias and, repeat Step 3 for all training data.

**Step 5.** Repeat Step 3 and Step 4 until the error is reduce to an acceptable value.

### 3.5.1 Training Algorithm

In this study, three BPL training algorithm were selected and studied, each having a different computational and storage requirement.

#### 3.5.1.1 Levenberg-Marquardt (LM)

The LM algorithm for network training is used in [53] and [54]. The LM algorithm belongs to the quasi-Newton method which is a modification of the classical Newton algorithm approximation of the inverse of the Hessian. The approximation can guarantee that the Hessian matrix is positive definite while maintaining a convergence rate of the order of two. The second-order convergence enables it to escape local minima in the error surface [55].

#### 3.5.1.2 Bayesian Regularization (BR)

BR training algorithm improves generalization and reduces the difficulty of determining the optimum network architecture. BR uses object functions such as MSE to improve generalisation that requires expensive computation to reach optimal level. A detailed discussion of BR is described in [56], [57] and [58].
3.5.1.3 Scaled Conjugate Gradient (SCG)

Training in a neural network is conducted by minimising a global error function, a multivariate function that depends on the weights in the network. Many of the training algorithms are based on gradient descent algorithms [59] such as LM and BR training algorithm. SCG design is based on the conjugate gradient method in numerical analysis [60]. Unlike other conjugate gradient methods, SCG was designed to avoid time consuming line search. It however requires more iteration to converge but the number of computation is reduced because line searches are avoided. This technique does not require any user specified parameters and its computation is faster and inexpensive. A detailed description of the algorithm can be found in [61].

3.5.2 Transfer Function

The Transfer function is used to calculate the net weight or calculate a layer’s output from its network input. Each layer has its own transfer function. The optimum forecast results are based on a combination of ANN learning algorithm and transfer functions [47][62][63]. Several combinations of the following transfer functions are widely used:
Table 3.1: Transfer function

<table>
<thead>
<tr>
<th>Transfer function</th>
<th>Graphical illustration</th>
<th>Mathematical equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td><img src="#" alt="Linear Graph" /></td>
<td>$y = x$</td>
</tr>
<tr>
<td>Sigmoid (Logistic)</td>
<td><img src="#" alt="Sigmoid Graph" /></td>
<td>$y = \frac{1}{1+e^{-x}}$</td>
</tr>
<tr>
<td>Hyperbolic Tangent Sigmoid</td>
<td><img src="#" alt="Hyperbolic Tangent Sigmoid Graph" /></td>
<td>$y = \frac{e^x - e^{-x}}{e^x + e^{-x}}$</td>
</tr>
</tbody>
</table>

3.5.3 Input Model

It is a common practice to split the available data into training dataset and validation dataset. It is important that the training and validation datasets are representative of the same population because ANNs are unable to extrapolate beyond the range of the data used for training [64]. Poor prediction will occur if the validation dataset contains values outside the range of the data used for training. It is often difficult to have a representative validation dataset when the amount of data available is limited. Techniques used to optimally divide the available data are reviewed in [65]. Other issues related to network training is overfitting. Overfitting occurs when the network has memorized the
training examples, but has not learned to generalize or adapt to new inputs [36]. On the other hand, to resolve the overfitting problem, the available dataset need to be divided into three sets, a training dataset, a testing dataset and a validation dataset [47][66]. The testing dataset is a cross-validation dataset for the implemented network. The training dataset is used to train a number of different ANN model configurations. The test set is used to decide when to stop training and also to determine which of the networks is the most accurate. Finally, the validation set is used to evaluate the chosen model against independent data [67].

3.6 Data Pre-processing for Meteorological Parameters

Common problem in meteorological study is missing data due to faulty data acquisition or instrument measurement error. Fundamentally, imprecise information can come from unreliable data containing noise that degrade the quality of data [68]. It is important to pre-process input data before ANN training in order to minimise uncharacteristic inputs which may limit the application of the developed model [69]. Data pre-processing consists of data cleaning and data normalisation. Data cleaning generally include assessing and removing uncharacteristic data, maybe interpolating to replace missing data and reducing noise in the data. Data cleaning attempts to obtain a simple systematic relationship between data input and output [14]. The available data should be standardised to ensure input variables measured on different scales will not dominate the training process because initial weights within a network are randomised to the same finite range [70]. In a general normalisation, data are rescaled to the intervals \([-1, 1]\) or \([0, 1]\) depending on the type of transfer function used in the network [36][70].
3.7 Fuzzy C-Means Clustering

Clustering analysis is a method of clustering data or patterns into groups based on shared characteristic. It attempts to organise an uncategorised input into clusters such that data points within a cluster are more similar to each other than those belonging to different clusters [71][72]. Fuzzy clustering is suitable for handling the issues related to patterns recognition, incomplete or noisy data, mixed media information and can provide approximate solutions faster [73]. FCM introduces the fuzziness for the similar property of each data and can retain more information about the dataset than hard clustering algorithms [74]. Hard clustering only produces crisp partitions directly where each data object is assigned to one cluster only [75]. In contrast, the fuzzy clustering technique assigns a data object to all clusters with different membership degrees [76]. Hard clustering analysis methods like K-means distribute each data to a unique group; however soft clustering algorithms such as FCM utilizes the membership values between 0 and 1 that show the degree of membership for each data to each group and data that belong to more than one cluster [77]. FCM uses fuzzy partitioning of dataset to position the data points that belong to multiple cluster groups with a correspondence fuzzy truth value between 0 and 1, the summation of all membership degrees of every data point to all clusters must be 1 [78][79]. Different membership values show the probability of each object belongs to different groups. The FCM algorithm is one of the most popular clustering methods based on the minimisation of a generalised least-squared errors function. FCM is an iterative and unsupervised algorithm initially developed in [80] and [81]. The use of a FCM approach in meteorology and climatology is presented in [82] to classify the precipitation series and identify the hydrological homogeneous groups. Rainfall data from all weather stations in Turkish basins was used. Regional homogeneity tests were implemented to identify homogeneity of six regions identified by cluster
analysis. The results showed that the performance of FCM method was better than the K-Means method for identification of homogeneous precipitation regions. According to the results, FCM method is recommended for classification of rainfall series [82].

3.8 Performance Indices

The performance of a network can be measured using Mean Absolute Error (MAE) [83], Root Mean Square Error (RMSE) [83] and correlation coefficient (R) [47][84] between observed and forecast values. ANN model evaluation using the magnitude of error has been used by many researches [6][54][69][85][86][70]. In equation (3.6), (3.7) and (3.8), \( y_i \) is the observed value, \( \hat{y}_i \) is forecasted value and \( n \) is the number of observations.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{3.6}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.7}
\]

The correlation coefficient is a measurement of the linear dependency and the strength of the relationship between two variables. If each variable has \( n \) scalar observations, then the Pearson correlation coefficient [87] is defined as:

\[
R = \frac{n \sum y_i \hat{y}_i - (\sum y_i)(\sum \hat{y}_i)}{\sqrt{n(\sum y_i^2) - (\sum y_i)^2} \sqrt{n(\sum \hat{y}_i^2) - (\sum \hat{y}_i)^2}} \tag{3.8}
\]

The \( R \) value will determine the linear relationship between the forecast and observed value. If \( R = 1 \), this indicates that there is an exact linear relationship. If \( R \) is close to zero, then there is no linear relationship between them.
3.9 Summary

This chapter has presented a general overview of ANN modelling and a review on its implementations. Previous studies on relevant topics of weather study have been presented. Standard ANN designs consist of three integral parts namely input, training networks and output or expected results. ANN approaching is susceptible of acquisition the correlation between rainfall condition and related meteorological data. Supervised learning methods are applied throughout the training to find appropriate weights that reduce the inaccuracies between real and forecast value. In this thesis, the recommended ANN is utilizing meteorological parameters like temperature, air pressure, dew point, humidness and wind speed data at the same point of measured rain quantities. Integrating pre-process and soft clustering method on those parameters possess the potential of increasing forecast accuracy. The algorithm of FCM soft clustering is applied to discover the correlation between the meteorological variables and rain intensity. Other similar study of adapting additional technique in ANN implementation is by integrating FCM clustering in daily rainfall forecast model [88]. Data pre-processing algorithms such as PCA-ALS and FCM soft clustering can be integrated with ANN in order to improve the reliability of the model.
Chapter 4

Artificial Neural Network in Weather Forecasting

4.1 Introduction

This study aims to use meteorological parameters to predict weather conditions. Meteorological data from North-West Malaysia is used in this research. The data and pre-processing method have been described in Chapter 2. Many techniques are used in weather forecasting ranging from simple observations to highly complex computerized mathematical models. The Numerical Weather Prediction (NWP) method is commonly used for weather forecasting in Malaysia [38]. NWP forecasting is suitable for large areas but not for localised forecasting [8]. Furthermore, the NWP model is adapted from non-tropical regions such as Europe and Japan [38].

Different ANN algorithms (LM, BR and SCG) have been used. LM, BR and SCG algorithms have been described in Chapter 3. The performance of each algorithm is determined by evaluating the magnitude of the error and the correlation coefficient value between observed and forecasted values. This
Chapter describes the ANN forecasting model to predict environmental parameters for one, three and six hours for localised point prediction.

