

**ESTIMATING MISSING DAILY RAINFALL  
DATA VIA ARTIFICIAL NEURAL NETWORK  
OVER PENINSULAR MALAYSIA**

**LOH WING SON**

**UNIVERSITI TUNKU ABDUL RAHMAN**

**ESTIMATING MISSING DAILY RAINFALL DATA VIA ARTIFICIAL  
NEURAL NETWORK OVER PENINSULAR MALAYSIA**

**LOH WING SON**

**A project report submitted in partial fulfilment of the requirements for the  
award of Master of Mathematics**

**Lee Kong Chian Faculty of Engineering and Science  
Universiti Tunku Abdul Rahman**

**Jan 2021**

## DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature :  \_\_\_\_\_

Name : \_\_\_\_\_ Loh Wing Son \_\_\_\_\_


ID No. : \_\_\_\_\_ 20UEM00089 \_\_\_\_\_

Date : \_\_\_\_\_ 18<sup>th</sup> April 2021 \_\_\_\_\_

## APPROVAL FOR SUBMISSION

I certify that this project report entitled **ESTIMATING MISSING DAILY RAINFALL DATA VIA ARTIFICIAL NEURAL NETWORK OVER PENINSULAR MALAYSIA** was prepared by **LOH WING SON** has met the requirement for the award of Master of Mathematics at Universiti Tunku Abdul Rahman.

Approved by,

Signature :  \_\_\_\_\_

Supervisor : Dr. Tan Wei Lun

Date : 19 April 2021

The copyright of this report belongs to the author under the terms of the copyright Act 1987 as qualified by Intellectual Property Policy of Universiti Tunku Abdul Rahman. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this report.

© 2021, Loh Wing Son. All right reserved.

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to deliver my highest gratitude to my research project supervisor, Dr. Tan Wei Lun upon all of the advices, guidance, inspirations and supervisions provided throughout this research project.

I would like to also offer my most sincere thankfulness to my family and course mates for the encouragements as well as being helpful and supportive at all times.

Finally, I wish to extend my gratitude towards the Lee Kong Chian Faculty of Engineering and Science under the Universiti Tunku Abdul Rahman for approving my research project, allowing me to explore and gain valuable knowledge from this research project.

## ABSTRACT

The presence of missing rainfall data has always known to be an obstacle for rain gauge stations to preserve a serially complete real time rainfall database. Various techniques were implemented in dealing with missing rainfall data in the past but artificial neural network (ANN) models have also gradually earned much renown due to its promising estimation results.

The Self-Organising Feature Map (SOFM), a type of ANN was proposed in this research to account for the missing daily rainfall values and the complex dynamics of rainfall over Peninsular Malaysia. SOFM was applied in two stages for which the first stage was to train the SOFM model using the complete daily rainfall data and the second stage was to apply the trained SOFM to estimate the missing daily rainfall data. The estimated results were then compared and contrast by setting up different proportion of 10%, 20%, and 30% for the missing daily rainfall data.

Ten different rainfall stations distributed over the Peninsular Malaysia were studied. The daily rainfall data for the North-East monsoon (NEM) season from the rainfall stations were obtained to assess the performance of the SOFM in describing the spatial relationship of the rainfall events as well as in estimating the missing daily rainfall data. The mean error (ME) and root mean square error (RMSE) were computed to evaluate the missing daily rainfall data estimated by the SOFM model.

The analysis for all of the three different proportion of missing daily rainfall data suggested that each of the rainfall stations possess distinctive rainfall patterns. The SOFM has also provided reasonable estimates for the missing daily rainfall data.

**Keywords:** missing daily rainfall data, North-East monsoon (NEM), artificial neural network (ANN), Self-Organising Feature Map (SOFM)

## TABLE OF CONTENTS

<b>DECLARATION</b>	<b>ii</b>
<b>APPROVAL FOR SUBMISSION</b>	<b>iii</b>
<b>ACKNOWLEDGEMENTS</b>	<b>v</b>
<b>ABSTRACT</b>	<b>vi</b>
<b>TABLE OF CONTENTS</b>	<b>vii</b>
<b>LIST OF TABLES</b>	<b>ix</b>
<b>LIST OF FIGURES</b>	<b>x</b>
<b>LIST OF SYMBOLS / ABBREVIATIONS</b>	<b>xi</b>
<b>CHAPTER</b>	
<b>1 INTRODUCTION</b>	<b>1</b>
<b>1.1 Research Background</b>	<b>2</b>
<b>1.2 Research Objectives</b>	<b>4</b>
<b>1.3 Significance of Research</b>	<b>5</b>
<b>2 LITERATURE REVIEW</b>	<b>6</b>
<b>2.1 Selection of Homogeneous Stations</b>	<b>8</b>
<b>2.2 Missing Rainfall Data Estimation Models</b>	<b>9</b>
<b>2.2.1 Supervised ANN Models</b>	<b>9</b>
<b>2.2.1.1 Feed Forward Neural Network (FFNN)</b>	<b>9</b>
<b>2.2.1.2 Radial Basis Neural Network (RBNN)</b>	<b>9</b>
<b>2.2.2 Unsupervised ANN Models</b>	<b>10</b>
<b>2.2.2.1 Self-Organising Feature Map (SOFM)</b>	<b>10</b>
<b>2.2.3 Non-ANN Models</b>	<b>10</b>
<b>2.2.3.1 Singular Imputation</b>	<b>10</b>
<b>2.2.3.2 Multiple Imputation</b>	<b>11</b>
<b>2.2.3.3 Euclidian Distance-Based Models</b>	<b>11</b>
<b>2.3 Model Assessments</b>	<b>12</b>



<b>3</b>	<b>RESEARCH METHODOLOGY</b>	<b>14</b>
	3.1 Missing Mechanism	14
	3.2 Self-Organising Feature Map (SOFM)	15
	3.2.1 Data Pre-Processing	15
	3.2.2 Data Training	15
	3.2.3 Missing Daily Rainfall Estimation	17
	3.3 Model Assessment	18
<b>4</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>19</b>
	4.1 Research Area	19
	4.2 Daily Rainfall Data	20
	4.3 Trained SOFM	23
	4.4 Performance of SOFM	28
<b>5</b>	<b>CONCLUSIONS</b>	<b>31</b>
	5.1 Conclusions	31
	5.2 Recommendations	31
	<b>REFERENCES</b>	<b>32</b>

## LIST OF TABLES

Table 2.01:	Summary of Some Recent Reviewed Literatures	7
Table 4.01:	Geographical Coordinates of Rainfall Stations Studied	20
Table 4.02:	Descriptive Statistics of the Daily Rainfall Data	21
Table 4.03:	Wet Days of the Daily Rainfall Data	23
Table 4.04:	Performance Measure of the SOFM	28

