

**DEVELOPMENT OF A COMBINED CLASSIFICATION DECISION
TREE AND ARTIFICIAL NEURAL NETWORK RAINFALL
GENERATOR FOR LANGAT RIVER BASIN**

By

LIAN CHAU YUAN

A dissertation submitted to the Department of Civil Engineering,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Master of Engineering Science.
January 2018

ABSTRACT

DEVELOPMENT OF A COMBINED CLASSIFICATION DECISION TREE AND ARTIFICIAL NEURAL NETWORK RAINFALL GENERATOR FOR LANGAT RIVER BASIN

Lian Chau Yuan

Climate change is the most concerned global issue as its consequences bring the significant impacts toward the environment and livelihood in the world. Thus, the development of statistical downscaling models has become crucial for climate change impact studies due to their ability in downscaling the future climate data. This study aimed to select a suitable two-stage rainfall generator for the Langat River Basin based on its performance in simulating the observed rainfall series for the period of 1976–2005. Therefore, the first challenge of this study is to model the observed rainfall occurrence using generalized linear model (GLM), non-homogeneous hidden Markov models (NHMM) and bootstrap aggregated classification tree (BACT) model. Overall, the BACT model exhibited good prediction ability in simulating the rainfall persistence and spell lengths distribution. Besides, the BACT model outperformed the GLM and NHMM with its higher probability of detection and critical success index in the range of 0.51–0.65 and 0.29–0.44, respectively. The BACT model also exhibited reasonably good prediction with indices of Pierce skill score and Heidke skill score greater than 0.15 at every station. Hence, the BACT model was selected as the suitable rainfall occurrence model and combined with artificial neural network (ANN) to form a rainfall generator. The second challenge is the

adoption of data pre-processing approach to improve the performance of traditional ANN model. The combined BACT-ANN model not only produced the root mean square errors smaller than ANN and NHMM, but also achieved the acceptability index greater than 91% in passing most of the Kolmogorov-Smirnov, Mann-Whitney U and squared-rank tests. Furthermore, they are capable of producing the monthly rainfall series significantly correlated to the observed series, resulting in 100% acceptability index for both Kendall's tau-b and Spearman's rho correlations. In conclusion, the combined BACT-ANN model is the recommended rainfall generator for the Langat River Basin.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Ir. Dr. Huang Yuk Feng, who has a great knowledge and expert in hydrology field. His precious advice, guidance and patience help me a lot throughout this study. He always motivated me when I was facing the problems in my study.

Besides, I would like to thank my co-supervisor, Dr. Goh Yong Kheng, for providing his useful suggestion and great knowledge in mathematical field. In addition, a special thanks to Dr. Ling Lloyd, for his inspiring comments to help me understand and solve the problems in my study.

At last, I would like to thank my family and friends for supporting and encouraging me when I was facing any problem or feeling depressed during my master study.

APPROVAL SHEET

This dissertation entitled **“DEVELOPMENT OF A COMBINED CLASSIFICATION DECISION TREE AND ARTIFICIAL NEURAL NETWORK RAINFALL GENERATOR FOR LANGAT RIVER BASIN”** was prepared by LIAN CHAU YUAN and submitted as partial fulfillment of the requirements for the degree of Master of Engineering Science at Universiti Tunku Abdul Rahman.

Approved by:

(Ir. Dr. Huang Yuk Feng)
Associate Professor/Supervisor
Department of Civil Engineering
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

Date: 24/01/2018

(Dr. Goh Yong Kheng)
Assistant Professor/Co-supervisor
Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

Date: 24/01/2018

LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

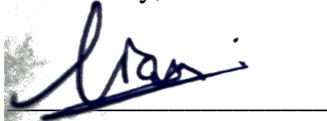
Date: 24/01/2018

SUBMISSION OF DISSERTATION

It is hereby certified that *Lian Chau Yuan* (ID No: *15UEM06657*) has completed this dissertation entitled “*DEVELOPMENT OF A COMBINED CLASSIFICATION DECISION TREE AND ARTIFICIAL NEURAL NETWORK RAINFALL GENERATOR FOR LANGAT RIVER BASIN*” under the supervision of Ir. Dr. Huang Yuk Feng (Supervisor) from the Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, and Dr. Goh Yong Kheng (Co-Supervisor) from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science.

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

Yours truly,



(Lian Chau Yuan)

DECLARATION

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



(LIAN CHAU YUAN)

Date 24/01/2018

TABLE OF CONTENTS

	Page
ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
APPROVAL SHEET	v
SUBMISSION SHEET	vi
DECLARATION	vii
LIST OF TABLES	xi
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xvi
CHAPTER	
1.0 INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Aim and Objectives	5
1.4 Significance of Study	6
1.5 Scope of Work	7
2.0 LITERATURE REVIEW	8
2.1 Climate Change	8
2.2 Homogeneity Tests	18
2.3 Downscaling Approaches	22
2.4 Statistical Downscaling	25
2.4.1 Regression-based Techniques	26
2.4.2 Weather Pattern Classification	34
2.4.3 Weather Generator	37
2.5 Classification and Regression Trees	39
2.6 Hybrid Models	44
2.7 Summary	47
3.0 METHODOLOGY	49
3.1 Study Area	49
3.2 Workflow of Study	52
3.3 Data Collection	53
3.3.1 Observed Rainfall Series	53
3.3.2 NCEP & NCAR Reanalysis Dataset	55
3.4 Normality Tests	56
3.4.1 The Anderson-Darling Test	57
3.4.2 The Lilliefors Test	58
3.4.3 The Jarque-Bera Test	59
3.5 Homogeneity Tests	59
3.5.1 The Standard Normal Homogeneity Test (SNHT)	61

3.5.2	The Buishand Range Test (BR)	62
3.5.3	The Pettitt Test (PET)	63
3.5.4	The Von Neumann Ratio Test (VNR)	64
3.6	Screening of Predictors	65
3.7	Rainfall Occurrence Models	66
3.7.1	Generalized Linear Model (GLM)	66
3.7.2	Bootstrap Aggregated Classification Tree Model (BACT)	68
3.7.3	Non-Homogeneous Hidden Markov Model (NHMM)	72
3.7.4	Goodness of Fit	75
3.7.4.1	Rainfall Persistence	75
3.7.4.2	Spell Lengths Distributions	76
3.7.4.3	Matching	76
3.8	Rainfall Amount Models	78
3.8.1	Artificial Neural Network (ANN)	78
3.8.2	Combined BACT-ANN model	83
3.8.3	Non-Homogeneous Hidden Markov Model (NHMM)	85
3.8.4	Goodness of Fit	89
3.8.4.1	Root Mean Square Error	89
3.8.4.2	Kolmogorov-Smirnov Test	90
3.8.4.3	Mann-Whitney U Test	91
3.8.4.4	Squared-Rank Test	92
3.8.4.5	Kendall's Tau-b Correlation	92
3.8.4.6	Spearman's Rho Correlation	93
3.8.4.7	Acceptability Index	94
4.0	RESULTS AND DICUSSION	95
4.1	Normality Tests	95
4.2	Homogeneity Tests	97
4.3	Predictor Selection	98
4.4	Out-of-Bag Classification Error for BACT model	101
4.5	Determination of Optimum Hidden States and Rainfall Distribution for NHMM	104
4.6	Goodness of Fit of Rainfall Occurrence Models	110
4.6.1	Rainfall Persistence	110
4.6.2	Spell Length Distribution	112
4.6.3	Matching	117
4.7	Data Pre-Processing Approach in Combined BACT-ANN Model	121
4.8	Structure of Neural Network for ANN and Combined BACT-ANN Model	121
4.9	Goodness of Fit of Rainfall Amount Models	123
4.9.1	Parametric Test	123
4.9.2	Non-parametric Tests	126
4.9.3	Quantile Plot	139

5.0	CONCLUSIONS	143
5.1	Summary	143
5.1.1	Rainfall Occurrence Model	144
5.1.2	Rainfall Amount Model	147
5.2	Recommendations	149
	REFERENCES	150
	APPENDICES	159
	APPENDIX A	159
	APPENDIX B	161
	APPENDIX C	162
	APPENDIX D	163
	APPENDIX E	165
	APPENDIX F	167
	APPENDIX G	169

LIST OF TABLES

Table		Page
2.1	Advantages and disadvantages of dynamical downscaling technique (Fowler et al., 2007)	24
2.2	Advantages and disadvantages of statistical downscaling technique (Fowler et al., 2007)	25
3.1	List of station code, name, coordinate and study period	54
3.2	Description of predictor variables from NCEP & NCAR reanalysis dataset	56
3.3	Classification of homogeneity tests' results	61
4.1	Normality tests results of observed daily rainfall series at each station	95
4.2	Statistical characteristics of observed daily rainfall series at each station	96
4.3	Homogeneity tests results of monthly rainfall series at each station	98
4.4	Description and lag-transformation of selected predictors at each station	100
4.5	Partial correlation coefficient and p-value of selected predictors at each station	100-101
4.6	Log-likelihood values of NHMM as a function of different hidden states number at each station	105
4.7	Bayesian information criterion (BIC) scores of NHMM as a function of different hidden states number at each station	106
4.8	Akaike information criterion (AIC) scores of NHMM as a function of different hidden states number at each station	107
4.9	Log-likelihood values of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station	108

4.10	Bayesian information criterion (BIC) score of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station	109
4.11	Akaike information criterion (AIC) score of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station	109
4.12	Performance of rainfall occurrence models in terms of their rainfall persistence and absolute difference with observed rainfall persistence during calibration (1976–1995) and validation periods (1996–2005)	111
4.13	Probability of detection (POD) and critical success index (CSI) of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods	118
4.14	False alarm rate of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods	119
4.15	Heidke skill scores (HSS) and Pierce skill scores (PSS) of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods	120
4.16	Percentage of observed wet days become dry days after data pre-processing approach	121
4.17	Structure of combined BACT-ANN and traditional ANN models	122
4.18	Root mean square error (RMSE) of rainfall amount models during calibration (1976–1995) and validation (1996–2005) periods	125-126
4.19	Kolmogorov-Smirnov (K-S) tests results of rainfall amount models at each station	127-128
4.20	Mann-Whitney U tests results of rainfall amount models at each station	129-130
4.21	Squared-rank tests results of rainfall amount models at each station	131-132

4.22	Acceptability index for the simulated monthly rainfall series in passing the Kolmogorov-Smirnov (K-S), Mann-Whitney U and squared-rank tests for all stations	133
4.23	Kendall's tau-b correlation coefficients of rainfall amount models at each station	135-136
4.24	Spearman's rho correlation coefficients of rainfall amount models at each station	137-138
4.25	Acceptability index for the simulated monthly rainfall series obtaining the significant coefficient in Kendall's tau-b and Spearman's rho correlations for all stations	139
4.26	Bias calculation for the over-prediction and under-prediction of monthly rainfall series at each station	142

LIST OF FIGURES

Figures	Page
2.1 Globally averaged greenhouse gases concentration (IPCC, 2014)	10
2.2 Globally anthropogenic CO ₂ emissions (IPCC, 2014)	10
2.3 Globally average combined land and ocean surface temperature anomaly (IPCC, 2014)	12
2.4 Typical architecture of ANN (Abdulkadir et al., 2012)	32
3.1 Langat River Basin in Peninsular Malaysia	51
3.2 Flowchart of work	52
3.3 Location of selected rainfall stations at the Langat River Basin, Malaysia	54
3.4 Structural layout of a single classification decision tree	70
3.5 Flow chart of rainfall occurrence models	74
3.6 Flow chart of ANN	82
3.7 Flow chart of combined BACT-ANN model	84
3.8 Flow chart of NHMM	88
4.1 The out-of-bag classification error produced by bootstrap aggregated classification tree models (BACT) as a function of the number of grown trees at (a) station 2815001, (b) station 2913001, (c) station 2917001, and (d) station 3118102	102-103
4.2 Log-likelihood values of NHMM as a function of different hidden states number at each station	105
4.3 Bayesian information criterion (BIC) scores of NHMM as a function of different hidden states number at each station	106

4.4	Akaike information criterion (AIC) scores of NHMM as a function of different hidden states number at each station	107
4.5	Observed and simulated spell lengths distribution at each station. (a) distribution of wet-spell length and (b) dry-spell length at station 2815001, (c) distribution of wet-spell length and (d) dry-spell length at station 2913001, (e) distribution of wet-spell length and (f) dry-spell length at station 2917001, (g) distribution of wet-spell length and (h) dry-spell length at station 3118102	113-116
4.6	Scatter plots of simulated (ANN, NHMM and combined BACT-ANN model) versus observed 95th percentile of monthly rainfall series at (a) station 2815001, (b) station 2913001, (c) station 2917001 and (d) station 3118102	140-142

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AR4	Fourth Assessment Report
AR5	Fifth Assessment Report
BACT	Bootstrap Aggregated Classification Tree
BIC	Bayesian Information Criterion
BR	Buishand Test
CART	Classification and Regression Trees
CCSM3.0	Third Generation of Community Climate System Model
CGCM3	Third Generation Coupled Global Climate Model
CIMP5	Coupled Model Intercomparison Project-phase 5
CLIGEN	Climate Generator
CSI	Critical Success Index
DID	Department of Irrigation and Drainage Malaysia
ECHAM5	Fifth Generation of European Centre Hamburg Model
FAR	False Alarm Rate
GCM	Global Circulation Model/ Global Climate Model
GLM	Generalized Linear Model
HadCM3	Third Generation of Hadley Centre Coupled Model
HMM	Hidden Markov Model
HSS	Heidke Skill Score
IPCC	Intergovernmental Panel on Climate Change
k-NN	k-Nearest Neighbors
LARS-WG	Long Asthon Research Station-Weather Generator
LMR	Linear Multiple Regression
LS-SVM	Least Squares Support Vector Machines
MPI-ESM	Max Planck Institute-Earth System Model
MIROC-ESM	Model for Interdisciplinary Research on Climate-Earth System Model
MLP	Multilayer Perceptron
MMD	Malaysian Meteorological Department

MRI	Meteorological Research Institute
MSDM	Multi-site Statistical Downscaling Model
MVNHMM	Multivariate Non-Homogeneous Hidden Markov Model
NCAR	National Centers for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NHMM	Non-Homogeneous Hidden Markov Model
OOB	Out-of-Bag
PET	Pettitt Test
PGSL	Probabilistic Global Search Algorithm
POD	Probability of Detection
PSS	Pierce Skill Score
<i>purelin</i>	Linear Transfer Function
RBF	Radial Basis Function
RCM	Regional Climate Model
RCP	Representative Concentration Pathway
RMSE	Root Mean Square Error
RF	Random Forest
SNHT	Standard Normal Homogeneity Test
SDSM	Statistical Downscaling Model
SVM	Support Vector Machine
<i>tan-sig</i>	Tan-Sigmoid transfer function
<i>traingd</i>	Gradient Decent training algorithm
<i>trainlm</i>	Levenberg-Marquardt backpropagation training algorithm
<i>trainscg</i>	Scaled Conjugate Gradient training algorithm
UNFCCC	United Nations Framework Convention on Climate Change
VNR	Von Neumann Ratio Test
WeaGETS	Weather Generator of Ecole de Technologies Superieure
WGs	Weather Generators

CHAPTER 1

INTRODUCTION

1.1 Background

Langat River Basin is one of the most urbanized river basins in Malaysia with an approximately 2,200 km² basin area, which cover the southern region of Selangor state. Langat River originates from the Titiwangsa Range and flows in a south-western direction into the Malacca Strait. Langat River Basin is the major source of raw water and other facilities to approximately 1.2 million people, which cover the important metropolises, such as Bangi, Cheras, Kajang, Putrajaya and Sepang. Furthermore, there are two main reservoirs within the basin and eight water treatment plants to serve the clean and safe water for the populations.

In recent decades, the surrounding areas of the basin are undergoing a high degree of urbanization; especially the land-use along the river basin has been converted from agricultural to industrial (Memarian et al., 2014; Saudi et al., 2015). The rapid urban development and population growth definitely increase the rate of surface runoff, which resulting in the rivers become shallow and triggering the occurrence of floods. This may lead to the destruction of physical structures and affecting the lives of residents, more so when more than

two thirds of the Selangor population reside in floodplain area (Juahir et al., 2011).

Global warming is one of the consequences of climate change, which has been and still is the critical issue for the past two decades, as it brings great impacts to the environment and livelihood in the world. Climate change is strongly believed to have a strong connection with human activities. As the technologies are getting more advanced over the years, the energy demand also increases with the population growth. This situation resulted in the rising combustion of fossil fuel and a continually rising level of carbon dioxide trapped in the atmosphere, thus the global warming condition is getting more severe over time. There is a high probability for the global surface temperature to be increased for more than 1.5 °C at the end of the 21st century, if the condition of global warming still remain the same or is getting more severe over the time.

The main reason for the development of statistical downscaling models is due to the rising interest in future weather simulation on the climate change studies. According to the Fifth Assessment Report (AR5), which is the latest report of the Intergovernmental Panel on Climate Change (IPCC), there are four emission scenarios provided through experiment protocol of CIMP5 according to the representative concentration pathways (RCPs), namely RCP 2.6, RCP 4.5, RCP 6 and RCP 8.5. All of them were named according to the consideration of different greenhouse gases concentration in every emission scenario. The ultimate aim of a statistical downscaling model is to establish a statistical relationship between the large-scale atmospheric variables and local climate

variables, then simulate the future local climate variables under different emission scenarios.

Rainfall contributes a significant role in the hydrological data for generating a weather generator to predict and analyse the future climate change. Malaysia is a tropical rainforest country and it experiences a type of tropical climate with no dry season. Malaysia has no winter or summer, and is typically hot or humid throughout the year. Hence, the simulation of future rainfall using a statistical downscaling model is relatively important for future needs, especially where the study areas located at those flood prone areas, are subjected to heavy rainfall and used as storage of reservoir for supplying the raw water for domestic usage or for recreational purposes.

1.2 Problem Statement

The demand and pressure on water resources are increasing tremendously over the years due to the population growth, urbanization, industrialization and the intensive expansion of agriculture. Since 1980s, the economics of Malaysia has been developed extensively, which resulted in the high water demands in commercial and developed areas. Being a foremost industrial state, Selangor is expected to have rapid population growth in 21st century, and one major issue is that the current demand of raw water has exceeded the availability of water. The Langat River Basin has been supplying

the raw water to approximate half of the population of Selangor for the past 40 years (Memarian, et al., 2012).

Based on the historical data, Malaysia experienced some cases of flood, which brought severe damages to lives and properties, especially the recent flood cases in Kelantan, Pahang and Terengganu during years 2014 and 2015. There was a huge number of victims and properties losses involved in that incident. Those impacts directly influenced the economic growth of country due to the high recovery cost after the occurrence of extreme events. Hence, the simulation and analysis of future rainfall using an efficient rainfall generator is a good precautionary measure, which can reduce the consequences brought by the unpredictable huge amount of rainfall intensity, especially at the flood prone areas if the prediction can be made on time and to allow sufficient time for disaster response efforts. The Langat River Basin is one of the flood prone areas in Malaysia that receives a huge amount and high frequency of rainfall during the monsoon periods. Other than water supply, the two main reservoirs within the basin also play an important role in mitigating the floods as the risk of flood is classified as high level in this basin.

There are numerous weather generators that have been developed and implemented for the analysis of future hydrological events in the climatology studies, but the majority of them focused on the countries with four seasons. However, an effective and reliable rainfall generator for the tropical climate countries still needs to be explored and developed based on the local tropical climate for effective simulation of future rainfall. In recent studies, the

inefficient performance of traditional individual rainfall generators have become a concerned issue to be overcome by improving their limitations. The two-stage approach, which combine two models to form a hybrid rainfall generator, has become an effective solution for this issue lately. However, the selection of a suitable model to be combined with the traditional model has become another challenge for the researchers.

1.3 Aim and Objectives

The ultimate aim of this research is to develop a combined bootstrap aggregated classification tree and artificial neural network rainfall generator for the Langat River Basin. Several specific objectives are:

- (i) To develop, evaluate and compare the rainfall occurrence models, namely generalized linear model, non-homogenous hidden Markov model and bootstrap aggregated classification tree;
- (ii) To develop and generate the rainfall series using an artificial intelligence feed forward back propagation network rainfall amount model; and
- (iii) To enhance the performance of rainfall amount model using data pre-processing approach.

1.4 Significance of Study

The developed combined rainfall generator in this study can be proceeded further to downscale the future rainfall occurrence and amount series under different emission scenarios, using the output from global circulation models (GCMs). These downscaled data are useful for hydrologists, engineers and researchers when designing or improving the hydrological structures. Besides, the high accuracy of generated rainfall series from the developed rainfall generator is relatively important for an effective planning and management of water resources, to ensure the water is supplied sufficiently for the usage in domestic, industrial and agricultural sectors. Therefore, the developed combined rainfall generator can be used for further study on the changes of future rainfall. This can help to establish a benchmark for making the better policy decision with regards to water projects in the Langat River Basin.

As part of the much needed comprehension on the future climate change, the future rainfall data produced by a reliable rainfall generator is important for the users to continue their further investigation. Besides, this study can provide a new finding in the development of rainfall generator for the tropical climate regions since the development of statistical downscaling models under the tropical climate is still lacking. The selected rainfall occurrence model in this study can also be an alternative model to determine its suitability for combining with the traditional individual model in other studies. Besides, the development of combined rainfall generator in this study may provide some ideas for the

future researches to combine two similar or different approaches as a whole model to improve the limitations and performance of a traditional individual model.

1.5 Scope of Work

This study only focus on four selected rainfall stations within the Langat River Basin, with the study period from years 1976 to 2005. The potential atmospheric variables are screened with the observed rainfall series in order to obtain the suitable set of variables for each station. The Generalized Linear Model (GLM) and Bootstrap Aggregated Classification Tree (BACT) model are developed in the Matlab R2015a platform to model the rainfall occurrence. Thereafter, the Artificial Neural Network (ANN) is also developed in the Matlab R2015a platform to model the rainfall amount. The Non-homogenous Hidden Markov Model (NHMM) is developed in a MVNHMM software toolkit to model the rainfall occurrence and rainfall amount.

The performance of rainfall occurrence models will be examined and compared in terms of their rainfall persistence, spell lengths distribution and matching. For rainfall amount models, parametric and non-parametric tests will be used to evaluate and compare their performance. At the end, a suitable rainfall generator is selected for Langat River Basin based on their overall performance.

CHAPTER 2

LITERATURE REVIEW

2.1 Climate Change

According to the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4), the climate change was defined as an identifiable change in the state of climate, and remains for a long period of time, which is measured in terms of decades or longer. The changes can be recognized through the mean or the variability of its properties in a statistical sense. Basically, these changes are either caused by the impact of human activities or the natural variabilities. There is another definition from the United Nations Framework Convention on Climate Change (UNFCCC). The composition of the global atmosphere is changed directly or indirectly by human activities or the natural climate variabilities over equivalent time periods are the situation of climate change. In short, the climate change can be simplified as the changing of climatic variables (precipitation, temperature, wind patterns, etc.) under a long time period.

Climate change, a resultant of rising of greenhouse gas (GHG) concentration retained at atmosphere, has been and still is, the main important concerned global issue for the world to tackle and be in terms with it. Climate change is believed to have a strong interdependence bond connected with the

human activities, but it would be much difficult in recognizing the impacts on climate caused by human activities. The identification on the flexibility of the projected rates of climate change on those impacts are also another challenging task. The connection between the human activity and climate change is strong and unbreakable since the early century, even until now the human activity still dominate the major cause on the change of climate state. However, the impacts of human activities on climate change can be mitigated by some remedy actions, and are already undertaken once the human realized the severity of the impact brought by the their activities on climate change, but still, there are no substantial reduction of impacts showed in the results. Thus, the improvement on the remedy action should be encouraged for coming decades with the minimum requirement of targeted maximum level of reduction to be realized.

The anthropogenic emissions of greenhouse gases from recent data showed the upmost among the historical data in the IPCC's Fifth Assessment Report (AR5). The influences caused by human activities on the climate change, bring the impacts toward the human life and natural systems, which are very clear and widespread. The globally averaged greenhouse gas concentration from 1850 to 2000, as shown in Figure 2.1, can be used to prove that the uptrend on the emission of greenhouse gases over the years, especially a rapid growth is achieved from years 1950 to 2000. The main reason for the sudden rapid growth of carbon dioxide (CO₂) during the period of 1950 to 2000, as shown in Figure 2.2, is due to the burning of carbon based fossil fuels, cement production and flaring were increased tremendously. As the technologies are getting more advanced over years, the energy demand also increase with the population

growth. In order to ensure that the energy is supplied sufficiently, the extended construction of power plants was resulted in the rising combustion of fossil fuel. As there is a continually rising level of the emitted carbon dioxide trapped in the atmosphere, the global warming condition would only get more severe over time.

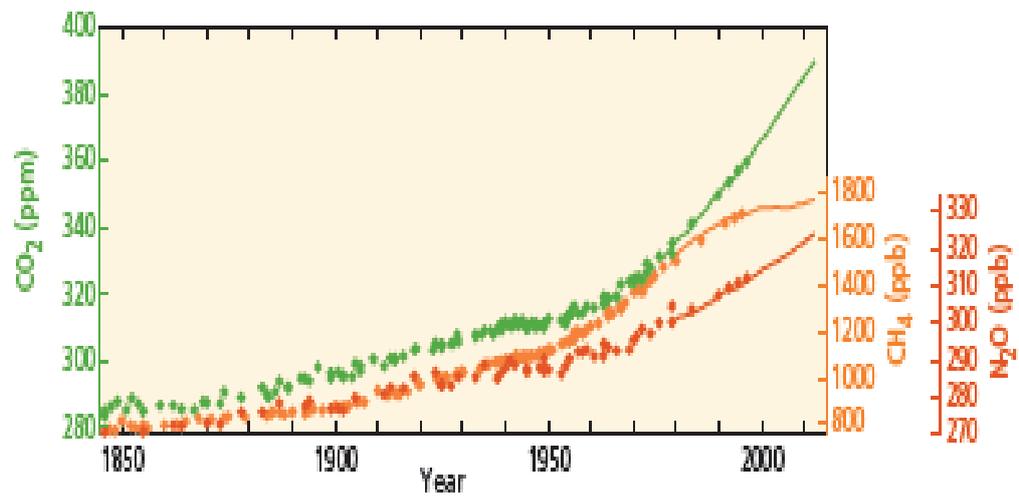


Figure 2.1: Globally averaged greenhouse gases concentration (IPCC, 2014)

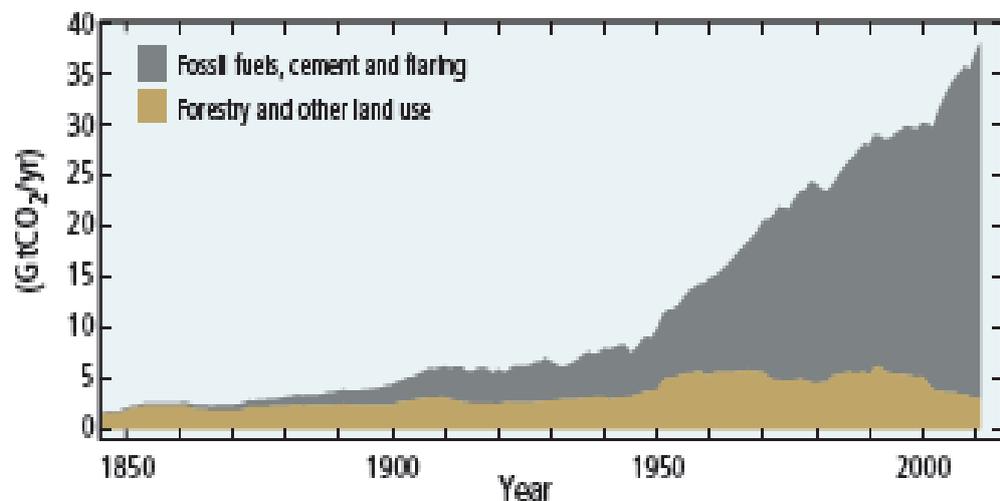


Figure 2.2: Globally anthropogenic CO₂ emissions (IPCC, 2014)

Figure 2.3 shows the average surface temperature with the combination of land and ocean in global scale, where the black line represents the abnormal surface temperature of land, orange colour represents the abnormal surface temperature of ocean, and blue line represents the average between both black and orange colours. The trend of temperature for land and ocean in Figure 2.3 is similar with the trend as shown in Figure 2.1, with an increasing trend from 1850 until 2000, especially the sudden rise during the period of 1950 to 2000. Besides the warming due to past anthropogenic emissions, the effect of future anthropogenic emission and natural climatic variable also play a significant role in influencing the future climate condition.

According to the projected changes done by IPCC (2014) on climate system corresponding to 1850-1900, the global surface temperature is deemed to be increased beyond the value of 1.5 °C with high confidence at the end of the 21st century. While, the temperature may reach up to 4.8 °C in the worst projected scenario in relative to 1986-2005. In fact, most of the land areas would experience more frequent hot and fewer cold temperature extremes as the increasing of global surface temperature, no matter in daily or seasonal timescale.

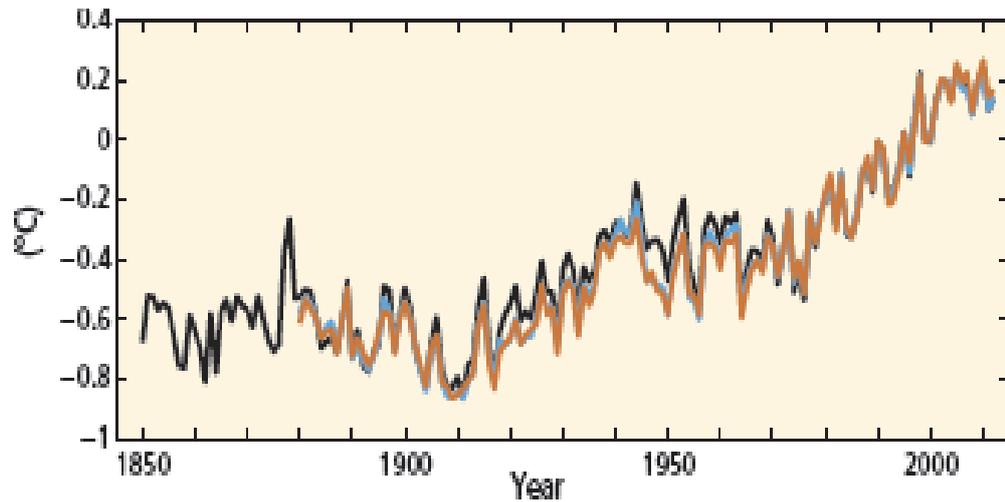


Figure 2.3: Globally average combined land and ocean surface temperature anomaly (IPCC, 2014)

The impacts of climate change on the natural systems have been proven by many evidences, which are strong and comprehensive. Other than the increasing of global surface temperature as stated at above, the melting of glaciers and the rising of global mean sea level due to ice melting are also indicating the scenarios of climate change. Due to climate change, the hydrological systems are altered by the changing of rainfall pattern or the melting of ice and snow in several regions. The quality and quantity of water resources are affected too. As the global surface temperature is increasing over the years, the living organisms on earth are facing the problem of adapting to the changed living environment. They might need to migrate to another place for surviving due to the destruction of their habitat caused by human activities. Even some rare species of plants might face the extinction because they can only survive in a certain area with specific temperature and humidity.

Generally, the climate change is triggered by two main factors, which are natural factors and human activities. Natural factors can be explained in terms of solar radiation and volcanic eruption. The reason of solar radiation contributes to climate change is because the incoming energy radiated from Sun has the influence on the greenhouse effect. The greenhouse gases in atmosphere are capable of absorbing and emitting the infrared radiation from the surface of earth, then part of the energy will be radiated out to outer space but some will be reflected back to earth surface. Thus, the heat is trapped by the greenhouse gases in the atmosphere, which cause the temperature of earth also increase.

There is a substantial emission of particles during the eruption of a volcano, which including the aerosol, and those particles are tend to be released and trapped in the atmosphere, be a part of the causes that contribute climate change. However, the occurrence of eruption is irregular and unpredictable. Thus, its effect may not be notable as it only bring a short term effect on climate change.

Human activities is one of the primary factors contribute to the climate change and its impact has a strong influence on climate. During the era of pre-industry, the concentration of gases (CO₂, methane and nitrous oxide) in atmosphere have been increased largely due to the emissions of anthropogenic greenhouse gases. The main driving forces that trigger those emissions are the economic and population growth. During the Industrial Revolution in 18th century, the energy demand was getting higher due to the rapid growth of population and economic, which resulting in the development of advanced

technologies for harvesting the natural resources. Therefore, the burning of fossil fuels for supplying the energy to industries and domestic purposes increased tremendously. At the end, the severe global warming effect was triggered by the greenhouse gases released from the combustion, thus increasing the average surface temperature of land and ocean.

The evidences for the impacts of human activity on climate system has been brought to attention since the IPCC Fourth Assessment Report (AR4), where the increasing of observed global average surface temperature was mainly caused by the concentrations of anthropogenic greenhouse gases and other anthropogenic forcing together. Besides, there are several consequences of the anthropogenic influences, such as the retreat of glaciers, surface melting of Greenland ice sheet, loss of Arctic sea-ice, increasing of the upper ocean heat content and mean sea level in global scale, and the imbalance of global water cycle. The consequences of climate change on the physical and biological systems in the Earth are expected to extend on the timescales from upcoming decade to the century (Sullivan and Huntingford, 2009)

Generally, precipitation can be explained as rainfall, snowfall and other form of water falling from the clouds. The occurrence of precipitation is strongly depending on the temperature and weather situation. First is the condensation of the precipitation in the form of water vapour, normally in the rising air that expands and thus cools. The warm air rising over the cooler water, colder air pushing under warmer air, convection take places from the local heating of surface, or other weather and cloud systems (Trenberth, 2011). When

the temperature of air is higher than freezing point, the precipitation turns into rain. This is the reason of why global circulation in terms of precipitation is an important element to the Earth's system, as it transports the heat from tropics to the higher latitudes.

According to the explanation from Intergovernmental Panel on Climate Change (IPCC), the hydrological cycle, which include the changes in precipitation, is influenced greatly by the increasing average global temperatures. The changes in the circulation patterns of atmosphere and the rising of water vapour and evaporation associated with warmer temperature, are the main factors to have direct impact on the precipitation. The changes in precipitation are in terms of amount, intensity, frequency and type of precipitation. Therefore, the current condition of global warming in this world causes an overall increase in precipitation with an unsure magnitude. This is because the increased heating of global temperature leads to the greater evaporation rate and causes the drying of surface, hence the intensity and duration of drought are increased. However, the atmospheric water vapour is also increased with the water holding capacity of air rises by about 7% per 1 °C warming, due to the increasing of global temperature.

The occurrence of extreme precipitation event takes place due to any type of storm supplied with increased moisture, thereby increasing the risk of flooding. Furthermore, the changes in precipitation amount have the direct influence to the corresponding regional changes in runoff, so the management of water supply also get affected, especially for the runoff and river flows in

semi-arid regions are sensitive and will be most susceptible to the changes in precipitation. Groundwater also is one of the sources in supplying water, while the recharge rates of groundwater are affected directly by the rainfall rate, so the changes in average rainfall is potentially impacting the water supply.

The changes of precipitation are expected to be different in every region, as some regions may become wetter or dryer when compared to previous conditions, and the probability for the region to remain unchanged is very low. However, according to the results obtained from most models, they showed that the most high-latitude regions will experience the increasing of precipitation, while the most subtropical areas will experience reducing of precipitation. For Equatorial regions, the prediction on the changes in precipitation contains a high level of uncertainty. According to the research of O’Gorman (2015), the projections of 21st century climate change with GCM, showed the intensity of extreme precipitation is generally increasing in most of the regions except for some regions in the subtropics.

The countries in Southeast Asian, which include the Myanmar, Thailand, Vietnam, Laos, Malaysia, Singapore, Indonesia and etc. Those countries are affected by the monsoon which can be defined as a great scale seasonal reversals of the wind regime. The northeast and southwest monsoons are the two main monsoons, which is occurring from November to March and from May to September, respectively, while the transition month from southwest to northeast monsoon is October. Hence, the monsoonal areas will receive the maximum of summer rainfall and the most of double rainfall maximum (Loo, et al., 2015).

Besides, the temperature is predicted to be increased in the late 21st century and early 22nd century, which will affect the frequency of changes and influence the monsoon precipitation. The precipitation during the monsoon periods can be up to 70% below normal levels (Schewe and Levermann, 2012), which is resulting the Indian summer monsoon and the onset monsoon over the Southeast Asia postponed up to 15 days.

Floods can be defined as the accumulation of precipitation in a huge amount until the surface of land is submerged, and they are occurring at a high frequency in Asian countries, especially Southeast Asian countries. Recently, the precipitation patterns have changed globally due to the climate change. The increasing of monsoon rainfall intensities become the major factor of triggering the occurrence of flood and landslide events in Malaysia and some Southeast Asia countries. The loss of lives and damage of properties are the consequences of the monsoon flooding in Southeast Asia countries. In the mid of December 2014, most regions of Malaysia had experienced the heavy seasonal rains and strong winds, this situation were continued until the early January 2015. Therefore, a severe flooding was triggered by the accumulation of this heavy rain in East Coast, and the affected areas covered the state of Terengganu, Pahang and Kelantan. Other than that, heavy rainfalls also impacted the four states in Peninsular Malaysia (Johor, Perak, Perlis and Selangor) and one state in East Malaysia (Sabah) to have the occurrence of floods.

According to ABC NEWS (26 December 2014), there were more than 100, 000 people evacuated, as the North-eastern of Peninsular Malaysia had been hit hard by the northeast monsoon. The number of evacuation rose sharply from 100,000 to 160,000 in one day (Reuters, 27 December 2014). On 31 December 2014, more than 21 people were killed and around a quarter of a million people were displaced because of the floods in Malaysia. However, the state that had the worst condition was Kelantan, with 14 deaths and at least 158, 000 people displaced (AFP, 31 December 2014). The total damage to property and infrastructure in all affected states were calculated as close to 1 billion ringgit. The occurrence of these severe floods had great impacts to Malaysia in the aspects of healthcare, education, economy and the lives of residents.