4.2 Meteorological Dataset

The dataset was divided into training and testing dataset. Chuping and MARDI Bukit Tangga data were selected for training, while Alor Setar weather station has been used for testing. Hourly meteorological data for pressure, dry bulb temperature, dew point, humidity, wind speed, rainfall amount and rainfall rate have been used. The pre-process dataset that was used as input to the model has been discussed in Chapter 2.

4.3 Weather Forecast Implementation Flowchart

Figure 4.1 represent the flowchart for the ANN weather forecasting implementation. The process is divided into three phases; data preparation, ANN implementation and forecast accuracy evaluation.
4.4 ANN Design

The design of the ANN followed the standard steps from data input, network training and finally, evaluation and application [89]. The implementation of multi-step ahead ANN forecast is used to find the applicable forecast accuracy. The architecture of ANN makes use of LM, BR and SCG as training algorithms. Linear (LT) and Hyperbolic Tangent Sigmoid (HT) are used for transfer functions. Multiple input single output (MISO) ANN
architectures have been used in this study. Meteorological parameters such as atmospheric pressure, temperature, dew point, humidity and wind speed were used to map input and output of MISO architecture. MISO network with single output target reduces the time of the training process and MISO architecture coupled with proper training algorithm would return a good convergence [90]. Additionally, it is easy to construct in parallel computation algorithm [91]. The following sections describe the implementation of MISO-ANN architecture for localised short-term weather forecasting.

4.4.1 ANN Preliminary implementation

In the design stage, several models were tested in order to find the optimal ANN for each algorithm. Data was divided into training, validation and testing set. In order to avoid overfitting and better generalisation, 70 % of the training data set was selected and used for cross validation. Overfitting problems has been studied by many researchers, the results show that large training dataset tends to yield optimal solution [92][93][94]. In this study, the minimum training data is set to 70 %. During the training process, it is important to monitor errors. When errors using the validation set increase, the training should be stopped because the point of best generalisation has been reached. The cross validation approach with split-sample training was adopted in this study. Three years of data with hourly observed parameters was used. Randomly selected data was divided into 70 % for training, 15 % for validation and 15 % for testing.

4.4.2 Data normalisation

ANN learns faster and yields better results if the input variables are pre-processed before the network is trained [10]. Data normalization is used to pre-process input and target data. Data normalization is the process of scaling data to fall within a set range. The advantage of scaling the data is to make all
weighted neurons to remain within a predictable range. Scaling of data to lie between -1.0 and 1.0 is adopted and the normalised data was derived using the equation:

\[
x_{\text{scale}} = 2 \left( \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \right) - 1
\]  

(4.1)

where \( x \) is the value before normalisation, \( x_{\text{max}} \) is the maximum value and \( x_{\text{min}} \) is the minimum value in the dataset.

### 4.4.3 Input layer, hidden layer and output layer

In the input layer, each neuron has a single input which is a meteorological data sample. The hidden layer and output layer are responsible for accepting an arbitrary number of inputs based on the type of the chosen interconnection of the neurons. In ANN approach, an issue of overfitting has occurred, fixing the number of a hidden neuron is important to overcome that issue. Number of hidden neurons should be fixed properly in order to provide good generalisation capabilities for the prediction [90]. MLP with BPL ANN architecture allows the potential of neural networks being universal approximator by adding hidden layer [95]. Two hidden layers with 20 and 10 fixed number of neurons have been used, respectively, in ANN training. The number of hidden layers and neurons is randomly selected. Linear (LT) and Hyperbolic Tangent Sigmoid (HT) transfer function have been used in this study (Table 4.1). In ANN, the forecasted value is the value that maps the parameters from the input and hidden layers to the output. This output is known as the target output during the training period. In this implementation, the forecasted values were the individual atmospheric parameters, namely atmospheric pressure, temperature, dew point, humidity and wind speed. The input and output applied in this study can be describes in a schematic
representation from Equation (4.2) and (4.3) in Section 4.4.6.

4.4.4 Training algorithm

In this study, supervised learning of LM, BR and SCG algorithms have been adopted. LM is widely used in the training of ANN-MLP-BPL model. LM iteratively adjusts estimates of model parameters to minimise residual errors between measured dependent variable outputs and predictions from a numerical model based on independent variable inputs [96]. LM minimises the error in order to ensure the validity of the linear approximation reached. This can be accomplished by modifying or updating the error function [90]. The BR training algorithm is a linear combination of a Bayesian method and ANN to automatically determine the optimal regularisation parameters [97]. SCG is a training algorithm that utilises the conjugate gradient method. The SCG training algorithm was developed to avoid the time consuming line search for faster convergence [98].

4.4.5 Transfer function

The approach of modelling weather forecasting using ANN-MLP-BPL is a regression approximation technique. LT and HT are the learning mechanisms capable of adjusting weights in order to minimize the error. It has been discussed in literature that LT is sometimes superior to the sigmoid (ST) functions. When LT is used in the output layer and the values are bounded within a small range (in this this study it is between -1.0 and +1.0 after normalization) the learning rate is improved [63]. The HT function will generates a value that is close to -1.0, and thus will maintain learning. In contrast, ST will generate a value that it close to 0.0 if the argument of the function is substantially negative. Consequently, the output (mostly if adopted in hidden layer) will be close to zero and ST also drops the learning rate for all subsequent weights. In this situation the regularisation of the BPL network will
have the potential to terminate the learning process. For this particular study, LM and HT are recommended to avoid early termination of the learning process. However, a non-linear transfer function is used in the first hidden layer instead of LM, otherwise it will end up with linear solutions. Table 4.1 shows the transfer function use in this study.

### Table 4.1: List of transfer function used in the experiment.

<table>
<thead>
<tr>
<th>Transfer function</th>
<th>Graphical illustration</th>
<th>Mathematical equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td><img src="image" alt="Linear Graph" /></td>
<td>$y = x$</td>
</tr>
<tr>
<td>Hyperbolic Tangent Sigmoid</td>
<td><img src="image" alt="Hyperbolic Graph" /></td>
<td>$y = \frac{e^x - e^{-x}}{e^x + e^{-x}}$</td>
</tr>
</tbody>
</table>

#### 4.4.6 Prediction setup

The input vectors of hourly pressure, temperature, dew point, humidity and wind speed were used for training. One to six hours lag of target vectors were used as an additional input for each forecast parameter. Each iteration find the optimum outcome for one to six hours forecast.

One, three and six hours forecast of pressure (P), temperature (T), dew point (DP), humidity (H) and wind speed (WS) were evaluated. The LM algorithm was used in the same ANN architecture. The accuracy of each hourly predicted value was analysed. In general the input parameters $P_{input}$ and output parameters $P_{output}$ for the ANN architecture can be describe as:
\[
P_{\text{input}} = [P_{t-n}, T_{t-n}, DP_{t-n}, H_{t-n}, WS_{t-n}, PF_{t-n}]
\]  
(4.2)

\[
P_{\text{output}} = [PF_{t}]
\]  
(4.3)

where PF is the metrological parameter observed at time \(t\) and \(n\) is the hour(s) before the forecast. In the initial setup for an hour forecast, \(n = 1, t = 2\). For 3 hours \(n = 3, t = 4\) and for 6 hours ahead \(n=6, t=7\).

The ANN model used was trained using the BPL-MLP training algorithm as shown in Figure 4.2.

![Figure 4.2: A four layer ANN model with three layer feed forwards ANN.](image)

### 4.4.7 Input parameters and selection

The choice of input parameters is very important in ANN model. Even though there are five meteorological parameters (temperature, pressure, dew point, humidity and wind speed), only parameters that have strong correlation value are applied to the ANN model. Wind direction was measured on the
ground and there is no data available for atmospheric wind direction at all of the weather stations. Ground based wind speed and wind direction observations are affected by the local topography [99], but the atmospheric wind direction data is needed as inputted to the forecast model. Single observations on the ground are sometimes poorly correlated to the general cloud direction, the wind direction on the ground estimated from several point measurements can be linked to precipitation occurrence [99]. Due to unavailability of atmospheric wind direction and the propose model is for localised prediction, therefore wind direction is excluded in this study.

There is no systematic approach in parameter selection but statistical analysis can be used to determine the relevant inputs. In this study, Pearson rank correlation is applied to all datasets to find the relevant inputs [87].

4.4.8 ANN Configuration

The structure of the ANN forecasting model can be summarized as follows:

- Input layer: N number of neurons where N > 0
- Hidden layer: Two hidden layer with 20 and 10 neurons.
- Output layer: One output layer, where the output for next one hour parameter forecast is obtained.
- Training functions: LM, BR, SCG
- Transfer functions: HT and LT.
- Training set: 78912 datasets.
- Test set: 26304 datasets.
- Training iterations: 1000 epochs.
- Performance function: MSE=0.0001.