## LIST OF FIGURES

Figure 4.01:	Geographical Map of Rainfall Stations Studied	19
Figure 4.02:	Kohonen Map of Trained SOFM (10% Missing Data)	24
Figure 4.03:	Kohonen Map of Trained SOFM (20% Missing Data)	24
Figure 4.04:	Kohonen Map of Trained SOFM (30% Missing Data)	25
Figure 4.05:	Heat Maps of Rainfall Stations (10% Missing Data)	26
Figure 4.06:	Heat Maps of Rainfall Stations (20% Missing Data)	27
Figure 4.07:	Heat Maps of Rainfall Stations (30% Missing Data)	27
Figure 4.08:	Comparison Plot Between Actual and Estimated Rainfall Observations (10% missing data)	29
Figure 4.09:	Comparison Plot Between Actual and Estimated Rainfall Observations (20% missing data)	30
Figure 4.10:	Comparison Plot Between Actual and Estimated Rainfall Observations (30% missing data)	30

## LIST OF SYMBOLS / ABBREVIATIONS

$x$	- Daily rainfall data
$\hat{x}_{MD}$	- Estimator for missing data
$z$	- Normalised daily rainfall data
$m$	- Number of input vectors
$M$	- Number of neurons in the SOFM model
$\eta$	- Learning rate
$w$	- Connecting weights between two neurons in the Kohonen map
$D$	- Euclidean distance between two neurons in the Kohonen map
$H$	- Neighbourhood function
$\delta$	- Radius of the neighbourhood function
$N$	- Neighbourhood region
$\theta$	- Best matching unit
$s$	- Number of iterations during training stage
$T$	- Maximum number of iterations during training stage
$R^2$	- Coefficient of determination
AANN	- Auto regressive neural network
AI	- Artificial intelligence
ANN	- Artificial neural network
BMU	- Best matching unit
CCW	- Correlation coefficient weighted
FFNN	- Feed forward neural network
IDW	- Inverse distance weighting
$k$ NN	- $k$ nearest neighbourhood
MAR	- Missing at random
MCAR	- Missing completely at random
ME	- Mean error
MNAR	- Missing not at random
NEM	- North-East monsoon

PCA	-	Principal component analysis
RBNN	-	Radial basis neural network
RK	-	Ridge kriging
RMSE	-	Root mean squared error
RT	-	Regression tree
SAM	-	Simple averaging method
SEM	-	South-East monsoon
SOFM	-	Self-organising feature map

## CHAPTER 1

### INTRODUCTION

The Peninsular Malaysia consists of 11 states and 2 federal territories with over hundreds of river systems which altogether contributes to approximately 97% of the total raw water supply.[1] Sitting on the Khatulistiwa line near the equator and due to proximity to water makes Malaysia a tropical terrain accompanied by only sunny and rainy days throughout each of the year. Due to its strategic location, the Peninsular Malaysia is also protected from most natural disasters such as volcano eruption and earthquakes but remains vulnerable to other disasters like haze, droughts, floods and landslides.[2] In addition, it faces two regimes of monsoon seasons affecting both the occurrence and intensity of rainfalls. The monsoon transition periods, South-West monsoon (SWM) and North-East monsoon (NEM) are known to take place from late May to early September and from early November to late February respectively. This causes frequent wet spells for these seasons with the latter contributing much higher rainfall. The monsoon rainfalls have been contributing high amount in the annual average rainfall reported to be 2420 mm over the Peninsular Malaysia.

In terms of hydrological process, the dependence on climate data are somewhat obvious and relevant predictive models such as rainfall-runoff modelling are highly involved in a diverse application research field and projects.[3] As such, rainfall in Malaysia reflects a major process of the hydrologic cycle. This can be seen by efforts made in numerous studies to account for the complex and non-linear relationships within rainfall related models.[4] In other words, failure in possessing a complete rainfall data set could cause modelled relationships being falsified and more severely, failure in obtaining any findings from the research.

## 1.1 Research Background

In statistical data analysis, dealing with missing data remains one of the most concerned issues as these missing data may conceal vital information that highly influence a particular system or model. Essentially, a serially complete and reliable set of observations from the studied population is required to accurately perform parameter estimations for a particular model. Since interpreting the synoptic circulations of the rainfall behaviour within Peninsular Malaysia served as one of the main motivations of possessing a complete data, an appropriate missingness mechanism for the rainfall data must be defined. The missing at random (MAR) is the mechanism applied in this research in accordance with the assumption of statistical relationships existence between rainfall values from different stations.[5]

Researchers might have learned and claimed, whether from past experience or through literature review that station homogeneity is determined by Euclidean distances between rainfall stations. The particular concern would be the validity of the selected homogenous stations under such assumptions. As homogenous stations were claimed to play a significant role towards describing the respective rainfall intensities (i.e. similarity in rainfall patterns), awkward situations may occur. For instance, closely distanced but were non-homogeneous stations being included in the study. Another scenario would be not taking the far distanced but homogenous stations into the study. Thus, based on the reviewed literatures, it was found that the selections of rainfall stations were limited to only incorporating nearby stations from pre-determined clusters. It is possible that closely distanced stations but are not homogenous were incorporated and qualified homogenous stations could be left behind.

Furthermore, most conventional models or strategies in estimating missing rainfall data such as regression-based models, inverse distance weighting (IDW) and many others were found providing fairly adequate results. As they carry different assumptions of their owns, it remains questionable on their abilities in complying with the complicated behaviour of rainfall and other climate factors due to the drastic changes in climate over the past few decades.

On the other hand, the perpetual development of artificial intelligence (AI) in this era contributes to a massive transformation towards the viewing perspective of different issues. Countless versions of statistical machine learning models proposed at present by various researches. One of the popular ones, artificial neural network (ANN), occasionally referred as the black-box model [6] consists of empirical models that needs no tangible pre-information from the original data, to seek for possible relationship developed among historical inputs and outputs.[7]

Additionally, ANNs were not just popular but also reported with superior results in existing works in the reviewed literatures.[4] More specifically, this research performs the estimation of the missing daily rainfall data over Peninsular Malaysia by employing a particular type of ANN, namely the self-organising feature map (SOFM).



## **1.2 Research Objectives**

The explicit aim of this research is to generate reliable, consistent, and accurate estimations of the missing daily rainfall data over Peninsular Malaysia. By possessing a serially complete daily rainfall data set for each rainfall stations, it is then available for rainfall modelling or other related hydrological processes for effective planning of water resources management and monitoring disasters.

This leads to the following objectives with the aim of providing solutions to the addressed research problems:

- (1) To assess the performance of ANN model in estimating missing daily rainfall data over Peninsular Malaysia.
- (2) To investigate the existence of spatial relationship of rainfall events shown by the ANN model.

### 1.3 Significance of Research

Understanding the behaviour of rainfalls and runoff processes is the major key for proper planning and management in many fields including disasters studies, fisheries, agricultures industries, construction projects, transportations, hydro-solar-wind power generations and the list continues.