2.2 Homogeneity Tests

The detection on the variability of collected rainfall data in terms of homogeneity is an important issue to ensure its reliability before it is used as an input in any hydrologic analysis model. Since 1990s, many studies had employed various statistical tests in detecting the non-homogeneity of precipitation and temperature series. There are two types of homogeneity tests for the time series data, which are the absolute method and the relative method. Both methods are applied under different conditions. The absolute method is used to check the time series data of each station, while the relative method is applied for the time series data of a station with respect to its neighbouring stations. The relative tests are deemed to be more effective than absolute method, especially for the time series data of two highly correlated stations (Wijngaard

et al, 2003). However, the absolute method is more appropriate for the area with less dense station network because of the location of stations are sparsely distributed, and the time series of each stations are hardly to be correlated with each other.

Four commonly employed tests in several studies (Costa and Soares, 2009; Sahin and Cigizoglu, 2010; Talaei, Kouchakzadeh and Some'e, 2014), are the standard normal homogeneity test, the Buishand range test, the Pettitt test, and the von Neumann ratio test. Wijngaard et al. (2003) was the first researcher applied these four tests all at once in their studies, to evaluate the non-homogeneity of daily temperature and precipitation series. These tests could be grouped as absolute tests, with the ability of handling the simultaneous changes in observational network as their advantage. In their study, a two-step approach was applied in evaluating the homogeneity of both series for the period of 1901 – 1999 and the sub-period of 1946 – 1999. The application of these four tests on the dataset of European Climate Assessment (ECA) project was the first step, followed by an overall evaluation of all tests under the significant level of 1%. The homogeneity of data series of individual station is classified into: 'useful', represent the null hypothesis was rejected by one or none test; 'doubtful', represent the null hypothesis was rejected by any two tests; and 'suspect', represent the null hypothesis was rejected by more than two tests.

Based on the study of Wijngaard et al. (2003), the number of temperature series under the class of 'suspect' were up to 92% and 54% for the periods of 1901 – 1999 and 1946 – 1999, respectively. Compared to

temperature series, the condition of precipitation series is much better with the number of precipitation series greater than 70% under the class of ‘useful’ for both periods. They explained that the repositioning of station or instruments, and the altering of observing and measuring methods, are the main reasons of causing the breaks in the temperature series. Besides, the breaks caused by climate variations in temperature series are less preferable to be detected by absolute tests. However, the absolute tests is more sensitive than relative tests in detecting the breaks in precipitation series, which are frequently caused by the simultaneous changes in observational routines. In considering the sparse distribution of station network, the absolute tests were preferred to be applied in their study, but the relative tests could become gradually useful when more station series are included in the ECA dataset.

Basically, there are no specific rules in selecting the best methods for detecting the non-homogeneity of time series in Malaysia, but the two-step approach proposed by Wijngaard et al. (2003) have been widely applied by many researchers in their studies (Suhaila et al., 2008; Kang and Yusof, 2012; Ahmad and Deni, 2013; Ng et al., 2015). According to the study of Suhaila et al. (2008), the two-step approach was applied to assess the non-homogeneity of daily rainfall series of 50 stations in Peninsular Malaysia from years 1975 to 2004. The annual amount series and annual number of wet days with the threshold value of 0.1 mm and 1 mm, were the two variables checked by all four homogeneity tests. They found that the results of both threshold values were similar and the majority of rainfall stations were classified under the class of ‘useful’, but the number of wet day series suspected to be non-homogeneous

was more than the rainfall amount series. In addition, they highlighted the non-homogeneous series should be rejected from any further trend analysis, by taking the consideration on the historical metadata was not included in this study to assess the detected break and make correction in non-homogeneous series.

However, different testing variables were used in the study of Kang and Yusof (2012), which include the mean, maximum and median of daily rainfall series in annual scale and under the significant level of 5%. The daily rainfall data with the relatively small percentage of missing data ($< 10\%$) from 33 stations in Damansara, Johor and Kelantan were obtained in the years of 1998–2007, 1996–2005, and 1998–2007, respectively. In comparison to all homogeneous series of annual mean and annual maximum, the number of stations under the class of ‘doubtful’ and ‘suspect’ were one and four, respectively, in respect to annual mean. In addition, they found that the missing values had no influences on the results of homogeneity. According to the study of Ahmad and Deni (2013), same testing variables as in the study of Suhaila et al. (2008) with the threshold value of 1mm, were checked using these four homogeneity tests under the significant level of 1%. The results showed that the rainfall data from 17 out of 83 stations were detected to be non-homogeneous, so those stations were rejected for further analysis, since there was no correction approach involved in their study.

The absolute homogeneity tests in the study of Ng et al. (2015), were applied to the rainfall data in monthly, yearly and seasonal scale, due to the sparse density of rainfall stations in their study area. The collected daily rainfall

data from 10 stations were transformed into 120 monthly series, 10 yearly series and 40 seasonal series, and those series were tested using four homogeneity tests under the significant level of 5%. They found that the number of monthly series classified as ‘doubtful’ and ‘suspect’ were six and one, respectively, but there were no yearly and seasonal series suspected to be non-homogeneous. At the end, they concluded that nearly all the tested series were homogeneous, but still need to be examined properly especially for those non-homogeneous series before they were used for further analysis.

In conclusion, the two-step approach proposed by Wijngaard et al. (2003) in assessing the non-homogeneity of time series, have been widely used by many researchers in Malaysia. According to their studies, the absolute methods were applied to their collected rainfall amount series, rather than using relative methods. The main reason is due to the sparsely distributed location of rainfall gauge stations in Malaysia at different regions, so the surrounding area and weather condition of each station may not be the same. Therefore, using the relative tests in respect to the series of neighbouring station might not be as powerful as using absolute tests, which test the variables of each station separately.

2.3 Downscaling Approaches

The General Circulation Models (GCMs), also known as global climate models, are the useful devices in expressing the physical processes in the atmosphere, land surface and ocean. The global climate system in response to

the raising of the greenhouse gases concentration, is able to be simulated by these GCMs. Hence, they have been widely used to project the future climate under different scenarios and these results may deliver the qualitative and quantitative information regarding to the issue of future climate change. However, GCMs are a three dimensional model, in which their resolution is considered to be too coarse if compared to the dimension of units revealed in the majority of impact assessments. These GCMs cover the horizontal resolution between 250 to 600 km, and 10 to 20 vertical layers in the atmosphere, but sometimes may achieve 20 layers in the oceans (Olsson et al., 2013). Thus, the variables produced from GCMs may not be identical to those found in observed data.

The downscaling approaches have been developed to act as a bridge for reducing the discrepancy and acquiring the prerequisite regional/local climatic variables. These downscaling approaches can be classified into two types, which are dynamical and statistical downscaling techniques. Dynamical downscaling is a technique of operating a climate model to produce higher resolution of time-varying atmospheric boundary conditions with the help of GCMs. Regional climate model (RCM) is the climate model, which can resolve the atmospheric features into regional scale by applying this technique. However, other than the forcings at regional scale, the biases from the running of GCM is another key of affecting the accuracy of model. The statistical downscaling is a technique of transforming the climate variable from coarse to local scale, by establishing a relationship between the climate at two different

spatial resolutions. The advantages and disadvantages of using these two downscaling techniques have been summarized in Table 2.1 and Table 2.2.

Table 2.1: Advantages and disadvantages of dynamical downscaling technique (Fowler et al., 2007)

Advantages	Disadvantages
<ul style="list-style-type: none"> • Provide responses based on existing physically processes • Provide higher resolution information from GCM-scale output, which can resolve atmospheric processes on a finer scale 	<ul style="list-style-type: none"> • Computationally intensive • Restricted number of scenario ensembles • Depend strongly on GCM boundary

Table 2.2: Advantages and disadvantages of statistical downscaling technique (Fowler et al., 2007)

Advantages	Disadvantages
<ul style="list-style-type: none"> • Comparative inexpensive and computationally efficient • Present point-scale climatic variables from GCM-scale outputs • Able to derive the variables not available in RCMs • Easily transferable to other regions • Application of standard and accepted statistical procedures • Observations can be combined into method directly 	<ul style="list-style-type: none"> • Require long and reliable observed data for training • Dependent on the choice of predictors • Relationship between predictor and predictand is non-stationary • Responses of climate system are not included • Dependent on GCM boundary forcing • Downscaling accuracy is affected by domain size, climatic region and season

2.4 Statistical Downscaling

Statistical downscaling is a technique generally used to predict the local climate variables through a robust relationship established between large scale atmospheric (predictors) and local climate (predictand) variables. The prediction of future climate variables is obtained using the future predictors projected by GCMs. The statistical downscaling method is currently deemed to be more preferable than dynamical downscaling, due to the high demand conditions of low cost and rapid impact assessments of local climate change. Another reason is the ability of this method to produce the climate projections at certain location, if compared to the spatial resolution of RCMs, which are computationally restricted to 20–50 km. However, there are two critical assumption made in statistical downscaling method (Fowler, et al, 2007). Firstly, the predictor variables should be physically significant and able to mirror the variability of climate in various timescales. Second, the relationship between predictor and predictand is assumed to be time-invariant, so it still remain constant under the climate change conditions (Yarnal et al., 2001). There are many developed statistical downscaling methods with various application and complexity, and they can be categorized into three types, namely the regression-based techniques, the weather pattern classification and the weather generators. However, the implementation of these methods are still considered to be simple as long as an adequate amount and high quality of observed data are given.

2.4.1 Regression-based Techniques

Regression-based models are commonly used for downscaling purpose by establishing a linear or nonlinear relationship between predictors and predictand. The development of multiple linear regression models using the atmospheric variables in grid cells as predictors for local climatic variables (precipitation, temperature, etc.), are the easiest method, if compared to those more complex methods include the application of artificial neural network (Snell al., 2000; Schoof and Pryor, 2001), canonical correlation analysis (Busuioc, et al., 2008; Lutz et al., 2012) and singular value decomposition (Widmann et al., 2003; Liu and Fan, 2013). However, other than those methods, generalized linear models with the application of logistic regression are one of the recent innovated regression-based method which have been used for downscaling in the studies of Buishand et al. (2004) and Kenabatho et al. (2012).

Statistical downscaling model (SDSM) is a model with the combination of weather generators and regression models. SDSM employs these robust statistical downscaling methods to act as a decision support tool during the impacts assessment of climate change and to generate the multiple ensembles of synthetic daily weather sequences. Besides, SDSM facilitates the rapid development of low-cost scenarios of daily weather variables at single and multiple sites under the regional climate forcing of current and future using the simulations of GCMs. There were over 170 published studies contributed by employing SDSM, since the toolbox of this model was released in the year of 2001(Wilby and Dawson, 2012). In the study of Souvignet et al. (2010), the

performance and ability of SDSM in simulating the extreme events in Upper-Elqui sub-basin were examined and thereafter, it was used to downscale the future temperature and precipitation under scenarios of A2a and B2a from Third Generation of Hadley Centre Coupled Model (HadCM3). The results showed the SDSM capable to simulate the maximum and minimum temperatures, and the extreme events accurately. However, the simulation results of precipitation were not as good as temperature, which concluded that the SDSM was not robust in simulating the extreme rainfall events.

Hassan et al. (2014) compared the SDSM and LARS-WG in simulating and downscaling the rainfall and temperature under emission scenarios of A2 and B2 in Peninsular Malaysia. They found that the SDSM exhibited better performance than LARS-WG in downscaling the daily and monthly time series of rainfall and temperature data. However, LARS-WG showed the better ability in simulating the dry- and wet-spell lengths of rainfall, when compared to SDSM, which under-predicted the wet spell length. In terms of maximum and minimum temperature, SDSM performed slightly better than LARS-WG. LARS-WG model is just a stochastic weather generator, which could not extract the outputs of GCM directly. However, SDSM can be applied together with GCM's outputs under different scenarios for downscaling. They concluded that SDSM under scenarios of HadCM3 A2 and B2, was found able to reproduce the observed monthly rainfall and temperature series. Therefore, these two scenarios are also assumed to be useful for the prediction under future climate period.

Linear and nonlinear regression methods have been used broadly in downscaling the rainfall with different capabilities of each method. The generalized linear model (GLM) can be defined as a flexible generalization of the ordinary linear regression. The use of maximum likelihood methods is one of the most important characteristics in GLM, which allow the response variables and error terms from the fitted models to have different types of distributions other than normal distribution. A GLM is typically consists of three components, first is the random component, which stating the distribution of response variable conditional on the predictor variables. The probability distribution of response variable come from the exponential family, which include normal, binomial, Poisson, gamma and negative binomial. Second is the systematic component, which represents the continuous or categorical predictors in the model. These predictors must be interrelated with each other. Last is the link function, such as identity link, log link and logit link, which act as a bridge to connect the random and systematic components.

Logit link is the common link function to be used in hydrological models, because it is applicable to binary data and logistic regression. In the study of Prasad, et al. (2010), the logistic regression was required due to the nature of response variable is in binary form, and the logit transformation ensures the generated probability lies between zero and one. They employed a multi-predictor logistic regression model to prediction the monthly rainfall and the model performed well in predicting the extreme rainfall years and the total rainfall in corresponding to such years. However, they pointed out two limitations of their developed empirical model, which are the model still

remains connected with the biases of GCMs in simulating the atmospheric variables and the non-stationary statistical relationship between predictors and predictand. Lee et al. (2011) also employed a multiple stepwise logistic regression to develop the downscaling models for rainfall occurrence prediction in Hong Kong. The occurrence of wet and dry days can be represented by the binary sequence, where one represents the wet day and zero represent the dry day (Beckmann and Buishand, 2002). Hence, the GLM with logistic regression are suitable in rainfall occurrence modelling. There are few studies that employed the GLM for rainfall occurrence modelling conditional on a range of large-scale atmospheric variables (Buishand et al., 2004; Hasan and Dunn, 2012; Abdellatif et al., 2013) and they showed the good performance in predicting the occurrence of rainfall at their study areas. After the rainfall occurrence model was developed, the rainfall amount was calculated by fitting the gamma distribution conditional on modelled wet days (Chandler and Wheeler, 2002; Yang et al., 2005; Beecham et al., 2014).

In the study of Fealy and Sweeney (2007), they employed a two-step generalized linear modelling approach to downscale the precipitation in Ireland. First, the rainfall occurrence model was modelled using the logistic regression, then followed by the modelling of rainfall amount using a log link function and a gamma distribution. Based on the simulation results of their study, they showed the derived models were considerably better than the reference prediction with the Heidke skill score greater than zero and capable to predict the interannual variability of precipitation. The derived models were then used to simulate the future changes in precipitation based on three different GCMs.

They realized that the application of different GCMs may arise the emission uncertainties, especially the magnitude and direction of changes are varied between the GCMs. Kenabatho et al. (2012) also employed the logistic regression and gamma distribution to simulate the rainfall occurrence and amount, respectively, in the Limpopo basin, Botswana. The simulation results showed the properties of observed rainfall were generally well predicted by the models, especially the interannual variability, after the involvement of external atmospheric variables. They concluded that the GLM is possible to be applied in other regions and climate change studies due to its good performance in simulating the multi-site rainfall.

Artificial neural network (ANN) is one of the artificial intelligence, which has been widely used for the development of model with downscaling technique. ANN can be simplified as a computational model similar to the structure and functions of biological neural networks. The arrangement of ANN is influenced by the flowing of information through the network, because the neural network is changed or learned into a different form based on the complex relationship between the inputs and outputs or the patterns are found. Therefore, ANN can be treated as a flexible nonlinear statistical data modelling tool, which resemble the properties of biological neural system. The reason that makes ANN so well-known is its ability of identifying the complex pattern for creating a linkage between input and output data. Besides, ANN is able to generate an optimum solution through the learning and generalization process, even when there are some missing or errors in input data. Therefore, the accurate prediction can still be achieved as long as the model is trained with relevant data.

Basically, ANNs are structured by a network of three layers, for instance input layer, hidden layer and output layer. They are connected to each other with many simple processing units. Multilayer perceptron (MLP) is the typical architecture of ANN, as illustrated in Figure 2.4. The function of input layer is to receive the information of inserted data, while the output layer is used to generate the output information. The parameters in input and output layers are the independent and dependent variables, respectively. The number of hidden layers to be constructed is depending on the user, but more hidden layers may reduce the training speed without any improvement on the efficiency of network due to the increment of parameters to be estimated.

The weights w_{ij} and w_{jk} , as shown in Figure 2.4, are used to transfer the information from input layer to output layer via the connection of nodes within different layers. The advantage of using ANN is that the desired output can be obtained for specific input by adjusting the weights of nodes. There might be a difficult and complicated task to calculate all the weights when the network consists of hundreds or thousands of nodes, therefore the implementation of various algorithms can help to improve this limitation.

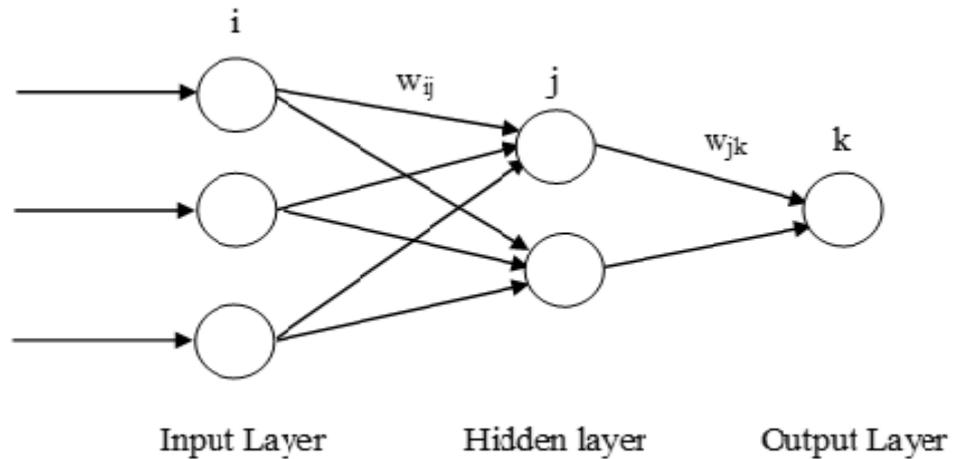


Figure 2.4: Typical architecture of ANN (Abdulkadir et al., 2012)

The MLP neural network has been used for weather prediction (Deshpande, 2012; Kumar and Jha, 2013) due to its flexibility and simplest structure of ANN. Besides, there are some recent studies that employed the ANN to establish a relationship between atmospheric variables and local climate variables for downscaling purpose (Cannon, 2008; Ahmed et al., 2015). In the study of Goyal and Ojha (2012), they employed the linear multiple regression (LMR) and ANNs to downscale the surface temperature of Pichola lake catchment in India. Other than the monthly maximum and minimum temperatures, the National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP&NCAR) reanalysis dataset were used to train both model, thereafter, they were used to downscale the simulation of Third Generation Coupled Global Climate Model (CGCM3) under the emission scenarios of A1B, A2, B1 and COMMIT. The ANNs performed better than LMR-based model in simulating the observed maximum and minimum temperatures. The downscaled results showed the increasing trend for both

maximum and minimum temperatures under A1B, A2 and B1 scenarios, but no trend is detected under COMMIT scenario.

Mendes et al. (2014) developed ANN in their study to downscale the precipitation for the trend analysis of extreme rainfall in Amazon Basin. ANNs were found able to simulate the observed precipitation with the high correlations of approximate 88.9% and 91.3% in the cities of Belén and Manaus, respectively. Besides, ANNs also showed the high similarity of spatial distribution with observed data in the correction process. Campozano et al. (2016) compared the performance of SDSM, ANNs and the least squares support vector machines (LS-SVM) approaches in downscaling the monthly rainfall in Paute River Basin. They utilized the neural toolbox of Matlab with the algorithm of Levenberg-Marquardt to optimize the neural network. Both ANN and LS-SVM models showed the overall better performance than SDSM, even though the SDSM performed better in certain months. They suggested the selection of synoptic predictors for certain months or season for further analysis and the combination of dynamical and statistical downscaling to obtain the characteristics, which might not be able to be signified by GCMs.

The use of feed forward back propagation in ANN was demonstrated in the study of Vu et al. (2016), to simulate the observed daily precipitation in Bangkok using the reanalysis data. Then, they downscaled the future precipitation using Fifth Generation of European Centre Hamburg Model (ECHAM5), Third Generation of Community Climate System Model (CCSM3.0) and Model for Interdisciplinary Research on Climate-Earth System

Model (MIROC-ESM) under CIMP3 scenario A1B, and Max Planck Institute-Earth System Model (MPI-ESM) under CIMP5 RCP 4.5. The results proved the ANN is capable to reproduce trend of observed long term precipitation with the correlation coefficient and Nash-Sutcliffe efficiency of 0.8 and 0.65, respectively. The future trend of precipitation was detected to be increasing, especially the extreme rainfall intensity until the end of twenty-first century.

2.4.2 Weather Pattern Classification

The approach of weather typing/classification is to group the local variables into different classes of large-scale atmospheric variables or “states”, in respect to their similarity in synoptic weather pattern. The future outputs simulated by GCMs, which are corresponding to their most similar historical atmospheric pattern, are used to predict the values of local variables in response to the future climate change (Bo  , et al., 2006). An important assumption is made in this approach, and that is the characteristics of the classes will still remain the same even in future (Brinkmann, 2000). Therefore, this method is appropriate for downscaling the non-normal distributed daily rainfall (Moron et al., 2008). However, the classification of possible weather patterns require a substantial amount of observed data, which is considered to be more computationally demanding than linear models.

There are several models employing the approach of weather pattern classification to predict and downscale the rainfall, for example the hidden Markov models (HMMs) (Mares et al., 2014) are able to reproduce the

characteristics of wet and dry spells by classifying the spatial rainfall patterns and inferring the corresponding weather pattern to predict the rainfall occurrence. While, the spatial and temporal variability are able to be predicted by non-homogeneous hidden Markov models (NHMMs) (Kioutsioukiset al., 2008) through the recognition of noticeable patterns in the multi-station and persistence in weather state, respectively. The NHMM is defined as a kind of double stochastic finite state machine. The reason to make the NHMM different from a regular Markov chain model is that the ‘hidden’ means the states in NHMM are not directly observed (Greene et al., 2011). After the predictors are incorporated into the NHMM, the transition probabilities no longer remain the same in between the hidden states, and they change in time along with the large scale exogenous atmospheric variables (Mehrotra et al., 2006; Fu et al., 2013).

There are few researches had employed the NHMM as downscaling model and they showed the good prediction ability in simulating the rainfall occurrence and amount (Bellone et al., 2000; Robertson et al., 2006; Greene et al., 2011). In addition, the NHMM has been proven in the study of Robertson et al. (2004); it is capable of simulating the characteristic of rainfall occurrence in the field of spell lengths distribution and deliver the majority interannual simulation skill of GCMs into daily rainfall sequences. Robertson et al. (2009) employed NHMM as a downscaling technique in their study to obtain the daily rainfall sequences at Indramayu, Indonesia, which was conditional on a set of seasonal predictions from ECHAM4.5. However, the results showed NHMM exhibited the highest and lowest performance in simulating the rainfall frequency and mean rainfall intensity, respectively. According to the study of

Verbist et al. (2010), the NHMM was found capable to simulate the daily rainfall sequences in central Northern Chile, based on the evaluation of seasonal amount, daily rainfall frequency and mean daily amount on wet days. The downscaled results were then adapted to drought prediction, by calculating the standardize precipitation index (SPI). The SPI was quite well predicted by NHMM with the high hit rates.

In China, Liu et al. (2011) evaluated the performance of NHMM and SDSM in simulating the daily precipitation at Tarim river basin. Both models showed the good prediction ability and stability with slight difference in simulating the wet- and dry-spell lengths. NHMM outperformed the SDSM in simulating the amount of wet day precipitation and monthly precipitation. The SDSM was inferior in predicting the extreme value of wet day precipitation amount and monthly precipitation for certain months. Furthermore, similar good simulation results of NHMM, were obtained by Liu et al., (2013) when they compared with another two multi-site statistical downscaling models, namely conditional resampling SDSM and generalized linear model for daily climate series (GLIMCLIM), in North China Plain.

In the study of Mares et al. (2014), they employed the Baum-Welch algorithm to acquire a set of parameters, which achieve the greatest conditional probability of the observed data in terms of likelihood. Seven hidden states were fitted into hidden Markov model (HMM), where the optimum hidden state number was determined based on the mean values of Bayesian Information Criterion (BIC). Thereafter, the NHMM was applied with the introduction of

atmospheric predictors. The performance of NHMM using four predictors to simulate the spring precipitation were evaluated separately. They concluded the daily sea level pressure and geopotential at 850 hPa were the suitable predictors, which can improve the performance of NHMM in simulating the precipitation at Danube Basin.

2.4.3 Weather Generator

Weather generators (WGs) are basically referred as stochastic weather generators, which are the tool used to produce numerous realizations of the same data or reproduce the missing data. The major application of stochastic weather generator is to produce a long time series of plausible weather data, according to the characteristics of observed weather data and random number of sampling. A typical stochastic weather generator is developed using a two-step process, with the modelling of precipitation occurrence through the Markov chain processes, then followed by the modelling of climate variables of interest conditional on precipitation occurrence. However, the major drawbacks of WGs are the requirement of long daily data sequences and sensitivity to missing or erroneous data in training set. Despite of that, the WGs have been adapted to perform the statistical downscaling of local climate variables by conditioning their parameters on the large-scale atmospheric variables. There are several WGs have been developed and widely used for downscaling purposes, namely Long Ashton Research Station-Weather Generator (LARS) (Hassan et al, 2014),

Weather Generator of Ecole de Technologies Superieure (WeaGETS) (Caron et al., 2008) and Climate Generator (CLIGEN) (Chen et al., 2011).

In the study of Khan et al. (2006), they employed three statistical downscaling models, such as SDSM, LARS and ANN to downscale the precipitation, maximum and minimum temperatures in daily scale. Their performances were evaluated and compared based on the assessments of different uncertainty analysis with the confidence intervals of 95%. However, the SDSM exhibited the better performance in reproducing the statistical characteristics of observed, followed by LARS-WG and ANN. Hassan and Harun (2013) utilized the LARS-WG to downscale the future rainfall data at Kerian, Malaysia and under the emission scenario of HadCM3 A2. They commented the direct selection of the GCM-variables in LARS-WG make it become less complicated than SDSM. However, the results showed the poor performance of LARS-WG in simulating observed data with small R value.

According to the study of Zhang et al. (2004), the CLIGEN was used to downscale the monthly predictions projected by GCM HadCM3 to daily weather series, and the potential impacts of future climate change on soil erosion and wheat production were examined. The results showed the CLIGEN is capable to reproduce the variance ratios of maximum temperature, but slightly over-prediction for minimum temperature. In addition, the variance of monthly precipitation was found to be over-predicted by CLIGEN, because more events with larger or smaller precipitation were produced due to the increase of

daily variance. However, the summation of larger and smaller values in monthly total would help to reduce the monthly variance.

In general, the low operation cost and rapid assessment of local climate change are the main advantages of statistical downscaling methods, which make them become preferable than dynamical downscaling in practical uses. A statistical downscaling methods can be categorized into three main types. Regression-based models are the most common method used for downscaling purpose by establishing a linear or nonlinear relationship between large scale atmospheric variables and local climate variables. Weather pattern classification is a method of developing relationship between large scale atmospheric variables and local climate variables according to the provided weather pattern schemes. The approach of weather generators utilizes the parameters conditional on the large scale atmospheric variables to simulate the local climate variables.

2.5 Classification and Regression Trees

In machine learning, classification is one of the supervised techniques, which can acquire an algorithm learned from training data. The main task of this technique is to predict the value from any input data after learning the relationship between pairs of input and target output data. Therefore, this technique is suitable for the use of statistical downscaling method due to the involvement of both local and atmospheric variables. CART is a model that refers to both classification and regression trees, which has been widely used in

predicting the rainfall occurrences and amounts, respectively, as a function of the coarse resolution atmospheric variables. There are several advantages of using CART model (Kannan and Ghosh, 2011; Mandal et al., 2016):

- (i) No prior statistical distribution assumption on predictor variables is made or followed;
- (ii) Predictor variables can be a combination of discrete, continuous and categorical;
- (iii) Good in recognizing and revealing the interactions in the dataset;
- (iv) Flexible and efficient with high dimensional data.
- (v) Invariant under monotonic transformation of independent variables;

Based on the present studies, the occurrence of daily rainfall state was first generated using a classification technique, such as K-means clustering, coupled with the CART model. Then the generation of daily rainfall amount conditioned on rainfall state was done using a nonparametric kernel regression model (Kannan and Ghosh, 2013) and a beta regression model (Mandal et al, 2016). Both models showed the good abilities in predicting the spatial and temporal variability of rainfall by establishing a statistical relationship between the atmospheric variables and the observed rainfall. Ingsrisawang et al. (2008) applied the decision tree in comparison with the ANN and SVM, to prediction the short term rainfall in the northeastern part of Thailand. Firstly, the rainfall occurrence was classified by a C4.5 decision-tree induction model and achieved up to 94.41% accuracy in overall during the cross validations. Thereafter, the model was used to classify the rainfall amount into three classes, and compared the performance with ANN and Support Vector Machine (SVM). The overall

accuracy of decision tree, ANN and SVM were 62.5%, 68.15% and 69.10%, respectively. Both ANN and SVM exhibited the better performance in classifying the rainfall amount, while the decision tree performed well in classifying the occurrence of rainfall. Tsai, et al. (2012) also combined the CART and ANN techniques to predict the water stages during typhoons in Tanshui river basin, Taiwan. The combined CART-ANN model consisted of two-step process, which the CART model was used to classify the river stages into three levels, followed by the ANN model to predict the water stages

Nevertheless, bagging and boosting are two popular ensemble machine learning techniques designed to overcome the problems of weak prediction, done by single decision tree. Bagging is also known as bootstrap aggregating, which is the method is proposed by Breiman (1996) to reduce the prediction error of learning machines. In the study of Gaitan, et al. (2014), bagging of classification trees (tree ensemble) were employed to model the rainfall occurrences in southern Ontario and Quebec. The tree ensembles outperformed other models in simulating rainfall occurrence with the highest Pierce skill score (PSS). The results also showed the application of bagging on classification tree to improve the performance of a single classification tree. In addition, the study of Li et al. (2010) proved the use of bagging on the SVM based downscale model solved the uncertainties of model estimation and reduced the variance of prediction based on just one set of parameters. There were only 100 times of bootstrap resampling adopted in their study instead of using 500 samples which did not remarkably reduced the variance given the validation period only ten years long.

Boosting is a sophisticated version of bagging by producing a linear combinations out of many models for supervised learning and each model is dependent on the preceding models (Elith et al., 2008). The similarity of both methods is to combine the outputs from different predictors, but the permutation of training data and the combination of predictions are the ways to differentiate them. Aggregated boosted trees are the extension of boosted trees, which consist of a collection of boosted trees generated on a cross-validation subset, able to reduce the prediction error corresponding to a single boosted tree (De'ath, 2007). Tisseuil et al. (2010) employed three nonlinear and one linear statistical downscaling models to downscale the streamflow data of 51 stations in the southwest of France. The simulation results of fortnightly flow showed the nonlinear models performed better than linear model. However, the aggregated boosted trees showed the higher stability with less variables and higher R^2 values among the nonlinear models in downscaling the hydrological variability.

Random forests (RF) are a non-parametric and ensemble learning algorithm with the combination of tree predictors in the way that each tree is depending on the values of a randomly chosen subset of input variable sampled independently. Breiman (2001) also emphasized the distribution of all trees in RF are similar. Besides, in considering the ability of RF in providing the high prediction accuracy and robust against the overfitting problem, it has been increasingly used in the studies of ecological, climatic and many other fields (Chan and Paelinckx, 2008; Ibarra-Berastegi et al., 2011; Stumpf and Kerle, 2011; Vincenzi et al., 2011; Bucklin et al., 2013). RF is generated based on the CART technique and deemed as an improvement over the bagged trees. Jing et

al. (2016) downscaled the monthly precipitation over north China using k-nearest neighbour (k-NN), SVM, CART and RF in their study. They found that SVM- and RF-based models performed the highest accuracy results, followed by CART and k-NN models, but the good performance of SVM-based model was exhibited after the application of residual correction. Besides, the errors obtained from downscaling were found to be increased with the monthly total precipitation, but RF-based model showed itself to be less affected by this proportional relationship.

In the recent studies, RF models have been employed to perform the spatial downscaling of monthly precipitation and land surface temperatures (Hutengs and Vohland, 2016). According to the study of Shi and Song (2015), they employed a non-parametric regression model with random forest algorithm to downscale the fine-spatial resolution precipitation at Tibetan Plateau. The proposed model exhibited the good performance with the R^2 value of 0.98 in simulating the observed precipitation. Besides, they realized that the implementation of disaggregation method with calibration, showed no obvious improvement on the simulation of monthly precipitation, but able to reduce the bias and mean absolute error for major seasons. He et al. (2016) proposed an adaptable random forests model to downscale the precipitation, by establishing a nonlinear relationship between local variable and large-scale/fine resolution predictors. They used 50 individual decision trees in their study for the purpose of enhancing the computational efficiency and stability of prediction, due to no further reduction of out-of-bag errors was detected after 20 trees and the involvement of additional trees was deemed to be unnecessary. However, the

poor performance of proposed model in under-predicting the spatial variability and temporal dependence, and the frequency of extreme rainfall rates, were due to the weaknesses of methodology and the insufficient predictors used to downscale the local variable.

2.6 Hybrid Models

Nowadays, the development and application of hybrid models have been arisen as a hydrological model to predict or downscale the rainfall due to inefficiency or poor performance of traditional individual models. Therefore, the hybrid models tend to overcome the problem through the combination with another model, which can overcome the limitation of conventional model. The statistical downscaling model (SDSM) is the best example of existed hybrid models, because this model is frequently used in downscaling application due to its good performance and low cost usage. SDSM unify the statistical downscaling methods, namely stochastic weather generator and multiple linear regression techniques in downscaling the climate variables. There are few studies in Malaysia showed that the performance of ANN was degraded when it was compared with other statistical downscaling models, such as SDSM (Amirabadizadeh et al., 2016) and genetic programming-based logistic regression model.

In the study of Amirabadizadeh et al. (2016), SDSM and ANN model were used to downscale and project the local climate variables under the emission scenario of A2 from CGCM3.1. Based on the results, both models

showed more ability in simulating the temperatures, when compared to their reasonably poor performance in predicting the variability of precipitation. However, the SDSM exhibited the better performance in reproducing the observed local climate variables with smaller errors than ANN model. In terms of precipitation, the ANN model over-predicted the wet-spell length and average monthly precipitation, while SDSM is capable of predicting the wet- and dry-spell lengths. Hence, the SDSM was then used to project the future temperature and precipitation over the Langat River Basin, Malaysia.

Harpham and Wilby (2005) used SDSM to generate the daily precipitation series for a single site and then extended to a multi-site simulation of precipitation through a conditional resampling approach. The performance of SDSM was compared with another two statistical downscaling models, namely radial basis function (RBF) ANN and multilayer perceptron (MLP) ANN, in downscaling the heavy precipitation occurrence and amounts. Another hybrid downscaling model was developed by Jeong et al. (2013) to downscale the daily precipitation using the global scale of GCM precipitation outputs in their study. The proposed hybrid model was a multi-site statistical downscaling model (MSDM), combining the first-order Markov chain and probability mapping approaches to reproduce the precipitation occurrences and amount, respectively. The single-site results analysis showed the MSDM sufficiently reproduced the precipitation occurrence lag-1 autocorrelation and the standard deviation of wet-day precipitation amounts. Besides, the MSDM was also capable of reproducing the daily precipitation occurrence and amount series.

SVM is a machine learning technique, which is similar to ANN, and able to describe the highly nonlinear relationship between global atmospheric variables and local climate variables. SVM exhibits better performance than the conventional linear regression-based downscaling model, but the overtraining of data is still one of the weaknesses of SVM. Ghosh (2010) proposed a hybrid model, using SVM combined with a probabilistic search technique, namely Probabilistic Global Search Algorithm (PGSL) for parameter selection, in order to minimize the overtraining to possible level. The proposed model achieved the Nash-Sutcliffe coefficient of 0.65, which was acceptable. However, they pointed out the major weakness of SVM is that their outputs are point estimate, which was unable to develop the conditional distribution of predicted variables with the given inputs.

There was a hybrid WNN model developed using the wavelet technique to combine with ANN for monthly rainfall prediction (Venkata Ramana et al., 2013). The proposed WNN model outperformed the ANN model with higher efficiency index. In the application of weather downscaling, wavelet transform is usually combined with weather generator to pre-process the input data for the purpose of improving the performance of model. For example, the wavelet transform and support vector machine hybrid model is developed for reservoir inflow prediction under GCM scenario, by decomposing the NCEP & NCAR reanalysis predictors (Halik et al., 2015). According to the study of Rashid et al. (2016), the continuous wavelet transform is applied to decompose and reconstruct the rainfall and NCEP reanalysis data before feeding into a downscaling model. The results from both examples showed the hybrid models

exhibited better performance in simulating the observed rainfall data, when compared to the original models without wavelet transformation.

There is another study of combining the Generalized Linear Model (GLM) and ANN to simulate the future precipitation in North Western England. The main reason that Abdellatif et al. (2013) proposed the application of GLM was to enhance the performance of traditional ANN model using a resampling scheme. The development of this hybrid GLM-ANN was consisted of two stage processes, where the GLM was used to model the rainfall occurrence, followed by ANN model used to downscale the rainfall amount with a Levenberg-Marquardt approach. The resampling scheme was applied on the observed rainfall data based on the rainfall occurrence model, to form a resampled observed rainfall series. Based on the results, the hybrid GLM-ANN exhibited the better simulation performance than the traditional ANN model due to the inclusion of additional rainfall occurrence model. Besides, Osman and Abdellatif (2016) declared that the approach of combining the prediction from different statistical downscaled models into ANN, not only outperformed the SDSM, but also improved the performance of traditional ANN in reproducing the observed rainfall.