4.5 Results and Discussions
The results in this section were used to evaluate the accuracy of LM, BR and SCG algorithms used in the ANN weather forecasting model. The emphasis was on predicting atmospheric pressure, temperature, dew point, humidity and wind speed values and the results show that LM is the best algorithm for forecasting with average of coefficient 0.9 for one hour, 0.8 for three hours and 0.5 for six hours. Results are summarized in Table 4.2, Table 4.3 and Table 4.4. LM yields low MAE and RMSE compared to BR and SCG algorithms except in a very few cases where BR produced better results. The correlation coefficient (R) values produced by LM is greater than 0.95 for one hour forecast. Figure 4.3 shows the training, validation, testing and overall regression plots for temperature forecast using LM algorithm for one hour forecast. LM offers the best accuracy, followed by BR and SCG. In terms of processing time, SCG is the faster [61] compared to LM and BR but it does not produce good convergence. BR algorithm takes more time compared to LM and SCG during training but it converges faster. Figure 4.4 to Figure 4.8 show the comparison between forecast and measure data results for one hour forecast using LM algorithm. For three and six hours forecasts, the accuracy of the prediction gradually degrades when the forecast interval is increased. In ANN architecture, the drawback of the multi-step forecast is that the network mapping capability is reduced when the forecast steps are increase [91].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ALGORITHM</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>LM</td>
<td>0.637</td>
<td>0.917</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.649</td>
<td>0.927</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.676</td>
<td>0.954</td>
<td>0.954</td>
</tr>
<tr>
<td>Pressure</td>
<td>LM</td>
<td>0.441</td>
<td>0.548</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.445</td>
<td>0.566</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.477</td>
<td>0.585</td>
<td>0.956</td>
</tr>
<tr>
<td>Dew Point</td>
<td>LM</td>
<td>0.067</td>
<td>0.132</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.077</td>
<td>0.189</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.141</td>
<td>0.245</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td>BR</td>
<td>SCG</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Humidity</td>
<td>2.368</td>
<td>3.444</td>
<td>0.959</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.409</td>
<td>3.486</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.682</td>
<td>3.707</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td>Wind Speed</td>
<td>0.047</td>
<td>0.068</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.068</td>
<td>0.096</td>
<td>0.994</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.052</td>
<td>0.074</td>
<td>0.996</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3: Three hour parameter forecast

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ALGORITHM</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>LM</td>
<td>1.185</td>
<td>1.577</td>
<td>0.869</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>1.190</td>
<td>1.583</td>
<td>0.868</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>1.279</td>
<td>1.661</td>
<td>0.854</td>
</tr>
<tr>
<td>Pressure</td>
<td>LM</td>
<td>0.782</td>
<td>0.966</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.787</td>
<td>0.970</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.833</td>
<td>1.014</td>
<td>0.862</td>
</tr>
<tr>
<td>Dew Point</td>
<td>LM</td>
<td>0.558</td>
<td>0.741</td>
<td>0.869</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.568</td>
<td>0.750</td>
<td>0.866</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.587</td>
<td>0.785</td>
<td>0.852</td>
</tr>
<tr>
<td>Humidity</td>
<td>LM</td>
<td>4.185</td>
<td>5.716</td>
<td>0.833</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>4.216</td>
<td>5.768</td>
<td>0.881</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>4.491</td>
<td>6.035</td>
<td>0.869</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>LM</td>
<td>0.474</td>
<td>0.601</td>
<td>0.708</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.480</td>
<td>0.609</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.682</td>
<td>0.622</td>
<td>0.682</td>
</tr>
</tbody>
</table>

Table 4.4: Six hour parameter forecast

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ALGORITHM</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>LM</td>
<td>2.178</td>
<td>2.739</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>2.182</td>
<td>2.734</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>2.400</td>
<td>2.890</td>
<td>0.422</td>
</tr>
<tr>
<td>Pressure</td>
<td>LM</td>
<td>1.279</td>
<td>1.629</td>
<td>0.581</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>1.289</td>
<td>1.637</td>
<td>0.575</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>1.389</td>
<td>1.727</td>
<td>0.506</td>
</tr>
<tr>
<td>Dew Point</td>
<td>LM</td>
<td>0.830</td>
<td>1.049</td>
<td>0.714</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.840</td>
<td>1.058</td>
<td>0.707</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.901</td>
<td>1.129</td>
<td>0.658</td>
</tr>
<tr>
<td>Humidity</td>
<td>LM</td>
<td>7.756</td>
<td>9.850</td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>7.736</td>
<td>9.854</td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>8.403</td>
<td>10.859</td>
<td>0.526</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>LM</td>
<td>0.596</td>
<td>0.732</td>
<td>0.509</td>
</tr>
<tr>
<td></td>
<td>BR</td>
<td>0.613</td>
<td>0.748</td>
<td>0.476</td>
</tr>
<tr>
<td></td>
<td>SCG</td>
<td>0.612</td>
<td>0.750</td>
<td>0.471</td>
</tr>
</tbody>
</table>
Figure 4.3: Regression plot for temperature forecast using ANN with LM algorithm.

Figure 4.4: One hour forecast results for pressure using LM training algorithm.
Figure 4.5: One hour forecast results for temperature using LM training algorithm

Figure 4.6: One hour forecast results for humidity using LM training algorithm
Figure 4.7: One hour forecast results for dew point using LM training algorithm

Figure 4.8: One hour forecast results for wind speed using LM training algorithm
Figure 4.9: Three hours forecast results for temperature using LM training algorithm

Figure 4.10: Six hours forecast results for temperature using LM training algorithm

### 4.6 Summary

The ANN weather parameter forecasting model has been trained using historical meteorological data from three weather stations. The relationship between meteorological parameters such as pressure, temperature, dew point, humidity and wind speed were used to develop the model. MISO-ANN with LR, BR and SCG training algorithms together with the HT and LT transfer functions have been used in developing the forecasting model. PCA-ALS imputation and pre-processing technique used to improve the model’s
performance. From the overall performance, the proposed and implemented technique is capable of capturing and predicting the dynamic behaviour of atmospheric pressure, temperature, dew point, humidity and wind speed for one to six hours ahead. The average of 1 hour prediction gives 0.95 correlation value, 3 hours prediction yields 0.84 correlation value and 0.62 correlation value for 6 hours prediction. However, one hour and three hours prediction results yield correlated values that are greater than 0.75. This study has initiated a new development in meteorological study and weather forecasting system for localised tropical climate. Forecast results from this study can be very useful in environmental condition prediction. In the next chapter, further investigation of ANN forecasting model that makes use of similar attributes to predict rain condition and rainfall rate is presented. Due to the non-linearity of the meteorological processes and dynamic weather condition in tropical climates, developing a rain forecasting model is a demanding and challenging task.
Chapter 5

ANN Rainfall Forecasting Model

5.1 Introduction

Rainfall forecasting is extremely important for efficient water management and effective flood forecasting. Flooding is a common natural hazard in tropical countries that has significant impact on life and property [100]. Reliable rainfall forecasting will reduce the damages resulting from flood events, allowing time for evacuation forming part of a flood management system. Rain in the tropics is characterized by high intensity over short periods of time within smaller rain cell sizes that are localised and maybe highly dynamic depending on wind condition. Due to these characteristics, localised flash flood occurrence is frequent. In order to predict these events, an accurate rainfall forecast model is crucial. Rainfall forecasting remains one of the most challenging topics in hydrological operation due to spatial and temporal variations of rainfall distribution, dynamic behaviour of tropical climates and the highly non-linear nature of rainfall events. The NWP method is commonly used for weather forecasting in Malaysia and it is unsuitable because it is has been adapted from non-tropical regions such as Europe and Japan [38].
The proposed ANN model in this study uses atmospheric pressure, temperature, dew point, humidity and wind speed data measured at the same point as rainfall measurements. The permutations of meteorological parameters, expert knowledge and empirical study have been used to process the data and develop the ANN model. This chapter presents a basic implementation of ANN rainfall forecast model and evaluates the accuracy.

5.2 Study Area and Meteorological Data

Three years of data from 01/01/2012 to 31/12/2014 that consists of 26304 dataset from Chuping weather station have been used. Figure 5.1 shows the geographical location of Chuping weather station. The minimum, maximum and average value of the atmospheric parameters obtained from the location of interest are summarize in Chapter 2. From the available data, 2974 dataset are rainfall events and the remaining 23330 dataset are non-rainfall events. Table 5.1 shows statistics of rainfall rate for Chuping in three years. The non-zero rain data is about 11.3 %. Figure 5.2 is a graph that indicates the monthly average rainfall rate. Over three years, the Chuping rainfall rate patterns are not uniformly distributed without clear distinction between dry and wet seasons. However, both dry and wet seasons happen twice a year based on the annual rain pattern.
Figure 5.1: Map of Malaysia and Chuping

Table 5.1: Hourly and monthly average rainfall rate statistic for Chuping.