Moreover, the accelerated change of climate associated with the monsoon season will consequently cause many problems such as destruction in agricultures, power supplies and buildings, loss of livestock and infection of waterborne diseases. A recent flood in the Peninsular Malaysia in December 2019 was seen crippling several states including Pahang, Kelantan, Terengganu and Johor with at least two deaths reported from this disaster.[8] Another recent case occurred in Indonesia, Malaysia's neighbour country, which suffered from the most powerful monsoon rain with a death toll at 66 reported by officials as of 6<sup>th</sup> January 2020, causing destruction of crops and stop in many construction sites, weakening the country's economy.[9]

These clear signals indicate that there is a need for a in depth study on the hydrological process in terms of water management, drought, and flood control. Consequently, the missing rainfall data turned out to be an obstacle for hydrological models to perform accurate and reliable prediction results. This is because a serially complete rainfall data is acquired in order to carry out studies such as climate variability studies, flow estimations, water resource management and identifying disasters occurrence (e.g. droughts, floods, and landslides).[10]

The employment of ANN in this research to estimate missing daily rainfall data also contributes to the review of ANN applications in hydrological models since there were not many published studies of estimating missing daily rainfall data in Malaysia via ANN models.

## **CHAPTER 2**

### **LITERATURE REVIEW**

It is evident that the presence of missing rainfall data could at some extent reduce the accuracy of the results of investigated models such as models applied in tropical cyclone forecast, rainfall-runoff analysis, drought, landslide conditions and flood monitoring. Usually in estimating missing rainfall data, the common assumption in mind is that homogenous stations towards the target station that consists missing rainfall data would be able to provide reliable, sufficient information. Their observations should then precisely approximate the corresponding missing rainfall values.

The latter depends on the criteria of selecting appropriate homogenous stations that are highly correlated and the geographical, climate factors including the Euclidian distance between stations, humidity, temperature, wind speed and atmospheric pressure. With these considerations, there exists a great deal of techniques to perform estimations and each technique possesses distinct strengths as well as limitations with respect to the accuracy of the estimated missing rainfall data. Table 2.01 showed the summary of the recent works in missing rainfall.

Table 2.01 : Summary of Some Recent Reviewed Literatures

Cited authors	Country	Climate zone	Data period	Station homogeneity tests / analysis	Estimation Models
[11] Googhari, <i>et al.</i>	Peninsular Malaysia	Tropical	1995 – 2006	-	<a href="#">FFNN</a>
[5] Ho, M.K. and Yusof, F.			1996 – 2004	Pettitt, Normal homogeneity, Von Neumann ratio	<a href="#">SOFM</a>
[12] Nastar, <i>et al.</i>	South-western Columbia		1983 – 2016	Non-linear PCA	Non-linear PCA network
[13] Madhuri, D. and Hardaha, M.K.	India	Sub-tropical	1901 – 2011	-	SAM, normal ratio, IDW, MLR, <a href="#">FFNN</a>
[14] Kim, J.W. and Pachepsky, Y.A.	North America		2001 – 2008	Regression trees	Regression trees, <a href="#">FFNN</a>
[15] Moradkhani, <i>et al.</i>	South-western America	Mediterranean	1989 – 1998	<a href="#">SOFM</a>	<a href="#">SOFM</a> , <a href="#">FFNN</a>
[16] Aieb, <i>et al.</i>	Algeria		1982 – 2015	PCA	Hot deck, <i>k</i> NN, weighted kNN, SAM, multiple imputation, LR
[17] Annalisa, D.P., <i>et al.</i>	Italy		1921 – 2004	-	RBF-spline, IDW, ordinary kriging, LR, GWR, <a href="#">FFNN</a>
[18] Nkuna, T.R. and Odiyo, J.O.	South Africa	Temperate	1950 – 2008	-	<a href="#">RBNN</a>
[19] Barrios, A., Trincado, G., and Garreaud, R.	Chile		1995 – 2012	Radius Euclidian distance	CCW, MLR, weighted IDW, <a href="#">FFNN</a>

## 2.1 Selection of Homogenous Stations

The incorporation of homogenous stations was emphasised and assumed to provide robust estimates.[5],[20] Non-homogenous stations were eliminated, subjected to a variety of climate and other relevant factors. They were claimed to have low significance relative to describing the rainfall intensities and distributions of targeted stations. As such, missing precipitation estimations from the target stations were restricted to the choice of relative neighbouring stations.

High homogeneity stations were extracted before inputting into the employed models by including the geographical and regional variabilities. Seasonal trends such as Meridional wind, humidity and temperature were also included in the homogeneity analysis. Precipitation intensity were found varied consistently based on the seasons except for temperature, that is not applicable for the four seasons. This implies studies on a seasonal trend appeared to be more sensible than annual trends,[21] thus suggesting studying seasonal rainfall period especially for tropical provinces to be more meaningful.

The regression tree (RT) technique in the work of Kim and Pachepsky partitioned input variables into groups based on their homogeneity level.[14] Alternatively, the employment of principal component analysis (PCA) technique in [12],[16] selects the corresponding relevant principal feature components that returns the lowest error. Missing precipitations will then be estimated accordingly based on the clusters visualised by PCA graphs.[16] Nkuna and Odiyo instead used a relatively simple selection rule of at least three neighbouring stations around the target station.[18]

## **2.2 Missing Rainfall Data Estimation Models**

### **2.2.1. Supervised ANN Models**

#### **2.2.1.1 Feed-Forward Neural Network (FFNN)**

The robust implementation of FFNN and the geo-statistical technique by Annalisa, *et al.* with spatial and elevation information attained optimal estimations.[17] Both studies by Barrios, *et.al.* and Madhuri have reported that FFNN effectively presented robust estimates than other standard imputation techniques without accounting for the elevation impact.[13],[19] Validity of results were also shown Googhari, *et al.* in forecasting daily reservoir flows.[11] The two-step reconstruction strategy (RT)+FFNN by Kim and Pachepsky returned lowest error.[14] Nastar, *et al.* implemented the non-linear PCA technique. It operates a type of FFNN, namely autoregressive associative neural network (AANN) where the identity mapping structure (inputs being reproduced at outputs) allows compression of numerous precipitation data from different stations.[22],[23] Ideal results were found under extractions of non-linear components in hierarchical order, mapped to the hidden units for non-linear transformation.[12]

#### **2.2.1.2 Radial Basis Neural Network (RBNN)**

RBNN's key design enables optimal parameters to be searched under the shuffled complex evolution optimisation approach. Pioneer implementation of estimating missing precipitation succeeded in South Africa, by Nkuna and Odiyo. Results were shown to be on a satisfactory level as estimations were consistent.[18] There may however be presence of biasness in context of this study as most available stations were dispersed on the West region.