2.7 Summary

The frequent occurrences of extreme hydrological events are one of the consequences triggered by climate change, which can bring the significant impacts on the productivity of crops, occurrence of floods and droughts, and the

supply of water resources and a host of calamities around the world. Therefore, the future climate change studies have become relatively important among the researchers to understand and minimize the impacts. The statistical downscaling models have been developed and utilized extensively in downscaling the future climate data, using the future variables projected by GCMs. There are two main reservoirs within the Langat River Basin supply the water resources in Selangor state. The level of floods in Langat River Basin is classified as high risk and more than two thirds of populations in Selangor state live in those flood prone areas. The simulation performance of a statistical downscaling model is very important, because this may directly affect the reliability of downscaled future data. To have full control of the future water resources and consider climate change vagaries, it is important that the prediction of temporal and magnitude as well as spatial parameters of the precipitation is accurate.

CHAPTER 3

METHODOLOGY

3.1 Study Area

Langat River Basin, the chosen study area for this research, is situated at the south of Klang Valley. The exact coordinates of this river basin are the latitudes from 2° 40' 152'' N to 3° 16' 15''N and the longitudes from 101° 19' 20''E to 102° 1' 10''E. The whole catchment area of Langat River Basin covers approximate 2,282km². Langat River is the main stream, which flows from the main range (Banjaran Titiwangsa) in south-western direction into the Malacca Strait. Langat River consists of several tributaries, with the Beranang River, Labu River and Semenyir River being the main one. The Langat River Basin is the important and most urbanized catchment area in Malaysia, with an approximate 1.2 million population.

There are two main reservoirs within the basin, namely the Langat Reservoir and the Semenyih Reservoir. The Semenyih Reservoir was constructed in year 1982, right after the construction of the Langat Reservoir in year 1981, with catchment area with 41 km² and 54 km², respectively. The main function of both reservoirs is supplying the water for industrial and domestic uses, and the areas covered include Cheras, Hulu Langat, Kuala Langat, Petaling, Putrajaya and Sepang. Both reservoirs are also important in controlling the flood

discharges, particularly at the downstream of Langat River. The Langat Reservoir also serves to supply a hydro power plant, supplying the power with moderate capacity for the population within Langat Valley.

The climate condition in Langat River Basin is characterized as equatorial monsoon. An equatorial climate can be defined as a climate condition with the characteristics of high rainfall, high average and uniform annual temperature, and high humidity. There are two major types of monsoon in Peninsular Malaysia, which are the Southwest monsoon and the Northeast monsoon. Generally, Langat River Basin experiences two major monsoon seasons (May to September and November to March) and two inter-monsoon seasons (April and October) every year. The climate of this area is strongly depending on the Southwest monsoon as the wind blows across the Straits of Malacca. Therefore, this area can be a flood prone area during the period of that monsoon as it receives huge amount and high frequency of rainfall.

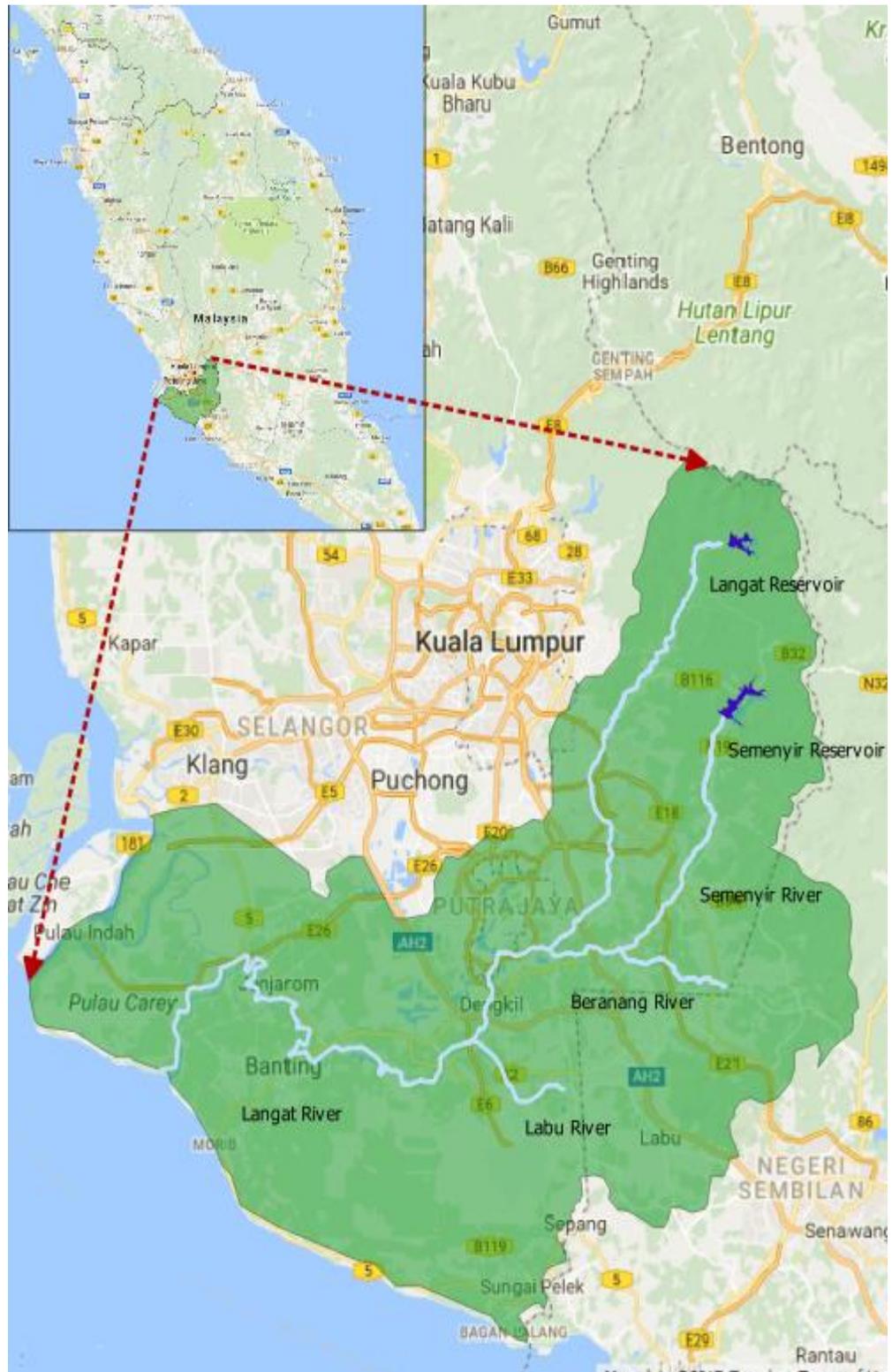


Figure 3.1: Langkat River Basin in Peninsular Malaysia

3.2 Workflow of Study

The flowchart of work in this study is presented in Figure 3.2.

Rainfall Occurrence Models

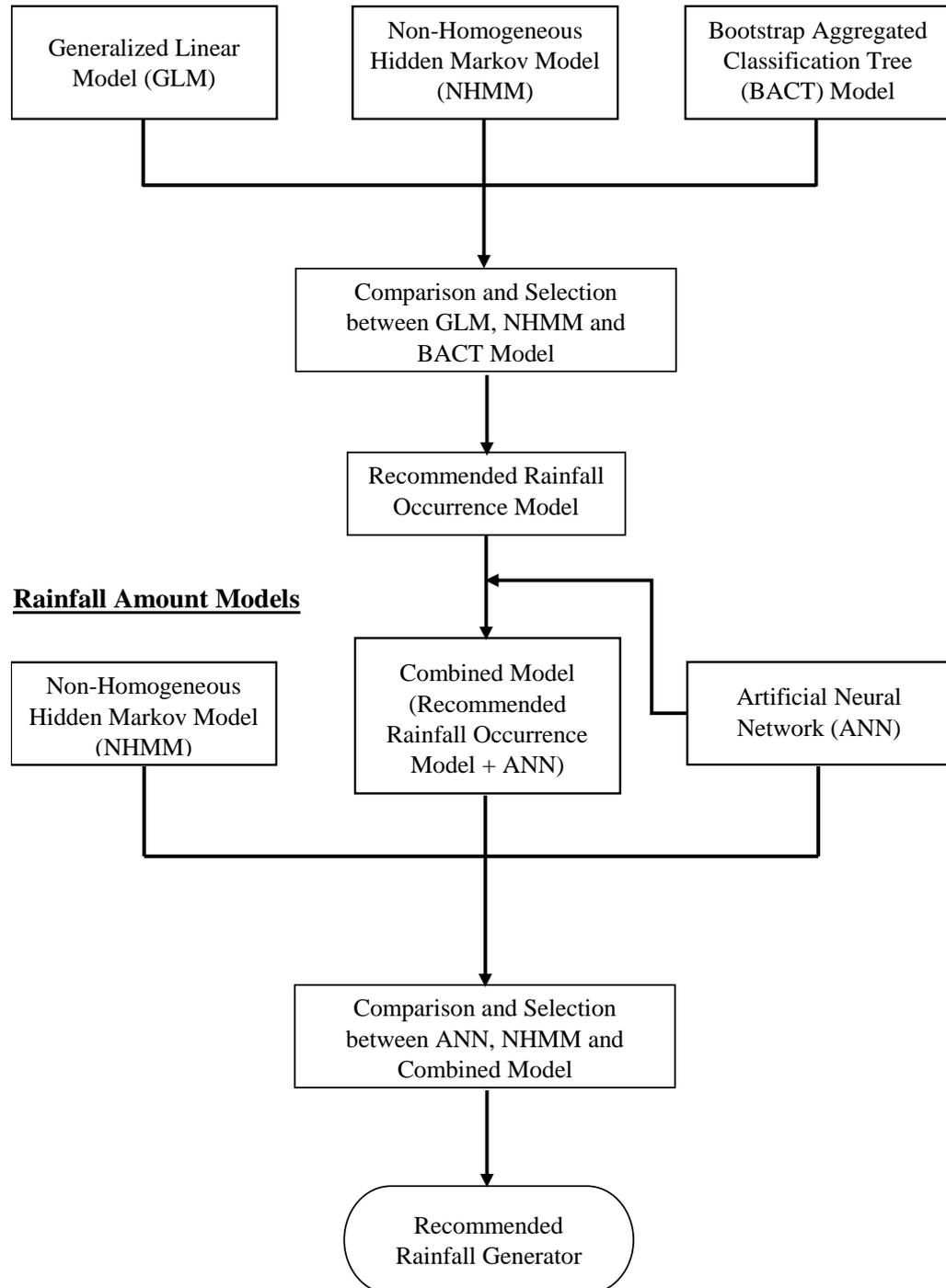


Figure 3.2: Flowchart of work

3.3 Data Collection

The NCEP & NCAR reanalysis dataset and the observed rainfall series, are the data required in this study for modelling. Both data are in global- and local-scales, respectively.

3.3.1 Observed Rainfall Series

The observed rainfall series of selected stations within the Langat River Basin were acquired from the Malaysian Meteorological Department (MMD) and the Department of Irrigation and Drainage Malaysia (DID). In order to achieve the accurate and reliable downscaling results, the minimum requirement for the duration of observed rainfall series is 30 years. Four rainfall stations (numbered as 2815001 2913001, 2917001 and 3118102) with the collected rainfall series from years 1975–2012, were selected in this study for further investigation and their location within Langat River Basin are presented in Figure 3.3. The details of selected rainfall stations with their station name, coordinates and the periods of collected rainfall series, were tabulated in Table 3.1.

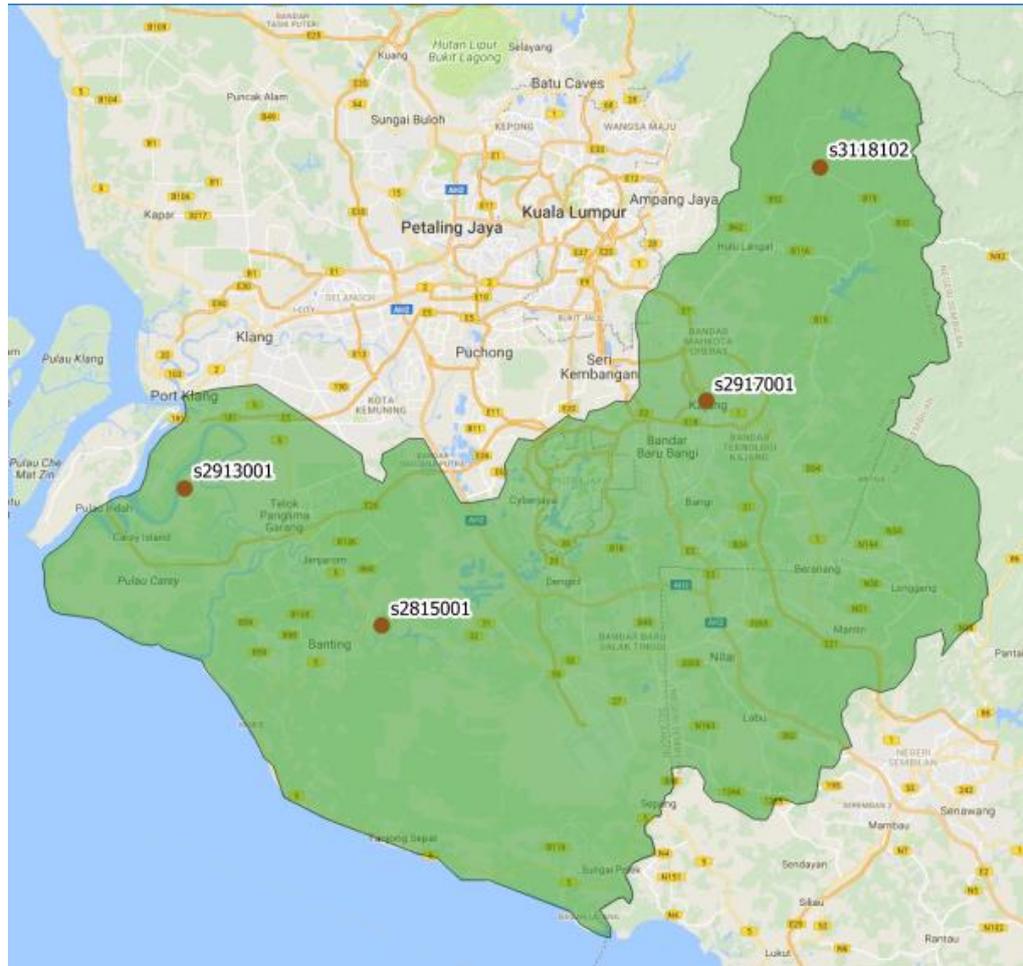


Figure 3.3: Location of selected rainfall stations at the Langkat River Basin, Malaysia

Table 3.1: List of station code, name, coordinate and study period

Station code	Station name	Longitude	Latitude	Study period (years)
2815001	Pejabat JPS Sg. Manggis	101°32'E	2°49'N	1975-2012 (38)
2913001	P/KWLN/S Telok Gong	101°23'E	2°55'N	1975-2012 (38)
2917001	RTM Kajang	101°47'E	2°59'N	1975-2012 (38)
3118102	Sek. Keb. Kg. Sg. Lui	101°52'E	3°10'N	1975-2012 (38)

3.3.2 NCEP & NCAR Reanalysis Dataset

Another type of large scale data required in this study, is the National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP & NCAR) reanalysis dataset. The NCEP & NCAR reanalysis dataset was acquired according to the location of selected stations based on the grids box showed on the map of Peninsular Malaysia. This reanalysis dataset represents the condition of the atmosphere, hence the purpose of obtaining this dataset is to establish a relationship between the large scale atmospheric variables and local rainfall series using a statistical downscaling model. These potential predictors are the daily values of twenty-six variables from years 1961–2005. Aside from the circulation variables (i.e. geopotential height and wind component), they also comprise of the temperature, precipitation and moisture variables (specific humidity).

The description of potential predictor variables from NCEP & NCAR reanalysis dataset are presented in Table 3.2. For consistency between the observed rainfall series and the reanalysis dataset, their daily data from years 1976–2005 were extracted out from their respective raw dataset for modelling. Based on the study of Amirabadizadeh et al. (2016), they showed the lagged predictors improved their correlations with the observed daily rainfall data in Langat River Basin. Therefore, each variable had undergone the lag-transformation process (from lag -9 to lag 9) for the purpose of selecting the suitable variables, which are highly correlated with the observed rainfall series, for each station.

Table 3.2: Description of predictor variables from NCEP & NCAR dataset

No	Predictor Variables	Description
1	msslpgl	Mean sea level pressure
2	p1_fgl	Geostrophic airflow velocity at 1000hPa
3	p1_ugl	Zonal velocity components at 1000hPa
4	p1_vgl	Meridional velocity componets at 1000hPa
5	p1_zgl	Vorticity at 1000hPa
6	p1thgl	Wind direction at 1000hPa
7	p1zhgl	Divergence at 1000hPa
8	p500gl	Geopotential heigh at 500hPa
9	p5_fgl	Geostrophic airflow velocity at 500hPa
10	p5_ugl	Zonal velocity components at 500hPa
11	p5_vgl	Meridional velocity componets at 500hPa
12	p5_zgl	Vorticity at 500hPa
13	p5thgl	Wind direction at 500hPa
14	p5zhgl	Divergence at 500hPa
15	p850gl	Geopotential heigh at 850hPa
16	p8_fgl	Geostrophic airflow velocity at 850hPa
17	p8_ugl	Zonal velocity components at 850hPa
18	p8_vgl	Meridional velocity componets at 850hPa
19	p8_zgl	Vorticity at 850hPa
20	p8thgl	Wind direction at 850hPa
21	p1zhgl	Divergence at 850hPa
22	prcpgl	Accumulated precipitation
23	s500gl	Specific humidity at 500hPa
24	s850gl	Specific humidity at 850hPa
25	shumgl	Near surface humidity
26	tempgl	Mean temperature at 2m

3.4 Normality Tests

There are few tests can be adopted for checking the normality of data. However, the Kolmogorov-Smirnov test is rejected in this study due to its high sensitivity to extreme values. Besides, the Shapiro-Wilk test was not recommended in this study, as this test is more appropriate for the samples less than 5,000 observations. Three common normality tests were adopted in this study, namely the Anderson-Darling test, the Lilliefors test and the Jarque-Bera

test. These normality tests were applied on the observed rainfall series under the significance level of 0.05, for the purpose of checking whether they are normally distributed at each station. The null hypothesis of these tests, which the sample follows a normal distribution, was accepted when the computed p-value is greater than 0.05. Otherwise, the null hypothesis can be rejected to indicate that the sample does not follow a normal distribution pattern.

3.4.1 The Anderson-Darling Test

The Anderson-Darling test is one of the normality tests, which can be considered as a modified version of the Kolmogorov-Smirnov test and is applicable to numerous distributions, such as normal, uniform, lognormal, exponential, Weibull and logistic distributions, under the condition of the parameters are not known and have to be estimated (Pettitt, 1997; Engmann and Cousineau, 2013). Besides, the Anderson-Darling test is a powerful and sensitive test, as it utilizes the specific distribution to compute the critical values, which can be expressed as:

$$AD = -n - \frac{1}{n} \sum_{i=1}^n (2i - 1) [\ln F(x_i) + \ln(1 - F(x_{n-i+1}))] \quad (3.1)$$

The n represents the sample size and F represents the cumulative distribution function. While, the x_i denotes the i^{th} sample arranged in ascending order.

3.4.2 The Lilliefors Test

The Lilliefors test is another improvement version of the Kolmogorov-Smirnov test to test for normality according to the mean and standard deviation of sample. Generally, there are few steps to conduct the Lilliefors test as follows:

- 1) Calculation of sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.2)$$

and standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n x_i - \bar{x}} \quad (3.3)$$

- 2) Calculation of normalized sample values, Z_i

$$Z_i = \frac{x_i - \bar{x}}{s}, i = 1, 2, \dots, n \quad (3.4)$$

- 3) Calculation of test statistic

$$T = \sup |F^*(x) - S(x)| \quad (3.5)$$

where, T denotes the supremum of the absolute difference between the $F^*(x)$ and $S(x)$, while $F^*(x)$ denotes the cumulative distribution function of a normal distribution, where the $\bar{x} = 0$ and the $s = 1$. $S(x)$ represents the empirical distribution function of Z_i .

3.4.3 The Jarque-Bera Test

Instead of comparing the differences between empirical distribution function and theoretical normal cumulative distribution function, the Jarque-Bera test prefers to test the normality based on the skewness and kurtosis of sample to match a normal distribution. The test statistic of Jarque-Bera test can be defined as:

$$JB = \frac{N}{6} \left(S^2 + \frac{(K-3)^2}{4} \right) \quad (3.6)$$

where, N represents the sample size, and S represents the sample skewness. While, K is denoting the sample kurtosis.

3.5 Homogeneity Tests

The detection on the variability of collected data in terms of homogeneity is an important issue to ensure its reliability before they are used as an input in the hydrologic analysis models. Generally, the instruments and environments are two main keys to ensure the data to be homogenous, as the measurements of data must be taken at a time with the constant instruments and environments. However, the changes in the techniques of measurement and the observational procedures, characteristic and structures of environment, and location of stations always are the main problems in affecting the homogeneity of observed rainfall data, which make the task of obtaining a set of rainfall data with good quality become more challenging. Hence, prior to the use of the collected data in any

hydrologic modelling process, homogeneity test is an essential step need to be carried out for ensuring the reliability of observed data.

There are several homogeneities tests have been used to detect the non-homogeneity of data, such as bivariate test, non-parametric Mann-Kendall test, non-parametric multi-response permutation procedure and others. In Malaysia, the application of homogeneity tests is an inevitable step in analysing the rainfall series. The checking on homogeneity of data have been done in the majority of studies using the tests proposed by Wijngaad et al. (2003). In this study, the observed monthly rainfall series were checked using four homogeneity tests, such as the standard normal homogeneity test (SNHT), the Buishand range test (BR), the Pettitt test (PET) and the Von Neumann ratio test (VNR). These tests were highly recommended in this study rather than using the relative test (Peterson-Easterling test), which require the strong correlation between the stations, due to the consideration of the sparse distribution of station network within the Langat River Basin. Besides, absolute tests is more sensitive than relative tests in detecting the breaks in rainfall series, which are frequently caused by the simultaneous changes in observational routines. So, all four homogeneity tests were utilized separately on the observed monthly rainfall data.

Under the null hypothesis created in these four tests, the data series are determined to be homogenous when the annual values of testing variables are independent and distributed equivalently. At the same time, the alternative hypothesis indicates the data series is non-homogeneous with the presence of break in the data. After the application of homogeneity tests, the criteria used to

evaluate the results of homogeneity tests in this study are modified from the study of Wijngaad et al. (2003) under the significance level of 0.05, as shown in Table 3.3.

Table 3.3: Classification of homogeneity tests' results

Category	Class	Condition	Remarks
Class 1	Useful	The null hypothesis is rejected by none or one test	Absence of non-homogeneity
Class 2	Doubtful	The null hypothesis is rejected by two tests	Presence of non-homogeneity
Class 3	Suspect	The null hypothesis is rejected by three or four tests	Presence of non-homogeneity

3.5.1 The Standard Normal Homogeneity Test (SNHT)

As mentioned in above, SNHT presumes the presence of a step-wise shift in the mean under the alternative hypothesis, however the breaks happened near the beginning and at the end of a series is more preferred to be detected by SNHT. The normal distribution of annual values of the testing variable is the assumption made in this test. The comparison between the mean of first y years and of last $(n - y)$ years was made in the statistic equation:

$$T_y = y\bar{z}_1 + (n - y)\bar{z}_2, y = 1, 2, \dots, n \quad (3.7)$$

where

$$\bar{z}_1 = \frac{1}{y} \sum_{i=1}^y \frac{(Y_i - \bar{Y})}{s} \quad \text{and} \quad \bar{z}_2 = \frac{1}{n-y} \sum_{i=y+1}^n \frac{(Y_i - \bar{Y})}{s} \quad (3.8)$$

The Y_i denotes the annual value of testing variable, and i denotes the number of years from 1 to n , while \bar{Y} and s indicate the mean and standard

deviation, respectively. When the break is detected to be present in the year of y , the maximum value of T_y will be obtained, where the statistic is expressed as:

$$T_0 = \max T_y \quad (3.9)$$

However, SNHT was applied under the significance level of 0.05 in this study. The null hypothesis is accepted when the computed p-value is greater than 0.05, which indicate the rainfall series is homogeneous. While, the null hypothesis is rejected when the computed p-value smaller than 0.05 and the rainfall series is treated as non-homogeneous.

3.5.2 The Buishand Range Test (BR)

The assumption of null hypothesis, alternative hypothesis and the normal distribution of annual values under the Buishand range test (BR) are similar to the SHNT. However, the only difference between both of them is the sensitivity in detecting the break, where the BR test is more efficient on the middle of a series, instead of either near beginning or at the end of a series. The adjusted partial sum under this test is explained as:

$$S_0^* = 0 \text{ and } S_y^* = \sum_{i=1}^y (Y_i - \bar{Y}), y = 1, 2, \dots, n \quad (3.10)$$

The homogeneity of series is determined according to the value of S_y^* , where the series is treated as homogenous when the value obtained by S_y^* is at the range close to zero. In contrast, the value is achieved at a maximum (negative shift) or minimum (positive shift), when there is a break in year y . After a break

in year y is detected, then rescaled adjusted range, R is applied in order to compute the difference between maximum and minimum value of the rescaled S_y^* .

$$R = \frac{\max_{0 \leq y \leq n} S_y^* - \min_{0 \leq y \leq n} S_y^*}{s} \quad (3.11)$$

Similar to SNHT, the null hypothesis is accepted when the computed p-value is greater than 0.05, which indicate the rainfall series is homogeneous. While, the null hypothesis is rejected when the computed p-value smaller than 0.05 and the rainfall series is treated as non-homogeneous.

3.5.3 The Pettitt Test (PET)

The computation of Pettitt test (PET) is based on the ranking order of data time series instead of making assumption on the normally distribution of annual values, which make this test differs from SNHT and BR tests. The sensitivity of PET in detecting the break in a time series is similar to BR test, which has the satisfactory results in detecting the break happened at the centre of time series. The less sensitivity to the outliers is the disadvantage of this test, when compared to other tests. The calculation of PET is expressed as:

$$X_y = 2 \sum_{i=1}^y r_i - y(n+1), y = 1, 2, \dots, n \quad (3.12)$$

The r_i is denoting the rank of i^{th} observation of the testing variable. When the value of X_y achieves the maximum or the minimum value, then the break is determined at the year of k , which can be explained by following equation:

$$X_k = \max_{1 \leq y \leq n} |X_y| \quad (3.13)$$

The rejection of null hypothesis of this test also depends on its computed p-value under the significance level of 0.05.

3.5.4 The Von Neumann Ratio Test (VNR)

Unlike the three homogeneity tests discussed in above, the Von Neumann ratio test (VNR) has the tendency to determine the random distribution of the series, rather than detect the breaks in the series. This test can be explained as the mean square successive difference divided by the variance, which can be expressed as:

$$N = \frac{\sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3.14)$$

Theoretically, the series is treated as homogeneous when the computed N value is two. However, the series is non-homogeneous and a break is detected when the computed N value smaller than two. The rapid variation is existed in the mean of series when the computed N value greater than two. In this study, the determination of series to be homogeneous is depending on the computed p-value in this PET test to accept or reject the null hypothesis.

3.6 Screening of Predictors

A complete set of the twenty-six potential predictors contains plenty of information and almost all of them are mutually correlated, but still part of them may be treated as excessive information since they can't provide any significant effect in the downscaling process. The selection of predictor variable(s) is one of the most challenging steps in developing a statistical downscaling model because the characteristics of the downscaled scenario are directly influenced by this decision. In order to prevent the problem of overfitting, the selected predictors should not be highly correlated, which has been proved by Liu et al. (2011) as the performance of model was generally enhanced when different types of variables were included. They found that the same type of variables with high inter-correlations were most likely to be selected if the selection criterion only focused on the explained variance of rainfall.

Thus, the selection criteria used in this study were the explained variance and the partial correlation analysis. The explained variance described the correlations between rainfall and potential predictors, while the partial correlation analysis not only represented the correlations between rainfall and predictors without the influence of other predictors, but also showed the associations between predictors themselves in term of p-value. Each variable and their lag-transformed variables (from lag -9 to lag 9) were screened with the observed daily rainfall series for selecting the most suitable set of predictors for each station. Firstly, the predictors were selected based on their correlations with the rainfall. Then, the selected predictors were proceeded to the partial

correlation analysis. The association among themselves was measured in term of p-value. In addition, the significance level was set to 0.05 in this study to explain the association between the variables. The higher correlation values indicate the higher degree of association, while the smaller p-values imply that associations are less likely to happen by chance.

3.7 Rainfall Occurrence Models

There were two rainfall occurrence models developed in this study, namely the generalized linear model (GLM) and the bootstrap aggregated classification tree (BACT) model. All the models were calibrated and validated using the required data in 20 years and 10 years, respectively.

3.7.1 Generalized Linear Model (GLM)

Logistic regression is an example of a generalized linear model, which has been extensively used in statistical downscaling studies for modelling the rainfall occurrence series using a set of suitable atmospheric predictors. The occurrence of wet and dry days in a series can be denoted as a binary series, which consists of the value of zero and one only.

In this study, the value of zero represents the dry days with rainfall amount equal to or smaller than 0.1 mm, while the value of one indicates the wet days with the rainfall amount greater than 0.1 mm. Then, the logistic regression

was utilized for the modelling of the probability of rainfall occurrence series. The probability P , which used to imply a day being a wet day, is the significant parameter in a logistic model. Let P_i indicates the probability of i^{th} day conditional on a vector X'_i , which could be expressed as:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = X'_i\beta \quad (3.15)$$

This can be revised in another form as below:

$$\left(\frac{P_i}{1 - P_i}\right) = e^{X'_i\beta} \quad (3.16)$$

where, e denotes the base of natural logarithms, while β denotes the estimated coefficients.

In this study, the GLM were developed in the Matlab platform using the code of *glmfit(binomial)* and *glmval(logit)*. There are several distribution parameters can be chosen for the development of GLM, such as normal, gamma, poisson and binomial distributions. However, the binomial distribution is adopted as the suitable distribution in this study due to the response variable (observed rainfall occurrence series) is in binary form. The normal distribution is suitable for the response variable with any real number and the gamma distribution for the response variable with any positive number. While, the poisson distribution is suitable for the response variable with any non-negative integer.

Both selected predictors and observed daily binary rainfall series at each station were fitted into GLM to model the occurrence of rainfall. The maximum likelihood estimation was used to select the parameters within logistic regression instead of using least squares estimation. Besides, there are several link function can be chosen to map between the mean and the linear predictors, namely identity (default for normal distribution), log (default for poisson distribution), logit (default for binomial distribution), probit and reciprocal (default for gamma distribution) links. The application of logit link was adopted in this study as it ensures the predicted value lies in the interval between zero and one, which is more interpretable and can be explained as the probability of a day to represent wet day. Since the predicted value is the probability of a day being wet day, the threshold value for the probability was set as 0.5 in this study. In other words, the day with the probability equal to or smaller than 0.5 was marked as zero to represent dry day, while the day with the probability greater than 0.5 was marked as one to represent wet day. The flowchart of development of GLM is presented in Figure 3.5.

3.7.2 Bootstrap Aggregated Classification Tree Model (BACT)

A decision tree is formed using a binary recursive partitioning algorithm, which is an iterative data splitting process, to split the data into child nodes with certain condition and repeat itself until certain rule is met or no further splitting is possible (Kannan and Ghosh, 2011). In this study, the classification decision tree was selected to model the rainfall occurrence instead of using regression

decision tree, as the rainfall occurrence state could be treated as a classification type problem, with the condition of predicting the value of response variable (rainfall state) from a combination of predictor variables.

Classification decision trees are a learning technique that splits the data from the combinations of predictors, which may be categorical and/or continuous, repeatedly into more homogeneous groups, to clarify the variations of a single response variable. Each group represents a typical value of the response variable, which is defined by the number of observations and the values of predictors (Tisseuil et al., 2010). The approach of this technique to perform accurate classification or prediction is based on the if-then logical conditions. A simple example of structural layout of a single classification decision tree used to predict the binary response, is given in Figure 3.4. The response variable (zero or one) is predicted follows the decisions made in the tree from the root node (beginning) down to another root node, until meet a leaf node, which contains the response.

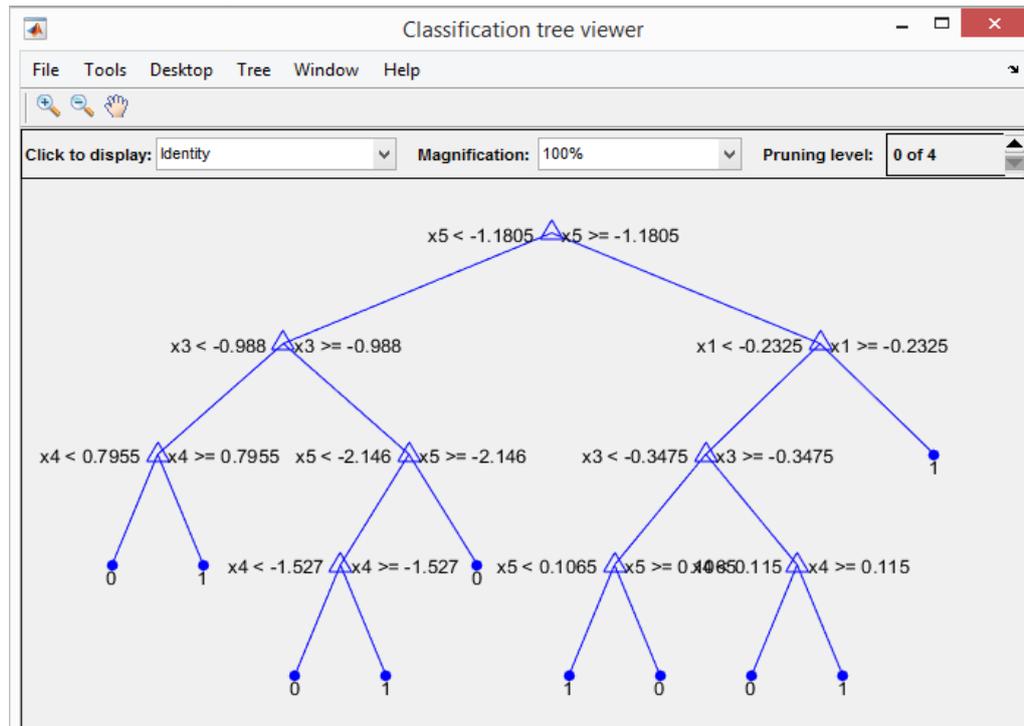


Figure 3.4: Structural layout of a single classification decision tree

However, the bootstrap aggregation (bagging) of classification tree was employed in this study to reduce the problem of overfitting and improve the generalization of an individual classification decision tree, with the classification results determined by the majority votes. The bootstrap aggregated classification tree models were developed in Matlab platform using the function code of *TreeBagger*, where each tree in the ensemble is produced on an independently drawn bootstrap imitation of input data. During the process of bootstrap aggregation, the number of predictor variables to be selected randomly for each node (*'NVarToSample'*) was set to *'all'* in function code to invoke the random forest algorithm, which can provide high prediction accuracy and high stability performance.

Basically, the random forest is similar to bagging with an additional randomness during the splitting of nodes (Breiman, 2001) and its algorithm can be defined as follows (Liaw and Wiener, 2002):

- (i) The *ntree* (number of trees) samples are randomly drawn from the original training sample with replacement.
- (ii) For each bootstrap sample, an unpruned classification tree is grown with the modification that splitting process of each node is based on the best among a subset of predictors randomly selected at that node, instead of using the best split among all variables.
- (iii) Prediction of new sample was made by taking the majority votes from all the individual classification trees.

In this study, there were a few parameters set in the process of building the model. For instance, the '*OObpred*' was set to '*on*' for determining the optimum number of trees grown in the BACT model. This can be achieved based on the classification error generated by out-of-bag (OOB) sample, which are the data not included in bootstrap sample. The OOB sample was predicted at each bootstrap iteration, using the tree grown with the bootstrap sample. Thereafter, the predictions of OOB sample were aggregated to compute the classification error. Therefore, the BACT model would be trained using the predetermined number of trees rather than using a larger number of tree, which may cause the intensive computation time. In addition, the prior probabilities for each class ('*prior*') was set to '*uniform*', where all class probabilities are equal. The nodes

were split (*'splitcriterion'*) based on their impurity using the formula of *'deviance'* for maximum deviance reduction:

$$D = - \sum_i p(i) \log p(i) \quad (3.17)$$

Where, $p(i)$ denotes the observed fraction of classes with the class i arrive at the node. Theoretically, a pure node has the deviance value equal to zero, or else, the value is positive.

Both selected predictors and observed daily binary rainfall series were fitted into the BACT model and trained with different positive integer value of *'minleaf'*, which can be defined as the minimum number of leaf node observation. Using the results of single classification tree as the references, the BACT model with the higher number of simulated wet days match with the observed, would be stored for further analysis. Thereafter, the well-trained BACT model were selected based on the measures of matching among the stored models during calibration (20 years) and validation (10 years) periods. The development of BACT model is presented in Figure 3.5.

3.7.3 Non-Homogeneous Hidden Markov Model (NHMM)

In this study, the NHMM was developed to simulate the observed rainfall amount series and the details of their development are explained in section 3.8. The final outputs obtained from the NHMM are the rainfall amount series, in which they are generated based on the type of fitted rainfall distribution and

conditioned on the states modelled by first-order Markov chain model and the selected predictors. In order to have further comparison on the performance of GLM and BACT model, the rainfall occurrence series of NHMM was visualized based on its simulated rainfall amount series. The days with rainfall values were marked as one to represent the wet days. Hence, the NHMM becomes third rainfall occurrence model for this study, as shown in Figure 3.5.