<table>
<thead>
<tr>
<th>Weather Station</th>
<th>Period</th>
<th>MIN ( \text{mmh}^{-1} )</th>
<th>MAX ( \text{mmh}^{-1} )</th>
<th>MEAN ( \text{mmh}^{-1} )</th>
<th>( \sigma )</th>
<th>MED</th>
<th>( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chuping</td>
<td>Month</td>
<td>0.1049</td>
<td>4.8564</td>
<td>2.8436</td>
<td>1.1474</td>
<td>2.7098</td>
<td>1.3166</td>
</tr>
<tr>
<td></td>
<td>Hour</td>
<td>0.0900</td>
<td>95.2800</td>
<td>2.9370</td>
<td>7.1997</td>
<td>0.5400</td>
<td>51.8355</td>
</tr>
</tbody>
</table>

\( \sigma \) = Standard Deviation, \( \sigma^2 \) = Variance

Figure 5.2: Monthly rainfall rate average for Chuping
5.3 ANN Design

The following flow diagram is the step by step implementation of the ANN [89]:

![Flow Diagram](image)

Figure 5.3: ANN step by step design

In the preliminary implementation, several models were tested in order to find the optimal ANN. Data were divided into training, validating and testing sets. In the input layer, each neuron receives a single input that is a meteorological parameter. However, the hidden layer and output layer accept an arbitrary number of inputs based on the type of chosen interconnection of
the neurons. Common practices to determine the number of hidden layers and neurons include trial and error methods, genetic algorithm and Bayesian network modelling [37]. Bayesian network modelling can be applied to determine the optimum neural network structure but this method is computationally complicated to implement [7]. There is also an input-output based guideline, which suggests that the number of neurons in a hidden layer conform to the following rule:

$$[2n^{1/2}, 2n+m]$$

(5.1)

where \( n \) is the number of inputs and \( m \) is the number of outputs in the neural network [55]. This study will use a combination of trial and error method and the range method from Equation (5.1).

The LM, BR and SCG training algorithms have been used in this study. In this basic ANN forecasting model several combinations of the following transfer functions are used:

<table>
<thead>
<tr>
<th>Transfer function</th>
<th>Graphical illustration</th>
<th>Mathematical equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td><img src="image" alt="Linear Illustration" /></td>
<td>( y = x )</td>
</tr>
<tr>
<td>Sigmoid (Logistic)</td>
<td><img src="image" alt="Sigmoid Illustration" /></td>
<td>( y = \frac{1}{1+e^{-x}} )</td>
</tr>
<tr>
<td>Hyperbolic Tangent Sigmoid</td>
<td><img src="image" alt="Hyperbolic Tangent Sigmoid Illustration" /></td>
<td>( y = \frac{e^x - e^{-x}}{e^x + e^{-x}} )</td>
</tr>
</tbody>
</table>

Table 5.2: Transfer function

The transfer function is used to calculate the net weight or calculate a layer's
output from its network input. Each layer in all setups has its own transfer function. The optimum forecast results are based on combination settings of ANN learning algorithm and transfer functions [101].

### 5.4 Rainfall Forecast ANN Architecture

The proposed ANN forecasting model is based on MLP with BPL architectures. Different neural network structures have been tested; Figure 5.4 is an illustration of the ANN forecasting model architecture. MISO implementation is used in the setup. In the standard ANN rainfall forecasting model, all meteorological parameter are selected and inputted into the network without any sanitation process.

![ANN architecture for rainfall forecasting model](image)

Figure 5.4: ANN architecture for rainfall forecasting model

The input parameters $P_{input}$ and output parameters $P_{output}$ for the neural network architecture can be described as follows:
\[ P_{\text{input}} = \{ P_{t-n}, T_{t-n}, D_{t-n}, H_{t-n}, W_{S_{t-n}}, R_{t-n} \} \]  \hspace{1cm} (5.2)

\[ P_{\text{output}} = \{ R_t \} \]  \hspace{1cm} (5.3)

where \( R \) is the rainfall rate observed at time \( t \) and \( n \) is the number of hours before the rain. The ANN architecture can be simplified to:

\[ [I, H_1, ..., H_n, O] \]  \hspace{1cm} (5.4)

where \( I \) is the input layer, \( H \) is the hidden layer, \( O \) is the output layer and \( n \) is the number of hidden layers. For example, in the preliminary network architecture which consists of 5 inputs, 1 output and 2 hidden layers with 20 and 10 neurons respectively, the architecture is 5-20-10-1.

### 5.5 Forecasting System Setup

The implementation was initially designed to predict the state (raining or not raining) and if it was raining, to predict the rain intensity. State forecast is a binary predictive target where 0 represents a non-rain event and 1 represents a rain event, while the value forecast is the model that forecasts the rainfall intensity. For the purposes of this study, the input vectors of atmospheric data are used for training and classification of the ANN forecasting model. Hourly meteorological parameters from January 2012 to December 2014 were used. The number of hidden units in the single hidden layer was set to 10 for each model. The number of neurons and hidden layers are based on trial and error method and Equation (5.1).

Each of the implementation finds the optimum for hourly period forecasts, evaluates the magnitude of error and correlation coefficient of the predicted output against the desired output. The initial numbers of iterations were performed using a basic ANN such as a single hidden layer, small amount
of validation checks and training iterations (epochs). Randomly selected data for training, validation and test sometimes result in overfitting or saturation [6][102]. ANN, by its nature, is biased towards the training dataset and will always produce high correlation values (high accuracy). In order to reduce overfitting, the training dataset was set to 70 % and an indexing dataset selection approach was implemented [92][93]. The performance of the model was evaluated using MAE, RMSE and the correlation coefficient value between predicted and measured values.

5.6 Results and Discussions

In general, the proposed ANN utilised MISO with 5 inputs and single output for both state and rain intensity forecasting. Presented in this section are the results for 1 to 6 hours forecasting based on correlation coefficient R, MAE and RMSE. Table 5.3 shows the results for ANN rainfall forecasting model.

<table>
<thead>
<tr>
<th></th>
<th>Transfer Function</th>
<th>Architecture</th>
<th>Training Function</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RAIN INTENSITY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.374</td>
<td>2.408</td>
<td>0.334</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.542</td>
<td>2.426</td>
<td>0.305</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.608</td>
<td>2.536</td>
<td>0.094</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.598</td>
<td>2.543</td>
<td>0.057</td>
</tr>
<tr>
<td><strong>STATE FORECAST</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.090</td>
<td>0.212</td>
<td>0.589</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.117</td>
<td>0.240</td>
<td>0.407</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.158</td>
<td>0.285</td>
<td>0.342</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,L</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.310</td>
<td>0.288</td>
<td>0.310</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.164</td>
<td>0.291</td>
<td>0.276</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.225</td>
<td>0.339</td>
<td>0.269</td>
</tr>
</tbody>
</table>

BR=Bayesian Regularization, LM= Levenberg Marquardt, SCG=Scaled Conjugate Gradient, T=Hyperbolic Tangent Sigmoid, L=Linear, S=Sigmoid

From the different learning algorithms, BR gives a good convergence compare to LM and SCG. In the basic ANN model implementation (Table 5.3), rainfall rate forecast gives less than 0.1 correlation coefficient value for more
than 2 hours forecast. The implementation of standard ANN suffers from poor convergence where most of the correlation results is less than 0.5 and it indicates that using a basic ANN for rainfall forecast is not reliable. The transfer function used in each layer has a significant impact on the forecasting process. Hyperbolic Tangent Sigmoid and Linear transfer functions work well in most ANNs. Logistic Sigmoid did not meet the expected accuracy for prediction. The number of hidden layers and their neurons were based on a trial and error approach. When the number of hidden layers is set to more than 2, the accuracy remained approximately the same but it slowed down the convergence process. Therefore, the maximum number of hidden layers was set to 2. Estimating the rainfall intensity used between 5 to 20 neurons based on Equation (5.1).

5.7 Summary

In tropical areas, rain dynamics are more challenging to predict because of their localised characteristic. The study area in this paper experiences high rain intensity over short durations. The highly dynamic and non-linear nature of the weather condition is a reflection of the complex interaction between the various processes at different scales. Therefore, the design of an effective system for prediction is challenging. This study uses ANN to develop a rainfall forecasting model. The accuracy of the forecast results based on correlation value is less than 0.5 in average. The available data to train the network for high rain intensity (rainfall rate greater than 31 $mm\cdot h^{-1}$) is about 5.4 % which compounds the challenge to develop an accurate model. The data for high rain intensity is low (5.4 % of recorded rain events) because of insufficient time resolution of weather station (rain gauges) to measure high rain intensity. The rainfall rate is averaged over an hour which obscures the high rain intensity over short time periods. That is a limitation of the data but one that does not limit the validity of any developed model in this study to estimate rain
intensities over hourly periods.
Chapter 6

PCM-FCM-ANN Rainfall Forecasting Model

6.1 Introduction

This chapter presents how the proposed pre-processing method and FCM soft clustering have been combined with ANN to implement a rainfall forecasting model. FCM clustering is applied in order to find the relationship between the environmental parameters and rainfall rate. Integrating additional processes or techniques in ANN recently attracted more attention in meteorological study. [103] applied PCA to the input vectors to improve rainfall-runoff forecast accuracy. Comparisons between rainfall data with and without PCA showed PCA enhanced accuracy. [88] applied FCM clustering technique was also used to cluster rainfall into three clusters in ANN training dataset. Both of the mentioned approaches were applied in different climate were rainfall is wide spread and less variable. This chapter determines the reliability of the clustering technique in forecasting rainfall rate by exploring the integration of FCM and the ANN model in cloud burst scenarios.
6.2 Meteorological Data

The PCA-ALS was applied to the hourly data from all weather stations. For rainfall rate forecasting, additional inputs of clustered value are added to the multiple inputs with single output structure of standard ANN after FCM clustering is applied.