## **2.2.2. Unsupervised ANN Model**

### **2.2.2.1 Self-Organising Feature Map (SOFM)**

Unlike the supervised learners, FFNN and RBNN where approximation and classification tasks will be performed; SOFM instead adapts an unsupervised learning algorithm, enabling close density estimations and non-parametric projections. The major advantage of the SOFM is that high-dimensional input patterns were reduced into the Kohonen map with its dimensional space being as low as two-dimensional space.[7] Additionally, the SOFM has the ability to preserve topological properties within the input space. Neurons in the Kohonen map are arranged based on the similarities in terms of responding towards the input patterns, signified by the weight vectors. The weights stored within the neurons provide a representation of the rainfall intensities of the input observations in an ordered fashion. The learning rate determines how much updates will be made towards the neighbourhood weights of a neuron upon a given new input. The updated neurons are selected as the best matching unit (BMU). The corresponding radius and neighbourhood function control the effect on the neighbours to the neuron, where two most commonly implemented neighbourhood functions are Gaussian and “bubble”. Ho, M.K. and Yusof, F. estimated missing rainfall data via SOFM only after predicting dry and wet spells[5] while Moradkhani, *et al.* instead made use of an innovative strategy by employing a hybrid structure for which RBNN parameters were extracted and tuned after clustering via SOFM.[15]. Results have clearly indicated the superiority of the SOFM compared to other models.

### **2.2.3. Non-ANN Models**

#### **Singular Imputation**

Singular imputation replaces missing data with a single evaluated statistical measure. For instance, simple averaging method (SAM) estimates the missing data by computing the arithmetic mean of the rainfall data from available stations.[13],[16] For the normal ratio imputation, average of the rainfall observations were extracted from the defined neighbouring stations. The incorporated stations were subjected to exceed the target station by a stipulated proportion.[24]

### 2.2.3.2 Multiple Imputation

Multiple imputation generates multiple individual estimates for the missing data and obtain the average as a point estimate. It could reduce uncertainty on the estimations by considering results from several imputations such as linear regression in Aieb, *et al.*[16]. Regression based imputations preserve the size and frame of precipitation data set. Once the parameters are estimated, the relationship between predictors is determined and missing observations can be estimated from the modelled equation.[25] Parameters of linear [16],[17] and of multiple linear regression [13],[19] but geostatistical weighted regression accounting for the geographical variability, treat parameters as variables. Having too few rainfall stations may however be an issue where rainfall stations are sparsely dispersed since it causes high standard error.[17] Large samples are commonly used for regression imputations to estimate missing precipitation data for higher precision under estimation of standard errors were common.

### 2.2.3.3 Euclidian Distance-Based Models

Euclidian distances between stations are among the essential factors that contribute to the dynamics of rainfall and inverse distance weighting (IDW) is an imputation technique that accounts for it.[13],[19] The underlying assumption treats neighbouring stations of smaller distances as better representation (i.e. high homogeneity), thus contributes greater weights in computations.[17] The drawback is to solely rely on Euclidian distance in estimating missing rainfall data because inconsistent trend of the closer stations will highly affect the result of estimations due to its biasness.

On the other hand,  $k$ -nearest neighbourhood ( $k$ NN) is not limited to only the geographic space.[26] It assigns the target missing value towards the nearest  $k \leq 4$  stations which best resembles it in terms of Euclidian distance and altitude.[16] Similarly, hot-deck is a specific case of  $k = 1$  (single nearest neighbourhood).[27] The weighted  $k$ NN also seen employed by Aieb, *et al.* computes the weighted-average for the  $k$  stations where the weight coefficients includes Euclidian distances and friction distances for the respective  $k$  stations.[16]



Coefficient of determination,  $R^2$  substitutes the Euclidian distance from IDW yields the correlation coefficient weighted (CCW) imputation technique.[19] In other words, this major assumption of linearity of the rainfall data between stations returns the weighted correlations of all neighbourhood stations as an estimation for the missing data. Unfortunately, CCW exhibits the worst estimations with highest biasness and comparatively low precision for all of the target stations.

### 2.3 Model Assessments

In general, the related works and investigations have applied a few similar statistical measures for stations selection and model assessments.[12] where the root mean square error (RMSE) and  $R^2$  is often used. Most works generally agreed that compared to other imputation techniques for missing precipitation data, ANN-related models generate results of higher precision where errors lie between acceptable margins.[29] The studies illustrated concrete evidence that ANN which also acts as a semi-parametric regression estimator,[21] has high degree of potentiality to distinguish non-linearity of the input stations and target stations,[10] as there are complex features for each station, due to climate factors and geographical features which would vary from time to time.[7]

Generalisation errors were verified by the calibration and validation phase which calibrates over few data sets and then validated, previewing the strength of the models and architectures for selection. Efforts were made in studying the effect of the training size on the model performance [30] because of the design of ANN which depends highly on the training set. Hence, it is essential to determine a minimum data size which is relatively insensitive to the chosen period. In context of missing rainfall data estimations, larger size of training sets was recommended as it enhances the performance of models especially ANN.[15]

In short, there is no single model that performed the best at all times and cases as different locations, geographical settings, climate, and other factors were involved in each of the study. For instance, the unevenly distributed rainfall discovered in the Lvuvhu river where most stations were scattered around on the western region suggests that it is not recommended for the implementation of IDW technique. This is because

there were too few rainfall stations to participate in analysing the spatial relationship between the targeted rainfall stations. [18]

Typically, quick, and less sophisticated computational algorithms were main reasons for employing conventional non-ANN models. Most of these models are simpler to work with but the results of estimations were found to be not necessarily robust and are of low accuracy. Nevertheless, ANN models were reviewed to be the most appropriate candidate among the existing imputation techniques as it adapts the non-linear complexity and spatial correlation of rainfall data among different stations associated with relatively low bias error. The high appraisal of the ANN in effectively estimating missing rainfall with high accuracy and precision motivates researchers to employ ANN models in their work.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Missing Mechanism

Common practice by researchers or even the default settings in some statistical software adapts the list-wise deletion strategy where the particular missing observation is discarded directly from the analysis. Although straightforward and easily implemented, such practices could easily possibly lead to a falsify relationship inferred from the remaining data because meaningful, high-influential data might be discarded. It should be noted that methods involving direct discarding data assume that missing observations are missing completely at random (MCAR). Consequently, it is only appropriate for very small proportion of missing data and is insufficient for many analyses due to the severe loss of data. Thus, researches that deals with missing data should clearly define the missing mechanism of the data before any analysis is performed.

Another missing mechanism known as missing not at random (MNAR) is defined when there are specific reasons (i.e. not random) that causes the data to be missing. MNAR however is not applicable in this context. This research emphasises on estimation of missing data in the context of daily rainfall. Thus, it is crucial to recognise that missing daily rainfall data carries the assumption that they are missing at random (MAR). This is so because of the existence of statistical dependence between rainfall values such that the probability of a particular rainfall data,  $x$  coming from the database consisting missing data,  $x_{MD}$  is assumed to depend on other available observable values,  $x_{OBS}$  but not on the missing observation itself.

$$\Pr(x_{MD}|x_{OBS}) = \Pr(x_{MD}|x) \quad (1)$$

After obtaining the daily rainfall data of the ten selected stations across Peninsular Malaysia, missing values were assigned. The assignment of missing data (i.e. data amputation) follows the proportion specified as 10%, 20% and 30% under the MAR mechanism.