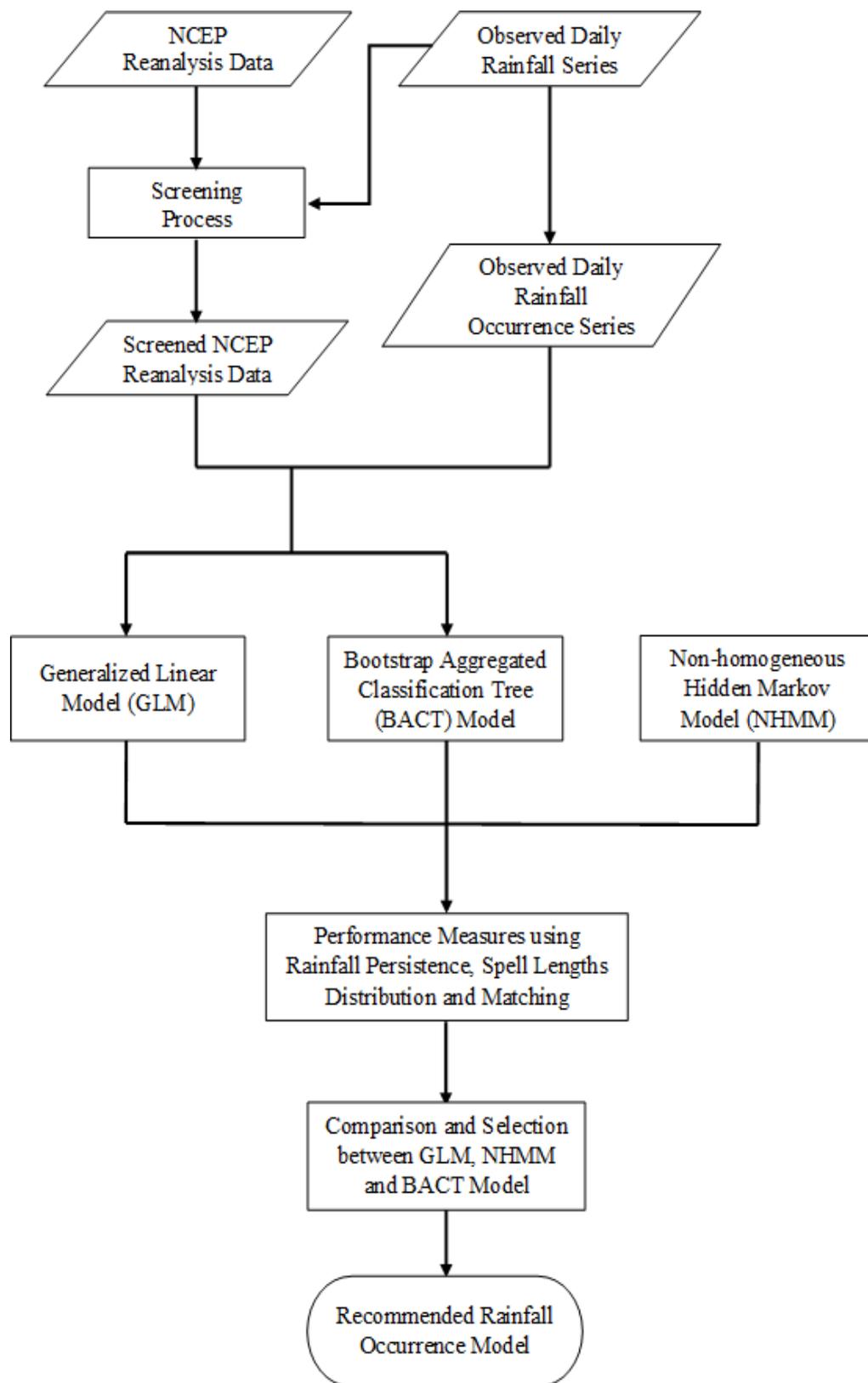


Figure 3.5: Flow chart of rainfall occurrence models

3.7.4 Goodness of Fit

In this study, the accuracy of GLM and BACT model were evaluated in terms of rainfall persistence, wet- and dry-spell length distribution, by comparing their simulated rainfall occurrence series with observed series. Both models were compared together with the performance measures of NHMM, which its rainfall occurrence series were visualized based on its simulated rainfall amount series. However, for a typical two-state rainfall generator, the generation of rainfall amount is conditional on the simulated wet days. Other than rainfall occurrence series, the development of these models in this study involved another set of predictors. The predictors corresponded to simulated wet days which would be extracted and further used in the modelling of rainfall amount model. Therefore, the matching between both simulated and observed wet days is very important, in order to extract the predictors correctly on the exacted wet days. The performance of these models were further evaluated and compared based on their matching between observed and simulated rainfall occurrence series.

3.7.4.1 Rainfall Persistence (RP)

Rainfall persistence (RP) can be simplified as the number of pair of consecutive rainy days divided by the total number of rainy days. In addition, the absolute differences between the observed and simulated rainfall persistence were computed to measure the accuracy of models during calibration (years

1976–1995) and validation periods (years 1996–2005). The advantage of this measure is to identify the ability of model in simulating the pair of observed consecutive rainy day. The smaller the absolute difference, the better is the simulation performance of a model.

$$RP = \frac{\text{number of pair of consecutive rainy days}}{\text{total number of rainy days}} \quad (3.18)$$

$$\text{Absolute Difference} = |RP_{\text{observed}} - RP_{\text{simulated}}| \quad (3.19)$$

3.7.4.2 Spell Lengths Distribution

The distribution of spell length is one of the common performance measures, which is very useful in evaluating the ability of a model in simulating the observed rainfall occurrence at a specific location. Wet-spell length can be explained as the number of consecutive rainy days, which can be used to examine the potential of flood occurrence. While, the dry-spell length is the number of consecutive non-rainy days and its distribution is the parameter specifically significant to the field of agriculture.

3.7.4.3 Matching

In this study, the performance of rainfall occurrence models were further evaluated in terms of the probability of detection (POD), false alarm rate (FAR), critical success index (CSI), Heidke Skill score (HSS) and Peirce Skill score

(PSS) indices. These indices were obtained from a 2x2 contingency table, with the following equations:

$$POD = \frac{a}{a + c} \quad (3.20)$$

The range of POD index is from 0 to 1, and one represents perfect score. POD is sensitive to hits, but ignore the false alarms.

$$FAR = \frac{b}{a + b} \quad (3.21)$$

The range of FAR index is from 0 to 1, and one represents perfect score. FAR is sensitive to false alarms, but ignore the misses.

$$CSI = \frac{a}{a + b + c} \quad (3.22)$$

The range of CSI index is from 0 for no wet days predictions are correct to 1 for all wet days predictions are correct. CSI sensitive to hits and penalizes both misses and false alarms.

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)} \quad (3.23)$$

The range of HSS is from -1 to 1, with the HSS = 1 indicates the perfect prediction and HSS = 0 indicates the prediction is random prediction. The prediction is worse and standard prediction is more accurate than prediction, when HSS < 1.

$$PSS = \frac{ad - bc}{(a + c)(b + d)} \quad (3.24)$$

The range of PSS is from -1 to 1. When $PSS > 0$, which mean the number of hits surpasses the false alarms and the prediction is good, and $PSS \leq 0$ indicates the prediction is random.

3.8 Rainfall Amount Models

Three rainfall amount models were developed in this study, a traditional ANN, the combine bootstrap aggregated classification tree-artificial neural network (BACT-ANN) model and the NHMM.

3.8.1 Artificial Neural Network (ANN)

The Multi-layered perceptron (MLP) is a widely used feed-forward neural network, which consists of one or more hidden layers between the input and output neurons. The number of neurons in input and output layers are strongly depending on the number of input and output data to be fitted into network. The function of input layer is to receive the input data and transfer them for further processing process in network. However, a nonlinear relationship existed between the data in input and output layers, so the neuron in hidden layer acts as a bridge to link both layers and connect them.

In this study, the ANN with feed-forward backpropagation algorithm was used to develop a relationship between selected predictors (NCEP & NCAR reanalysis dataset) and predictand (observed rainfall series). Theoretically, the

mechanism of backpropagation algorithm can be simplified as two flows. First flow is the forward pass, where the information of input data at input layer propagates through the network layer by layer, then a set of output data is produced and used as the actual response of the network. During this flow, all the weights used between the layers are fixed with a random number. Then, the second flow is the backward pass. The main purpose of this flow is to adjust the weights used in previous flow in corresponding to the rule of error correction. An error signal is generated based on the difference between the actual response of network and the desired response. After that, the error signal is propagated in the direction opposite to the linked connection of network, to make the adjustments on the weights, therefore the actual response of network can become even closer to its desired response.

Compared to the scaled conjugate gradient algorithm (*trainscg*), The Levenberg-Marquardt backpropagation algorithm (*trainlm*) was preferred as the training algorithm of neural network in this study, as it is designed to speed up the training process. This algorithm is also considered to be more efficient and skilful than the training algorithm of gradient descent (*traingd*). The gradient descent suffers from various convergence problems. Besides, the log-sigmoid transfer function was not recommended to be used in hidden layer as it constrains the output of a network in the range of 0 to 1, which is suitable for the pattern recognition problems. For a multilayer network, the suitable notation is that the tan-sigmoid transfer function (*tansig*) and linear transfer function (*purelin*) to be used in the hidden layer and output layer, respectively. This multilayer network able to approximate any function with a finite number of discontinuities

arbitrarily well, which is suitable for linking the predictors and predictand together.

Another critical task in training the ANN is to determine the number of hidden layer and neurons in hidden layer. Basically, the determination of most suitable number of hidden neurons is problem dependent, and there is no systematic method to be used in this process. Based on the previous studies (Goyal and Ojha , 2012; Mendes et al., 2014; Campozano et al., 2016) , the number of hidden neurons in a successful ANN model was determined using the method of trial and error, therefore the model will be tested with different number of hidden neurons until a satisfactory result to be obtained. Theoretically, the complex problem with numerous data sets and variables can be solved by increasing the number of hidden layers to establish a relationship between input and output layers, but the major disadvantage of using higher number of hidden layers is time consuming in running the ANN model. However, any measureable functional relationship between input and output can still be approximated accurately by only one hidden layer, as long as sufficient large number of neurons are given in the hidden layer.

In this study, one hidden layer was used in ANN for the purpose of minimizing the complexity of network, which can also reduce the time for running the network. Since the number of hidden layer was fixed, the only variable need to be determined was the number hidden neurons using the method of trial and error. The training of neural network was started with five hidden neurons. Once the results exhibited by neural networks were repeated or did not

improved, the number was increased with one hidden neurons for further training. Every neural network with a certain number of hidden neuron would be retrained about 100 times for the purpose of minimizing the effect of random weights on the training process of network, since the weights are randomly initialized for every time of network training. Using the results of the linear regression model as reference, the neural networks exhibited the smaller RMSE would be stored for further investigation. At the end, the neural network which exhibited the lowest RMSE among the stored neural networks in monthly rainfall series, was selected as the well-trained ANN in this study. The flowchart of ANN in this study is presented in Figure 3.6.

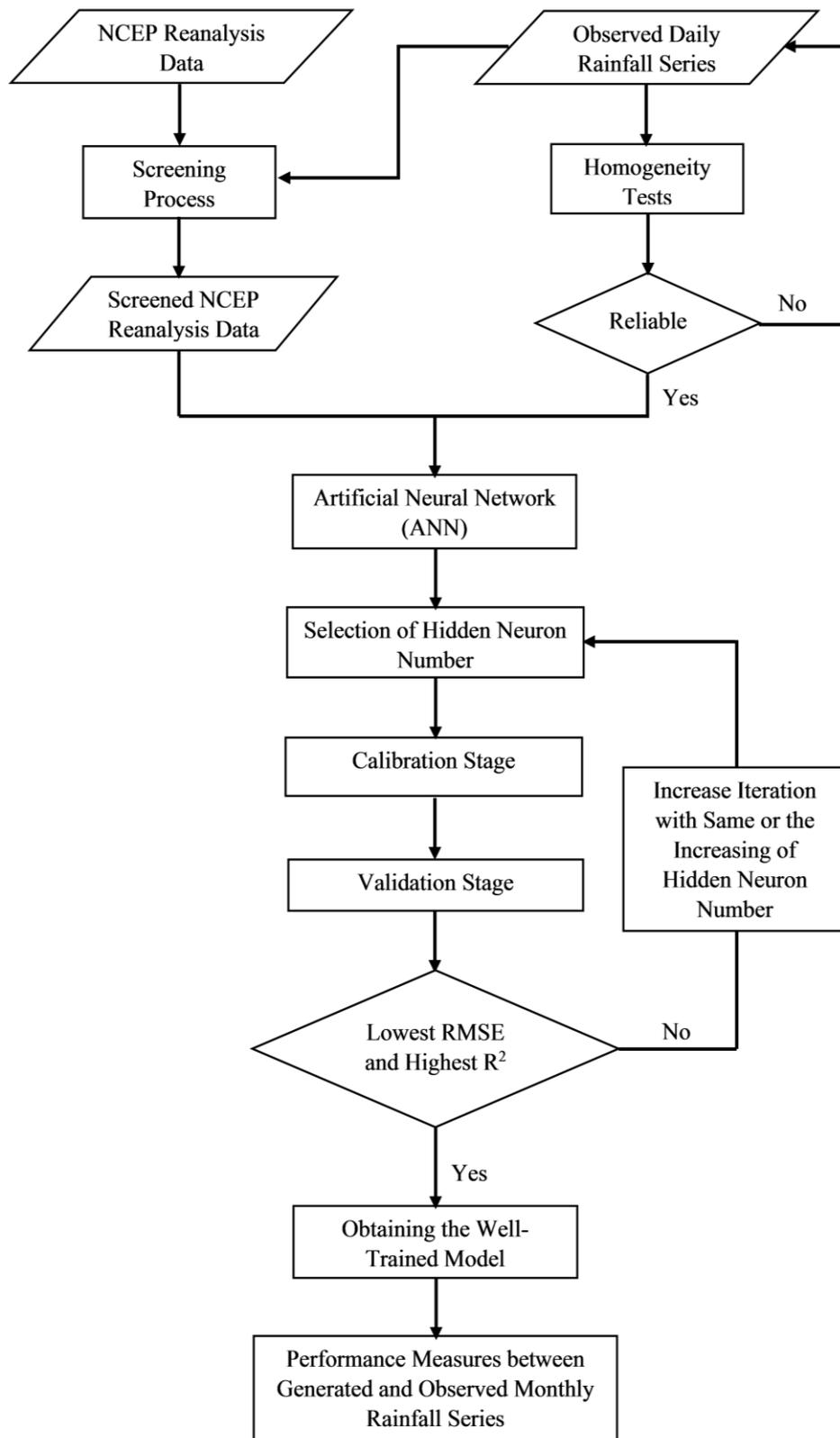


Figure 3.6: Flow chart of ANN

3.8.2 Combined BACT-ANN model

In this study, the approach of combining BACT and ANN was inspired by the studies of Abdellatif, et al. (2013) and Osman and Abdellatif (2016), to improve the performance of traditional ANN model. The excellent performance of the combining approach model is promising to be applied to improve the single model in downscaling the rainfall at a single site. Firstly, the bootstrap aggregated classification tree (BACT) used to model the rainfall occurrence, then combined with the ANN to model the rainfall amount condition on the simulated wet days. In addition, the approach of data-pre-processing was employed in this study to extract the selected predictors corresponding to the wet days simulated by BACT model and to ensure the ANN model the rainfall amount conditional on those wet days. Before the rainfall series was fitted into the ANN, it was pre-processed using the developed BACT model. Some of them may become zero amount even though the original rainfall series indicate them are wet days in reality. Thereafter, the whole wet days series (years 1976–2005) and their corresponding predictors were fitted into ANN model with the same steps as described in the development of a traditional ANN model, as shown in Figure 3.7. A well-trained combined BACT-ANN model was also selected based on the model which exhibited the lowest RMSE among the stored models in monthly rainfall series.

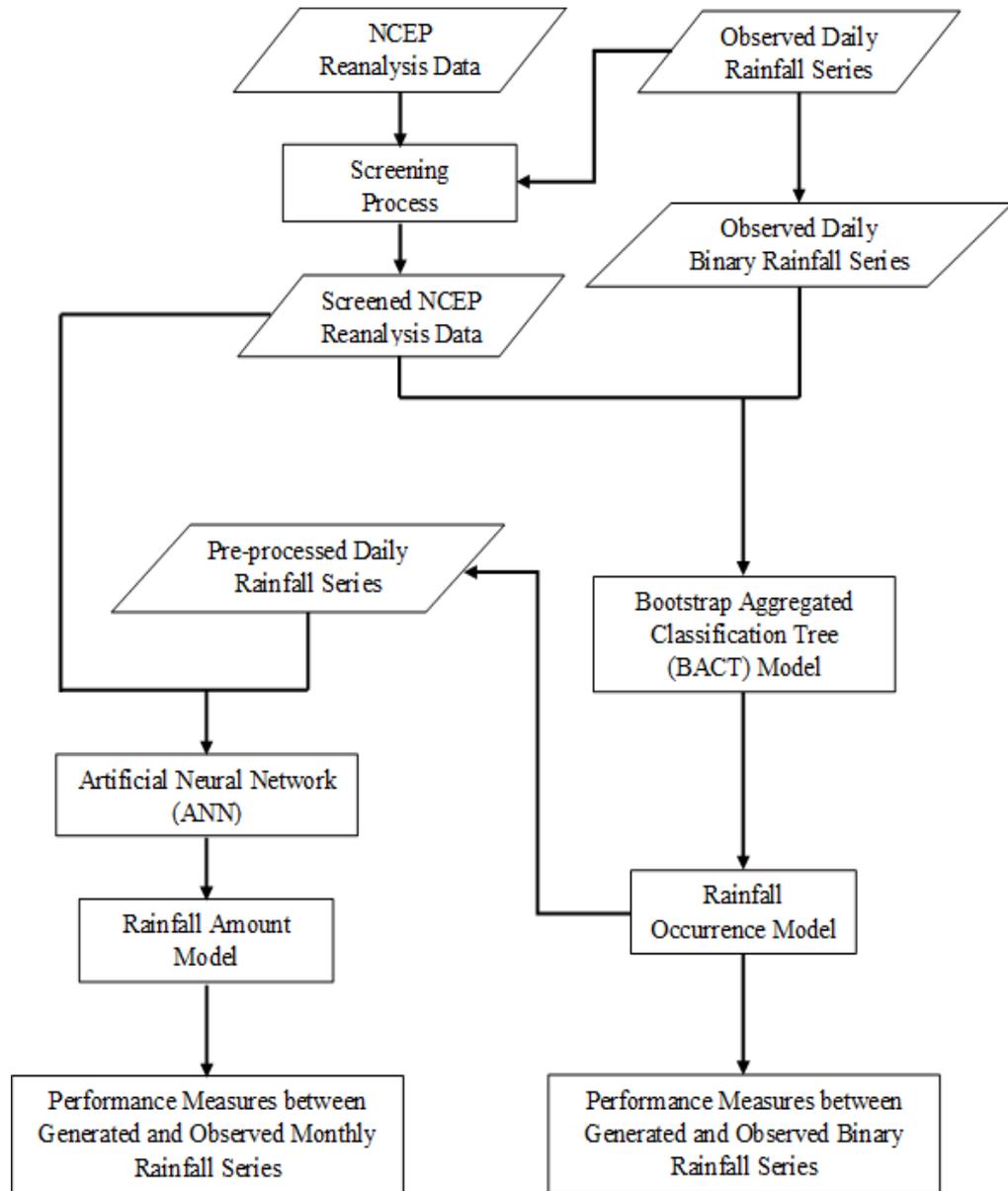


Figure 3.7: Flow chart of combined BACT-ANN model

3.8.3 Non-Homogeneous Hidden Markov Model (NHMM)

In this study, the NHMM was developed using the available software toolkit, namely the multivariate non-homogeneous hidden Markov model (MVNHMM). This software toolkit can be downloaded at <http://www.stat.purdue.edu/~skirshne/MVNHMM/>. The development of NHMM for this study, is present in Figure 3.8. Basically, the NHMM is the modification of hidden Markov model (HMM) by introducing the exogenous atmospheric predictors to simulate the rainfall. Firstly, the observed daily rainfall occurrence series were fitted into NHMM for the purpose of determining the optimum number of hidden state, which can adequately describe the observed rainfall state. The rainfall distribution type was set to independent and Bernoulli in advanced options, due to the rainfall at each station is independent of other station and the data used is binary data.

Theoretically, there are two conditional independence assumptions made in the model. First is the multivariate rainfall data R_t at day t , are unrelated to all other variables, but conditional on the weather state S_t at day t , which can be expressed as:

$$P(R_t|S_{1:t}, R_{1:t-1}) = P(R_t|S_t) \quad (3.25)$$

The next assumption is the application of first-order Markov chain model in NHMM to model the hidden state process, $S_{1:t}$, which can be defined as:

$$P(S_t|S_{1:t-1}) = P(S_t|S_{t-1}) \quad (3.26)$$

Besides, the transition probabilities between the states in this process do not change with time, which mean they are homogeneous in time. The optimum number of hidden states was determined by fitting the NHMM with 1 – 7 hidden states.

After the number of optimum hidden state is determined, the selected predictors and observed daily rainfall series were introduced into NHMM for rainfall amount modelling. With the X_t denotes a column vector of predictors at day t , and $X_{1:T}$ denotes the sequence of X_1, \dots, X_T . The Eq. (3.26) is then replaced by:

$$P(S_t|S_{1:t-1}, X_{1:T}) = P(S_t|S_{t-1}, X_t) \quad (3.27)$$

Therefore, the hidden state on the day t not only conditional on the predictor vector X_t at day t , but also on the hidden state S_{t-1} at day $t - 1$. Since the X_t changes with time, the transition probabilities also no longer remain homogenous in between the hidden state due to the changes in X , hence this model is called as non-homogeneous model.

In this study, the iterative expectation maximization (EM) algorithm also known as Baum-Welch algorithm, was used to estimate the parameter of NHMM in terms of maximum likelihood. However, the limitation of EM algorithm is that it can be trapped at local maxima and consequently fails to reach the global maxima. To avoid the high possibility of EM algorithm converge to local maxima, this algorithm was initialized 10 times with the generated randomly starting point. There was a numerical test using a larger number of random starting points and it showed the use of 10 initializations provide a good

settlement between the computation time and the estimation of global maximum value in the log-likelihood (Pineda and Willems, 2016). For each starting point, the EM was proceeded to convergence and the parameters achieved the highest log-likelihood over all 10 runs was selected as the best estimate of the global maximum in terms of the log-likelihood (Robertson et al., 2004; Robertson et al., 2006).

Furthermore, the NHMM were fitted with different rainfall distribution types during the rainfall amount modelling, namely the delta-gamma, the delta-exponential with one component and the delta-exponential with two-component. The log-likelihood value, Bayesian information criterion (BIC) and Akaike information criterion (AIC) scores were used as the guidelines in both selection of the optimum number of hidden state and the suitable rainfall distribution for rainfall amount modelling.

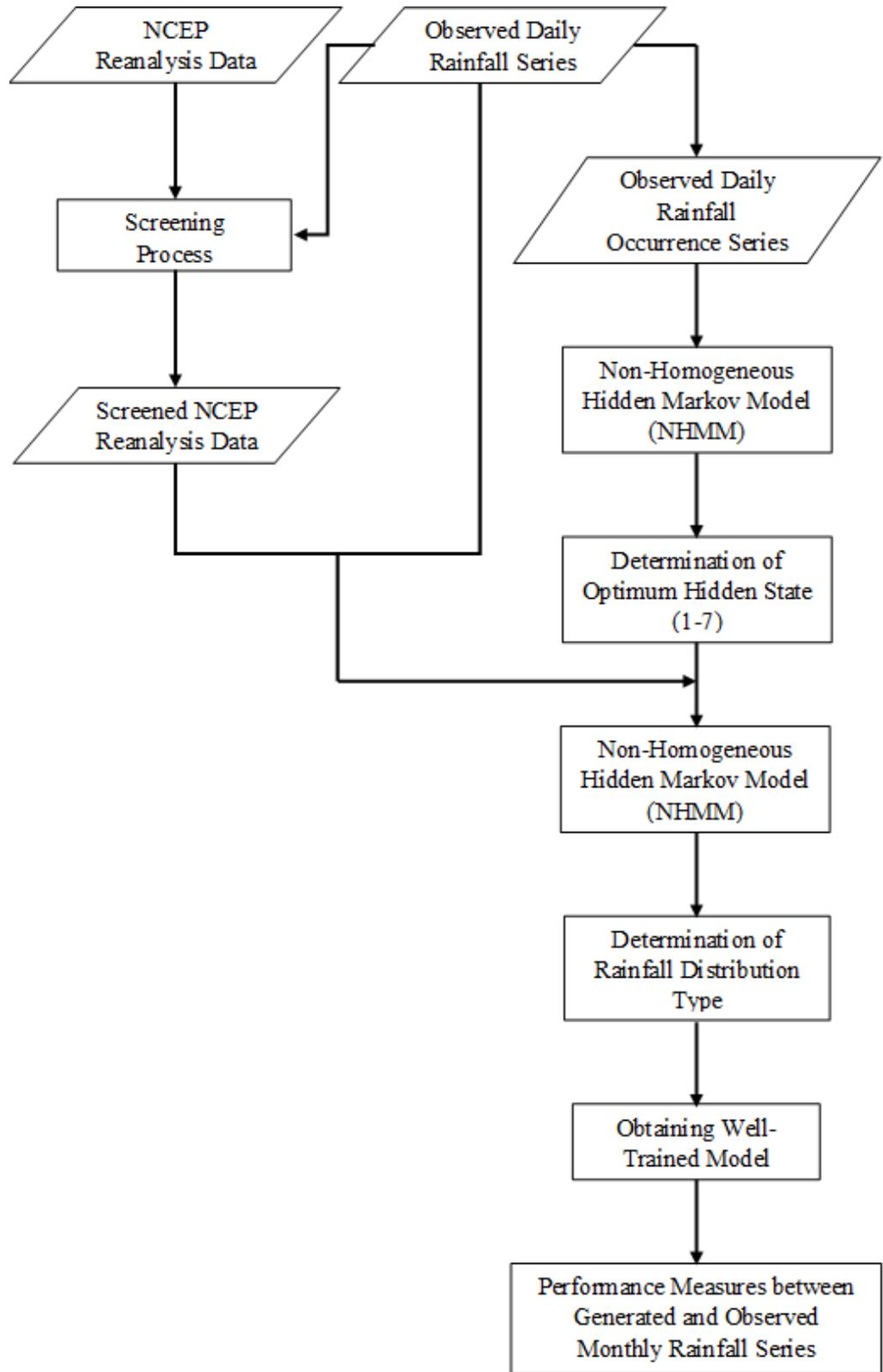


Figure 3.8: Flow chart of NHMM

3.8.4 Goodness of Fit

The accuracy of rainfall amount models (ANN, NHMM and combined BACT-ANN model) were measured by comparing the observed and simulated monthly rainfall series. The monthly rainfall series is obtained through the summation of daily rainfall series. These models were evaluated and compared in term of their results of parametric test and non-parametric tests.

3.8.4.1 Root Mean Square Error

The most common parametric tests, which has been widely used to measure the differences between the actual and predicted values, is the root mean square error (RMSE). The equation of RMSE can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (3.28)$$

where, y_i denotes the observed value for i^{th} observation and \hat{y}_i denotes the predicted value. While, n is the total number of sample. The value of RMSE ranges from zero to infinity. This measure is useful when large errors are particularly undesirable, as it give a relative high weight to the large error.

3.8.4.2 Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (K-S) test is one of the non-parametric tests used to compare the distributions of two independent samples, which is similar to the one-sample Kolmogorov-Smirnov test. The test statistic of K-S test can be expressed as:

$$D = \sup |S(x) - F_o(x)| \quad (3.29)$$

where, D denotes the supremum over all x of the absolute difference between $S(x)$ and $F_o(x)$. While, $S(x)$ represents the sample distribution function and $F_o(x)$ represents the observed cumulative distribution function.

For a two-sided K-S test, the null hypothesis is that the two testing samples come from the same distribution. However, if the likelihood of samples being from different distributions exceeds the demanded confidence level, then the null hypothesis is rejected to support the alternative hypothesis, where two samples are from different distribution. In this study, the K-S test was performed using the XLSTAT statistical analysis software, in which the p-value is given for determining the acceptance or rejection of null hypothesis under the significance level of 0.05.

3.8.4.3 Mann-Whitney U test

The Mann-Whitney U test is another form of the Wilcoxon Rank-Sum test, used to determine the equivalence of independent samples based on their ranks. Besides, this test does not assume any assumption to the distribution, so it is a non-parametric test (Jung et al., 1971; Zimmerman, 2014). The logic behind this Mann-Whitney U test is to assign the scores with their respective ranking from smallest to largest, for instance; the lowest score obtains a rank of “1”, and then next lowest score obtains a rank of “2”. Two or more scores obtain the average of their respective ranks under “tie” condition with identical scores to each other. The “U” in the Mann-Whitney U test displays the difference between the two rank total, which can be defined as:

$$U = N_1N_2 + N_x \frac{(N_x + 1)}{2} - T_x \quad (3.30)$$

where, N_1 and N_2 denote the number of scores in the sample, while N_x denotes the number of scores in the sample with the larger rank total, T_x .

Similar to the K-S test, the results of Mann-Whitney U test were evaluated based on their computed p-values under significance level of 0.05. For a two-sided Mann-Whitney U test, the null hypothesis is that the difference of location between the samples is equal to zero, while the alternative hypothesis is the difference of location between samples is different from zero.

3.8.4.4 Squared-Rank Test

The squared-rank test is another distribution free test used to examine the equality of variance between the independent samples. In this study, the squared-rank tests were performed using the StatsDirect3 statistical software under the significance level of 0.05. The two-tailed p-values were computed and displayed in the test results. However, the null hypothesis is accepted when the p-value is greater than significance level, which indicate the samples are from identical distribution. Otherwise, the alternative hypothesis, which the samples are from different distribution, is accepted when p-value is smaller than the significance level.

3.8.4.5 Kendall's Tau-b Correlation

The Kendall's tau-b correlation is a non-parametric test used to measure the strength of association between two samples based on their ranks. For a two-sided Kendall's tau-b correlation, the null hypothesis is both samples are independent to each other, while the alternative hypothesis is the association between both samples is significantly different from zero. The range of Kendall's tau-b correlation coefficient is from -1 to 1. The test statistic of this correlation is given by:

$$\hat{t} = \frac{S}{n(n-1)/2} \quad (3.31)$$

where, S denotes the difference between the number of concordant pairs and the number of discordant pairs. While, n represents the sample size. In this study, the Kendall's tau-b correlation was performed using the XLSTAT statistical analysis software under the significance level of 0.05.

3.8.4.6 Spearman's Rho Correlation

The Spearman's rho correlation is well-known non-parametric test, which has been widely used to compute the degree of association between two samples. The null hypothesis of this correlation is both samples are independent to each other, which is similar to Kendall's tau-b correlation. However, the alternative hypothesis is that both samples are either directly or inversely related to each other. The test statistic of Spearman's rho correlation is defined as:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3.32)$$

where

$$\sum d_i^2 = \sum_{i=1}^n [R(X_i) - R(Y_i)]^2 \quad (3.33)$$

The $R(X_i)$ denotes the rank of the i th values of X , which sorted from smallest to largest in order of magnitude. While, $R(Y_i)$ denotes the rank of the i th values of Y , which sorted in the similar way with $R(X_i)$. The n represents the sample size. The coefficient of Spearman's rho correlation, r_s ranges from -1 to 1.

Besides, the Spearman's rho correlation also performed in XLSTAT under significant level of 0.05.

3.8.4.7 Acceptability Index

In this study, the acceptability index (*AI*) proposed by Fodor et al. (2010) was used to measure the efficiency of the rainfall generators according to their non-parametric tests results, which can be expressed as:

$$AI = 100\left(1 - \frac{TS}{MS}\right) \quad (3.34)$$

where, *TS* denotes the total scores computed from the statistical non-parametric tests, while the *MS* denotes the maximum score. The value of zero or one was assigned to indicate whether the difference between observed and the simulated monthly rainfall series is non-significant or significant in every test. Hence, the *MS* of each test for monthly rainfall series is 48 (4 stations x 12 months). The higher the *AI*, the better the efficiency of the rainfall generator.

CHAPTER 4

RESULTS AND DICUSSION

4.1 Normality Tests

The normality of observed daily rainfall series was checked using three tests, namely Anderson-Darling, Lilliefors and Jarque-Bera tests, with the significance level of 0.05. Table 4.1 shows the results of normality tests at every station and evaluated in terms of p-value. The rejection of either null hypothesis or alternative hypothesis made in the test was determined based on the computed p-value. However, the computed p-value of each test at every station was smaller than the significance level alpha, $\alpha = 0.05$, as shown in Table 4.1. Therefore, the null hypothesis was rejected, which indicate the observed daily rainfall series in Langat River Basin are not normally distributed. Instead of using the parametric tests, the non-parametric tests would be more appropriate in evaluating the performance of rainfall amount models in this study.

Table 4.1: Normality tests results of observed daily rainfall series at each station

Station	p-value		
	Anderson-Darling	Lilliefors	Jarque-Bera
2815001	< 0.0001	< 0.0001	< 0.0001
2913001	< 0.0001	< 0.0001	< 0.0001
2917001	< 0.0001	< 0.0001	< 0.0001
3118102	< 0.0001	< 0.0001	< 0.0001

- Significance level alpha, $\alpha = 0.05$.

The statistical characteristics (mean, standard deviation, maximum, skewness coefficients and kurtosis coefficients) of observed daily rainfall series at each station, are presented in Table 4.2. Skewness can be defined as a measure used to assess the asymmetry of probability distribution, while the kurtosis provides the information on the tails of a distribution. Based on the results in Table 4.2, the computed Pearson's and Fisher's coefficients of skewness were greater than zero at every station, which mean the distributions of rainfall series are highly skewed to the left. The high positive Pearson's and Fisher's kurtosis values also indicated the distributions have the heavier tails and a sharper peak than the normal distribution. The observed rainfall series at station 2913001 obtained the highest skewness and kurtosis coefficients among all the stations due to the presence of highest maximum rainfall amount.

Table 4.2: Statistical characteristics of observed daily rainfall series at each station

Station	Mean	Std. dev.	Max	Skewness		Kurtosis	
				Pearson	Fisher	Pearson	Fisher
2815001	4.18	10.30	127.20	3.93	3.93	20.44	20.45
2913001	4.30	12.00	522.50	10.66	10.67	334.94	335.09
2917001	5.83	13.11	194.00	3.97	3.97	24.25	24.27
3118102	5.32	12.55	207.60	4.22	4.22	27.97	27.98

- Std. dev. represented standard deviation;

- Max represented maximum.

4.2 Homogeneity Tests

The homogeneity test results of 48 monthly rainfall series using the Standard Normal Homogeneity Test (SNHT), the Buishand Range (BR) test, the Pettitt (PET) test and the Von Neumann Ratio (VNR) test are shown in Table 4.3. Based on the evaluation criteria proposed by Wijngaard et al. (2003), 45 out of 48 monthly rainfall series were classified as “useful”, while another two monthly rainfall series and another series were under the class “doubtful” and “suspect”, respectively. Approximate 94% of monthly rainfall series of the selected stations within Langat River Basin are homogeneous.

Based on the results in Table 4.3, only the rainfall series at station 2815001 and 2917001 were homogeneous for all 12 months. The “doubtful” rainfall series was found at station 2913001 on October and station 3118102 on July, while the rainfall series at station 3118102 on March was detected to be non-homogeneous with four homogeneity tests failed. The historical metadata was not included in this study to assess the detected break and make correction to the specific non-homogeneous series. Only one out of twelve monthly rainfall series was non-homogeneous. Therefore, the homogeneity of rainfall series at station 3118102 was considered to be acceptable and included in this study for further analysis.

Table 4.3: Homogeneity tests results of monthly rainfall series at each station

Station	Test	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2815001	SHNT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	BR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	PET	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	VNR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2913001	SHNT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	BR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
	PET	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	VNR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
2917001	SHNT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	BR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	PET	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
	VNR	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
3118102	SHNT	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
	BR	✓	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓
	PET	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
	VNR	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓

- “✓” represented the tested series was homogenous;

- “✗” represented the tested series was non-homogeneous.

4.3 Predictor Selection

The NCEP reanalysis data with and without lag transformation (lag -9 to lag 9) were screened with the observed rainfall series, in order to select the suitable set of predictors for every station. Firstly, the predictors with the highest explained variance, which mean they exhibited the strong relationship with observed rainfall series, were selected. Then, those selected predictors were further screened among themselves using partial correlation analysis. Table 4.4 shows the description and lag-transformation of selected predictors, which are suitable and highly correlated with observed rainfall series at each station. In this study, the observed rainfall series of selected stations were sensitive to the

near surface specific humidity (shumgl), and zonal velocity component at different geopotential heights (p5_ugl and p8_ugl), as shown in Table 4.4.

The partial correlation coefficients of selected predictors were in the range of -0.095 to 0.059, as shown in Table 4.5. This situation indicated the observed rainfall series tended to be fluctuating and less predictable based on the large-scale predictors. According to the study of Amirabadizadeh et al. (2016), both temperature and precipitation of their selected stations within Langat River Basin were found to be sensitive to near surface specific humidity. Langat River Basin experience the tropical rainforest climate, which is hot and humid, throughout the year. Therefore, it is notable that the near surface specific humidity (shumgl) was still included as one of the predictors at stations 2913001 (p value = 0.2481) and 3118102 (p value = 0.3403), even their p-values were greater than the significance level of 0.05.

The results in Table 4.5 showed the p-value of geostrophic airflow velocity (p8_fgl at station 2815001 (p value = 0.0597)), zonal velocity component (p8_ugl at station 2913001 (p value = 0.3572)) and p5_ugl at station 3118102 (p value = 0.2307)) and vorticity (p8_zgl at station 2913001 (p value = 0.1138)) also greater than 0.05. However, the exclusion of any one of these predictors would reduce the partial correlation coefficient and increase the p-value of near surface specific humidity, hence those variables were selected to combine with near surface specific humidity as a suitable combination set of predictors for each station.