6.3 Implementation Framework of PCA-FCM-ANN Forecasting Model

The proposed model starts with data pre-processing by eliminating noise and assigning new data for missing or corrupted data. FCM clustering is then applied for rain and not rain conditions. Environmental parameters and rainfall rate were clustered into several subsets. Input vectors of clustered dataset together with pre-process environmental data are used as input to the PCA-FCM-ANN forecasting model. Figure 6.1 show the flowchart of the forecasting model.
Figure 6.1: Flowchart of forecasting for modular PCA-FCM-ANN implementation.
6.4 PCA-ALS Imputation Method

In Chapter 2, PCA-ALS imputation has been presented. A step by step imputation process based on the PCA-ALS algorithm was described in 2.4 and depicted in Figure 6.1. The processed datasets are then clustered using FCM soft clustering before they are inputted in the PCA-FCM-ANN forecasting model.

6.5 FCM Clustering

FCM soft clustering is a process of discovering an appropriate group of a data. Patterns of data can provide a crisp representation of the behaviour of a process by falling into groups that have similar characteristics and defined different clusters that have characteristics which are not identical. The aim is to reduce the non-linearity of the raw data by grouping similar segments of the training data. It attempts to organise uncategorised input into clusters where data points within a cluster are similar rather than those belonging to different clusters [104][105].

In the proposed PCA-FCM-ANN rainfall forecasting model, the training data is portioned into five clusters. The number of cluster is obtained using Partition Coefficient indexing [106]. The soft clustering approach has been used to identify the substantial input (environmental parameters) for hourly rainfall prediction. Figure 6.2 depicted the schematic illustration of the PCA-FCM-ANN implementation. FCM clustering is applied separately for no rain and rain datasets. Then the clustered parameters are combined to create input and output vectors. The clustering is important when there are similar class patterns
between input parameters for rain and not rain conditions. The clustered input parameters are the property of several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. From this situation, the relationships between input-output parameters are improved. Furthermore it can reduce overfitting during network training [107]. FCM is an iterative algorithm that will define the centres of the clusters (centroids) from the data reciprocal distances in order to minimise dissimilarity and works iterative until the termination criterion is satisfied [108] [109].

![Figure 6.2: Schematic illustration of FCM clustering integration into ANN rainfall forecasting model](image)

Assuming the matrix of dataset is $U$, FCM randomly initializes the membership matrix of the dataset based on Equation (6.1)

$$\sum_{j=1}^{n} u_{ij} = 1, \forall j = 1, ..., n \quad (6.1)$$

The dissimilarity function which is used in FCM is given by Equation (6.2)
where $u_{ij}$ is degree of membership of $x_i$ which takes a value between 0 and 1 (for a given data point $j$, the sum of the membership values for all clusters is 1), $c_i$ is the centroid of the cluster $i$, $n$ is the number of clusters, $d_{ij}$ is the Euclidian distance between $i^{th}$ centroid($c_i$) and $j^{th}$ data point, and $m$ is a weighting exponent (where $1 \leq m \leq \infty$) [69]. An iterative optimisation algorithm is used to minimize the dissimilarity functions through calculating the centre vectors using Equation (6.3) in which $x_i$ is the $i^{th}$ data point. This is followed by an update of the matrix ($U^k$ to $U^{k+1}$) where $k$ is the iteration step in Equation (6.4).

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_i}{\sum_{j=1}^{n} u_{ij}^m} \quad (6.3)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{kj}}{d_{ij}} \right)^{2/(m-1)}} \quad (6.4)$$

The iteration will stop when $\{\|u_{ij}^{k+1} - u_{ij}^k\|\} \leq \varepsilon$ where $\varepsilon$ is a termination criterion that has been defined.

The FCM algorithm can be summarised as follows:

**Step 1.** Randomly initialize the membership matrix ($U=\{u_{ij}\}$) by using Equation (6.1) Initialize the number of clusters ($c$), weighting exponent ($m$), termination criterion ($\varepsilon$) and iteration limit.

**Step 2.** Calculate the centroids ($c$) using Equation (6.3).

**Step 3.** Compute the dissimilarity between centroids and data points using Equation(6.2)

**Step 4.** Update $U^k$ to $U^{k+1}$ using Equation (6.4).

**Step 5.** If $\{\|u_{ij}^{k+1} - u_{ij}^k\|\} \leq \varepsilon$ then the iteration will terminate otherwise recalculating
the centroids \((c_i)\) using (6.3) in step 2.

FCM iteratively updates after refining the centroids and the membership value for each data point within a group of data. However, FCM does not ensure that it converges to an optimal solution because the cluster centres are randomly initialised in matrix \(U\). The performance depends on the initial centroids selected and for a robust approach, FCM should execute several times with different initial centroids. FCM requires users to specify the number of clusters; the cluster validity criterion is usually performed to find the optimal number of clusters for the dataset [110]. Cluster validity is the process of validating the cluster algorithm and determines the amount of cluster group. Partition Coefficient of validating indexing is used in this study [106]. FCM was also used to carry out pre-classification in data reduction [69].

After imputation and clustering of the raw data (environmental parameters), the next step is to determine the ANN architecture and training strategies.

### 6.6 ANN Design

An identical setup used in Chapter 5 (section 5.3) has been applied in the forecasting model development, comprising data normalisation, input layer, hidden layer, output layer, training algorithm and transfer function configurations. Figure 6.3 shows the ANN design flowchart that was integrated with pre-processing and clustering method.
Figure 6.3: ANN flowchart implementation

6.7 ANN Architecture
The proposed model is based on MLP with BPL different structures have been evaluated. Figure 6.4 is an illustration of the ANN forecasting model architecture. In the standard ANN rainfall forecasting model, all meteorological parameter are selected and inputted into network without any sanitation. The proposed PCA-FCM-ANN model takes advantage of pre-processing in model input selection and additional improvement of the accuracy through clustering implementation [78][111].

![Diagram of PCA-FCM-ANN architecture for rainfall forecasting model](image)

Figure 6.4: PCA-FCM-ANN architecture for rainfall forecasting model

In general, the input parameters $P_{input}$ and output parameters $P_{output}$ for the neural network architecture can be described as follows:

$$P_{input} = [(P,T,D,H,WS,R)_{FCM}, P_{t-n}, T_{t-n}, D_{t-n}, H_{t-n}, WS_{t-n}, R_{t-n}]$$  \hspace{1cm} (6.5)

$$P_{output} = [R_t]$$  \hspace{1cm} (6.6)

where $R$ is the rainfall rate observed at time $t$ and $n$ are the number of hours before the rain, $P,T,D,H$ and $WS$ is the environmental parameters which is pressure, temperature, dew point, humidity and wind speed, respectively.
\((P, T, D, H, WS)_{rcm}\) is the cluster value for each environmental parameter. The ANN architecture can be simplified to:

\[ [I,h_1,...h_n,O] \] (6.7)

where \(I\) is the input layer, \(h\) is the hidden layer, \(O\) is the output layer and \(n\) is the number of hidden layers. The cluster values are also added to input vectors in PCA-FCM-ANN forecasting model.

### 6.8 Results and Discussions

#### 6.8.1 PCA-ALS Imputation

The accuracy of the substituted values can be determined using KDE. PDF between raw and imputed data for atmospheric parameters are presented in Figure 6.5(a),(b),(c),(d) and (e). The identical distribution of the estimated values shows that imputation using PCA-ALS is suitable. The differences are not visible from the KDE plots for other meteorological parameters that have less amount of missing data. It shows that the PCA-ALS approach in pre-processing and interpolating data is robust.

![PDF distribution](image)

(a)
(b) 

(c) 

(d)
6.8.2 FCM clustering

The proposed FCM clustering for optimal partition of data was applied to all datasets (Chuping, Alor Setar and MARDI Bukit Tangga). Each dataset consisting of 26304 instances comprising weather parameters (with six attributes: pressure, temperature, dew point, humidity, wind speed and rainfall rate) are clustered using FCM algorithm. The initial number of clusters is set to five and the maximum number of iterations is set to 100. The termination criterion is set to $10^{-4}$. A general early stopping method is applied by testing the error on the validation and training set when the validation error reduced for a number of successive epochs [112]. Since the data are divided into rain (wet) and not rain (dry) condition, the proposed method partitions the data into two FCM cluster according to the rain condition (refer to Figure 6.1 FCM Clustering flowchart). Once the training data are prepared, the next step is to use a supervised ANN to capture the relationship between these input-output parameters.
6.8.2.1 FCM Clustering for wet condition

Figure 6.6, Figure 6.7, Figure 6.8 and Figure 6.9 show the results for each cluster and represent the rain condition. Five clusters are labelled and represent homogenous groups for all input parameters. Most of the cluster groups overlap because the dynamic nature of environmental parameters in tropical climate [113]. For example, the same temperature value can be obtained during wet and dry conditions. There are no distinctive attributes that will specifically determine the rainfall condition. Therefore, the extension of the relationship among others parameters is crucial in forecast determination.