## 3.2 Self-Organising Feature Map (SOFM)

With growing interest in hydrological applications such as the satellite imagery classifications, rainfall estimations, runoff modelling and other related analyses, SOFM was reported to be a promising tool in visualising pattern projections because of its topologically ordered mapping within the output that clusters similar neurons in a lattice structure, illustrating the distribution function while preserving topological information within projection space.[31] In this research, the SOFM was employed in two folds. The SOFM is trained using complete daily rainfall observations to obtain the Kohonen layer which is a discrete map of dimension 2. The Kohonen layer stores all computed weight vectors and exhibits the rainfall intensities for the rainfall stations.

### 3.2.1 Data Pre-Processing

Rainfall data were normalised by rescaling each input vectors before fitting into the SOFM to avoid any overweighted calculations on observations during training stage.

$$z_{ij} = \frac{x_{ij} - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}, 0 \leq z_{ij} \leq 1 \quad (2)$$

### 3.2.2 Data Training

Define the training set with normalised input vectors of  $\mathbb{R}^m$ , the following algorithm with three stages is applied:

(I) Initialisation stage

$\forall i = 1, 2, \dots, m, j = 1, 2, \dots, M$ ; initialise randomly at  $s = 0$ :

The connection weights between each of the  $m$  input neurons and the  $M$  neurons in the Kohonen layer,

$$w_{ij}(0) \sim \text{Uniform}(0, 1) \quad (3)$$

The learning rate of the network,

$$\eta(0) \sim \text{Uniform}(0, 1) \quad (4)$$

(II) Competition stage

For each training iteration  $s$ , each connected neuron is competed to obtain the best matching unit (BMU),  $\theta$  based on the minimised distance metrics:

$$D_j^{(\theta)} = \underset{1 \leq j \leq M}{\operatorname{argmin}} \|z_i(s) - w_{ij}(s)\| \quad (5)$$

$$\text{where } D_j = \sqrt{\sum_{i=1}^m (z_i(s) - w_{ij}(s))^2} \quad (6)$$

(III) Cooperative update and learning stage

For each iteration step  $s$ :

Define the radius of the neighbourhood region,  $N_\theta$ ,

$$\delta(s) = \delta(0)e^{-\frac{s}{T}} \quad (7)$$

Define the function  $H_\theta(s)$  for the size of  $N_\theta$  around the BMU  $\theta$ ,

$$H_\theta(s) = e^{-\frac{1}{2} \left( \frac{\|\theta - \theta^*\|}{\delta(s)} \right)^2} \quad (8)$$

where  $T$  is the maximum iteration number, and  $\theta^*$  is the closest neuron towards the BMU.

Define the learning rate function  $\eta(s)$ ,

$$\eta(s) = \frac{\eta(0)}{s} \quad (9)$$

Update the weights of the winning nodes (i.e. BMU) and its neighbourhood neurons are activated whilst the non-updated nodes are deactivated.

$$\begin{aligned}
& \mathbf{if } j \in H_{\theta}(s) , \\
& \quad \mathbf{then } w_{ij}(s + 1) = w_{ij}(s) + \eta(s)H_{\theta}(s)[z_i(s) - w_{ij}(s)], \\
& \quad \mathbf{else } w_{ij}(s + 1) = w_{ij}(s)
\end{aligned} \tag{10}$$

For the case where input vectors consisting missing rainfall value are identified at iteration  $s$ , the competitive stage is not computed for these particular vectors. The corresponding weights will not be updated as shown in Equation (11).

$$w_{ij}(s + 1) = w_{ij}(s) \tag{11}$$

The algorithm runs for each iteration  $s$ , until the training converges for all complete daily rainfall data from each station. The neighbourhood function used is the Gaussian function to describe the weights adjustments where closer nodes towards the BMU is updated more frequently. Both the linear learning rate and exponential decaying neighbourhood radius function are monotonically decreasing functions as intended over the iteration process.[32] Together with the fully updated, adjusted weights in the discrete lattice structure of the Kohonen map, a clear illustration of the distribution function will be exhibited. Similar clusters can be identified by the physically closely arranged neurons and individual component planes or heat maps visualisation helps in the individual breakdown of rainfall intensities for each station where corresponding influential of neurons are indicated by the colour intensities.

### 3.2.3 Missing Daily Rainfall Estimation

The Gaussian neighbourhood function within the Kohonen map enables a posteriori estimation due to the asymptotic convergence behaviour towards the average or mean value of the clustered class. (i.e estimating using weights of the BMU). This process will be repeated for all of the  $\alpha$  values.

$$\hat{x}_{MD} = \bar{x}^{(\theta)} \text{ in } N_{\theta} \tag{12}$$

### 3.3 Model Assessment

$$\text{Mean error, } ME = \frac{\sum_{k=1}^N (\hat{x}_k - x_k)}{N} \quad (13)$$

$$\text{Root mean square error, } RMSE = \sqrt{\frac{\sum_{k=1}^N (\hat{x}_k - x_k)^2}{N}} \quad (14)$$

where  $\hat{x}_k$  are the  $N$  estimated daily rainfall values,  $x_k$  are the  $N$  estimated daily rainfall observations and  $\bar{x}$  is the mean of estimated daily rainfall observations.

## CHAPTER 4

### RESULTS AND DISCUSSIONS

#### 4.1 Research Area

The Peninsular Malaysia is known to consist of 11 states and 2 federal territories with over hundreds of river systems. In addition, massive number of mountains ranges could also be discovered over the Peninsular Malaysia, leading to the disparate mountain elevations followed by unique geographical settings in different areas. Figure 4.01 illustrates the map of Peninsular Malaysia and the location of the 10 selected rainfall stations. The corresponding geographical coordinates were also provided in Table 4.01.