Table 4.4: Description and lag-transformation of selected predictors at each station

Station	Predictors	Description	Lag
2815001	p8_fgl	geostrophic airflow velocity at 850hPa height	2
	p8_ugl	zonal velocity component at 850hPa height	0
	s500gl	specific humidity at 500hPa height	1
	s850gl	specific humidity at 850hPa height	-1
	shumgl	near surface specific humidity	0
2913001	p8_ugl	zonal velocity component at 850hPa height	-2
	p8_zgl	vorticity at 850hPa height	1
	s500gl	specific humidity at 500hPa height	1
	shumgl	near surface specific humidity	-1
2917001	p1_vgl	meridional velocity component	0
	p5_ugl	zonal velocity component at 500hPa height	0
	p8_fgl	geostrophic airflow velocity at 850hPa height	-1
	p8_zgl	vorticity at 850hPa height	1
	shumgl	near surface specific humidity	-1
3118102	mslpgl	mean sea level pressure	2
	p5_ugl	zonal velocity component at 500hPa height	0
	p8zhgl	divergence at 850hPa height	0
	s850gl	specific humidity at 850hPa height	0
	shumgl	near surface specific humidity	0

Table 4.5: Partial correlation coefficient and p-value of selected predictors at each station

Station	Predictors	Partial r	p-value
2815001	p8_fgl(2)	0.033	0.0597
	p8_ugl	-0.074	0.0000
	s500gl(1)	0.04	0.0234
	s850gl(-1)	-0.04	0.0224
	shumgl	0.059	0.0004
2913001	p8_ugl(-2)	-0.015	0.3572
	p8_zgl(1)	0.029	0.1138
	s500gl(1)	0.046	0.0096
	shumgl(-1)	0.021	0.2481
2917001	p1_vgl	-0.063	0.0000
	p5_ugl	-0.049	0.0014
	p8_fgl(-1)	-0.095	0.0000
	p8_zgl(1)	0.053	0.0006
	shumgl(-1)	0.049	0.0017

- Bracket "(x)" denoted the optimum lag-transformation.

Table 4.5 continued: Partial correlation coefficient and p-value of selected predictors at each station

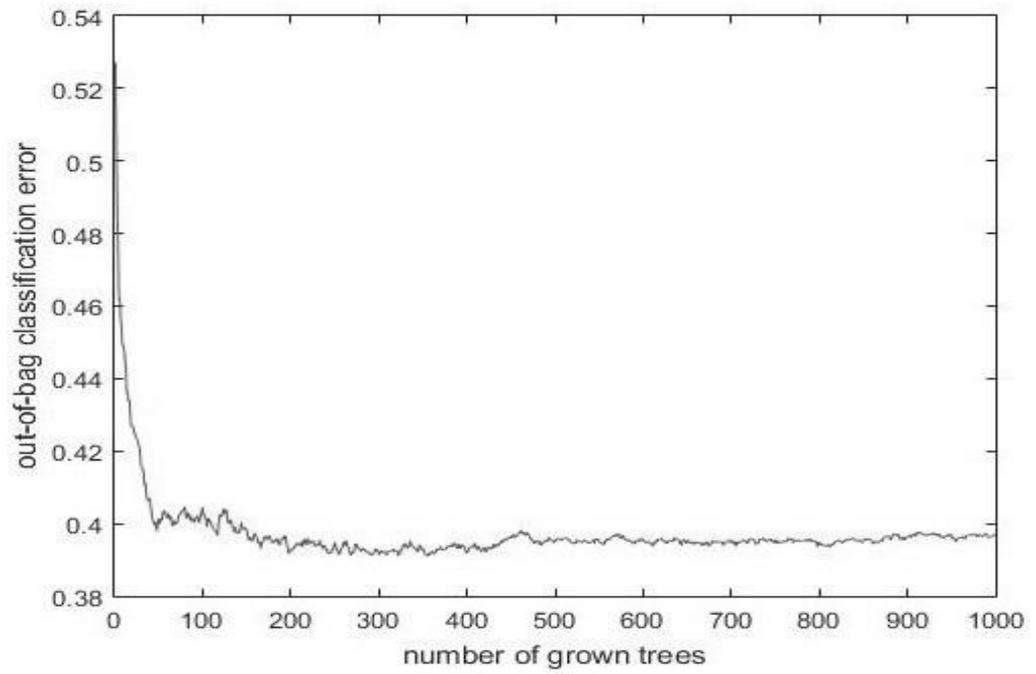
Station	Predictors	Partial r	p-value
3118102	msslpgl(2)	0.033	0.0465
	p5_ugl	-0.02	0.2307
	p8zhgl	-0.037	0.0217
	s850gl	0.042	0.0087
	shumgl	0.015	0.3403

- Bracket "(x)" denoted the optimum lag-transformation.

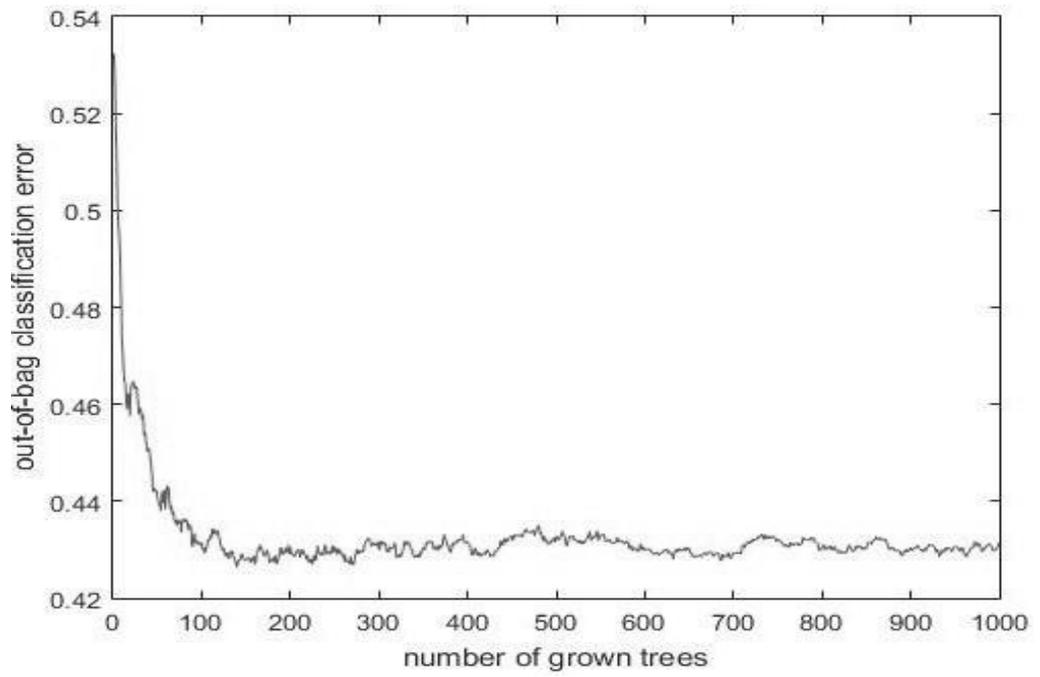
4.4 Out-of-Bag Classification Error for BACT Model

Out-of-bag (OOB) samples are defined as the data which are not included during the splitting process of each bootstrap sample. The prediction strength of BACT model can be measured based on the average prediction error generated by each OOB sample over the total grown trees in the model. Basically, if a sufficient large number of trees is given, the determination of number of trees is considered to be less critical. This is because there is no any further reduction of error to be achieved after reaching the optimum number of grown trees in the model. In this study, 1000 trees were first used to determine the optimum number of grown tree in BACT model based on their generated OOB classification error. As illustrated in Figure 4.1, the OOB classification error decreased with the number of grown trees. However, in considering the computational efficiency and stability of prediction, the number of grown trees was set to 150 in BACT model for every station. This is because the OOB classification errors is stabilized after 150 trees at all stations and adding more trees is unnecessary as this will only cause the intensive computational time of model.

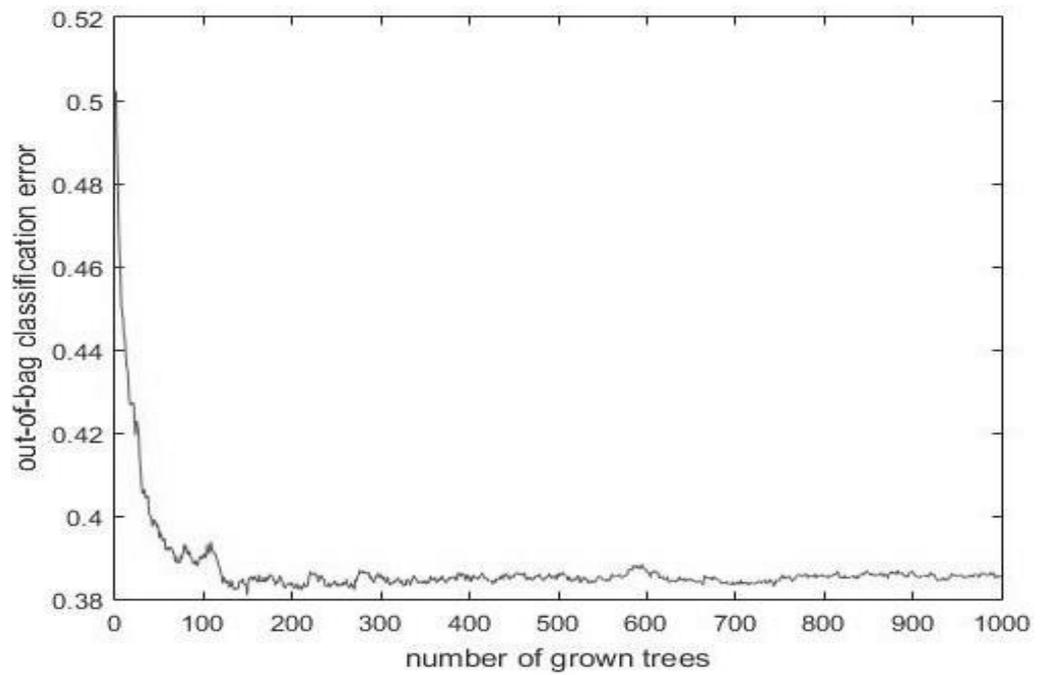
(a) Station 2815001



(b) Station 2913001



(c) Station 2917001



(d) Station 3118102

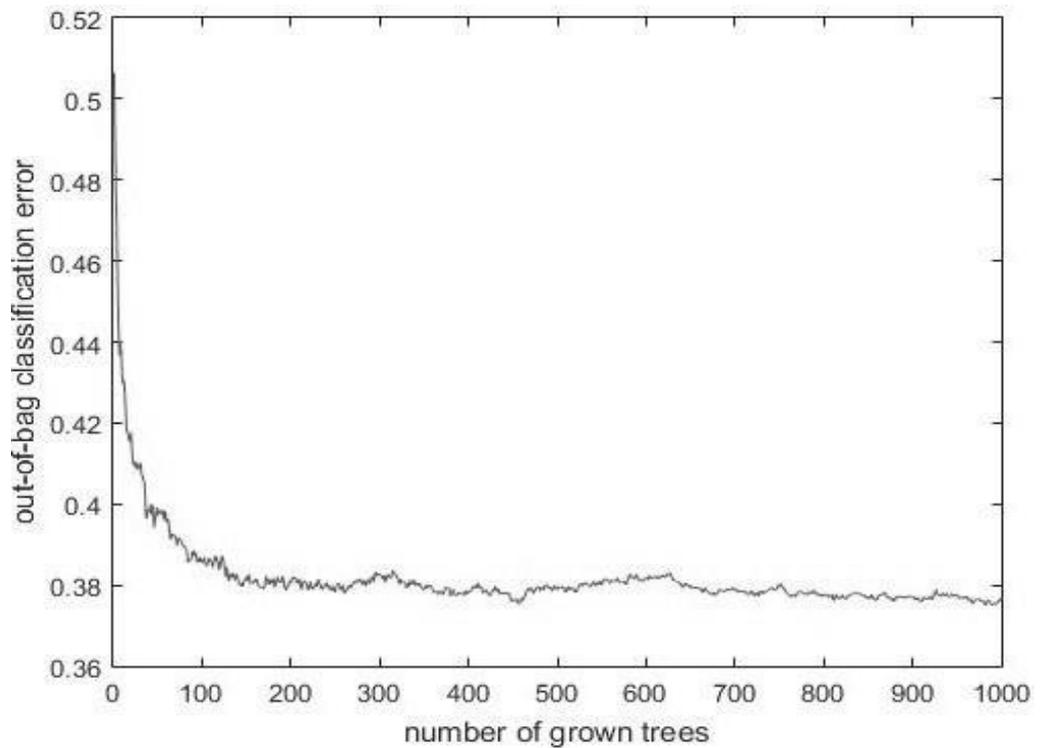


Figure 4.1: The out-of-bag classification error produced by bootstrap aggregated classification tree (BACT) models as a function of the number of grown trees at (a) station 2815001, (b) station 2913001, (c) station 2917001, and (d) station 3118102.

4.5 Determination of Optimum Hidden States and Rainfall Distribution for NHMM

The number of hidden states in NHMM is required to sufficiently express the rainfall state of observed rainfall series at selected station. Therefore, the optimum number of hidden states were determined prior to the process of fitting predictors into the model. The NHMM with different number of hidden states (from state 1 to state 7) were fitted to the training data. In this study, the log-likelihood values were computed to evaluate the quality of fitted models as a function of different number of hidden states, as shown in Table 4.6. However, the log-likelihood values obtained the decreasing trend with the number of hidden states increased from 1 to 7, as shown in Figure 4.2. This situation reflects that the determination of the optimum number of hidden states became more challenging.

Therefore, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) scores were included in this study for the purpose of penalizing the complexity of fitted model, which are presented in Table 4.7 and Table 4.8. The difficulty in the task of selecting optimum number of hidden state was reduced by BIC and AIC scores, because they did not show the decreasing trend all the way similar to log-likelihood values. As illustrated in Figure 4.3 and Figure 4.4, both BIC and AIC scores achieved their minimum points at 3 hidden states at station 2815001 and station 2913001. Besides, their log-likelihood values also decreased substantially from hidden states 1 to 3, then started to stabilize, as shown in Figure 4.2. Thus, three hidden states were deemed to be sufficient to predict the daily rainfall occurrence at station

2815001 and station 2913001. For station 2917001 and station 3118102, they obtained the minimum BIC and AIC scores at two hidden states, which is further justified by the stabilization of their log-likelihood values after a significant reduction from hidden states 1 to 2.

Table 4.6: Log-likelihood values of NHMM as a function of different hidden states number at each station

Hidden State	Log-likelihood			
	2815001	2913001	2917001	3118102
1	-4776.25	-4779.08	-5011.54	-4958.56
2	-4348.95	-4295.13	-4663.99	-4182.09
3	-4292.29	-4207.34	-4659.69	-4216.22
4	-4289.17	-4204.67	-4637.80	-4196.69
5	-4286.82	-4204.77	-4629.95	-4157.53
6	-4287.32	-4202.15	-4620.78	-4220.89
7	-4286.14	-4202.94	-4629.61	-4177.27

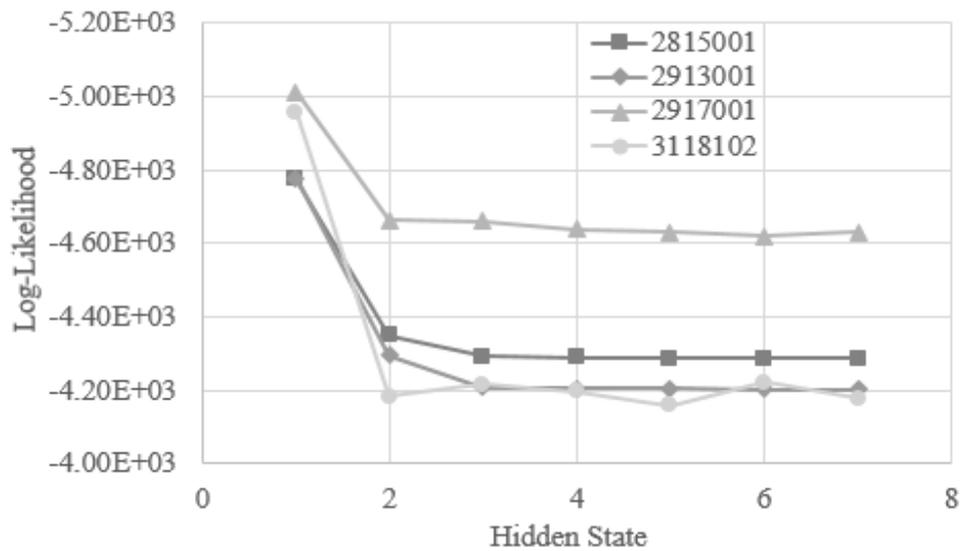


Figure 4.2: Log-likelihood values of NHMM as a function of different hidden states number at each station

Table 4.7: Bayesian information criterion (BIC) scores of NHMM as a function of different hidden states number at each station

Hidden State	BIC score			
	2815001	2913001	2917001	3118102
1	9561.39	9567.06	10031.97	9926.01
2	8742.39	8634.73	9372.46	8408.65
3	8682.43	8512.53	9417.23	8530.30
4	8747.36	8578.35	9444.62	8562.39
5	8831.61	8667.51	9517.87	8573.03
6	8939.37	8769.03	9606.28	8806.49
7	9061.54	8895.14	9748.48	8843.79

- Grey shaded value indicated the minimum BIC score.

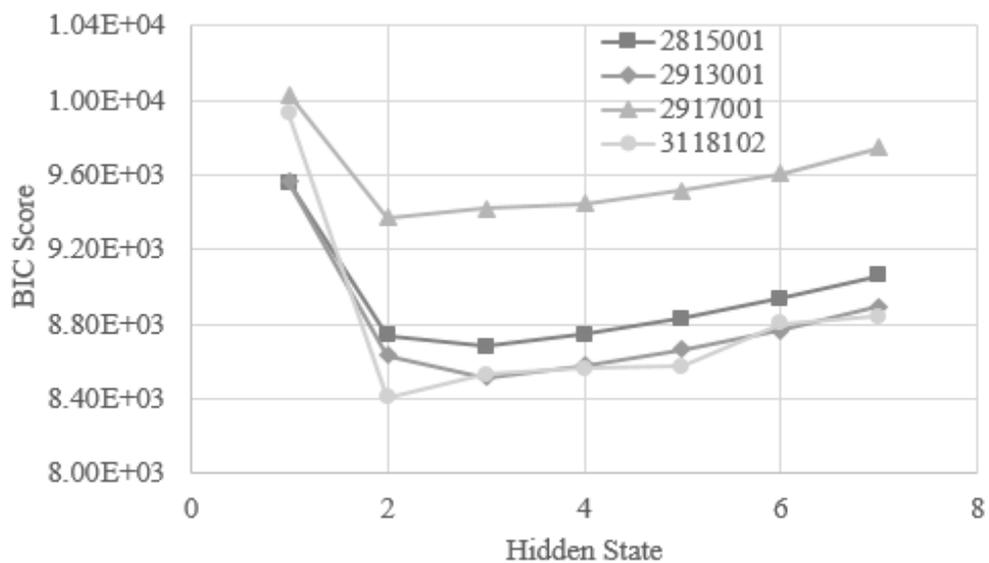


Figure 4.3: Bayesian information criterion (BIC) scores of NHMM as a function of different hidden states number at each station

Table 4.8: Akaike information criterion (AIC) scores of NHMM as a function of different hidden states number at each station

Hidden State	AIC score			
	2815001	2913001	2917001	3118102
1	9554.49	9560.17	10025.08	9919.12
2	8707.91	8600.25	9337.98	8374.17
3	8606.58	8436.67	9341.38	8454.45
4	8616.34	8447.34	9313.60	8431.38
5	8631.64	8467.54	9317.89	8373.06
6	8656.65	8486.30	9323.56	8523.77
7	8682.28	8515.88	9369.22	8464.53

- Grey shaded value indicated the minimum AIC score.

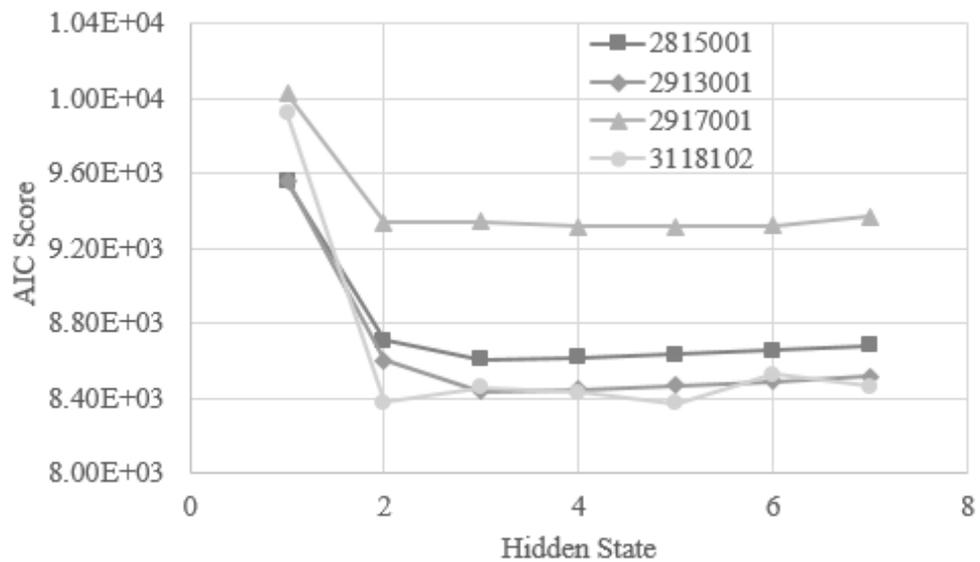


Figure 4.4: Akaike information criterion (AIC) scores of NHMM as a function of different hidden states number at each station

The optimum number of hidden state determined in NHMM at each station was fixed and further used with the introduction of predictors. There were three types of models used for fitting the rainfall amount. Each model consisted of a delta function to represent the days without rainfall amount, together with a gamma, a single-exponential or a two-component mixed-exponential function to model the rainfall intensity. Similarly, the log-

likelihood value, BIC and AIC scores were used as the criteria to determine the suitable rainfall distribution of each station. Based on the results in Table 4.9, the gamma models obtained the smaller log-likelihood value at each station, when compared to single-exponential and two-component mixed-exponential models. Besides, these results were further supported by their minimum AIC and BIC scores, as presented in Table 4.10 and Table 4.11. Hence, the gamma was selected as the suitable distribution in NHMM to model the rainfall intensity of selected rainfall stations within the Langat River Basin, due to its better performance over the single-exponential and two-component mixed-exponential distributions.

Table 4.9: Log-likelihood values of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station

Station	Hidden State	Log-Likelihood		
		Gamma	1-Exponential	2-Exponential
2815001	1	-13716.41	-13796.66	-13636.24
	2	-13242.72	-13307.27	-13185.28
	3	-12177.38	-13088.82	-13134.63
2913001	1	-14012.54	-14096.05	-13951.30
	2	-12579.90	-13534.64	-13427.80
	3	-12270.10	-13290.27	-13274.77
2917001	1	-16321.10	-16438.60	-16223.60
	2	-14517.39	-15800.39	-15755.45
3118102	1	-16227.16	-16402.99	-16098.05
	2	-14232.77	-15546.35	-15367.08

- Grey shaded value indicated the rainfall distribution with the smaller log-likelihood value.

Table 4.10: Bayesian information criterion (BIC) score of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station

Station	Hidden State	BIC Score		
		Gamma	1-Exponential	2-Exponential
2815001	1	27459.50	27611.11	27308.07
	2	26609.99	26721.29	26512.90
	3	24594.95	26391.13	26536.13
2913001	1	28051.76	28209.89	27938.18
	2	25275.45	27167.13	26989.04
	3	24762.60	26776.24	26798.62
2917001	1	32668.89	32894.99	32482.79
	2	29159.33	31707.53	31653.24
3118102	1	32481.01	32823.77	32231.69
	2	28590.08	31199.44	30876.49

- Grey shaded value indicated the rainfall distribution with minimum BIC score.

Table 4.11: Akaike information criterion (AIC) score of rainfall distributions (gamma, 1-exponential and 2-exponential) as a function of different hidden states number at each station

Station	Hidden State	AIC Score		
		Gamma	1-Exponential	2-Exponential
2815001	1	27432.82	27593.31	27272.49
	2	26485.45	26614.55	26370.57
	3	24354.77	26177.64	26269.26
2913001	1	28025.08	28192.10	27902.59
	2	25159.81	27069.28	26855.61
	3	24540.20	26580.53	26549.55
2917001	1	32642.20	32877.20	32447.20
	2	29034.79	31600.78	31510.91
3118102	1	32454.32	32805.98	32196.11
	2	28465.54	31092.70	30734.16

- Grey shaded value indicated the rainfall distribution with minimum AIC score.

4.6 Goodness of Fit of Rainfall Occurrence Models

Just to reiterate, the performances of rainfall occurrence models (GLM, NHMM and BACT model) were evaluated and compared in terms of their rainfall persistence, spell lengths distribution and matching.

4.6.1 Rainfall Persistence

Rainfall persistence is the probability of rainy day given the previous day is rainy day for a particular station. This evaluation was measured between the simulated and observed rainfall occurrence series during calibration and validation periods. Based on the results in Table 4.12, the GLM showed the higher absolute difference of rainfall persistence in the range of 0.09–0.22 during validation period, when compared to NHMM and BACT model at every station. The total number of rainy days and of two consecutive rainy days were simulated well by NHMM, which resulted the smaller difference with observed rainfall persistence. Therefore, NHMM outperformed the GLM and BACT model in simulating the observed rainfall persistence with smaller difference (range from 0.02 to 0.07) during validation period.

The main reason for the NHMM capable of simulating the observed rainfall persistence is because of the second assumption made during the development of NHMM. The transition probabilities between the hidden states are defined by a first-order Markov chain model, in which they are not only

conditional on the set of predictors on that day, but also consider the condition of state on previous day. However, GLM worked with the approach of logistic regression, while the BACT model employed the approaches of bagging and classification with if-then logical condition. Both models made no assumption on the condition of rainfall occurrence state on previous day unlike the NHMM. Therefore, the NHMM was able to simulate the number of two consecutive rainy days similar to the observed value, with the smaller difference when compared to GLM and BACT model at each station.

Table 4.12: Performance of rainfall occurrence models in terms of their rainfall persistence and absolute difference with observed rainfall persistence during calibration (1976–1995) and validation (1996–2005) periods

Station	Model	Rainfall Persistence				Absolute Difference	
		Observed		Simulated		Calib	Valid
		Calib	Valid	Calib	Valid		
2815001	GLM	0.56	0.51	0.71	0.73	0.15	0.22
	NHMM	0.56	0.51	0.51	0.58	0.05	0.07
	BACT	0.56	0.51	0.60	0.70	0.04	0.19
2913001	GLM	0.56	0.55	0.64	0.69	0.07	0.14
	NHMM	0.56	0.55	0.55	0.53	0.01	0.02
	BACT	0.56	0.55	0.59	0.62	0.03	0.06
2917001	GLM	0.59	0.62	0.77	0.81	0.19	0.19
	NHMM	0.59	0.62	0.56	0.56	0.03	0.05
	BACT	0.59	0.62	0.63	0.78	0.04	0.16
3118102	GLM	0.64	0.63	0.70	0.72	0.06	0.09
	NHMM	0.64	0.63	0.58	0.59	0.06	0.03
	BACT	0.64	0.63	0.65	0.71	0.01	0.08

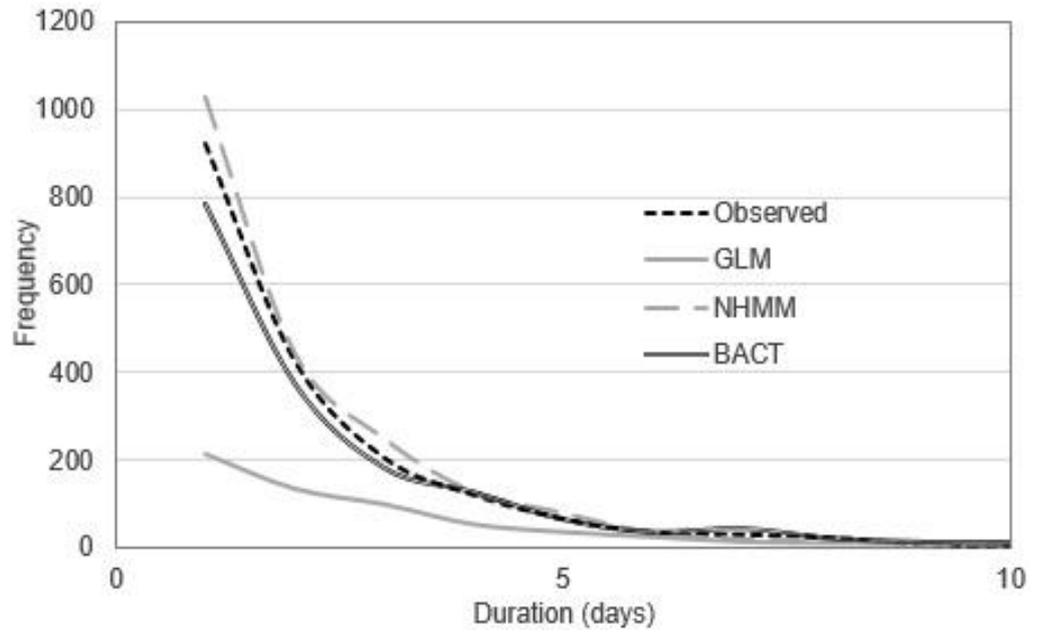
- Calib represented calibration;
- Valid represented validation;
- Grey shaded value indicated the model with the smaller difference.

4.6.2 Spell Lengths Distributions

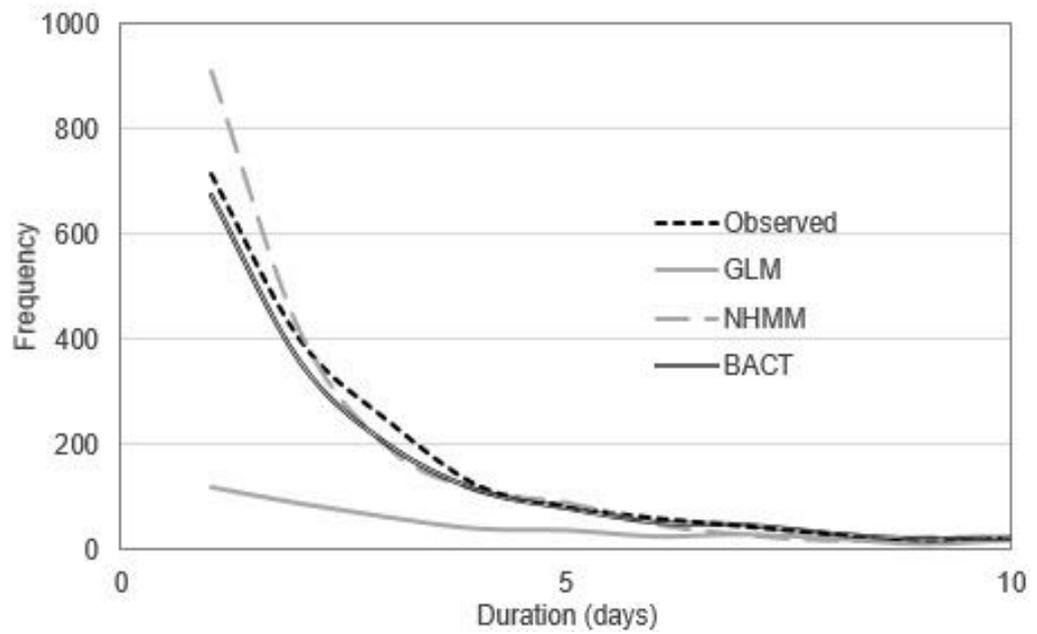
The distributions of wet- and dry-spell lengths were used to evaluate the performance of rainfall occurrence models at every station during the whole study period (years 1976– 2005). As illustrated in Figure 4.5, both NHMM and BACT model exhibited overall good performance in capturing the observed wet- and dry-spell lengths. The spell lengths distribution of both models showed the similar trend with observed by following approximately geometric distribution. However, NHMM tended to overestimate the wet- and dry-spell lengths at every station, especially a significant difference of dry spell length at station 2815001 (Figure 4.5(b)) and of wet spell length at station 2917001 (Figure 4.5(e)) and station 3118102 (Figure 4.5(g)). In addition, the BACT model also showed the overall under-prediction of both spell lengths at every station, except for the slight over-prediction at station 2913001.

Compared to NHMM and BACT model, the GLM also tended to follow the geometric distribution, but it showed the substantial under-prediction of frequency especially for the low spell durations at every station, as shown in Figure 4.5. Thus, both NHMM and BACT model showed the better prediction ability than GLM in simulating the observed spell lengths distribution at every station, by considering the huge difference obtained from GLM when compared to observed spell lengths distribution.