Figure 6.6: FCM Clustering for Pressure and Temperature during raining condition
Figure 6.7: FCM Clustering for Dew Point and Humidity during raining condition

Figure 6.8: FCM Clustering for Humidity during raining condition
6.8.2.2 FCM Clustering for dry condition

The dry condition or not rain is the dominant condition. Out of 78912 samples in the dataset, 5611 (7%) of them were rain events. Environmental parameters commonly have similar characteristic in rain and not rain conditions. A reduced dataset will increase the relationship between input and output parameters for the no rain condition. During rain events, humidity varies from 65% to 100%. Thus; if the humidity is less than 65% it is most probably it is not raining. For temperature greater than 33 °C and dew point greater than 28 °C, no rain is highly probable. For pressure and wind speed, no clear distinction between rain and not rain events can be found. Only a specific range of parameter values that contribute to the dry condition were selected and clustered. Figure 6.10 (a),(b),(c),(d) and (e) shows the scatter plot for five parameters that have been used in this study; pressure, temperature, dew point, humidity and wind speed. Parameters that have the same characteristic and overlapped were excluded. In Figure 6.10(e), the scattered plots shows that no distinctive characteristic exist between rain and not rain conditions from wind speed data. Figure 6.11 and
Figure 6.12 show the results of FCM clustering for pressure, temperature, humidity and dew point for the no rain condition. Each cluster group has its own cluster centre, also known as a centroid. Cluster centre or centroid is a vector containing one number for each variable, where each number is the mean of a variable of the observations in that cluster [114].
Figure 6.10: Wet and dry comparison for (a) pressure, (b) temperature, (c) dew point, (d) humidity and (e) wind speed.

Figure 6.11: Dew Point and Temperature clustering for dry condition.
After successfully clustering the environmental parameters for rain and not rain conditions, the next step is to apply the results as inputs to the rainfall forecasting model.

### 6.8.3 Forecast result

FCM clustering was applied to all three weather stations (Chuping, Alor Setar and MARDI Bukit Tangga). The main purpose for testing the data from all weather stations was to investigate the generalisation and capability of the proposed model in different locations with similar climatic conditions. The conventional ANN model was compared with the proposed PCA-FCM-ANN model. Results presented in this section are for 1 to 6 hours forecasts, state forecast (comparing rain versus not rain prediction - a binary forecast) and rain intensity forecast (rainfall rate).

#### 6.8.3.1 Results for Chuping weather station

Results for Chuping weather station is presented in Table 6.1 and Table 6.2. In
the basic ANN model implementation (Table 6.1), the rainfall intensity forecast has less than 0.1 correlations coefficient value (R) for more than 2 hours prediction. The correlation coefficient result is better in the PCA-FCM-ANN model and it shows that the pattern of the rainfall forecasts yield more than 0.6 correlations when FCM clustering is applied to the input vectors. For magnitude of error index, the average values of RMSE for PCA-FCM-ANN model is 1.98 for rain intensity prediction and 0.25 for state forecast, compared to basic ANN implementation the average values of RMSE is 2.45 for rain intensity forecast and 0.23 for state predictions. In contrast, MAE for the ANN basic model yields better results but the correlations are poor. The average value of MAE for the basic ANN is 0.52 for rain intensity and 0.17 for state prediction. The average value of MAE for the PCA-FCM-ANN implementation is 0.85 for rain intensity and 0.18 for state prediction. The PCA-FCM-ANN model produces a better solution for short-term forecast based on the correlation and RMSE results.
<table>
<thead>
<tr>
<th>Model</th>
<th>Transfer Function</th>
<th>Architecture</th>
<th>Training Function</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.374</td>
<td>2.408</td>
<td>0.334</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.542</td>
<td>2.426</td>
<td>0.305</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.608</td>
<td>2.536</td>
<td>0.094</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.598</td>
<td>2.543</td>
<td>0.057</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.090</td>
<td>0.212</td>
<td>0.589</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.164</td>
<td>0.291</td>
<td>0.276</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAIN INTENSITY FORECAST</th>
<th>Transfer Function</th>
<th>Architecture</th>
<th>Training Function</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.412</td>
<td>1.117</td>
<td>0.808</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.812</td>
<td>1.691</td>
<td>0.705</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.971</td>
<td>2.004</td>
<td>0.680</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>1.212</td>
<td>2.238</td>
<td>0.528</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>1.397</td>
<td>2.443</td>
<td>0.491</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>1.555</td>
<td>2.592</td>
<td>0.423</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STATE FORECAST</th>
<th>Transfer Function</th>
<th>Architecture</th>
<th>Training Function</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-10-1</td>
<td>LM</td>
<td>0.077</td>
<td>0.195</td>
<td>0.920</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.133</td>
<td>0.257</td>
<td>0.856</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.176</td>
<td>0.297</td>
<td>0.802</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.215</td>
<td>0.327</td>
<td>0.755</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.254</td>
<td>0.357</td>
<td>0.698</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.272</td>
<td>0.369</td>
<td>0.672</td>
</tr>
</tbody>
</table>

BR=Bayesian Regularization, LM=Levenberg Marquardt, SCG=Scaled Conjugate Gradient, T=Hyperbolic Tangent Sigmoid, L=Linear, S=Sigmoid
(a) Observed vs Forecast Chuping state forecast (1HOUR)

Regression $R=0.9204$, MAE=0.0767, RMSE=0.1947

(b) Observed vs Forecast Chuping state forecast (3HOUR)

Regression $R=0.8021$, MAE=0.1766, RMSE=0.2970
Figure 6.13: Comparison between observed and forecast for (a) one hour, (b) three hours and (c) six hours state forecast using PCA-FCM-ANN for Chuping weather station.
Comparison: Observed and Forecast

Dataset Tested

Regression $R=0.8082$, MAE=0.4116, RMSE=1.1166

(a)

Comparison: Observed and Forecast

Dataset Tested

Regression $R=0.6804$, MAE=0.9712, RMSE=2.0038

(b)
Figure 6.14: Comparison between observed and forecast rain intensity for Chuping, (a) one hour, (b) three hours and (c) six hours.

Figure 6.15: Regression plot for (a) one, (b) three and (c) six hour rain intensity forecast for PCA-FCM-ANN model for Chuping.

6.8.3.2 Results for Alor Setar weather station

The values of MAE, RMSE and correlation coefficient (R) with PCA-FCM-ANN results for Alor Setar are given in Table 6.3. The tables, mainly based on R
values (which are more than 0.6 correlation coefficients) indicate that the PCA-FCM-ANN model produces better forecast results for localised short-term rainfall. The binary prediction in the state forecast achieve 0.7 correlation for 6 hours prediction, 0.8 correlation for 3 hours prediction and 0.9 correlation for 1 hour prediction. High accuracy for state prediction can be achieved because the ANN was trained from multiple input into two possible outputs, either 0 (not rain) and 1 (rain). Figure 6.16, Figure 6.17, Figure 6.18 and Figure 6.19 show a forecast comparison and scatter plot for Alor Setar.

Table 6.3: PCA-FCM-ANN rainfall forecasting model for Alor Setar

<table>
<thead>
<tr>
<th>Model</th>
<th>Transfer Function</th>
<th>Architecture</th>
<th>Training Function</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAIN INTENSITY FORECAST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>1.425</td>
<td>3.655</td>
<td>0.826</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>1.468</td>
<td>3.956</td>
<td>0.799</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,L</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>1.861</td>
<td>4.740</td>
<td>0.661</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>2.110</td>
<td>4.888</td>
<td>0.520</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>2.355</td>
<td>5.674</td>
<td>0.468</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>2.543</td>
<td>5.855</td>
<td>0.438</td>
</tr>
<tr>
<td>STATE FORECAST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HOUR</td>
<td>T,T,T</td>
<td>5-10-1</td>
<td>BR</td>
<td>0.007</td>
<td>0.026</td>
<td>0.995</td>
</tr>
<tr>
<td>2HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.074</td>
<td>0.163</td>
<td>0.921</td>
</tr>
<tr>
<td>3HOUR</td>
<td>T,T,T</td>
<td>5-10-5-1</td>
<td>BR</td>
<td>0.150</td>
<td>0.264</td>
<td>0.846</td>
</tr>
<tr>
<td>4HOUR</td>
<td>T,T,T</td>
<td>5-10-10-1</td>
<td>BR</td>
<td>0.196</td>
<td>0.307</td>
<td>0.784</td>
</tr>
<tr>
<td>5HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.231</td>
<td>0.337</td>
<td>0.732</td>
</tr>
<tr>
<td>6HOUR</td>
<td>T,T,T</td>
<td>5-20-10-1</td>
<td>BR</td>
<td>0.254</td>
<td>0.355</td>
<td>0.695</td>
</tr>
</tbody>
</table>

BR=Bayesian Regularization, LM= Levenberg Marquardt, SCG=Scaled Conjugate Gradient, T=Hyperbolic Tangent Sigmoid, L=Linear, S=Sigmoid
Observed vs Forecast: Alor Setar state forecast

(a)

Regression R=0.9951, MAE=0.0073, RMSE=0.0260

Observed vs Forecast Alor Setar state forecast (3 HOUR)

(b)

Regression R=0.8459, MAE=0.1497, RMSE=0.2639
Figure 6.16: Comparison between observed and forecast for (a) one hour, (b) three hours and (c) six hours state forecast using PCA-FCM-ANN for Alor Setar.