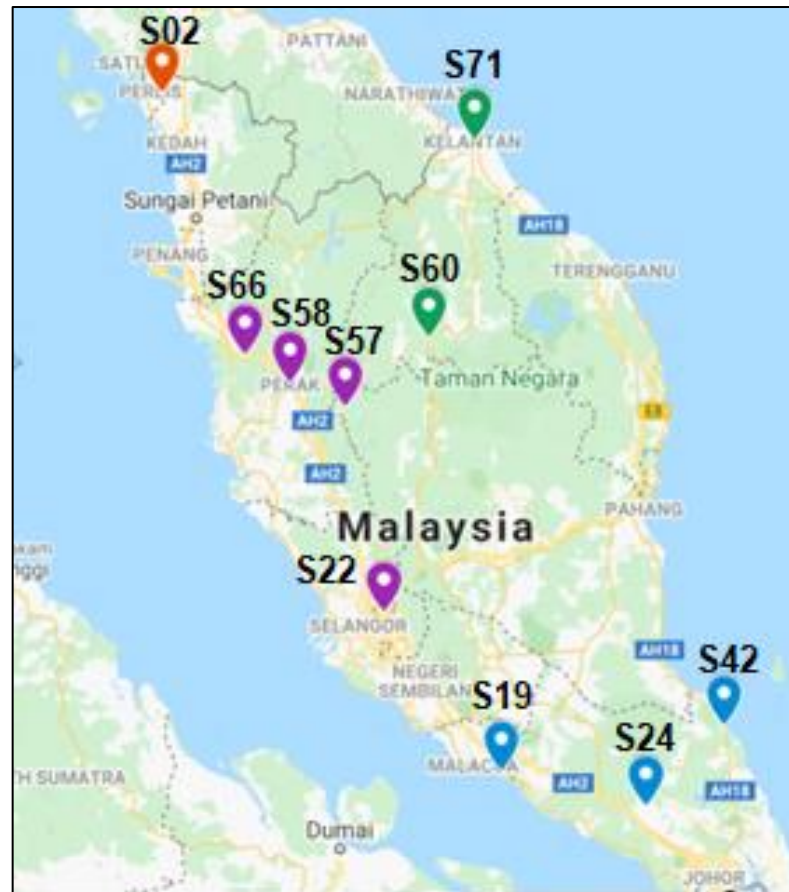


Figure 4.01 Geographical Map of Rainfall Stations Studied



Table 4.01 Geographical Coordinates of Rainfall Stations Studied

Station Code	Station Name	Latitude	Longitude
S02	Arau	6.42970	100.26984
S19	Pekan Merlimau	2.14637	102.43089
S22	JPS Wilayah Persekutuan	3.15564	101.68183
S24	Ladang Benut Rengam	1.92377	103.33382
S42	Mersing	2.43091	103.83611
S57	Ladang Boh	4.44773	101.43433
S58	Ipoh	4.59747	101.09010
S60	Gua Musang	4.88427	101.96817
S66	Bukit Berapit	4.77811	100.79567
S71	Kota Bharu	6.12478	102.25438

#### 4.2 Daily Rainfall Data

The NEM is driven by the land-sea temperature differences which is caused by the heat from solar radiation. The monsoon surge is formed due to the consistent upheaval of cool air that reacts with the lower atmospheric pressure and cyclonic vortices around the equator. This then contributed to the strong NEM winds associated with relatively higher rainfall intensities, ranging from the South China Sea and extended towards the Peninsular Malaysia. The East coast regions of the Peninsular Malaysia were being firstly reached by the NEM.

In this research, a 10-year period (1999-2008) of daily rainfall data for the NEM were studied. The respective daily rainfall data from the 10 selected stations were obtained from the Department of Irrigation and Drainage (DID) Malaysia. The specific date range of the data used for each period of year corresponds to the NEM, commences on 1<sup>st</sup> November, and ends at the 28<sup>th</sup> February (29<sup>th</sup> February for a leap year). The total number of daily rainfall observations were 1203 days, obtained from each of the 10 stations respectively. Based on the length of the studied data period, there were a total of 3 leap years which resulted in the additional 3 daily rainfall observations.

Table 4.02 shows the descriptive statistics, including the mean, standard deviation and the maximum observation of the daily rainfall data that is obtained from the selected rainfall stations. It is apparent that rainfall stations located at or nearer to the East coast regions in general have higher mean of daily rainfall such as the rainfall stations Mersing (S42) and Kota Bharu (S71). The lowest mean daily rainfall was from Arau (S02), located in the North-West most region among all stations. It follows that the total rainfall amount was least from Arau and the highest ones were from Mersing and Kota Bharu. Additionally, the standard deviations of the daily rainfall data were mostly consistent with the mean where the higher the mean, the higher the standard deviation. The largest standard deviation was again from Kota Bharu with the highest mean value. Correspondingly, the variation of daily rainfall data may be explained by the maximum rainfall within the data. The highest maximum rainfall observation of 327.3 mm from Kota Bharu for instance had contributed a correspondingly large value for the standard deviation of 27.8068. Stations with lower maximum rainfall observation tend to have smaller variation, shown by the lower standard deviations.

Table 4.02 Descriptive Statistics of the Daily Rainfall Data

<b>Station Code</b>	<b>Station Name</b>	<b>Mean (mm)</b>	<b>Standard Deviation (mm)</b>	<b>Maximum Rainfall (mm)</b>	<b>Total Rainfall (mm)</b>
S02	Arau	4.0832	12.4428	180.0	4912.1
S19	Pekan	4.3584	10.4217	83.0	5243.2
	Merlimau				
S22	JPS Wilayah Persekutuan	8.6453	18.0359	289.0	10400.3
S24	Ladang Benut Rengam	6.4601	15.9647	210.0	7771.5
S42	Mersing	10.8300	26.6662	312.1	13028.6
S57	Ladang Boh	6.9612	10.5502	75.0	8374.3
S58	Ipoh	8.0344	15.4409	132.8	9665.4
S60	Gua Musang	6.1745	12.8749	123.5	7427.9
S66	Bukit Berapit	7.2029	15.3481	140.0	8665.1
S71	Kota Bharu	11.0979	27.8068	327.3	13350.8

Apart from considering the daily rainfall values, it is also crucial to understand the condition of wet days of the data studied. The main reason is that daily rainfall values and occurrence of rainfall are different measures with distinct significance which neither of the two could be fully dominated by another. The frequency of wet days and the relative probability of the occurrence of a wet day were shown in Table 4.03. The Arau station remained having the lowest wet days count with the lowest probability of wet day occurrence. The three stations, Ladang Benut Rengam (S24), Pekan Merlimau (S19), and Bukit Berapit (S66) exhibited moderate frequency of wet days. The other stations have their wet days exceeded at least half of the total number of days studied (i.e. 1023 daily observations). These results provided evidence that wet days occurrence does not necessarily affect the daily rainfall values since any station can have low count of wet days but possesses a high rainfall daily rainfall value for each of its wet days.

Furthermore, the inconsistency between the wet days and the daily rainfall intensities could be due to the increased number of dry days at the end of the NEM season. Hence, near the end of NEM during February, the impact of NEM would be weakened due to the weaker winds contributed by the NEM. Consequently, a greater number of dry days were included in the study that diminishes the total wet days frequency. This results in the lower relative probability for the occurrence of wet days.

Followed by the reduced impact by the NEM during February, the warmer weather begins to take place where rainfall amounts were expected to decline. Humidity of cities remained to be higher, leading to more short tropical showers instead of long and heavy rainfalls. The reason is that the warmer and higher humidity allows more moisture (i.e. water content) to be held in the atmosphere, which then causes the increase in evaporation rate. This leads to the contribution of more wet days but low rainfall values for these particular wet days in the city states. The stations of JPS Wilayah Persekutuan and Ipoh were good examples that illustrates this particular phenomenon. The highest frequency of wet days can be seen from the Ladang Boh station (S57) located at a higher altitude. The higher altitude creates a lower atmospheric pressure causing the total water vapour to be held in the atmosphere to be decreased. This resulted in higher probability of wet days but with low rainfall amount, corresponding to the least maximum rainfall observation of 75.0 mm.