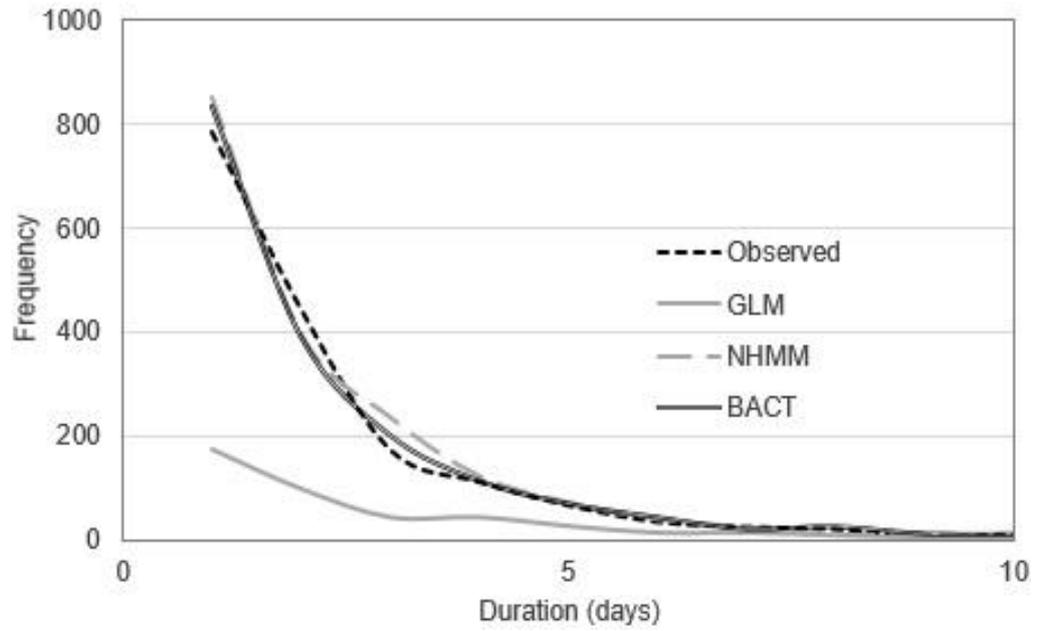
(a) Distribution of wet-spell length at station 2815001



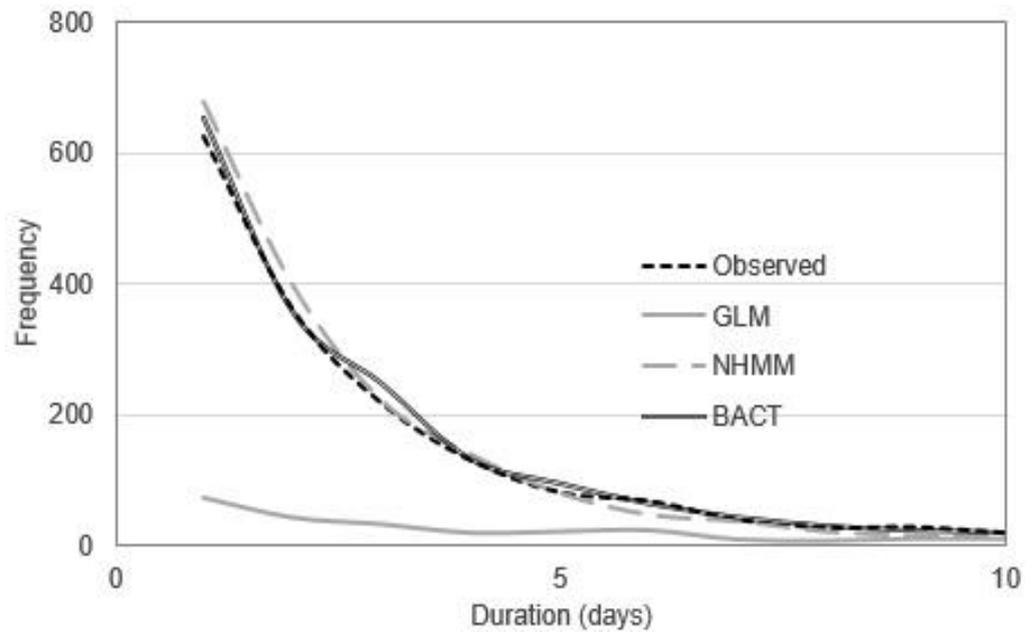
(b) Distribution of dry-spell length at station 2815001



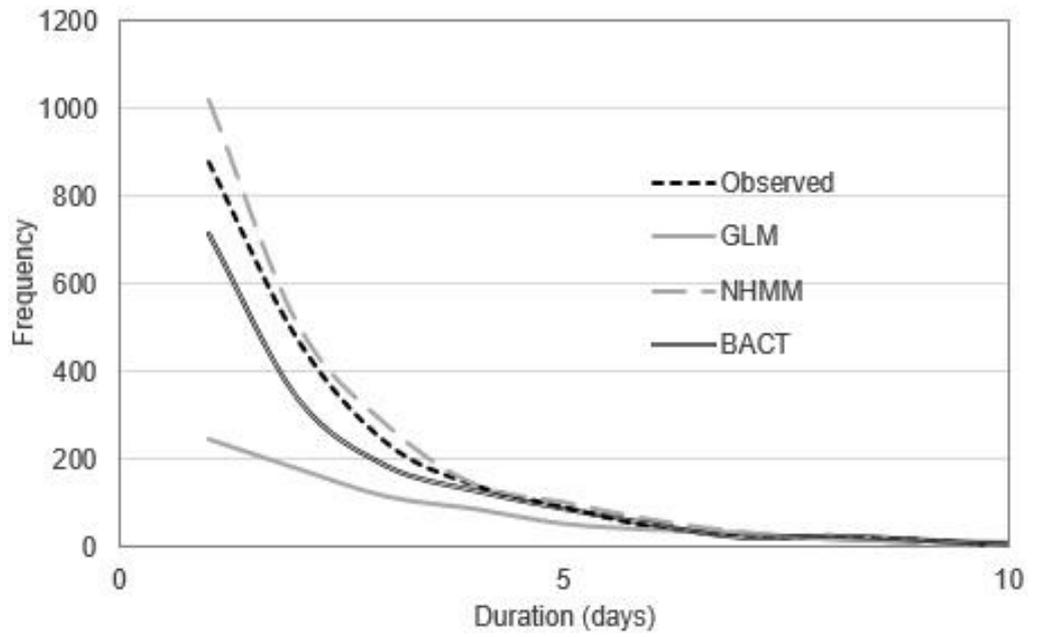
(c) Distribution of wet-spell length at station 2913001



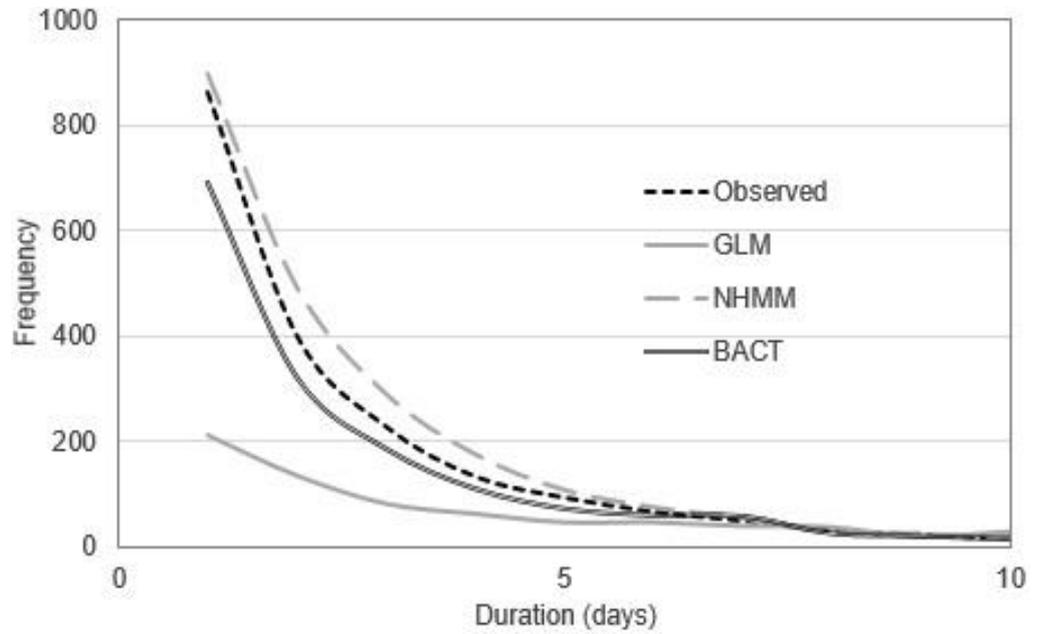
(d) Distribution of dry-spell length at station 2913001



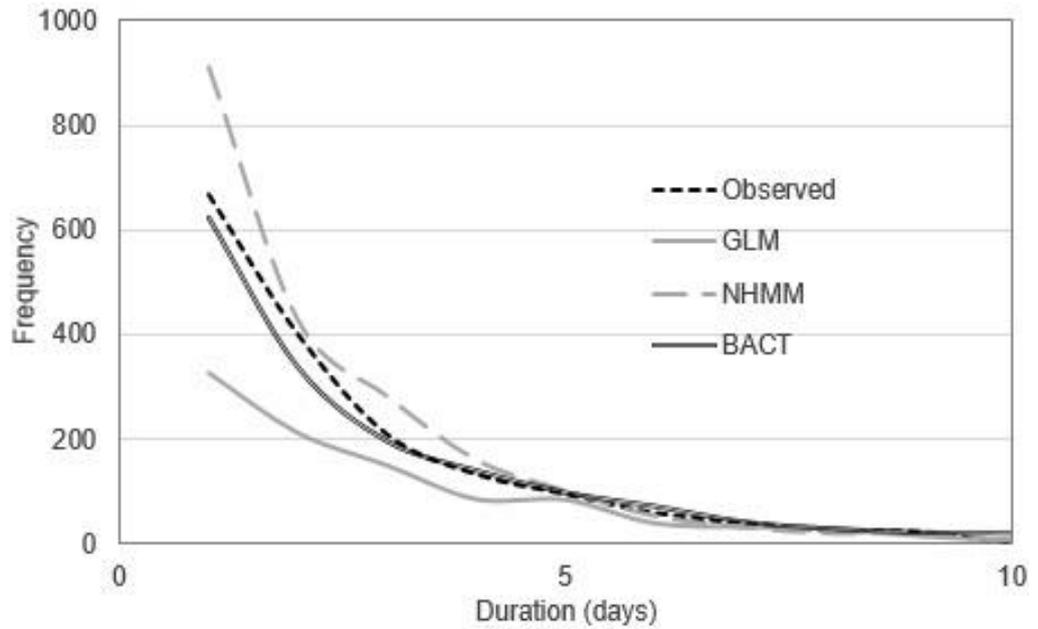
(e) Distribution of wet-spell length at station 2917001



(f) Distribution of dry-spell length at station 2917001



(g) Distribution of wet-spell length at station 3118102



(h) Distribution of dry-spell length at station 3118102

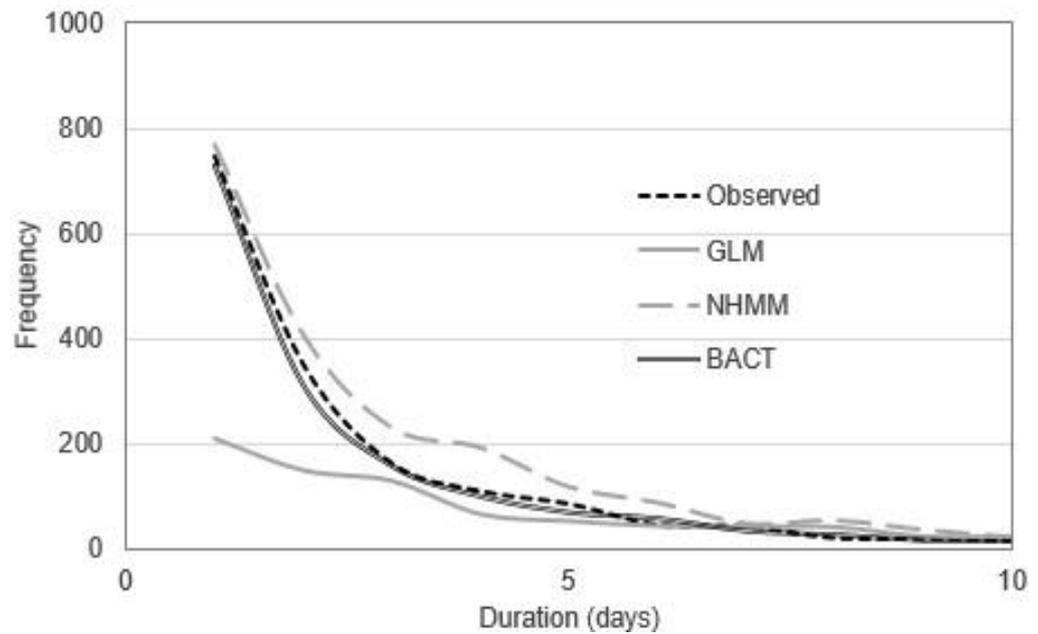


Figure 4.5: Observed and simulated spell lengths distribution at each station. (a) Distribution of wet-spell length and (b) dry-spell length at station 2815001, (c) Distribution of wet-spell length and (d) dry-spell length at station 2913001, (e) Distribution of wet-spell length and (f) dry-spell length at station 2917001, (g) Distribution of wet-spell length and (h) dry-spell length at station 3118102

4.6.3 Matching

The performance of rainfall occurrence models were compared and evaluated based on their probability of detection (POD), false alarm rate (FAR), critical success index (CSI), Heidke Skill score (HSS) and Peirce Skill score (PSS) indices. The ability of a good rainfall occurrence model is to simulate the observed rainfall occurrence as accurate as possible, thereafter, the rainfall values are generated by a rainfall prediction model conditional on the simulated wet days. Both indices of POD and CSI indicate how well a model simulates the wet days match with the number of observed wet days. The higher the index value, the better performance is the model.

Based on the results in Table 4.13, the BACT model achieved the better performance than GLM and NHMM during validation period with the higher indices of POD and CSI in the range of 0.51–0.65 and 0.29–0.44, respectively. NHMM obtained the POD and CSI indices smaller than 0.5 and 0.3, respectively, at every station, while, the GLM are capable of obtaining the POD and CSI indices in the range of 0.22–0.54 and 0.18–0.40, respectively. The indices of CSI are smaller than those of the POD for the fact that CSI included the number of events, which are predicted but actually not occurred, in their computation.

In general, FAR should be used in conjunction with POD and it is used to measure the probability wet day was simulated by a model, but in reality the rain actually did not occur on that day. Based on the results in Table 4.14, the

GLM exhibited the better performance with the lower FAR indices in the range of 0.38–0.52 during validation period, followed by BACT model and NHMM. The BACT model obtained the FAR indices in the range of 0.41–0.60, but the NHMM showed the FAR indices greater than 0.53 at every station. Hence, it can be concluded that BACT model possess the tendency of over-predicting the number of wet days, so they exhibited the better results in POD and CSI, but higher FAR indices when compared to GLM’s.

Table 4.13: Probability of detection (POD) and critical success index (CSI) of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods

Station	Model	POD		CSI	
		Calibration	Validation	Calibration	Validation
2815001	GLM	0.28	0.35	0.23	0.25
	NHMM	0.38	0.40	0.23	0.23
	BACT	0.98	0.60	0.91	0.34
2913001	GLM	0.17	0.22	0.15	0.18
	NHMM	0.38	0.35	0.23	0.21
	BACT	0.97	0.51	0.87	0.29
2917001	GLM	0.51	0.54	0.38	0.40
	NHMM	0.48	0.44	0.31	0.29
	BACT	0.96	0.64	0.89	0.44
3118102	GLM	0.43	0.45	0.34	0.34
	NHMM	0.46	0.44	0.29	0.29
	BACT	0.97	0.65	0.92	0.43

Table 4.14: False alarm rate of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods

Station	Model	FAR	
		Calibration	Validation
2815001	GLM	0.43	0.52
	NHMM	0.62	0.64
	BACT	0.07	0.56
2913001	GLM	0.42	0.51
	NHMM	0.63	0.66
	BACT	0.11	0.60
2917001	GLM	0.40	0.38
	NHMM	0.54	0.54
	BACT	0.08	0.41
3118102	GLM	0.40	0.41
	NHMM	0.57	0.55
	BACT	0.06	0.43

Other than that, the measures of skill score is a way to summarize the prediction performance of a model by determine the accuracy of prediction compared to the reference predictions (random chance, persistence, or climatology). In this study, the common skill scores, such as Heidke Skill score (HSS) and Pierce Skill score (PSS) were used to compare and evaluate the prediction ability of developed rainfall occurrence models. HSS is defined as the ratio of correct prediction to an random prediction that are statistically independent from the observations, while the PSS is another measure of skill with the false alarm been taken into account and the reference hit rate in the denominator is random and more likely to be an unbiased prediction. Both measures of skill scores show the perfect prediction with the score of one, prediction equivalent to the reference prediction with the score of zero, and prediction worse than reference prediction with the negative score. However, a prediction with the HSS index greater than 0.15, can be treated as reasonably good prediction (Kannan and Ghosh, 2011).

Based on the results in Table 4.15, both HSS and PSS indices showed the prediction of NHMM more likely to be a random prediction as they obtained the scores nearly equal to zero at every station. Both GLM and BACT model showed the reasonably good prediction with the scores greater than 0.15 at every station, except for station 2913001. In spite of that, BACT model still outperformed the GLM with the higher scores during calibration and validation periods. Therefore, the results in Table 4.13 and Table 4.14 were further supported by the HSS and PSS indices, which suggested that BACT model exhibited the better prediction ability than GLM and NHMM, and provided the higher accuracy in matching the number of observed wet days even though they showed the slight over-prediction over the GLM with the higher FAR indices.

Table 4.15: Heidke skill scores (HSS) and Pierce skill scores (PSS) of rainfall occurrence models during calibration (1976–1995) and validation (1996–2005) periods

Station	Model	HSS		PSS	
		Calibration	Validation	Calibration	Validation
2815001	GLM	0.17	0.16	0.15	0.15
	NHMM	0.00	0.02	0.00	0.02
	BACT	0.92	0.18	0.94	0.20
2913001	GLM	0.12	0.12	0.10	0.10
	NHMM	0.01	0.02	0.01	0.02
	BACT	0.89	0.13	0.91	0.14
2917001	GLM	0.25	0.25	0.24	0.25
	NHMM	0.03	-0.02	0.03	-0.02
	BACT	0.89	0.25	0.89	0.26
3118102	GLM	0.23	0.21	0.22	0.20
	NHMM	0.03	0.00	0.03	0.00
	BACT	0.93	0.25	0.93	0.25

4.7 Data Pre-processing Approach in Combined BACT-ANN Model

Prior to the development of ANN model, the rainfall data was pre-processed using the developed BACT model, which some days may become zero amount even though the observed rainfall series indicate them are wet days in reality. Therefore, the prediction ability of BACT model in simulating the observed rainfall occurrence may also affect the performance of ANN model. Based on the results in Table 4.16, the incident of observed wet days became dry days after data pre-processing approach was less than 2% during calibration period. However, the percentage of this incident was in the range of 13.5%– 17% during validation period, which mean there were 505–615 wet days from observed rainfall series became dry days.

Table 4.16: Percentage of observed wet days become dry days after data pre-processing approach

Station	Calibration	Validation
2815001	0.60	13.84
2913001	0.92	16.16
2917001	1.77	16.85
3118102	1.08	15.62

4.8 Structure of Neural Network for ANN and Combined BACT-ANN Model

The combined model in this study was developed using bootstrap-aggregated classification decision tree with the random forest algorithm (BACT model) to model the rainfall occurrence, thereafter combined with ANN to model the rainfall amount conditional on simulated rainfall occurrence. The comparison of structure between rainfall amount model in combined BACT-

ANN model and another individual traditional ANN model used in this study, is presented in Table 4.17. The structures of both models consisted of same neurons number in input and output layers, which are directly depending on the number of predictors and observed rainfall series, respectively. The only difference between them was the number of hidden neuron numbers used to link the predictors and observed rainfall series. Based on the results in Table 4.17, the number of hidden neuron used in combined BACT-ANN model was greater than traditional ANN model in the range of 5–40 neurons. The accuracy of model is generally contributed by the larger optimum number of hidden neuron numbers to sufficiently describe the relationship between predictors and predictand. This increased accuracy was presented by combined BACT-ANN model due to the application of pre-processing approach, which resulted in the increase of the number of wet days to be trained in the model.

Table 4.17: Structure of combined BACT-ANN and traditional ANN models

Station	ANN	BACT-ANN
2815001	5-40-1	5-60-1
2913001	5-45-1	5-85-1
2917001	5-50-1	5-85-1
3118102	5-55-1	5-60-1

- Structure of neural network = (number of selected predictors)-(number of optimum hidden neurons)-(number of predictand)

4.9 Goodness of Fit of Rainfall Amount Models

The performance of rainfall amount models developed in this study were evaluated and compared using the parametric (RMSE), non-parametric tests (Kolmogorov-Smirnov test, Mann-Whitney U test, squared-rank test, Kendall's tau-b correlation and Spearman's rho correlation) and quantile plots by comparing the observed and simulated monthly rainfall series. Besides, the performance of combined BACT-ANN model were further compared with traditional ANN and NHMM. It is noteworthy that the observed rainfall series was proved to be skewed and not normally distributed, so the non-parametric tests were main evaluation criteria to be focused in this study.

4.9.1 Parametric Test

Root mean square error (RMSE) is the one of the measures commonly used to evaluate the performance of a statistical downscaling model in terms of error. Both traditional ANN and combined BACT-ANN model, which exhibited the lowest RMSE during validation period, were selected as the well-trained models in this study. The performance of selected traditional ANN and BACT-ANN model were further evaluated and compared together with NHMM in terms of their RMSE, as shown in Table 4.18. Generally, the lower RMSE value is indicating the accuracy of model is higher with the smaller error is produced.

Overall, the combined BACT-ANN model showed the smaller errors than ANN and NHMM in the majority of 12 months during validation period. According to the results in Table 4.18, the combined BACT-ANN model, traditional ANN and NHMM produced the RMSE in the range of 38.75–144.10 mm, 49.53–155.23 mm, and 69.35–230.22 mm, respectively. Hence, the combined BACT-ANN model exhibited the better performance in this study with the smaller range of RMSE produced, followed by the traditional ANN and NHMM.

In other word, the application of combined BACT-ANN model reduced the maximum RMSE produced from traditional ANN and NHMM model up to 11.13 mm and 86.12 mm, respectively. However, the measure of RMSE belongs to one of the parametric tests, which is used to analyse the normal distributed data, but the observed rainfall series in this study as reported much earlier herein, were proved to be not normally distributed. Therefore, the performance of these three models were further evaluated and compared using the non-parametric tests.

Table 4.18 : Root mean square error (RMSE) of rainfall amount models during validation (1996–2005) period

Station	Month	Validation		
		ANN	NHMM	BACT-ANN
2815001	Jan	110.21	106.81	39.49
	Feb	70.74	84.38	42.36
	Mar	73.76	91.30	42.74
	Apr	72.42	113.31	52.57
	May	63.68	99.14	38.75
	Jun	57.53	93.80	66.85
	Jul	87.49	114.13	87.33
	Aug	54.27	88.73	59.08
	Sep	89.27	83.56	66.71
	Oct	91.64	77.72	53.74
	Nov	106.46	142.34	63.97
	Dec	77.96	69.35	61.53
2913001	Jan	96.02	125.14	76.71
	Feb	76.60	139.96	58.51
	Mar	88.32	165.49	46.06
	Apr	95.75	130.51	58.89
	May	49.53	95.11	51.33
	Jun	66.52	82.28	67.45
	Jul	72.09	136.02	80.99
	Aug	75.98	137.48	65.97
	Sep	84.72	94.64	93.04
	Oct	97.56	143.12	102.00
	Nov	106.83	134.08	78.33
	Dec	92.01	141.01	57.17
2917001	Jan	74.60	93.93	103.67
	Feb	84.50	131.78	122.36
	Mar	114.88	132.71	82.33
	Apr	139.92	183.62	89.20
	May	79.46	115.36	52.66
	Jun	80.99	93.25	50.58
	Jul	69.47	102.29	91.16
	Aug	121.53	140.81	130.26
	Sep	67.50	80.89	59.15
	Oct	106.92	154.43	118.07
	Nov	155.23	230.22	128.72
	Dec	79.53	159.28	65.86

- Grey shaded value indicated the model with the smaller RMSE.

Table 4.18 continued: Root mean square error (RMSE) of rainfall amount models during validation (1996–2005) period

Station	Month	Validation		
		ANN	NHMM	BACT-ANN
3118102	Jan	59.72	117.38	57.17
	Feb	77.26	161.90	100.07
	Mar	121.69	102.65	144.10
	Apr	97.38	124.84	109.17
	May	89.96	113.91	74.77
	Jun	93.91	69.97	84.82
	Jul	101.83	135.41	59.56
	Aug	96.34	112.28	66.14
	Sep	70.29	90.75	60.02
	Oct	97.79	107.48	69.85
	Nov	148.32	156.31	134.10
	Dec	100.78	132.72	95.66

- Grey shaded value indicated the model with the smaller RMSE.

4.9.2 Non-parametric Tests

Kolmogorov-Smirnov (K-S) test is one of the non-parametric statistical hypothesis tests used to determine whether two independent samples of data come from the same or different distributions. In this study, the K-S test was used to evaluate the significance difference between the observed and simulated monthly rainfall distributions under the significance level of 0.05. The computed monthly p-values of traditional ANN, NHMM and combined BACT-ANN model at each station, are present in Table 4.19. For instance, the computed p-value of combined BACT-ANN model in January was 0.952 at station 2815001, which is greater than the significance level. Therefore, the null hypothesis cannot be rejected and there is up to 95.2% to indicate the null hypothesis is true.

Based on the results in Table 4.19, both NHMM and combined BACT-model exhibited the better performance than ANN with the p-values exceed significance level of 0.05 in the majority of twelve months. Especially for station 2913001, both NHMM and combined BACT-ANN model able to simulate the distribution of observed rainfall series with the p-values greater than 0.05 in every month. However, the combined BACT-ANN model showed the higher number of months (eleven out of twelve months) passed the significant level of 0.05 than the NHMM (ten out of twelve months) at stations 2917001 and 3118102. In overall, the combined BACT-ANN model produced the p-values in the range of 0.035–0.998, which were higher than ANN (<0.001–0.586) and NHMM (0.003–0.952).

Table 4.19: Kolmogorov-Smirnov (K-S) tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0.035	0.388	0.952
	Feb	0.134	0.134	0.952
	Mar	0.003	0.071	0.236
	Apr	0.016	0.586	0.134
	May	0.016	0.799	0.035
	Jun	0.003	0.799	0.799
	Jul	0.001	0.586	0.586
	Aug	0.071	0.586	0.799
	Sep	0.001	0.035	0.236
	Oct	0	0.134	0.071
	Nov	0.016	0.016	0.035
	Dec	0.236	0.134	0.134

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

Table 4.19 continued: Kolmogorov-Smirnov (K-S) tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2913001	Jan	0.134	0.799	0.952
	Feb	0.035	0.134	0.236
	Mar	0.003	0.071	0.952
	Apr	0.003	0.236	0.586
	May	0.035	0.071	0.236
	Jun	0.001	0.388	0.998
	Jul	0.003	0.134	0.586
	Aug	0.003	0.952	0.952
	Sep	0.134	0.586	0.799
	Oct	0.003	0.236	0.799
	Nov	0.003	0.071	0.388
	Dec	0.071	0.952	0.799
2917001	Jan	0.035	0.236	0.586
	Feb	0.071	0.071	0.236
	Mar	0.035	0.388	0.952
	Apr	0.007	0.799	0.388
	May	< 0.001	0.003	0.035
	Jun	0.007	0.134	0.799
	Jul	0.035	0.071	0.998
	Aug	0.071	0.071	0.799
	Sep	0.586	0.388	0.799
	Oct	0.236	0.388	0.586
	Nov	0.071	0.016	0.586
	Dec	0.035	0.998	0.236
3118102	Jan	0	0.035	0.799
	Feb	0.035	0.134	0.799
	Mar	0.007	0.388	0.952
	Apr	0.016	0.236	0.799
	May	0.134	0.236	0.586
	Jun	0	0.035	0.388
	Jul	0.134	0.586	0.035
	Aug	0.071	0.236	0.586
	Sep	0.035	0.586	0.236
	Oct	0.134	0.236	0.388
	Nov	0.007	0.035	0.236
	Dec	0.007	0.586	0.388

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

In addition, Mann-Whitney U test is another parametric test used to compare two independent samples and determine whether the samples are identical to each other or not based on their ranks. In this study, the Mann-Whitney U test was used to evaluate equality between the observed and simulated monthly rainfall series relative to their ranks. If the computed p-value exceeded the significance level of 0.05, then the null hypothesis is accepted with the difference of location between the samples is equal to zero. All three models exhibited the good results with the computed p-value exceeded the significance level of 0.05 in the majority on twelve months, as shown in Table 4.20. Besides, the p-values produced from combined BACT-ANN model in the range of 0.012–0.971, was relatively higher than other models and indicating the higher probability of both simulated and observed rainfall series to be identical with each other.

Table 4.20: Mann-Whitney U tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0.068	0.171	0.971
	Feb	0.387	0.018	0.888
	Mar	0.096	0.059	0.154
	Apr	0.935	0.663	0.11
	May	0.059	0.371	0.102
	Jun	0.167	0.186	0.544
	Jul	0.167	0.228	0.297
	Aug	0.663	0.326	0.877
	Sep	0.05	0.038	0.307
	Oct	0.07	0.115	0.012
	Nov	0.075	0.001	0.014
	Dec	0.994	0.167	0.099

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

Table 4.20 continued: Mann-Whitney U tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2913001	Jan	0.888	0.923	0.706
	Feb	0.673	0.085	0.339
	Mar	0.018	0.018	0.501
	Apr	0.201	0.267	0.544
	May	0.039	0.042	0.387
	Jun	0.145	0.234	0.9
	Jul	0.149	0.141	0.391
	Aug	0.234	0.549	0.947
	Sep	0.923	0.326	0.982
	Oct	0.201	0.22	0.784
	Nov	0.052	0.048	0.46
	Dec	0.819	0.679	0.383
2917001	Jan	0.252	0.048	0.234
	Feb	0.492	0.112	0.258
	Mar	0.171	0.695	0.959
	Apr	0.026	0.959	0.212
	May	< 0.0001	0.004	0.191
	Jun	0.064	0.02	0.652
	Jul	0.333	0.044	0.695
	Aug	0.246	0.096	0.511
	Sep	0.877	0.641	0.929
	Oct	0.141	0.395	0.819
	Nov	0.473	0.002	0.367
	Dec	0.105	0.706	0.717
3118102	Jan	0	0.013	0.706
	Feb	0.684	0.333	0.882
	Mar	0.865	0.842	0.842
	Apr	0.326	0.141	0.544
	May	0.446	0.122	0.391
	Jun	0.01	0.021	0.252
	Jul	0.971	0.539	0.038
	Aug	0.326	0.511	0.315
	Sep	0.935	0.673	0.145
	Oct	0.61	0.212	0.196
	Nov	0.171	0.012	0.367
	Dec	0.234	0.717	0.351

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

Furthermore, the squared-rank test was another non-parametric test used to assess the equality of variance between observed and simulated monthly rainfall series with the null hypothesis of the samples are from the identical distribution. For station 2815001, the p-value obtained by combined BACT-ANN model was 0.967, which indicated the distributions of observed and simulated rainfall series in January were identical with no significant difference between their variances. Both NHMM and combined BACT-ANN model exhibited the better performance than ANN with the p-values exceeded the significance level of 0.05 in the majority of twelve months. Based on the results in Table 4.21, the ANN basically produced the p-value lesser than 0.001 in most of the months, so the simulated rainfall series from ANN was inferred to be significance different from observed rainfall series regarding to their variances.

Table 4.21: Squared-rank tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0.019	0.676	0.967
	Feb	0.000	0.827	0.199
	Mar	0.000	0.909	0.072
	Apr	< 0.001	0.523	0.179
	May	< 0.001	0.964	0.016
	Jun	< 0.001	0.188	0.705
	Jul	< 0.001	0.554	0.947
	Aug	< 0.001	0.343	0.401
	Sep	< 0.001	0.149	0.534
	Oct	< 0.001	0.450	0.702
	Nov	0	0.978	0.738
	Dec	0.004	0.798	0.797

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

Table 4.21 continued: Squared-rank tests p-values results of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2913001	Jan	< 0.001	0.401	0.475
	Feb	< 0.001	0.639	0.090
	Mar	< 0.001	0.543	0.646
	Apr	< 0.001	0.147	0.350
	May	< 0.001	0.088	0.039
	Jun	< 0.001	0.816	0.537
	Jul	< 0.001	0.541	0.963
	Aug	< 0.001	0.953	0.910
	Sep	< 0.001	0.576	0.354
	Oct	< 0.001	0.141	0.720
	Nov	< 0.001	0.421	0.334
	Dec	< 0.001	0.450	0.803
2917001	Jan	< 0.001	0.305	0.581
	Feb	< 0.001	0.141	0.043
	Mar	< 0.001	0.014	0.277
	Apr	< 0.001	0.148	0.388
	May	0.049	0.102	0.012
	Jun	< 0.001	0.931	0.070
	Jul	< 0.001	0.299	0.610
	Aug	< 0.001	0.455	0.150
	Sep	0.006	0.467	0.553
	Oct	0.076	0.343	0.390
	Nov	< 0.001	0.015	0.304
	Dec	0.004	0.920	0.276
3118102	Jan	0.031	0.897	0.495
	Feb	< 0.001	0.230	0.479
	Mar	< 0.001	0.004	0.749
	Apr	< 0.001	0.034	0.228
	May	< 0.001	0.167	0.755
	Jun	< 0.001	0.161	0.431
	Jul	< 0.001	0.487	0.212
	Aug	< 0.001	0.032	0.728
	Sep	< 0.001	0.190	0.792
	Oct	< 0.001	0.015	0.962
	Nov	< 0.001	0.025	0.073
	Dec	< 0.001	0.004	0.577

- Grey shaded value indicated the p-value of model exceeded the significance level of 0.05.

In considering that the main objective of this study is to select the most suitable rainfall generator for Langat River Basin, and the computed p-values in Tables 4.19–4.21 could not provide the significance information for the selection of the model, for this reason, the acceptability index (*AI*) proposed by Fodor et al. (2010) was used for further examination in summarizing the tests performance of these models. The computation of *AI* is related to the total score of model in passing the significance level of statistical tests. The higher the *AI*, the better the performance of model. As illustrated in Table 4.22, the combined BACT-ANN model achieved the better results by passing most of the K-S, Mann-Whitney U and squared-rank tests, thus it produced the *AI* higher than 91% in all three tests. In other words, the combined BACT-ANN model is capable of reproducing the distribution of observed monthly rainfall series with the lesser significance differences, when compared to ANN and NHMM. Also, the inclusion of rainfall occurrence model in this study did enhance the performance of traditional ANN model with the increase of approximate 60% *AI* in K-S tests, 8% *AI* in Mann-Whitney U tests and 89% *AI* in squared-rank tests.

Table 4.22: Acceptability index for the simulated monthly rainfall series in passing the Kolmogorov-Smirnov (K-S), Mann-Whitney U and squared-rank tests for all stations

	Acceptability Index (%)		
	K-S tests	Mann-Whitney U tests	Squared-rank tests
ANN	31.25	85.42	2.08
NHMM	85.42	70.83	83.33
BACT-ANN	91.67	93.75	91.67

In addition, there were two measures of non-parametric rank correlations used in this study to assess the statistical associations between observed and simulated monthly rainfall series, namely Kendall's tau-b correlation and Spearman's rho correlation. Both correlation coefficients in the range of -1 to 1, where the positive correlation implies that the ranks of both variables are increasing, while the negative correlation implies that the ranks of both variables are moving in opposite direction. These analyses can be used in assessing the correlation between two variables in hypothesis testing, with the null hypothesis of there is no correlation between two series.

Overall, the combined BACT-ANN model achieved the better performance than ANN and NHMM, with its computed p-values of Kendall's tau-b correlation coefficient smaller than significance level of 0.05 in every month. For example, the correlation coefficient of combined BACT-ANN model at station 2815001 was 0.707 in January, with the p-value smaller than 0.05. Thus, the null hypothesis is rejected and the obtained correlation coefficient is significantly different from zero. As presented in Table 4.23, the correlation coefficients obtained by combined BACT-ANN model were ranged between 0.287–0.737 and significantly different from zero correlation. Both ANN and NHMM showed the relatively low correlation coefficients with some computed p-values exceeded the significance level of 0.05 in certain months.

Table 4.23: Kendall's tau-b correlation coefficients of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0.29	-0.11	0.707
	Feb	0.211	-0.058	0.594
	Mar	0.115	0.124	0.737
	Apr	0.148	-0.088	0.569
	May	0.41	-0.299	0.494
	Jun	0.171	-0.125	0.509
	Jul	0.064	0.032	0.406
	Aug	0.295	0.087	0.525
	Sep	-0.148	0.037	0.375
	Oct	0.03	0.269	0.641
	Nov	0.108	-0.214	0.655
	Dec	0.329	0.136	0.545
2913001	Jan	0.198	-0.056	0.659
	Feb	0.354	-0.144	0.636
	Mar	0.15	-0.127	0.622
	Apr	0.383	-0.051	0.593
	May	0.192	-0.155	0.515
	Jun	0.197	0.192	0.45
	Jul	0.177	-0.137	0.341
	Aug	0.215	-0.083	0.651
	Sep	-0.064	0.087	0.355
	Oct	0.078	-0.295	0.551
	Nov	0.014	0.177	0.569
	Dec	0.314	-0.132	0.588
2917001	Jan	0.2	0.375	0.428
	Feb	0.444	0.053	0.407
	Mar	0.14	0.09	0.503
	Apr	0.108	-0.136	0.582
	May	0.37	0.039	0.563
	Jun	0.147	0.17	0.603
	Jul	0.343	0.067	0.287
	Aug	0.375	0.186	0.406
	Sep	0.393	0.287	0.434
	Oct	0.31	-0.025	0.338
	Nov	0.338	0.03	0.595
	Dec	0.448	0.053	0.674

- Grey shaded value indicated the coefficient with the p-value smaller than the significance level of 0.05.

Table 4.23 continued: Kendall’s tau-b correlation coefficients of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
3118102	Jan	0.346	0.009	0.541
	Feb	0.343	0.182	0.494
	Mar	0.267	0.198	0.382
	Apr	0.283	0.002	0.329
	May	0.255	0.14	0.577
	Jun	0.23	0.148	0.671
	Jul	0.094	-0.136	0.506
	Aug	0.251	0.067	0.54
	Sep	0.285	0	0.554
	Oct	0.14	0.113	0.674
	Nov	-0.053	-0.195	0.48
	Dec	0.291	0.065	0.611

- Grey shaded value indicated the coefficient with the p-value smaller than the significance level of 0.05.

Basically, the Spearman’s rho correlation analysis usually obtain the coefficients greater than Kendall’s tau-b correlation analysis. Besides, this analysis is much more sensitive to error and discrepancies in data. Based on the results in Table 4.24, the coefficients of Spearman’s rho were apparently larger than Kendall’s tau-b, but the evaluation results remained the same with the combined BACT-ANN model outperformed ANN and NHMM by exhibiting the higher significant correlation coefficients in the range of 0.418–0.876. Both ANN and NHMM obtained the some positive and negative correlation coefficients, which were not significantly different from zero in certain months.

Table 4.24: Spearman’s rho correlation coefficients of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0.355	-0.158	0.875
	Feb	0.32	-0.061	0.779
	Mar	0.176	0.168	0.876
	Apr	0.26	-0.116	0.73
	May	0.57	-0.406	0.658
	Jun	0.286	-0.199	0.685
	Jul	0.111	0.076	0.575
	Aug	0.408	0.098	0.699
	Sep	-0.211	0.076	0.509
	Oct	0.048	0.408	0.788
	Nov	0.11	-0.278	0.845
	Dec	0.449	0.203	0.72
2913001	Jan	0.295	-0.08	0.844
	Feb	0.495	-0.19	0.825
	Mar	0.251	-0.178	0.813
	Apr	0.526	-0.046	0.759
	May	0.284	-0.213	0.664
	Jun	0.296	0.285	0.578
	Jul	0.277	-0.191	0.488
	Aug	0.29	-0.121	0.839
	Sep	-0.11	0.123	0.506
	Oct	0.108	-0.416	0.727
	Nov	0.019	0.236	0.734
	Dec	0.395	-0.2	0.761
2917001	Jan	0.303	0.501	0.528
	Feb	0.616	0.102	0.578
	Mar	0.219	0.127	0.712
	Apr	0.178	-0.191	0.753
	May	0.507	0.068	0.748
	Jun	0.247	0.215	0.765
	Jul	0.491	0.096	0.418
	Aug	0.464	0.278	0.545
	Sep	0.573	0.378	0.626
	Oct	0.498	-0.018	0.466
	Nov	0.503	0.05	0.778
	Dec	0.656	0.08	0.846

- Grey shaded value indicated the coefficient with the p-value smaller than the significance level of 0.05.

Table 4.24 continued: Spearman’s rho correlation coefficients of rainfall amount models at each station

Station	Month	ANN	NHMM	BACT-ANN
3118102	Jan	0.496	0.016	0.701
	Feb	0.475	0.248	0.653
	Mar	0.343	0.299	0.513
	Apr	0.36	0.003	0.44
	May	0.38	0.237	0.771
	Jun	0.289	0.258	0.793
	Jul	0.123	-0.163	0.644
	Aug	0.387	0.093	0.708
	Sep	0.489	-0.013	0.771
	Oct	0.175	0.161	0.855
	Nov	-0.101	-0.257	0.621
	Dec	0.428	0.087	0.815

- Grey shaded value indicated the coefficient with the p-value smaller than the significance level of 0.05.

Similar to the K-S tests, the acceptability index (*AI*) was also employed to evaluate the overall performance of models in Kendall’s tau-b and Spearman’s rho correlations analysis. Based on the results in Table 4.25, the combined BACT-ANN model is capable to produce the monthly rainfall series, which were significantly correlated to the observed series of selected rainfall stations within Langat River Basin, thus resulting in 100% *AI* for both Kendall’s tau-b and Spearman’s rho correlations. On the other hand, the ANN and NHMM obtained the *AI* smaller than 50% in both correlations, which mean their simulated and observed monthly rainfall series were less likely correlated to each other. Besides, the strength of coefficients obtained by combined BACT-ANN model in Spearman’s rho correlation were in the range of “moderate” to “very strong”. Hence, the monthly rainfall series simulated by combined BACT-ANN model is more reliable than ANN and NHMM.