Figure 6.17: Scatter plot of observed and forecast for (a) one hour, (b) three hours and (c) six hours state forecast using PCA-FCM-ANN for Alor Setar.
6.8.3.3 Results for MARDI Bukit Tangga weather station

The values of MAE, RMSE and correlation coefficient (R) with PCA-FCM-ANN results for MARDI Bukit Tangga weather station are given in Table 6.4
6.8.4 Summary of forecasting result

One hour state forecast using the PCA-FCM-ANN model gives 0.92 correlation coefficient values and rain intensity forecast gives 0.80. The rainfall rate forecast for 1 hour using the basic ANN model yielded lower correlation value which is 0.33 but only slightly worse MAE and RMSE for 1 to 4 hours forecast. Results show that increasing the period of rain prediction will reduce the forecast accuracy. Six hours forecast can achieve 0.42 correlation values for rainfall rate forecast and 0.67 correlation value for state forecast using the PCA-FCM-ANN. The state forecast provided better accuracy because ANN training adapts well to the binary output.

MAE and RMSE are used to evaluate the quantitative error in each model (Basic ANN and PCA-FCM-ANN) outcome. Figure 6.20 and Figure 6.21 show the performance of the correlation coefficient between observed and forecasted values for state and rainfall rate forecasting. State forecasting yields better results compared to rainfall rate forecasting. One hour forecast in both state and rainfall intensity yield good results. As the prediction time increases, the forecast performance decreases. The MAE and RMSE values of each of the prediction models are plotted in Figure 6.22 and Figure 6.23, some of the
forecast results using basic ANN forecasting model outperformed the proposed PCA-FCM-ANN model. The average value of MAE for the basic ANN is 0.52 for rain intensity and 0.17 for state prediction. The average value of MAE for the PCA-FCM-ANN implementation is 0.85 for rain intensity and 0.18 for state prediction. In performance validation, the forecast results of correlation values and RMSE for the PCA-FCM-ANN is outperforming the basic ANN model.

Figure 6.20: Performance comparison of forecasting model for state prediction

Figure 6.21: Performance comparison of forecasting model for rainfall rate
Figure 6.22: MAE performance comparison of forecasting model for state prediction

Figure 6.23: RMSE performance comparison of forecasting model for state prediction

6.9 Summary

Rainfall is traditionally a very difficult weather phenomenon to predict even using a large datasets spanning many decades [66]. In tropical area, rain dynamics are more challenging to predict because of their localised characteristic. The study area in this paper experiences high rain intensity over short durations that are convective in nature. The highly dynamic and non-
linear nature of this weather condition is a reflection of the complex interaction between the various processes at different scales. Therefore, the design of an effective system for rain prediction is difficult to realise. This study used ANN to develop a rainfall forecasting model. The available data to train the network for high intensity rain patterns is 5.4 % of all rain events which compounds the challenge to develop an accurate model. The best forecast model is one hour forecast for state prediction using PCA-FCM-ANN that yields 0.92 correlation. Using FCM clustering to classify the ANN input parameters produces increase in accuracy in the ANN training. Comparative analysis between the basic ANN and the PCA-FCM-ANN implementation indicates that the performance level of PCA-FCM-ANN is better than that of a basic ANN model based on the correlation and RMSE results.
Chapter 7

Conclusions

The overall aim of the research described in this thesis was to develop a reliable system for the purpose of forecasting localised rain event occurrences and rain intensity with realistic lead time for tropical climates. The proposed rainfall forecasting system can be used as part of a water resource management system as well as flood prediction system. One of the key challenges of this research has been the limited dataset. Hourly meteorological data spanning of three years period was used in the study. Care needed to be exercised to avoid overfitting and underfitting problems in network training processes. Data imputation was used as a pre-processing approach in order to ensure the reliability of the dataset. An integration of soft clustering with ANN was developed and used. Along with the rainfall forecasting, a system to forecast weather parameters such as temperature, atmospheric pressure, dew point, humidity and wind speed was also developed.

In order to achieve the research aims, the objectives of the research undertaken had to be successfully fulfilled. The five objectives were:
• Obtain and study weather data such as temperature, atmospheric pressure, dew point, humidity, wind speed, rain event occurrences, rainfall amount and rainfall rate from North-West of Malaysia.

• Investigate ANN and Fuzzy Logic techniques for possible application in forecast models.

• Develop environmental parameter prediction model (temperature, atmospheric pressure, dew point, humidity and wind speed) using ANN and Fuzzy Logic at hourly interval.

• Develop rainfall forecast model for state prediction and rainfall rate prediction model.

• Evaluate and validate the developed model(s).

The overview of study area and the imputation method used to create a reliable dataset was described in Chapter 2. The reliability of data imputation technique to prepare input data for rainfall forecasting system has been proven to be accurate using KDF results. Chapter 4 described the ANN model implemented to predict environmental parameters for one, three and six hours. Atmospheric pressure, temperature, dew point, humidity and wind speed data for each weather station was used. Different algorithms such as LM, BR and SCG algorithms were used in the ANN forecasting model. The performance of each algorithm was assessed based on the magnitude of the error and the correlation coefficient value between observed and forecasted values. Chapter 5 presented the basic ANN model implemented for rainfall forecasting. The poor forecast from the basic ANN imply that the ANN fed with original meteorological data is not viable in the present case. Chapter 6 presented the main findings of this study where PCA-FCM-ANN forecasting model was proposed. This model takes advantage of pre-processing method and soft clustering to generalise the system. Performance comparisons have been made between basic ANN and
PCA-FCM-ANN system forecasted results. All materials presented from Chapter 2 to Chapter 6 are aligned with the five objectives presented in Chapter 1 of this thesis.

PCA-FCM-ANN forecasting model offers a better convergence. In order to achieve optimal forecasting, pre-processing coupled with soft computing approach shows better forecasting results compared to the basic ANN model. Imputation method using PCA-ALS and input-output mapping between environmental parameters with FCM clustering also improves network training and forecasting results. One hour PCA-FCM-ANN prediction model for rain or not rain occurrences gives 0.92 correlation value and rainfall rate prediction gives 0.80. Rainfall rate forecast using ANN model without applying FCM produced lower correlation value which is the rainfall intensity forecast has less than 0.1 correlation value for more than 2 hours prediction. The results also show that increasing the forecast time will reduce the convergence result. Six hour prediction can reach 0.67 correlations and 0.42 correlations using PCA-FCM-ANN for state and rain intensity forecast respectively. State forecast gained higher correlations value as a result of ANN learning process simply adapts to binary prediction of rain or no rain events. The proposed model for rainfall forecasting is, therefore, ANN integrated with FCM clustering and PCA-ALS pre-processing method. However, there is still considerable area that can be further developed to improve the prediction.

7.1 Research Contribution

The major contribution of this research is integrating PCA-ALS pre-processing method and FCM soft clustering into ANN forecasting model. This approach has been assessed for point prediction of rain state and intensity forecast. For time between one to six hours forecast interval, results shows that
PCA-FCM-ANN forecasting models yields good results.

Another contribution is the used of fuzzy clustering and neural network which minimised the drawback of having a small dataset, especially for rainfall intensity prediction. Even though the dataset was small, the integration of fuzzy clustering ensured that network training did not suffer from overfitting or underfitting problem.

Part of the contribution in this study is the cost effective localised rain prediction. Affordable environmental sensor devices can be used to measure atmospheric pressure, temperature, dew point and humidity and used with the developed system to provide localised weather prediction. This simple weather station can also record data and transmit to a receiving device where they can be used in forecasting system to provide high resolution weather prediction. In agriculture, especially in tropical area where the weather condition is normally uncertain, deploying this type of micro weather station could be used to schedule crop irrigation systems or water resource management in general.

### 7.2 Research Limitation

Although the study has achieved the objectives that have been outlined, there were some limitations or constraints in this study. In general, limitations in this study are divided into two main area; the limitation of data availability and methodology. This study used ANN to develop a rainfall forecasting model. The available data to train the network for high intensity rain patterns is low which compounds the challenge to develop an accurate model to predict high rain intensity. The poor time resolution of the weather stations meant that the transition from no rain to rain event was not always captured and hence difficult to predict or take account of in the model. This creates uncertainty, for examples in state forecast with binary prediction where the system may be
unable correctly predict the right state in the transition period.

Rainfall prediction system for one to six hours gives satisfactory results despite the variation in the conditions between different weather stations. Essentially, based on correlation results of greater than 0.6, it is indicate that the PCA-FCM-ANN model produces better forecast results for localised short-term rainfall. In average, the binary prediction in the state forecast achieve 0.7 correlation for 5 to 6 hours prediction, 0.8 correlation for 3 hours prediction and 0.9 correlation for 1 hour prediction. As the prediction interval increases, the forecast performances decrease rapidly. The selection of the number of neurons in the hidden layers is a very important part of the overall neural network architecture. There are many methods that could be used to determine the correct number of neurons in the hidden layers but there is no specific method to identify the optimal number of neurons and the number of hidden layers. There were similar problems with the selection of a suitable transfer function and learning algorithm. Since neural network architecture is comprised of many attributes, choosing the right combination of the number of neurons, number of hidden layers, transfer function and learning algorithms is a challenging problem to solve and the solution may not be optimal in all cases.

Initially, pre-processing technique was employed prior to clustering the data to reduce noise. FCM algorithm limitation is its sensitivity to noisy data. In FCM clustering a point will have partial membership of all the clusters. The limitation is that due to the influence of partial membership of all the data members, the cluster centre tend to move towards the centre of all the data points and these sometimes give unrealistic results.