Table 4.03 Wet Days of the Daily Rainfall Data

<b>Station Code</b>	<b>Station Name</b>	<b>Wet Days Frequency</b>	<b>Relative Probability</b>
S02	Arau	326	0.2710
S19	Pekan Merlimau	463	0.3849
S22	JPS Wilayah Persekutuan	694	0.5769
S24	Ladang Benut Rengam	404	0.3358
S42	Mersing	673	0.5594
S57	Ladang Boh	761	0.6326
S58	Ipoh	691	0.5744
S60	Gua Musang	642	0.5337
S66	Bukit Berapit	493	0.4098
S71	Kota Bharu	642	0.5337

### 4.3 Trained SOFM

The Kohonen map of the trained SOFM and the individual component planes for each station are convenient tools to visualise the relationship between the observed rainfall intensities and clusters of similar properties. Figure 4.02, Figure 4.03, and Figure 4.04 showed the Kohonen map of the trained SOFM with the proportion of missing daily rainfall observations 10%, 20% and 30% respectively. The structures of all Kohonen maps were hexagonal. By trial and error, the most appropriate map size was decided to be of dimension 6×6.

The area of individual coloured sectors inside each of the neuron represents the rainfall intensity of the studied rainfall stations respectively. For example, observations possessing the highest rainfall intensity from the Arau station that dominates all other rainfall intensities was indicated by the largest red sector in the neuron, located in (row 5, column 4) in Figure 4.02, (row 3, column 3) in Figure 4.03, and (row 2, column 2) in Figure 4.04. Lower rainfall intensity observations were indicated by smaller size of the sectors in a neuron. Neurons located in (row 6, column 6) in Figure 4.02 and (row 6, column 1) in Figure 4.03 and Figure 4.04 showed examples of observations of very low or totally no rainfall for all of the rainfall stations.

By trial and error using the trained SOFM, a total of 5 cluster regions was used to divide the Kohonen map based on their similarities between input and the neurons. Neurons that were of different clusters are clearly separated by the thick black boundary lines whereby same cluster regions were indicated with the same background colour in the neurons. In other words, observations that were classified into any of the neurons from the same cluster region is believed to share some degree of similarity in terms of the pattern of daily rainfall values. Though boundaries were constructed to obtain the 5 clusters, one could examine that the patterns of rainfall intensities among the rainfall stations illustrated by the colourful sectors in the neurons of the three respective Kohonen maps were mostly distinct from each and another.

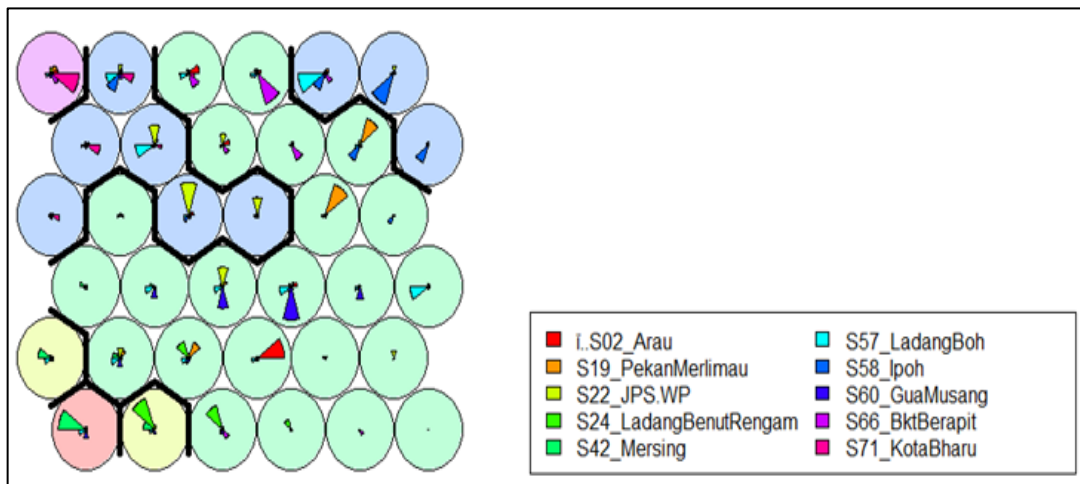


Figure 4.02 Kohonen Map of Trained SOFM (10% Missing Data)

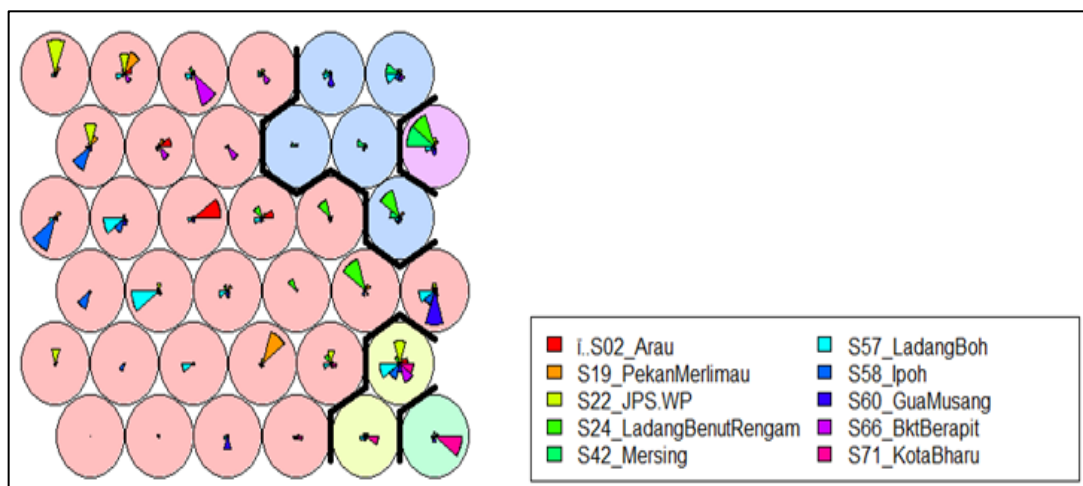


Figure 4.03 Kohonen Map of Trained SOFM (20% Missing Data)