Table 4.25: Acceptability index for the simulated monthly rainfall series obtaining the significant coefficient in Kendall's tau-b and Spearman's rho correlations for all stations

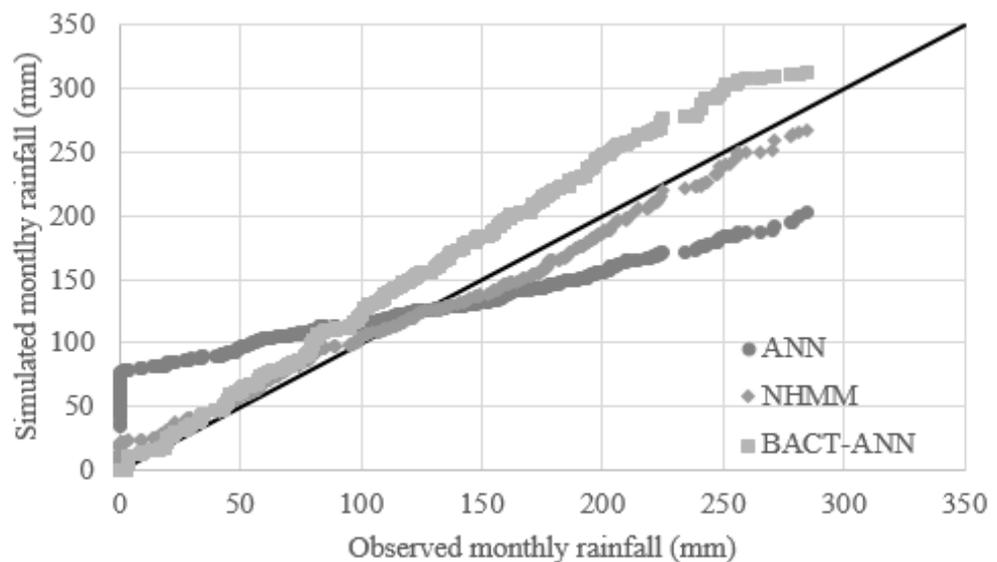
	Acceptability Index (%)	
	Kendall's tau-b	Spearman's rho
ANN	43.75	41.67
NHMM	10.42	10.42
BACT-ANN	100	100

4.9.3 Quantile Plot

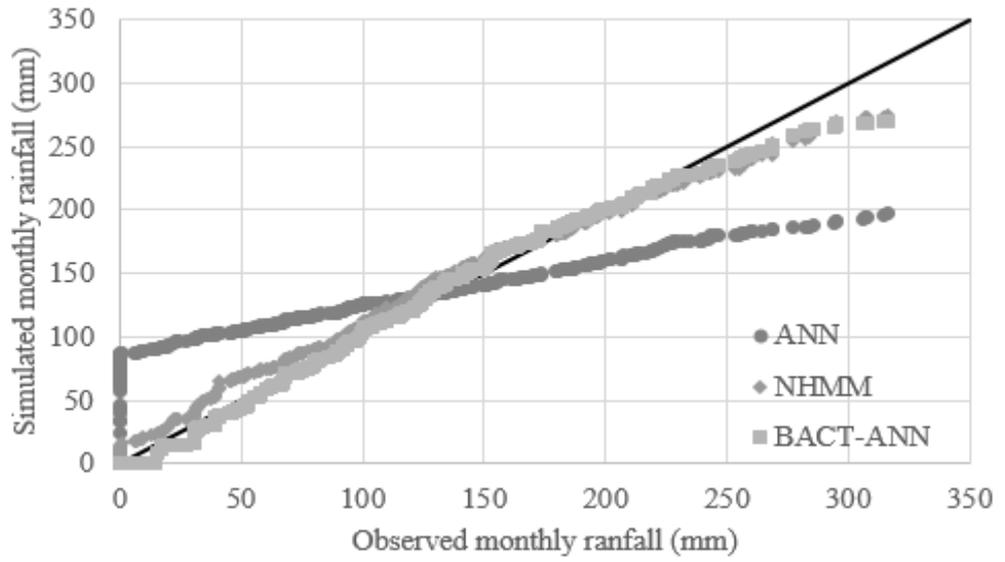
Other than the RMSE and non-parametric tests, there was another diagnostic test used in this study for the reproduction of observed monthly rainfall series. The scatter plots of simulated (ANN, NHMM and combined BACT-ANN model) versus observed 95th percentile of monthly rainfall series for all stations during study periods (years 1976–2005), are illustrated in Figure 4.6. Both ANN and NHMM showed the unstable prediction abilities in simulating the observed rainfall series at each station, with the over-prediction of small rainfall amount and under-prediction of huge rainfall amount. Based on the results in Figure 4.6, both ANN and NHMM achieved the threshold values of 129.2 mm, 140.9 mm, 224 mm and 165.6 mm at station 2815001, station 2913001, station 2917001 and station 3118102, respectively. These threshold values indicated that both models have the over-prediction tendency at the rainfall value smaller than threshold value and under-prediction tendency at the rainfall value higher than threshold value. In addition, both ANN and NHMM also obtained the higher bias values in both over-prediction and under-prediction, as shown in Table 4.26.

The combined BACT-ANN model exhibited the good performance in simulating the monthly rainfall distribution close to 1:1 line at station 2913001 (Figure 4.6(b)) and station 2917001 (Figure 4.6(c)). However, the combined BACT-ANN model showed the consistent prediction ability in over-predicting the monthly rainfall series with the distribution above 1:1 line, as shown in Figure 4.6(a) and Figure 4.6(d). This situation can be justified by the bias calculation results in Table 4.26. Specifically for station 2815001, the combined BACT-ANN model achieved a very high bias value in over-prediction but a low bias value very near to zero in under-prediction. Overall, the combined BACT-ANN exhibited better performance than ANN and NHMM with the smaller bias values in the over-prediction and under-prediction of rainfall.

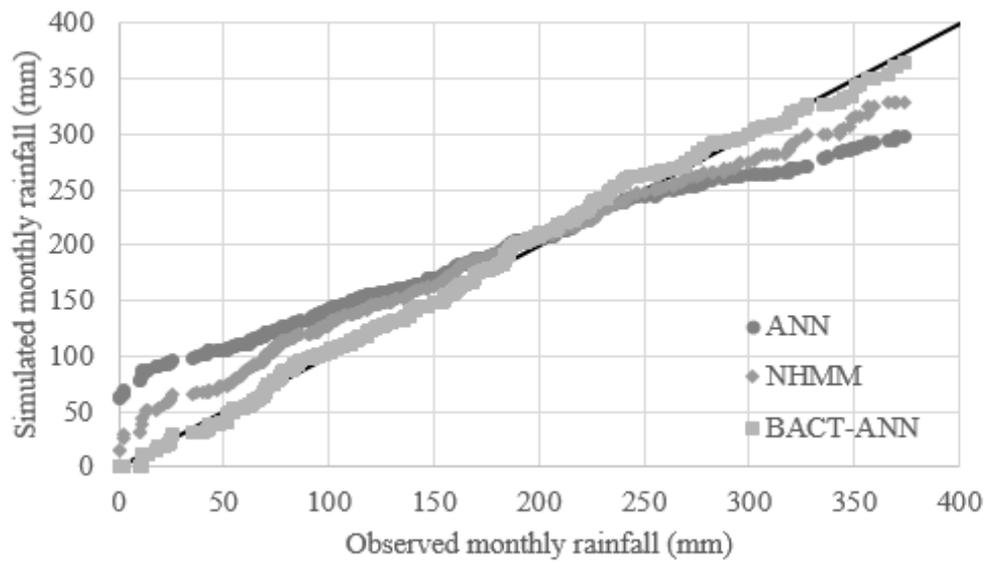
(a) Station 2815001



(b) Station 2913001



(c) Station 2917001



(d) Station 3118102

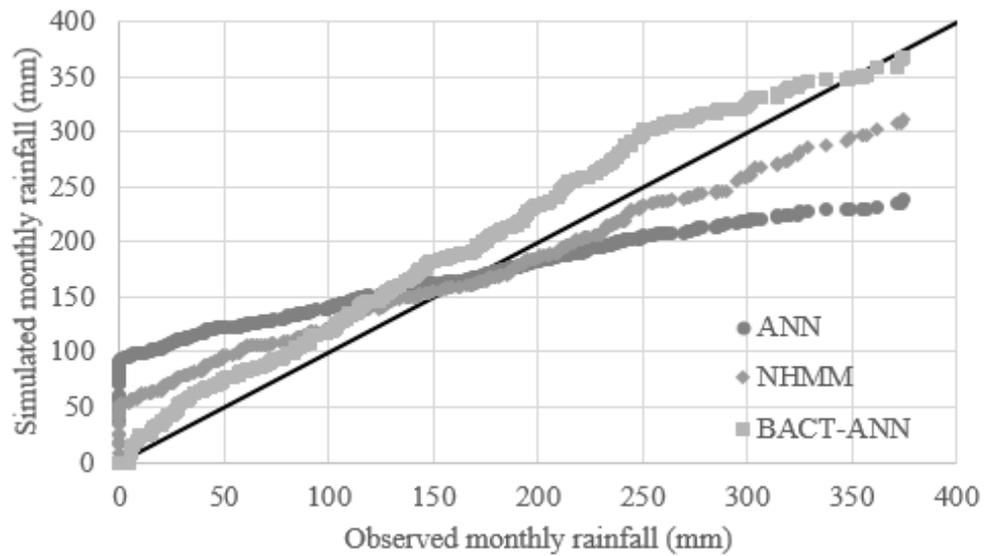


Figure 4.6: Scatter plots of simulated (ANN, NHMM and combined BACT-ANN model) versus observed 95th percentile of monthly rainfall series at (a) station 2815001, (b) station 2913001, (c) station 2917001 and (d) station 3118102

Table 4.26: Bias calculation for the over-prediction and under-prediction of monthly rainfall series at each station

Model	Over-prediction			Under-prediction		
	ANN	NHMM	Combined BACT- ANN	ANN	NHMM	Combined BACT- ANN
2815001	38.2	8.9	28.6	-42.7	-11.6	-1.4
2913001	43.7	11.4	5.0	-56.7	-19.2	-14.1
2917001	31.0	20.1	6.2	-52.0	-27.8	-9.3
3118102	54.0	31.3	27.0	-57.5	-29.1	-20.6

CHAPTER 5

CONCLUSIONS

5.1 Summary

Statistical downscaling is a technique which has been widely used to predict the local climate variables through a robust statistical relationship established between the large scale atmospheric (predictors) and local climate (predictand) variables. However, the condition of insufficient efficiency or poor performance of traditional individual models, led to the development of hybrid models as a hydrological model to predict or downscale the future rainfall data. There was a two-stage approach employed in this study for the purpose of improving the performance of a traditional ANN. The first stage is the application of bootstrap aggregated classification tree (BACT) model with the random forest algorithm to model the daily rainfall occurrence and, followed by the second stage, which is the modelling of rainfall amount using ANN model with the Levenberg-Marquardt approach conditional on the simulated wet days.

The Langat River Basin is the selected study area and the observed daily rainfall series (predictand) from four selected stations within the basin were collected from MMD and DID. Besides, there was another large scale data required, namely the NCEP & NCAR reanalysis dataset (predictors), which was obtained from the official website of the Canadian Climate Data and Scenarios.

Three normality tests (the Anderson-Darling, the Lilliefors and the Jarque-Bera tests) and four absolute homogeneity tests (SNHT, BR, PET and VNR tests) were applied on the observed rainfall series to determine whether they are normally distributed and homogenous or not. Thereafter, the potential predictors were screened with the predictand for selecting the suitable set of predictors for the modelling of rainfall occurrence and amount.

5.1.1 Rainfall Occurrence Model

In this study, the threshold value of 0.1 mm was used to define a wet day; a larger threshold value may cause the under-prediction of rainfall occurrence in Langat River Basin. A three-state NHMM was chosen for station 2815001 and station 2913001, based on the consideration of log-likelihood values, Bayesian information criterion (BIC) and Akaike information criterion (AIC) scores. Unlike the BIC and AIC scores, the normalized log-likelihood decreases as the number of hidden state increase (from state 1 to state 7). The purpose of considering both BIC and AIC scores is to reduce the complexity of the model by selecting the optimum hidden state number. However, a two-state NHMM was chosen and adequately described the observed rainfall occurrence at station 2917001 and station 3118102. The advantage of using NHMM is the application of the first-order Markov chain model, which has been proven by many studies regarding to its ability in predicting the weather persistence.

Furthermore, unlike the simple Markov chain model with the state is visible and the transition probabilities between the states are the only parameters, the 'hidden' in NHMM is referring to the sequence of states through which the model passes, but not to parameters of model. They are not directly visible, but the output generated based on hidden state is visible. Compared to NHMM, both GLM and BACT model showed the simplicity in modelling the rainfall occurrence using the algorithm of logistic regression and random forests, respectively, without the application of any hidden state.

The main advantages of using logistic regression are that the predictors need not to be normally distributed and it ensures the generated output lies between 0 and 1, which can be interpreted as a probability. For the BACT model, the random forests algorithm was used to improve the performance of single classification tree by reducing the variance and increasing the stability of model. However, in considering the computational efficiency and stability of prediction, the number of grown trees was set to 150 in BACT model based on their generated OOB classification errors.

The performance of rainfall occurrence models (GLM, NHMM and BACT model) were evaluated and compared in terms of their rainfall persistence, spell lengths distribution and matching. As predicted, the NHMM exhibited the better performance with the smaller absolute difference of rainfall persistence during validation period, followed by BACT model and GLM. This is because of the transition probabilities between the hidden states are defined by a first-order Markov chain model, in which they are conditional on the predictors on

that day and the rainfall state on previous day. Besides, both NHMM and BACT model showed good prediction ability in simulating the observed wet- and dry-spell lengths with slight under-prediction and over-prediction at some stations. However, the GLM showed the poor prediction ability with the significant under-prediction at every station when compared to them.

Further to this, the performance of developed rainfall occurrence models were evaluated using a 2x2 contingency table, which can precisely compare the matching between observed and simulated rainfall occurrence series. The BACT model outperformed the GLM and NHMM with the higher POD and CSI indices in the range of 0.51–0.65 and 0.29–0.44, respectively. Nevertheless, they over-predicted the number of wet days with the FAR higher than GLM's but lower than NHMM's. For the comparison in terms of skill score, the prediction ability of NHMM was more likely to be equivalent to a random prediction with the HSS and PSS indices nearly equal to zero at every station. Both GLM and BACT model exhibited a reasonably good prediction with the skill scores greater than 0.15 at every station, except for station 2913001. However, the BACT model still outperformed the GLM with the higher skill scores during calibration and validation periods. Based on the results in rainfall persistence, it can be concluded that the NHMM performed well in predicting the rainfall occurrence only if the time frame is not taken into account. Eventually, the BACT model was selected and recommended as the suitable rainfall occurrence model for this study due to its overall good simulation performance.

5.1.2 Rainfall Amount Model

In this study, the suitable rainfall distribution type in NHMM was determined based on the log-likelihood values, Bayesian information criterion (BIC) and Akaike information criterion (AIC) scores. The gamma distribution was selected as the suitable rainfall distribution to model the rainfall intensity of selected rainfall stations within the basin, due to its outstanding performance over the single-exponential and two-component mixed-exponential distributions. Besides, the number of hidden neurons used in combined BACT-ANN model was greater than the traditional ANN model due to the application of pre-processing approach, which resulted in the increasing number of wet days.

The performance of rainfall amount models (ANN, NHMM and combined BACT-ANN model) were evaluated and compared using the parametric (RMSE) and non-parametric (K-S tests, Mann-Whitney U tests, squared-rank tests, Kendall's tau-b correlation and Spearman's rho correlation) tests. According to the results in parametric tests, the combined BACT-ANN model produced the smaller range of RMSE during validation period, when compared to ANN and NHMM. Besides, the application of combined BACT-ANN model reduced the maximum RMSE produced from ANN and NHMM model up to 11.13 mm and 86.12 mm, respectively.

In addition, the combined BACT-ANN model achieved the *AI* greater than 91% in passing most of the K-S, Mann-Whitney U and squared-rank tests. Other than that, the combined BACT-ANN model was also capable to produce

the monthly rainfall series which are significantly correlated to the observed series, thus resulted in 100% *AI* for both Kendall's tau-b and Spearman's rho correlations. The strength of coefficients obtained by combined BACT-ANN model in Spearman's rho correlation were in the range of "moderate" to "very strong". However, the ANN and NHMM obtained the *AI* smaller than 50% in both correlations. Therefore, the combined BACT-ANN model was able to reproduce the distribution of observed monthly rainfall series with the lesser significance differences and more reliable prediction, when compared to ANN and NHMM.

Based on the results in quantile plots, the combined BACT-ANN model also exhibited the good simulation results with the monthly rainfall distribution close to 1:1 line at station 2913001 and station 2917001. However, it over-predicted the monthly rainfall amount at station 2815001 and station 3118102. This can be justified by the bias analyses, where the high bias value is obtained in the over-prediction of rainfall amount, but very low bias value in the under-prediction of rainfall amount. Furthermore, both ANN and NHMM showed the inconsistent prediction stability in simulating the observed monthly rainfall series with the higher bias values in both over-prediction and under-prediction of rainfall amount at each station. The combined BACT-ANN model not only reduced the errors produced by ANN but also improved their performance with the increase of approximate 60% *AI* in K-S tests, 8% *AI* in Mann-Whitney U tests and 89% *AI* in squared-rank tests. For the reasons delineated, the combined BACT-ANN model outperformed both the NHMM and ANN, and is recommended as the suitable rainfall generator for this study.

5.2 Recommendations

In order to have consistent periods between predictors and predictand, the study period used in this study was 30 years (years 1976–2005) only, which is considered short with 20 years used for calibration and 10 years used for validation. It must be reiterated here that this period is the longest period where the rainfall data is considered reliable. Therefore, the longer study period is recommended in future research in order to improve the performance of BACT model in simulating the observed rainfall occurrence series, especially reduce the over-prediction of number of wet days during validation period (The current percentage of over-prediction in this study is in the range of 20%–27%). Besides, longer study period also provide more information for ANN to establish more robust relationship between large scale atmospheric and local climate variables. It must be noted that another decade (2006–2016) can be included. This was not pursued due to research funding constraint.

The combined BACT-ANN model developed in this model can be used for further research in downscaling the future rainfall occurrence and amount using the future atmospheric variables projected by GCMs under different emission scenarios. However, the scaling method proposed by Abdellatif, et al. (2013) is recommended to prevent the bias in the case of utilizing the GCM outputs to downscale the future rainfall.

REFERENCES

- ABC NEWS, 2014. Malaysia floods: More than 100,000 evacuated after country's worst monsoon floods in decades, 26 December. Available at <http://www.abc.net.au/news/2014-12-26/northern-malaysia-hit-by-worst-monsoon-floods-in-decades/5989052> [Accessed: 8 December 2015].
- Abdellatif, M., Atherton, W. and Alkhaddar, R., 2013. A hybrid generalised linear and Levenberg–Marquardt artificial neural network approach for downscaling future rainfall in North Western England. *Hydrology Research*, 44(6), pp. 1084–1101.
- Abdulkadir, T. S., Salami, A. W. and Kareem, A. G., 2012. Artificial neural network modeling of rainfall in Ilorin, Kwara State, Nigeria. *Journal of Research Information in Civil Engineering*, 9(1), pp. 108–120.
- Agence France-Presse, 2014. Floods kill 21 in Malaysia, waters recede, 31 December. Available at <http://reliefweb.int/report/malaysia/floods-kill-21-malaysia-waters-recede> [Accessed: 8 December 2015].
- Ahmad, N. H. and Deni, S. M., 2013. Homogeneity test on daily rainfall series for Malaysia. *Matematika*, 29(1), pp. 141–150.
- Ahmed, K., Shahid, S., Haroon, S. and Wang, X.J., 2015. Multilayer perceptron neural network for downscaling rainfall in arid region: A case study of Baluchistan, Pakistan. *Journal of Earth System Science*, 124(6), pp. 1325–1341.
- Amirabadizadeh, M., Ghazali, A.H., Huang, Y.F. and Wayayok, A., 2016. Downscaling daily precipitation and temperatures over the Langat River Basin in Malaysia: A comparison of two statistical downscaling approaches. *International Journal of Water Resources and Environmental Engineering*, 8(10), pp. 120–136.
- Beckmann, B. R. and Buishand, T. A., 2002. Statistical downscaling relationships for precipitation in the Netherlands and North Germany. *International Journal of Climatology*, 22(1), pp. 15–32.
- Beecham, S., Rashid, M. and Chowdhury, R. K., 2014. Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalized Linear Model. *International Journal of Climatology*, 34(14), pp. 3654–3670.
- Bellone, E., Hughes, J. P. and Guttorp, P., 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Research*, 15(1), pp. 1–12.

- Boe, J., Terray, L., Habets, F. and Martin, E., 2006. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *Journal of Geophysical Research: Atmospheres*, 111(D23).
- Breiman L., 1996. Bagging predictors. *Machine Learning*, 24, pp. 123 – 140.
- Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp. 5–32.
- Brinkmann, W. A. R., 2000. Modification of a correlation-based circulation pattern classification to reduce within-type variability of temperature and precipitation. *International Journal of Climatology*, 20(8), pp. 839–852.
- Bucklin, D. N. et al., 2013. Climate downscaling effects on predictive ecological models: a case study for threatened and endangered vertebrates in the southeastern United States. *Regional Environmental Change*, 13(S1), pp. 57–68.
- Buishand, T. A., Shabalova, M. V. and Brandsma, T., 2004. On the choice of the temporal aggregation level for statistical downscaling of precipitation. *Journal of Climate*, 17(9), pp. 1816–1827.
- Busuioc, A., Tomozeiu, R. and Cacciamani, C., 2008, Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia-Romagna region. *International Journal of Climatology*, 28(4), pp. 449–464.
- Campozano, L., Tenelanda, D., Sanchez, E., Samaniego, E. and Feyen, F., 2016. Comparison of statistical downscaling methods for monthly total precipitation: case study for the Paute River Basin in Southern Ecuador. *Advances in Meteorology*, 2016, pp. 1–13.
- Cannon, A. J., 2008. Probabilistic multisite precipitation downscaling by an expanded bernoulli-gamma density network. *Journal of Hydrometeorology*, 9(6), pp. 1284–1300.
- Caron, A., Leconte, R. and Brissette, F., 2008. An improved stochastic weather generator for hydrological impact studies. *Canadian Water Resources Journal*, 33(3), pp. 233–256.
- Chan, J. C.-W. and Paelinckx, D., 2008. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6), pp. 2999–3011.
- Chandler, R. E. and Wheeler, H. S., 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, 38(10), pp. 1-11.
- Chen, J., Brissette, F. P. and Leconte, R., 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. *Journal of Hydrology*, 401(3–4), pp. 190–202.

- Costa, A. C. and Soares, A., 2009. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Mathematical Geosciences*, 41(3), pp. 293–305.
- De'ath, G., 2007. Boosted trees for ecological modeling and prediction', *Ecology*, 88(1), pp. 243–251.
- Deshpande, R. R., 2012. On the rainfall time series prediction using multilayer perceptron artificial neural network. *International Journal of Emerging Technology and Advanced Engineering*, 2(1), pp. 148–153.
- Elith, J., Leathwick, J. R. and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), pp. 802–813.
- Engmann, S. and Cousineau, D., 2013. Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Komolgorov-Smirnov test. *Journal of Applied Quantitative Methods*, 6(3), pp. 1–17.
- Fealy, R. and Sweeney, J., 2007. Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. *International Journal of Climatology*, 27(15), pp. 2083–2094.
- Fodor, N., Dobi, I., Mika, J. and Szeidl, L., 2010. MV-WG: a new multi-variable weather generator. *Meteorology and Atmospheric Physics*, 107(3–4), pp. 91–101.
- Fowler, H. J., Blenkinsop, S. and Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12), pp. 1547–1578.
- Fu, G., Charles, S. P. and Kirshner, S., 2013. Daily rainfall projections from general circulation models with a downscaling nonhomogeneous hidden Markov model (NHMM) for south-eastern Australia. *Hydrological Processes*, 27(25), pp. 3663–3673.
- Gaitan, C. F., Hsieh, W. W. and Cannon, A. J., 2014. Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. *Climate Dynamics*, 43(12), pp. 3201–3217.
- Ghosh, S., 2010. SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *Journal of Geophysical Research*, 115(D22), p. D22102.
- Goyal, M. K. and Ojha, C. S. P., 2012. Downscaling of surface temperature for lake catchment in an arid region in India using linear multiple regression and neural networks. *International Journal of Climatology*, 32(4), pp. 552–566.

- Greene, A. M., Robertson, A.W. and Kirshner, S., 2011. Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Quarterly Journal of the Royal Meteorological Society*, 137(655), pp. 347–359.
- Halik, G., Anwar, N., Santosa, B. and Edijatno, 2015. Reservoir Inflow prediction under GCM scenario downscaled by wavelet transform and support vector machine hybrid models. *Advances in Civil Engineering*, 2015, pp. 1–9.
- Harpham, C. and Wilby, R. L., 2005. Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology*, 312(1–4), pp. 235–255.
- Hasan, M. M. and Dunn, P. K., 2012. Understanding the effect of climatology on monthly rainfall amounts in Australia using Tweedie GLMs. *International Journal of Climatology*, 32(7), pp. 1006–1017.
- Hassan, Z. and Harun, S., 2013. Impact of climate change on rainfall over Kerian, Malaysia with Long Ashton Research Station Weather Generator (Lars-Wg). *Malaysian Journal of Civil Engineering*, 25(1), pp. 33–44.
- Hassan, Z., Shamsudin, S. and Harun, S., 2014. Application of SDSM and LARS-WG for simulating and downscaling of rainfall and temperature. *Theoretical and Applied Climatology*, 116(1–2), pp. 243–257.
- He, X.G., Chaney, N.W., Schleiss, M. and Sheffield, J., 2016. Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, 52(10), pp. 8217–8237.
- Hutengs, C. and Vohland, M., 2016. Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, 178, pp. 127–141.
- Ibarra-Berastegi, G. et al, 2011, Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression. *Hydrology and Earth System Sciences*, 15(6), pp. 1895–1907.
- Ingsrisawang, L., Ingsriswang, S., Somchit, S., Aungsuratana, P. and Khtiyanan, W., 2008. Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern Part of Thailand. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 2(5), pp. 1422–1427.
- IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.

- Jeong, D. I., St-Hilaire, A., Ouarda, T.B.M.J. and Gachon, P., 2013. A multi-site statistical downscaling model for daily precipitation using global scale GCM precipitation outputs. *International Journal of Climatology*, 33(10), pp. 2431–2447.
- Jing, W.L., Yang, Y.P, Yue, X.F. and Zhao, X.D., 2016. A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over North China', *Remote Sensing*, 8(10), pp. 835–851.
- Jung, S.M., Lipe, D. and Quirk, T.J., 1971. An alteration of program u test to determine the direction of group differences for the Mann-Whitney U test. *Educational and Psychological Measurement*, 31, pp. 269–273.
- Ng, J.L., Aziz, S., Huang, Y.F, Wayayak, A. and Kamad, M.R., 2015. Homogeneity analysis of rainfall in Kelantan, Malaysia. *Jurnal Teknologi*, 76(15), pp. 1–6.
- Juahir, H. et al., 2011. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment*, 173(1–4), pp. 625–641.
- Kang, H. M. and Yusof, F., 2012 Homogeneity tests on daily rainfall series in Peninsular Malaysia. *International Journal of Contemporary Mathematical Science*, 7(1), pp. 9–22.
- Kannan, S. and Ghosh, S, 2011. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. *Stochastic Environmental Research and Risk Assessment*, 25(4), pp. 457–474.
- Kannan, S. and Ghosh, S., 2013. A nonparametric kernel regression model for downscaling multisite daily precipitation in the Mahanadi basin. *Water Resources Research*, 49(3), pp. 1360–1385.
- Kenabatho, P. K., McIntyre, N.R., Chandler, R.E. and Wheeler, H.S., 2012. Stochastic simulation of rainfall in the semi-arid Limpopo basin, Botswana. *International Journal of Climatology*, 32(7), pp. 1113–1127.
- Khan, M. S., Coulibaly, P. and Dibike, Y., 2006. Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319, pp. 357–382.
- Kioutsioukis, I., Melas, D. and Zanis, P., 2008. Statistical downscaling of daily precipitation over Greece. *International Journal of Climatology*, 28(5), pp. 679–691.
- Kumar, N. and Jha, G. K., 2013. A time series ANN approach for weather forecasting. *International Journal of Control Theory and Computer Modeling*, 3(1), pp. 19–25.
- Lee, T.C., Chan, K.Y. and Chan, H.S., 2011. Projections of extreme rainfall in Hong Kong in the 21st century. *Acta Meteorologica Sinica*, 25(6), pp. 691–709.

- Li, P. H., Kwon, H.H., Sun, L.Q., Lall, U. and Kao, J.J., 2010. A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan. *International Journal of Climatology*, 30(8), pp. 1256–1268.
- Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R News*, 2(3), pp. 18–22.
- Liu, W.B., Fu, G.B., Liu, C.M. and Charles, S.P., 2013. A comparison of three multi-site statistical downscaling models for daily rainfall in the North China Plain. *Theoretical and Applied Climatology*, 111(3–4), pp. 585–600.
- Liu, Y. and Fan, K., 2013. A new statistical downscaling model for autumn precipitation in China. *International Journal of Climatology*, 33(6), pp. 1321–1336.
- Liu, Z.F., Xu, Z.X., Charles, S.P., Fu, G.B. and Liu, L., 2011. Evaluation of two statistical downscaling models for daily precipitation over an arid basin in China. *International Journal of Climatology*, 31(13), pp. 2006–2020.
- Loo, Y. Y., Billa, L. and Singh, A., 2015. Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia. *Geoscience Frontiers*, 6(6), pp. 817–823.
- Lutz, K. et al., 2012. Comparison and evaluation of statistical downscaling techniques for station-based precipitation in the Middle East. *International Journal of Climatology*, 32(10), pp. 1579–1595.
- Mandal, S., Srivastav, R. K. and Simonovic, S. P., 2016. Use of beta regression for statistical downscaling of precipitation in the Campbell River basin, British Columbia, Canada. *Journal of Hydrology*, 538, pp. 49–62.
- Mares, C. et al., 2014. A hidden Markov model applied to the daily spring precipitation over the Danube Basin. *Advances in Meteorology*, 2014, pp. 1–11.
- Mehrotra, R., Srikanthan, R. and Sharma, A., 2006. A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology*, 331(1–2), pp. 280–292.
- Memarian, H., Balasundram, A.K., Talib, J.B., Sood, A.M. and Abbaspour, K.C., 2012. Trend analysis of water discharge and sediment load during the past three decades of development in the Langat basin, Malaysia. *Hydrological Sciences Journal*, 57(6), pp. 1207–1222.
- Memarian, H. et al., 2014. SWAT-based hydrological modelling of tropical land-use scenarios. *Hydrological Sciences Journal*. Taylor & Francis, 59(10), pp. 1808–1829.
- Mendes, D., Marengo, J.A., Rodrigues, S. and Oliveira, M., 2014. Downscaling statistical model techniques for climate change analysis applied to the Amazon region. *Advances in Artificial Neural Systems*, 2014, pp. 1–10.

- Moron, V., Robertson, A.W., Ward, M.N. and Ndiaye, O., 2008. Weather types and rainfall over Senegal. Part I: Observational Analysis. *Journal of Climate*, 21(2), pp. 266–287.
- O’Gorman, P. A., 2015. Precipitation extremes under climate change. *Current Climate Change Reports*, 1(2), pp. 49–59.
- Osman, Y. Z. and Abdellatif, M.E., 2016. Improving accuracy of downscaling rainfall by combining predictions of different statistical downscale models. *Water Science*, 30(2), pp. 61–75.
- Pettitt, A. N., 1997. Testing the normality of several independent samples using the Anderson-Darling Statistic. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(2), pp. 156–161.
- Pineda, L. E. and Willems, P., 2016. Multisite downscaling of seasonal predictions to daily rainfall characteristics over Pacific–Andean River Basins in Ecuador and Peru Using a nonhomogeneous hidden Markov model. *Journal of Hydrometeorology*, 17(2), pp. 481–498.
- Prasad, K., Dash, S. K. and Mohanty, U. C., 2010. A logistic regression approach for monthly rainfall predictions in meteorological subdivisions of India based on DEMETER retrospective predictions. *International Journal of Climatology*, 30(10), pp. 1577–1588.
- Rashid, M. M., Beecham, S. and Chowdhury, R. K., 2016. Statistical downscaling of rainfall: a non-stationary and multi-resolution approach. *Theoretical and Applied Climatology*, 124(3–4), pp. 919–933.
- Reuters, 2014. More Than 160,000 Evacuated From Worst-Ever Floods In Malaysia, 27 December. Available at http://www.huffingtonpost.com/2014/12/27/malaysia-floods_n_6384172.html [Accessed: 8 December 2015].
- Robertson, A. W., Kirshiner, S., Smyth, P., Charles, S.P. and Bates, B.C., 2006. Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland. *Quarterly Journal of the Royal Meteorological Society*, 132(615), pp. 519–542.
- Robertson, A. W., Kirshner, S. and Smyth, P., 2004. Downscaling of daily rainfall occurrence over Northeast Brazil Using a hidden Markov model. *Journal of Climate*, 17(22), pp. 4407–4424.
- Robertson, A. W., Moron, V. and Swarinoto, Y., 2009. Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia. *International Journal of Climatology*, 29(10), pp. 1449–1462.
- Sahin, S. and Cigizoglu, H. K., 2010. Homogeneity Analysis of Turkish Meteorological Data Set. *Hydrological Processes*, 24(8), pp. 981–992.

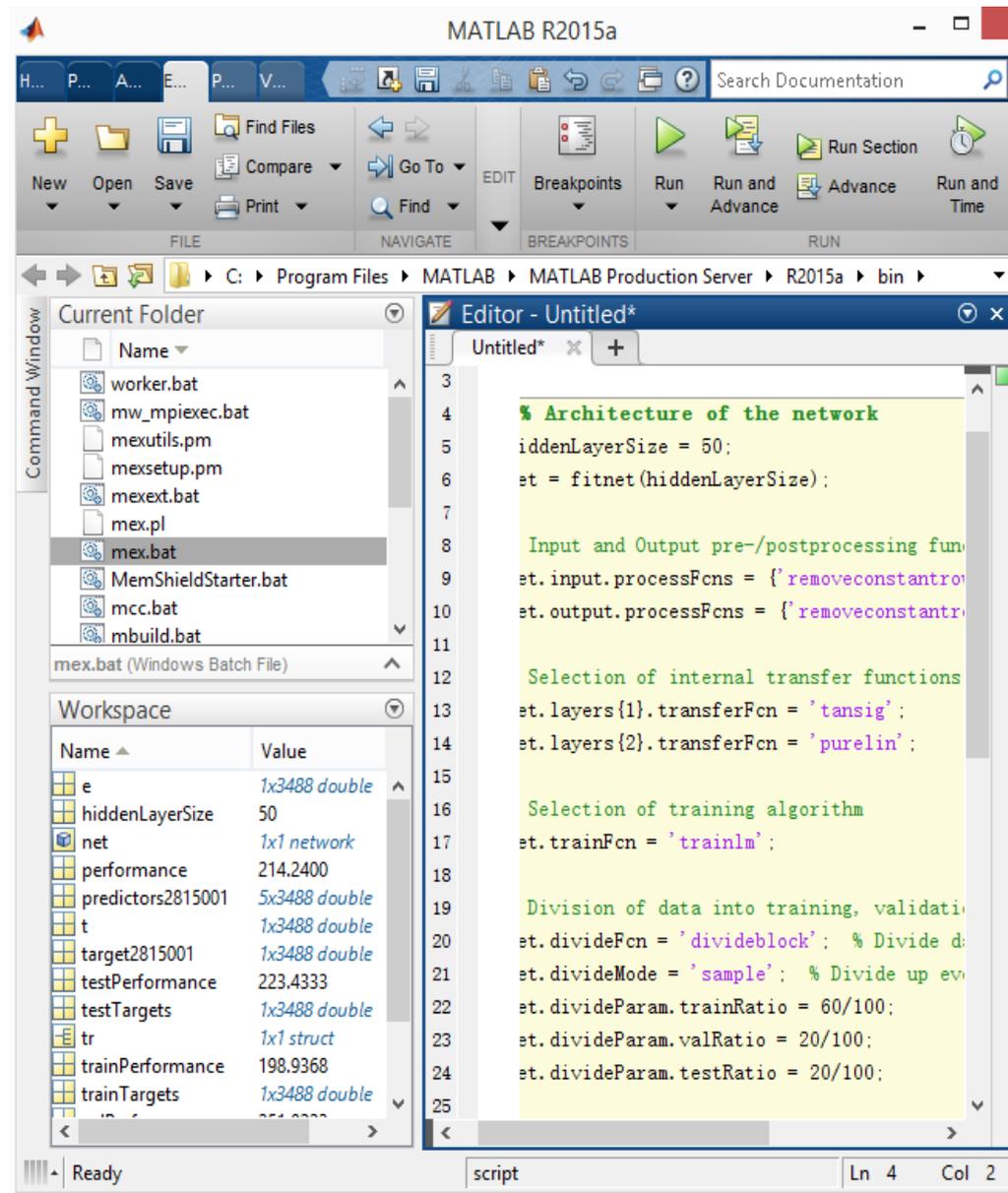
- Saudi, A. S. M. et al., 2015. Flood risk pattern recognition by using environmetric technique: A case study in langat river basin. *Jurnal Teknologi*, 77(1), pp. 145–152.
- Schewe, J. and Levermann, A., 2012. A statistically predictive model for future monsoon failure in India. *Environmental Research Letters*, 7(4), p. 44023.
- Schoof, J. T. and Pryor, S. C., 2001. Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. *International Journal of Climatology*, 21(7), pp. 773–790.
- Shi, Y. and Song, L., 2015. Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the Tibetan Plateau from 2001 to 2012. *Mountain Research and Development*, 35(2), pp. 180–194.
- Snell, S. E., Gopal, S. and Kaufmann, R. K., 2000. Spatial interpolation of surface air temperatures using artificial neural networks: Evaluating their use for downscaling GCMs. *Journal of Climate*, 13(5), pp. 886–895.
- Souvignet, M., Gaese, H., Ribbe, L., Kretschmer, N. and Oyarzun, R., 2010. Statistical downscaling of precipitation and temperature in north-central Chile: an assessment of possible climate change impacts in an arid Andean watershed. *Hydrological Sciences Journal*, 55(1), pp. 41–57.
- Stumpf, A. and Kerle, N., 2011. Object-oriented mapping of landslides using random forests. *Remote Sensing of Environment*, 115(10), pp. 2564–2577.
- Suhaila, J., Deni, S. M. and Jemain, A. A., 2008. Detecting inhomogeneity of rainfall series in Peninsular Malaysia. *Asia-Pacific Journal of Atmospheric Sciences*, 44(4), pp. 369–380.
- Sullivan, C. A. and Huntingford, C., 2009. Water resources, climate change and human vulnerability. *18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation: Interfacing Modelling and Simulation with Mathematical and Computational Sciences*, 13 -17 July 2009 Cairns, Australia, pp. 3984–3990.
- Talaei, P. H., Kouchakzadeh, M. and Some'e, B. S., 2014. Homogeneity Analysis of Precipitation Series in Iran. *Theoretical and Applied Climatology*, 118(1–2), pp. 297–305.
- Tisseuil, C., Vrac, M., Lek, S. and Wade, J., 2010. Statistical downscaling of river flows. *Journal of Hydrology*, 385(1–4), pp. 279–291.
- Trenberth, K., 2011. Changes in precipitation with climate change. *Climate Research*, 47(1), pp. 123–138.