7.3 Future Work

There are many possible ways that further research can be carried out to...
advance the work reported in this thesis. The future works are outlined below:

- **Training algorithm**

  Effective training algorithm capability in neural network architecture is considered necessary in future work. As mentioned earlier, BR, LM and SCG learning algorithms are some of the approaches by the researchers in training the network for better adaptability of ANN. In literature, the relative performance of the algorithms is greatly influenced by the particular problems. An investigation of other algorithms that may be suitable and stable for any problem might be another future direction.

- **High intensity rainfall forecast**

  As indicated by the analysis, the data time resolution is not sufficient for high intensity rain. Therefore, new methods should be explored to improve the forecast of high intensity rain.

- **Shorter interval forecast**

  The proposed technique should also be extended to forecast smaller time window such as ten minutes forecast interval.

- **Output fusion**

  In this study, multiple inputs with single output form of ANN was implemented. Output fusion works by combining or aggregating decisions from multiple targets. Therefore, multiple inputs with multiple output forms of ANN could be adapted for future work.
References


[112] F. Giannini, V. Laveglia, A. Rossi, D. Zanca, and A. Zugarini, “Neural


Appendix A. Publications


Appendix B. Monthly Rainfall Rate Average

Appendix B-1. Chuping Station

Appendix B-2. Alor Setar Station

Appendix B-3. MARDI Bukit Tangga Station
## Appendix C. Missing Data.

### Appendix C-1. Missing and faulty data for all station

<table>
<thead>
<tr>
<th>STATION: Alor Setar</th>
<th>PERIOD</th>
<th>NOTE : DATA NOT AVAILABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Hourly MSL Pressure</td>
<td>2012 - 2014</td>
<td></td>
</tr>
<tr>
<td>b. Hourly Dry Bulb Temperature</td>
<td>2012 - 2014</td>
<td></td>
</tr>
<tr>
<td>c. Hourly Dew Point</td>
<td>2012 - 2014</td>
<td>2012: Feb(3,13)</td>
</tr>
<tr>
<td>d. Hourly Relative Humidity</td>
<td>2012 - 2014</td>
<td>2012: Feb(3,13)</td>
</tr>
<tr>
<td>e. Hourly Mean Surface</td>
<td>2012 - 2014</td>
<td>2013: Oct(7,23) Apr(5)</td>
</tr>
<tr>
<td>Wind Direction and Speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and Amount</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STATION: Chuping</th>
<th>PERIOD</th>
<th>NOTE : DATA NOT AVAILABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind Direction and Speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and Amount</td>
<td>2014: Oct(8) Nov(22,23) Dec(5)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STATION: MARDI</th>
<th>PERIOD</th>
<th>NOTE : DATA NOT AVAILABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Oct(1-3) Nov(9-30) Dec(1-11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013: May(18,20) Jun(18) Sep(15) Oct(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2014: May(23-31) Jun(1-24) Nov(1-3,11,12,29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oct(1-3) Nov(9-16,20-30) Dec(1-11)</td>
</tr>
</tbody>
</table>
c. Hourly Dew Point 2012 - 2014

2013: May(20) Sep(15) Oct(2)
2014: May(23-31) Jun(1-24) Nov(1-3,11,12,29)

d. Hourly Relative Humidity 2012 - 2014

2013: May(20) Sep(15) Oct(2)
2014: May(23-31) Jun(1-24) Nov(1-3,11,12,29)

e. Hourly Mean Surface Wind Direction and Speed 2013 - 2014

2013: May(20) Sep(15) Oct(2)

f. Hourly Rainfall Duration and Amount 2012 - 2014

2013: Jan(3) May(20) Sep(15) Oct(2)
Appendix D. Weather Station Instruments

Appendix D-1. Automatic Weather Stations (AWS) managed by Malaysia Meteorology Department

The AWS measures precipitation (amount of rainfall), atmospheric pressure, temperature, humidity, wind speed and direction and global solar radiation, updating the data every minutes, 24 hours a day without human intervention.
### Appendix D-2. Hardware or instruments used by Malaysia Meteorology Department

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tipping Bucket Rain Gauge</strong></td>
<td>A tipping bucket rain gauge has a receiving funnel leading to two small metal collectors (buckets). When a bucket accumulates 0.2 mm of rain water, the weight of the water causes it to tip and empty itself. Each time a bucket tips, an electrical contact is made, thereby enabling recording or rainfall amount and intensity with time. The maximum detectable rainfall rate is 200 mm/hr.</td>
</tr>
<tr>
<td><img src="image1" alt="Tipping Bucket Rain Gauge" /></td>
<td></td>
</tr>
<tr>
<td><strong>Atmospheric Pressure</strong></td>
<td>The pressure sensor is a pressure capsule or a solid state capacitive device which outputs voltage which is converted into digitally encoded values of atmospheric pressure.</td>
</tr>
<tr>
<td><img src="image2" alt="Atmospheric Pressure" /></td>
<td></td>
</tr>
</tbody>
</table>
| **Temperature Sensor**             | The temperature sensing system uses integrated circuit technology in combination with an accurate resistance thermometer element to allow reliable measurements.  

The dry and wet bulb thermometers are placed vertically on a support inside the Stevenson screen. The bulb of the wet bulb thermometer is wrapped with muslin and is tied up with a wick. The wick is then dipped inside a container which contains distilled water. |
| ![Temperature Sensor](image3)     |                                                                                                                                               |
| **Solarimeter / Pyranometer**      | The solarimeter measures routine global solar radiation on a plane or level surface. It has a thermocouple junction-sensing element. The sensing element is coated with a highly stable carbon based non organic coating, which delivers excellent spectral absorption and long-term stability characteristics. The sensing element is housed under two concentric fitting glass domes. |
| ![Solarimeter / Pyranometer](image4) |                                                                                                                                               |
| **Wind Speed and Direction Sensor**| Wind direction is the direction from which the wind is blowing. It is expressed in degrees measured clockwise from geographical north. Wind vanes do not respond to changes in wind direction when the wind speed is less than one metre per second or two knots.  

Wind speed is measured in metres per second or knots. Calm is reported when the wind speed is less than 0.5 metres per second or less than one knot. Instruments used for measuring the surface wind speed are called anemometers, the most common of which is the cups mounted symmetrically at right angle to a vertical shaft. The difference in wind pressure from one side of the cup to the other causes the cup to spin about the shaft. The rate at which they rotate is directly proportional to the speed of wind. |
| ![Wind Speed and Direction Sensor](image5) |                                                                                                                                               |
# Appendix E. Certificate of Ethics Review

## Certificate of Ethics Review

<table>
<thead>
<tr>
<th>Project Title:</th>
<th>Application of Soft Computing and Sensor Fusion for environmental monitoring, prediction and control</th>
</tr>
</thead>
<tbody>
<tr>
<td>User ID:</td>
<td>680382</td>
</tr>
<tr>
<td>Name:</td>
<td>Noor Zuraidin Bin Mohd Safar</td>
</tr>
<tr>
<td>Application Date:</td>
<td>29/01/2015 08:39:25</td>
</tr>
</tbody>
</table>

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative for the School of Engineering is Giles Tewkesbury

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

**SchoolOrDepartment:** ENG  
**PrimaryRole:** PostgraduateStudent  
**SupervisorName:** Dr. David Ndzi  
**HumanParticipants:** No  
**PhysicalEcologicalDamage:** No  
**HistoricalOrCulturalDamage:** No  
**HarmToAnimal:** No  
**HarmfulToThirdParties:** No  
**OutputsPotentiallyAdaptedAndMisused:** No  
**Confirmation-ConsideredDataUse:** Confirmed  
**Confirmation-ConsideredImpactAndMitigationOfPotentialMisuse:** Confirmed  
**Confirmation-ActingEthicallyAndHonestly:** Confirmed

**Certificate Code:** 202F-3689-8C68-0D76-7119-98C2-6129-D1BF  
Page 1
Appendix F. Research Certificate Review Checklist

FORM UPR16
Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the
Postgraduate Research Student Handbook for more information)

Postgraduate Research Student (PGRS) Information

Student ID: UP680382

PGRS Name: NOOR ZURAIDIN BIN MOHD SAFAR

Department: Engineering

First Supervisor: DR. DAVID NDZI

Start Date: 25 March 2013

(or progression date for Prof Doc students)

Study Mode and Route:

Part-time □ MPhil □ MD □

Full-time ☑ PhD ☑ Professional Doctorate □

Title of Thesis: Integration of Principal Component Analysis, Fuzzy C-Means and Artificial Neural Networks for Localised Environmental Modelling of Tropical Climate

Thesis Word Count: 23985

(excluding ancillary data)

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University’s Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study.

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?

b) Have all contributions to knowledge been acknowledged?

c) Have you complied with all agreements relating to intellectual property, publication and authorship?

d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?

e) Does your research comply with all legal, ethical, and contractual requirements?

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):

202F-3889-8C69-0D76-7119-98C2-6129-D1BF

If you have not submitted your work for ethical review, and/or you have answered ‘No’ to one or more of questions a) to e), please explain below why this is so:

UPR16 – August 2015
<table>
<thead>
<tr>
<th>Signed (PGRS):</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>06/11/2017</td>
</tr>
</tbody>
</table>