- Tsai, C.-C., Lu, M.-C. and Wei, C.-C., 2012. Decision tree-based classifier combined with neural-based predictor for water-stage predictions in a river basin during typhoons: A case study in Taiwan. *Environmental Engineering Science*, 29(2), pp. 108–116.
- Venkata Ramana, R., Krishna, B., Kurmar, S.R. and Pandey, N.G., 2013. Monthly rainfall prediction using wavelet neural network analysis. *Water Resources Management*, 27(10), pp. 3697–3711.
- Verbist, K., Robertson, A.W., Cornelis, W.M. and Gabriels, D., 2010. Seasonal predictability of daily rainfall characteristics in Central Northern Chile for dry-land management. *Journal of Applied Meteorology and Climatology*, 49(9), pp. 1938–1955.
- Vincenzi, S. et al., 2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecological Modelling*, 222(8), pp. 1471–1478.
- Vu, M. T., Aribarg, T., Supratid, S., Raghavan, S.V. and Liong, S.Y., 2016. Statistical downscaling rainfall using artificial neural network: significantly wetter Bangkok?. *Theoretical and Applied Climatology*, 126(3–4), pp. 453–467.
- Widmann, M., Bretherton, C. S. and Salath é E. P., 2003. Statistical precipitation downscaling over the northwestern united states using numerically simulated precipitation as a predictor. *Journal of Climate*, 16(5), pp. 799–816.
- Wijngaard, J. B., Klein Tank, A. M. G. and Konnen, G. P., 2003. Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*, 23(6), pp. 679–692.
- Wilby, R. L. and Dawson, C. W., 2012. The statistical downscaling model: insights from one decade of application. *International Journal of Climatology*, 33(7), pp. 1707–1719.
- Yang, C., Chandler, R.E. and Isham, V.S., 2005. Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, 41(11), pp. 1–13.
- Yarnal, B., Comrie, A.C., Frakes, B. and Brown, D.P., 2001. Developments and prospects in synoptic climatology. *International Journal of Climatology*, 21(15), pp. 1923–1950.
- Zhang, X.-C., Nearing, M.A., Garbrecht, J.D. and Steiner, J.L., 2004. Downscaling monthly predictions to simulate impacts of climate change on soil erosion and wheat production. *Soil Science Society of America Journal*, 68(4), pp. 1376–1385.
- Zimmerman, D.W., 2014. Comparative power of Student T test and Mann-Whitney U test for unequal sample sizes and variances. *The Journal of Experimental Education*, 55(3), pp. 171–174.

APPENDICES

APPENDIX A

ANN in Matlab R2015a



Neural Network Training (nntraintool)

Neural Network

Input (5) → Hidden (50) → Output (1)

Algorithms

Data Division: Block (divideblock)
 Training: Levenberg-Marquardt (trainlm)
 Performance: Mean Squared Error (mse)
 Calculations: MATLAB

Progress

Epoch:	0	11 iterations	1000
Time:		0:00:02	
Performance:	1.75e+04	187	0.00
Gradient:	8.41e+04	91.8	1.00e-07
Mu:	0.00100	0.100	1.00e+10
Validation Checks:	0	6	6

Plots

Performance (plotperform)
 Training State (plottrainstate)
 Error Histogram (ploterrhist)
 Regression (plotregression)
 Fit (plotfit)

Plot Interval: epochs

Validation stop.

APPENDIX B

Computed p-value of homogeneity tests results

Station	Tests	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2815001	SHNT	0.77	0.85	0.72	0.41	0.39	0.47	0.85	0.27	0.62	0.16	0.54	0.89
	BR	0.83	0.86	0.77	0.46	0.18	0.21	0.74	0.14	0.72	0.43	0.82	0.50
	PET	0.79	0.10	0.53	0.49	0.67	0.75	0.77	0.78	0.71	0.22	0.46	0.63
	VNR	0.86	0.80	0.40	0.43	0.17	0.34	0.50	0.07	0.53	0.19	0.45	0.72
2913001	SHNT	0.68	0.70	0.74	0.70	0.62	0.71	0.87	0.32	0.85	0.11	0.33	0.41
	BR	0.75	0.93	0.54	0.77	0.53	0.75	0.71	0.35	0.68	0.04	0.87	0.21
	PT	0.05	0.75	0.39	0.59	0.33	0.26	0.48	0.08	0.15	0.06	0.19	0.56
	VNR	0.76	0.64	0.22	0.11	0.44	0.43	0.85	0.52	0.09	0.01	0.73	0.39
2917001	SHNT	0.18	0.78	0.46	0.44	0.56	0.52	0.98	0.89	0.82	0.37	0.20	0.13
	BR	0.11	0.29	0.66	0.31	0.32	0.63	0.98	0.91	0.50	0.10	0.25	0.09
	PET	0.08	0.55	0.39	0.91	0.30	0.98	0.02	0.01	0.45	0.70	0.57	0.16
	VNR	0.80	0.53	0.59	0.03	0.02	0.54	0.84	0.59	0.16	0.47	0.56	0.18
3118102	SHNT	0.20	0.59	0.03	0.36	0.44	0.68	0.24	0.12	0.41	0.14	0.33	0.20
	BR	0.09	0.90	0.04	0.71	0.04	0.15	0.02	0.18	0.35	0.28	0.11	0.26
	PET	0.15	0.56	0.02	0.53	0.42	0.64	0.96	0.25	0.82	0.16	0.70	0.28
	VNR	0.61	0.60	0.05	0.65	0.06	0.76	0.02	0.56	0.28	0.78	0.04	0.13

APPENDIX C

2x2 Contingency table

		Observed		
		Yes	No	
Prediction	Yes	a	b	$a + b$
	No	c	d	$c + d$
		$a + c$	$b + d$	$n = a + b + c + d$

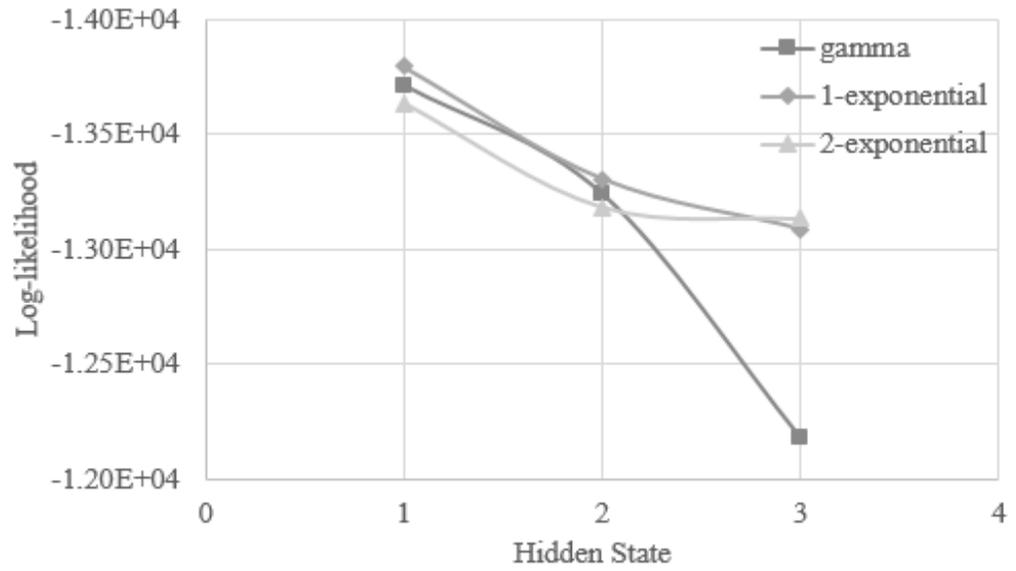
Four possible outcomes:

- (i) Number of events which are predicted and actually occurred (a).
- (ii) Number of events which are predicted but actually not occurred (b).
- (iii) Number of events which are not predicted but actually occurred (c).
- (iv) Number of events which are not predicted and actually not occurred (d).

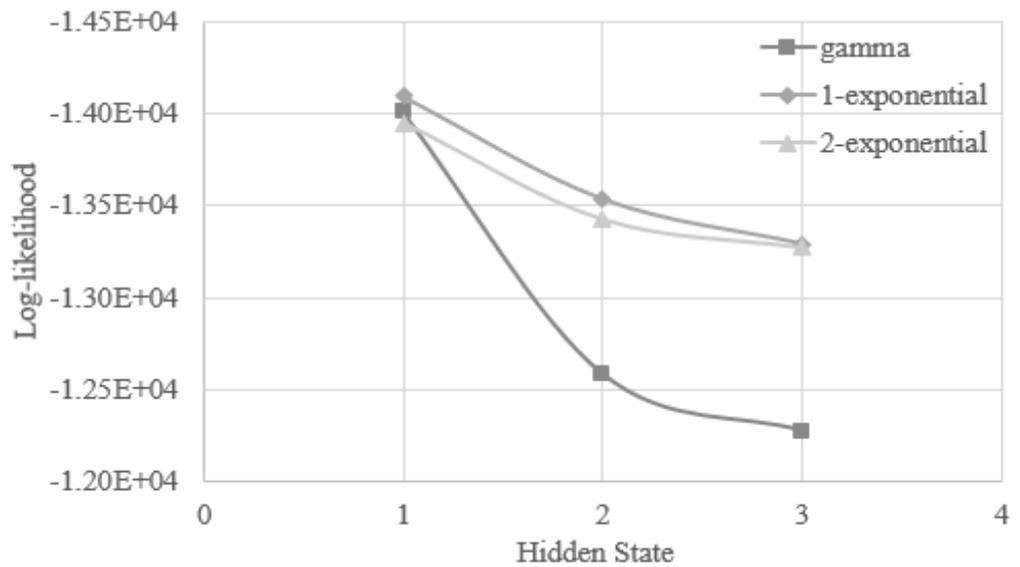
APPENDIX D

Graphs of log-likelihood values of rainfall distributions in NHMM

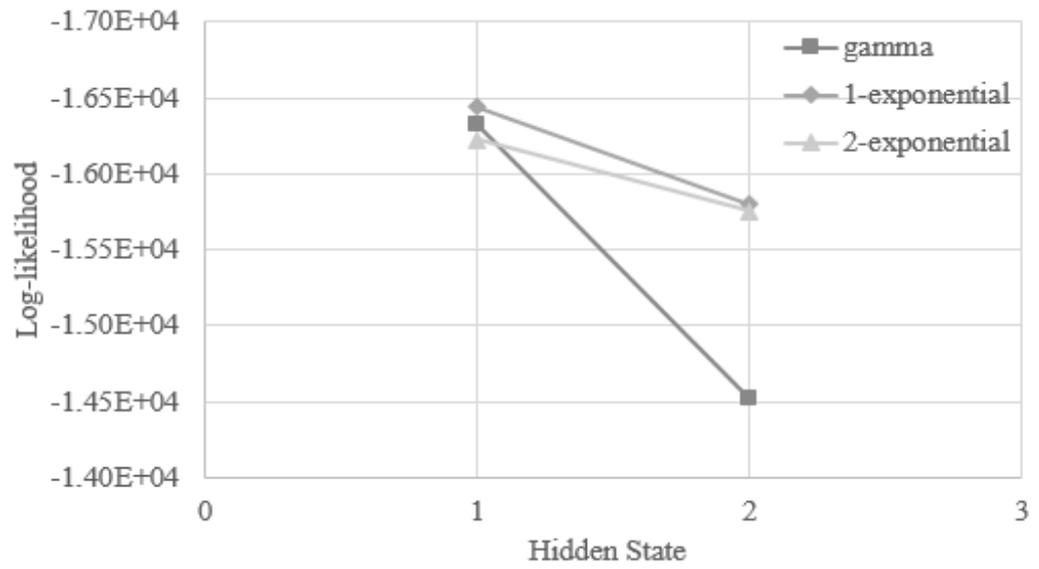
(a) station 2815001



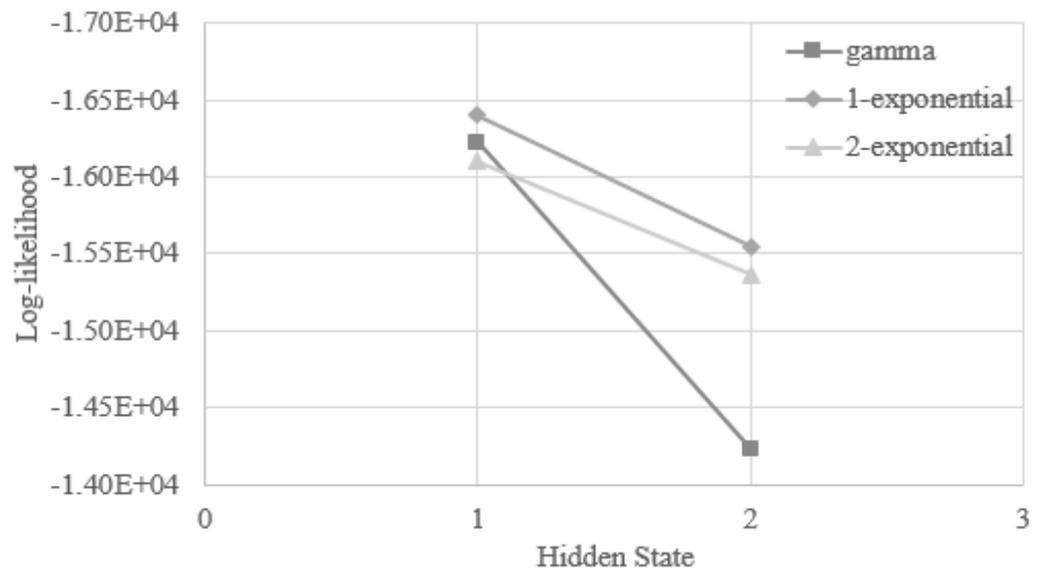
(b) station 2913001



(c) station 2917001



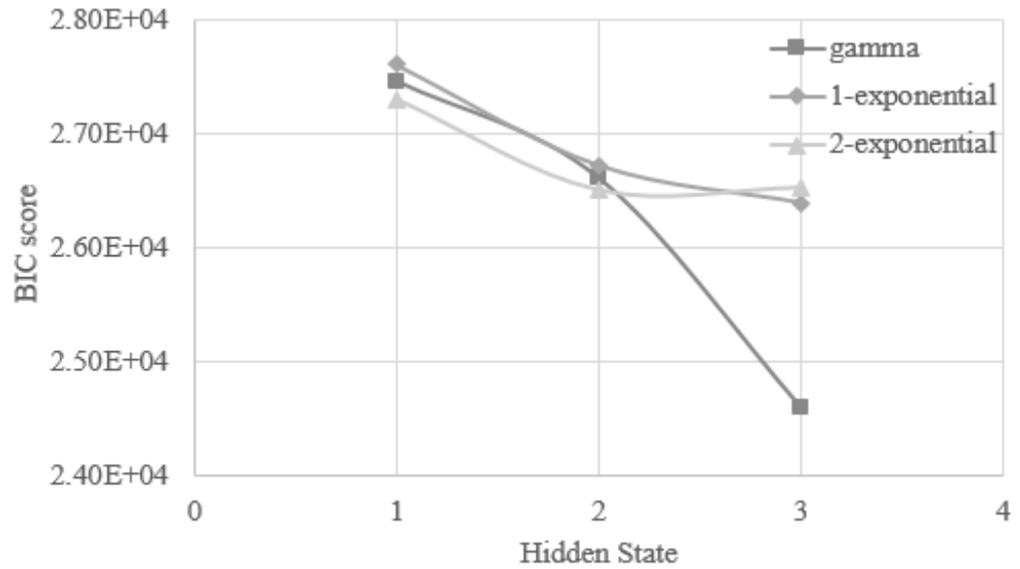
(d) station 3118102



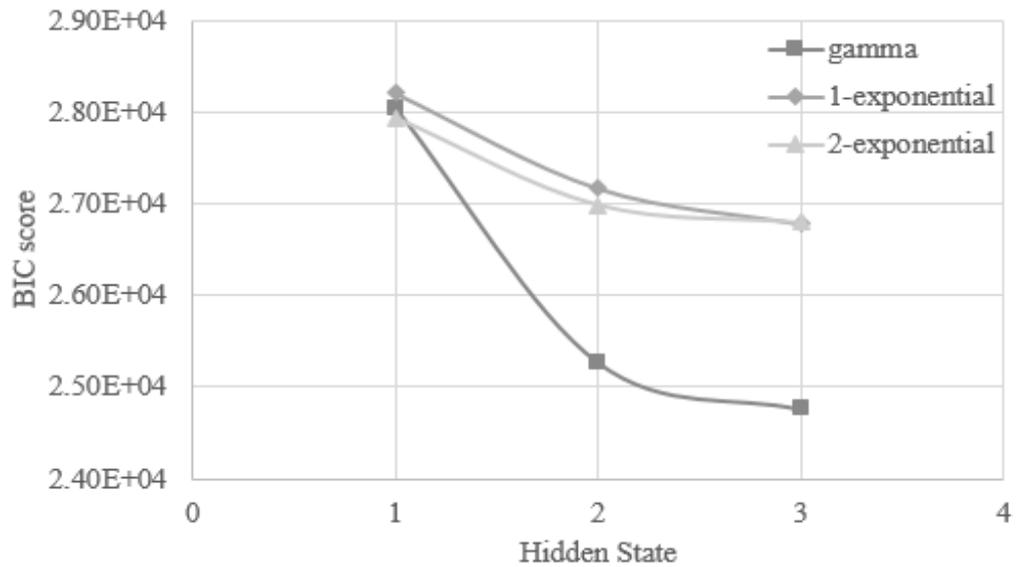
APPENDIX E

Graphs of BIC scores of rainfall distributions in NHMM

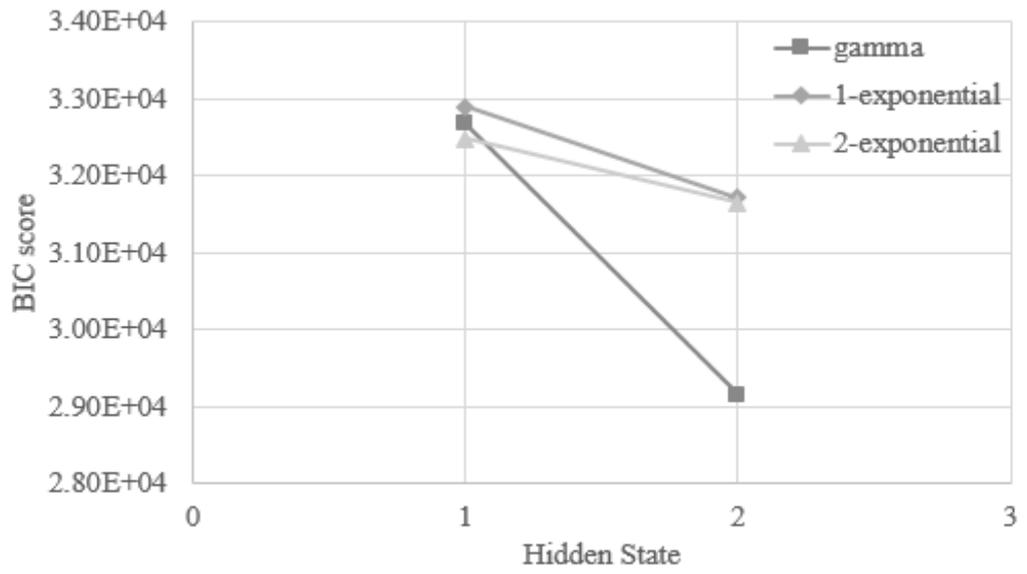
(a) station 2815001



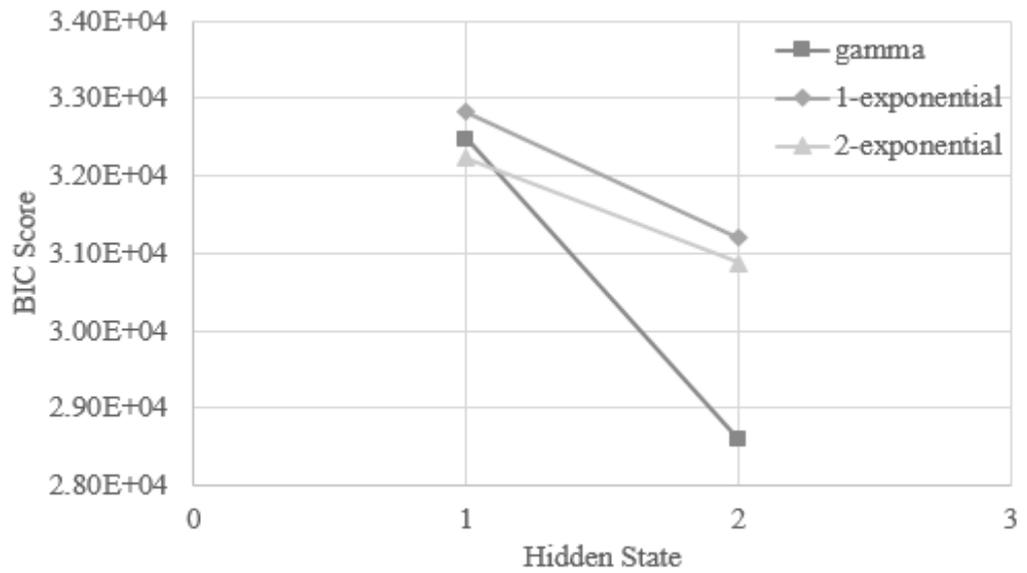
(b) station 2913001



(c) station 2917001



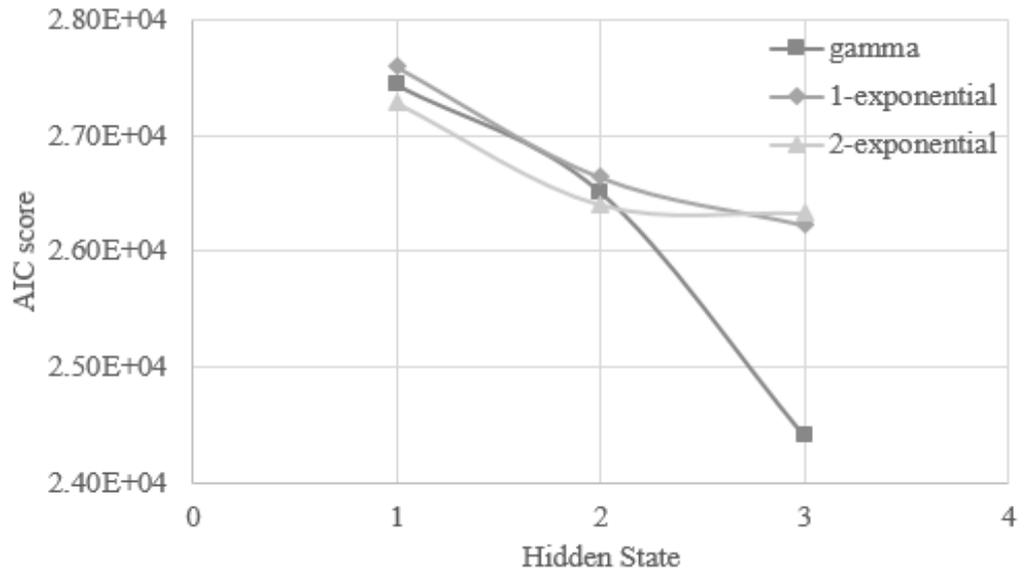
(d) station 3118102



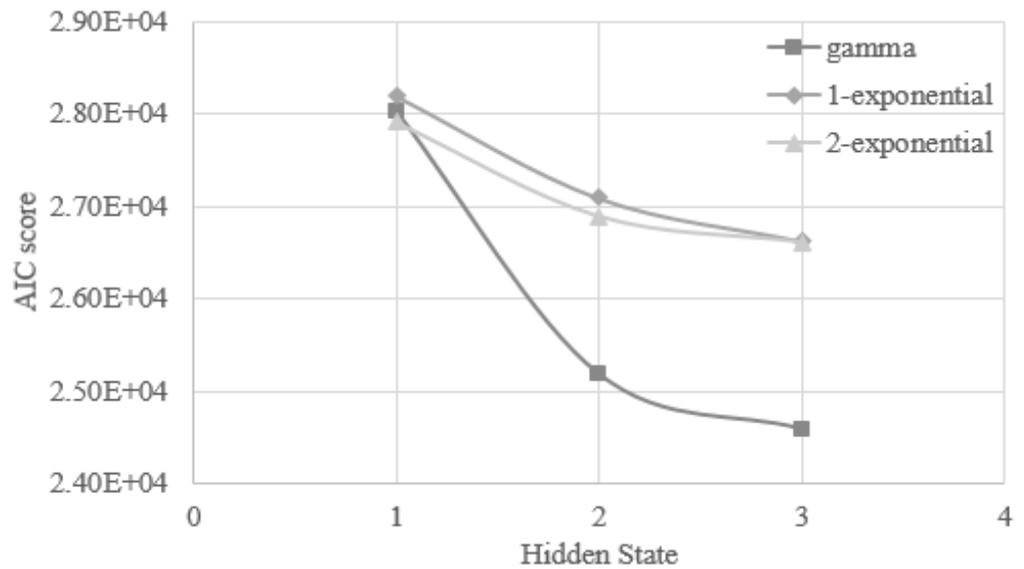
APPENDIX F

Graphs of AIC scores of rainfall distributions in NHMM

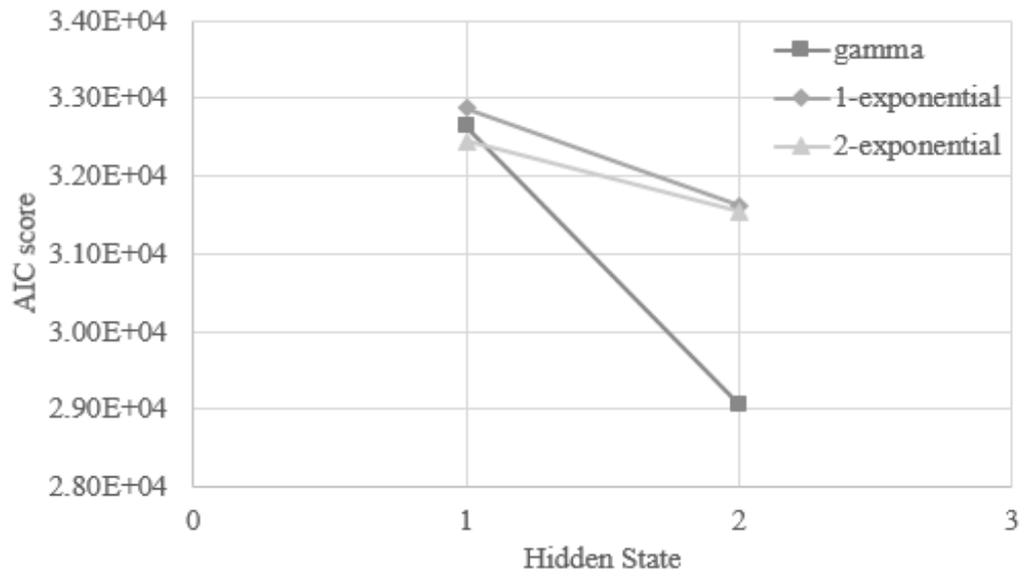
(a) station 2815001



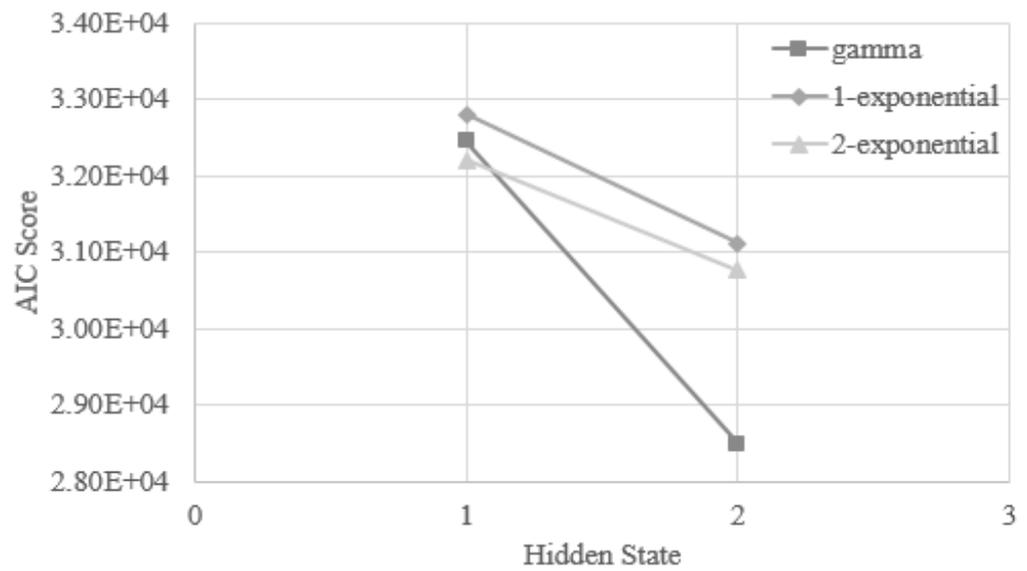
(b) station 2913001



(c) station 2917001



(d) station 3118102



APPENDIX G

Scores of non-parametric tests for calculating the total score in acceptability index

Kolmogorov-Smirnov test

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	1	0	0
	Feb	0	0	0
	Mar	1	0	0
	Apr	1	0	0
	May	1	0	1
	Jun	1	0	0
	Jul	1	0	0
	Aug	0	0	0
	Sep	1	1	0
	Oct	1	0	0
	Nov	1	1	1
	Dec	0	0	0
2913001	Jan	0	0	0
	Feb	1	0	0
	Mar	1	0	0
	Apr	1	0	0
	May	1	0	0
	Jun	1	0	0
	Jul	1	0	0
	Aug	1	0	0
	Sep	0	0	0
	Oct	1	0	0
	Nov	1	0	0
	Dec	0	0	0
2917001	Jan	1	0	0
	Feb	0	0	0
	Mar	1	0	0
	Apr	1	0	0
	May	1	1	1
	Jun	1	0	0
	Jul	1	0	0
	Aug	0	0	0
	Sep	0	0	0
	Oct	0	0	0
	Nov	0	1	0
	Dec	1	0	0

	Jan	1	1	0
	Feb	1	0	0
	Mar	1	0	0
	Apr	1	0	0
	May	0	0	0
3118102	Jun	1	1	0
	Jul	0	0	1
	Aug	0	0	0
	Sep	1	0	0
	Oct	0	0	0
	Nov	1	1	0
	Dec	1	0	0

- “0” indicated the non-significant difference between observed and simulated series;
- “1” indicated the significant difference between observed and simulated series.

Mann-Whitney U test

Station	Month	ANN	NHMM	BACT-ANN
	Jan	0	0	0
	Feb	0	1	0
	Mar	0	0	0
	Apr	0	0	0
	May	0	0	0
2815001	Jun	0	0	0
	Jul	0	0	0
	Aug	0	0	0
	Sep	1	1	0
	Oct	0	0	1
	Nov	0	1	1
	Dec	0	0	0
	Jan	0	0	0
	Feb	0	0	0
	Mar	1	1	0
	Apr	0	0	0
	May	1	1	0
2913001	Jun	0	0	0
	Jul	0	0	0
	Aug	0	0	0
	Sep	0	0	0
	Oct	0	0	0
	Nov	0	1	0
	Dec	0	0	0
2917001	Jan	0	1	0

	Feb	0	0	0
	Mar	0	0	0
	Apr	1	0	0
	May	1	1	0
	Jun	0	1	0
	Jul	0	1	0
	Aug	0	0	0
	Sep	0	0	0
	Oct	0	0	0
	Nov	0	1	0
	Dec	0	0	0
	Jan	1	1	0
	Feb	0	0	0
	Mar	0	0	0
	Apr	0	0	0
	May	0	0	0
3118102	Jun	1	1	0
	Jul	0	0	1
	Aug	0	0	0
	Sep	0	0	0
	Oct	0	0	0
	Nov	0	1	0
	Dec	0	0	0

- “0” indicated the non-significant difference between observed and simulated series;
- “1” indicated the significant difference between observed and simulated series.

Squared-rank test

Station	Month	ANN	NHMM	BACT-ANN
	Jan	1	0	0
	Feb	1	0	0
	Mar	1	0	0
	Apr	1	0	0
	May	1	0	1
2815001	Jun	1	0	0
	Jul	1	0	0
	Aug	1	0	0
	Sep	1	0	0
	Oct	1	0	0
	Nov	1	0	0
	Dec	1	0	0
2913001	Jan	1	0	0
	Feb	1	0	0

	Mar	1	0	0
	Apr	1	0	0
	May	1	0	1
	Jun	1	0	0
	Jul	1	0	0
	Aug	1	0	0
	Sep	1	0	0
	Oct	1	0	0
	Nov	1	0	0
	Dec	1	0	0
<hr/>				
	Jan	1	0	0
	Feb	1	0	1
	Mar	1	1	0
	Apr	1	0	0
	May	1	0	1
2917001	Jun	1	0	0
	Jul	1	0	0
	Aug	1	0	0
	Sep	1	0	0
	Oct	0	0	0
	Nov	1	1	0
	Dec	1	0	0
<hr/>				
	Jan	1	0	0
	Feb	1	0	0
	Mar	1	1	0
	Apr	1	1	0
	May	1	0	0
3118102	Jun	1	0	0
	Jul	1	0	0
	Aug	1	1	0
	Sep	1	0	0
	Oct	1	1	0
	Nov	1	1	0
	Dec	1	1	0

- “0” indicated the non-significant difference between observed and simulated series;

- “1” indicated the significant difference between observed and simulated series.

Kendall's tau-b correlation

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	0	1	0
	Feb	1	1	0
	Mar	1	1	0
	Apr	1	1	0
	May	0	0	0
	Jun	1	1	0
	Jul	1	1	0
	Aug	0	1	0
	Sep	1	1	0
	Oct	1	0	0
	Nov	1	1	0
	Dec	0	1	0
2913001	Jan	1	1	0
	Feb	0	1	0
	Mar	1	1	0
	Apr	0	1	0
	May	1	1	0
	Jun	1	1	0
	Jul	1	1	0
	Aug	1	1	0
	Sep	1	1	0
	Oct	1	0	0
	Nov	1	1	0
	Dec	0	1	0
2917001	Jan	1	0	0
	Feb	0	1	0
	Mar	1	1	0
	Apr	1	1	0
	May	0	1	0
	Jun	1	1	0
	Jul	0	1	0
	Aug	0	1	0
	Sep	0	0	0
	Oct	0	1	0
	Nov	0	1	0
	Dec	0	1	0
3118102	Jan	0	1	0
	Feb	0	1	0
	Mar	0	1	0
	Apr	0	1	0
	May	1	1	0
	Jun	1	1	0
	Jul	1	1	0

Aug	1	1	0
Sep	0	1	0
Oct	1	1	0
Nov	1	1	0
Dec	0	1	0

- "0" indicated the significant correlation between observed and simulated series;

- "1" indicated the non-significant correlation between observed and simulated series.

Spearman's rho correlation

Station	Month	ANN	NHMM	BACT-ANN
2815001	Jan	1	1	0
	Feb	1	1	0
	Mar	1	1	0
	Apr	1	1	0
	May	0	0	0
	Jun	1	1	0
	Jul	1	1	0
	Aug	0	1	0
	Sep	1	1	0
	Oct	1	0	0
	Nov	1	1	0
	Dec	0	1	0
2913001	Jan	1	1	0
	Feb	0	1	0
	Mar	1	1	0
	Apr	0	1	0
	May	1	1	0
	Jun	1	1	0
	Jul	1	1	0
	Aug	1	1	0
	Sep	1	1	0
	Oct	1	0	0
	Nov	1	1	0
	Dec	0	1	0
2917001	Jan	1	0	0
	Feb	0	1	0
	Mar	1	1	0
	Apr	1	1	0
	May	0	1	0
	Jun	1	1	0
	Jul	0	1	0
	Aug	0	1	0

	Sep	0	0	0
	Oct	0	1	0
	Nov	0	1	0
	Dec	0	1	0
<hr/>				
	Jan	0	1	0
	Feb	0	1	0
	Mar	1	1	0
	Apr	1	1	0
	May	0	1	0
3118102	Jun	1	1	0
	Jul	1	1	0
	Aug	0	1	0
	Sep	0	1	0
	Oct	1	1	0
	Nov	1	1	0
	Dec	0	1	0

-
- “0” indicated the significant correlation between observed and simulated series.;
 - “1” indicated the non-significant correlation between observed and simulated series.