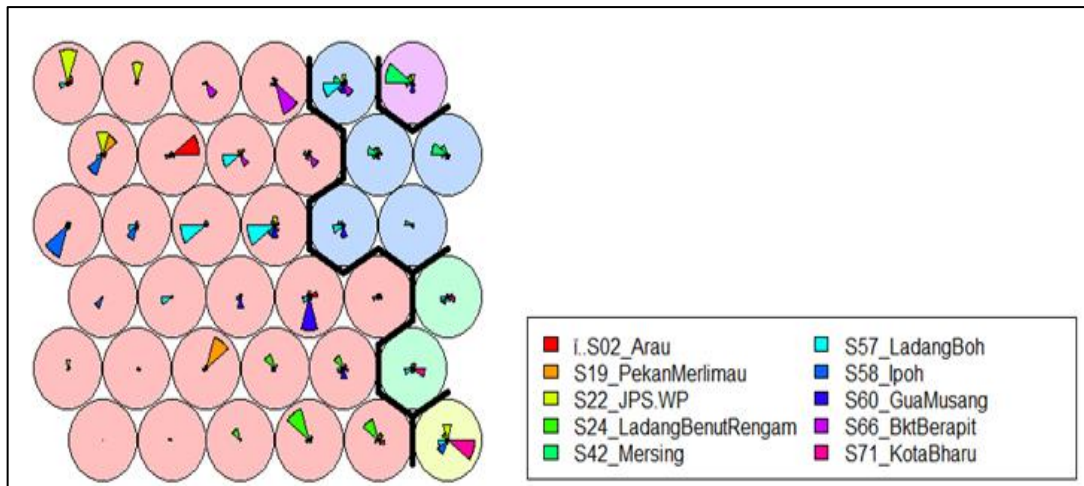


Figure 4.04 Kohonen Map of Trained SOFM (30% Missing Data)

The individual breakdown of the rainfall intensity distribution for each of the ten rainfall stations is illustrated using the heat maps as shown in Figure 4.05, Figure 4.06, and Figure 4.07 for each of the missing proportion 10%, 20%, and 30% respectively. The colour that is filled by the neurons in the heat maps are based on the spectral band that indicates the rainfall intensity. The colour of the band ranges from red, orange, yellow to white. A darker colour closer to the colour of red indicates a low rainfall intensity whereas a lighter colour that is closer to the colour of white indicates a higher rainfall intensity from the spectral band. For instance, the white neuron located at (row 5, column 4) of the heatmap from Arau station in Figure 4.05 corresponding to the same neuron position in the Kohonen map in Figure 4.02, indicates the location where observations of highest rainfall value from Arau stations will be classified.

Once again from the heat maps, it could be seen that the rainfall stations do not share much similarity in terms of the rainfall pattern. It should however be noted that the heat maps of the rainfall stations, Mersing (S42) and Kota Bharu (S71) contains observations of high daily rainfall values reflected by their wider range of spectral band. The remaining stations share a lower, but similar range of spectral band. This phenomenon can be explained by the geographical locations of Mersing and Kota Bharu. Both of the rainfall stations sit on the border of the West coast of the Peninsular Malaysia, facing high rainfall intensities associated by strong winds of the North-East monsoon season. The EWM wind may have weaker impact towards the other rainfall stations of the west regions due to the existence of numerous high mountains such as

the well-known Titiwangsa range, acting as a shield for these stations, lowering the convection and rainfall intensities. Since the geographical conditions for the rainfall stations are quite different from one another, the resulting rainfall distributions is then sparse and unequal.

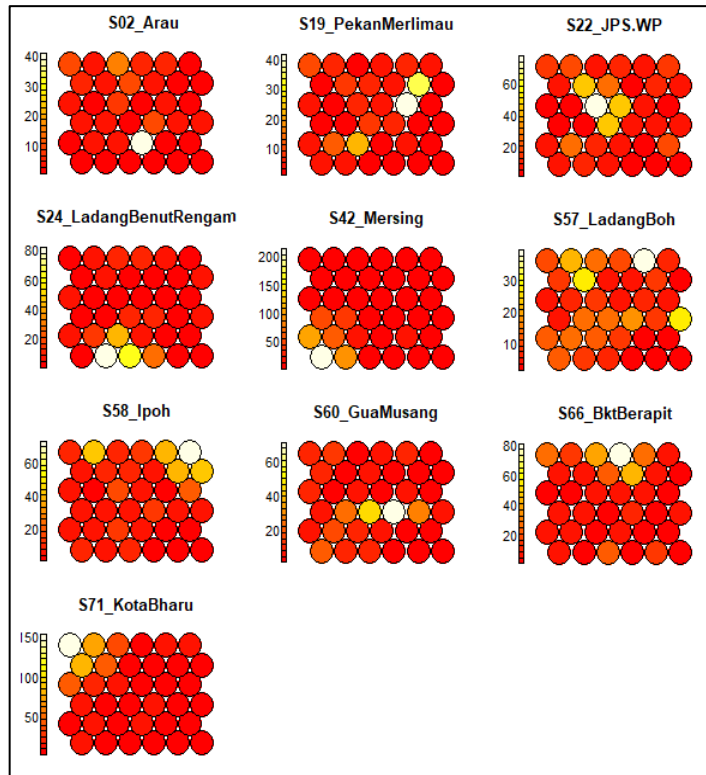


Figure 4.05 Heat Maps of Rainfall Stations (10% Missing Data)

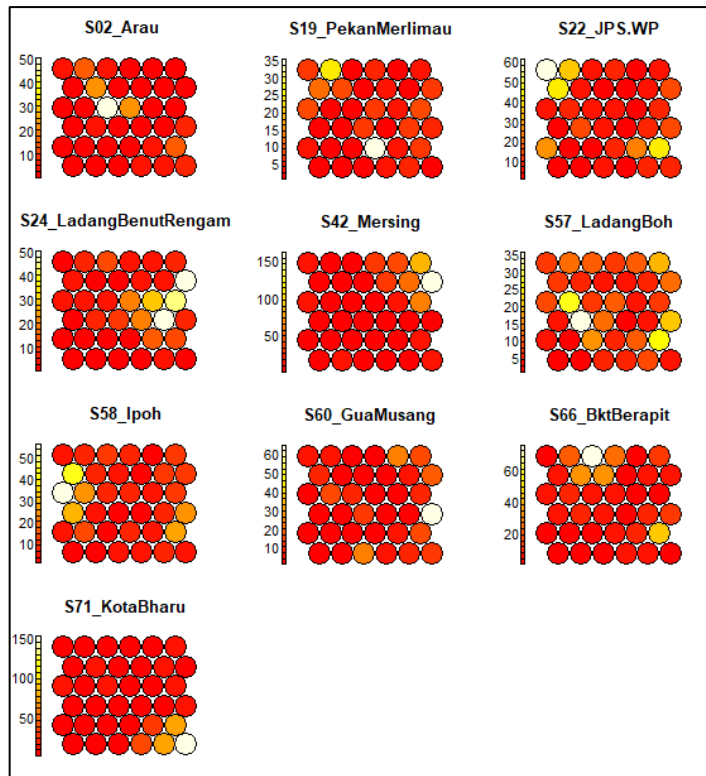


Figure 4.06 Heat Maps of Rainfall Stations (20% Missing Data)

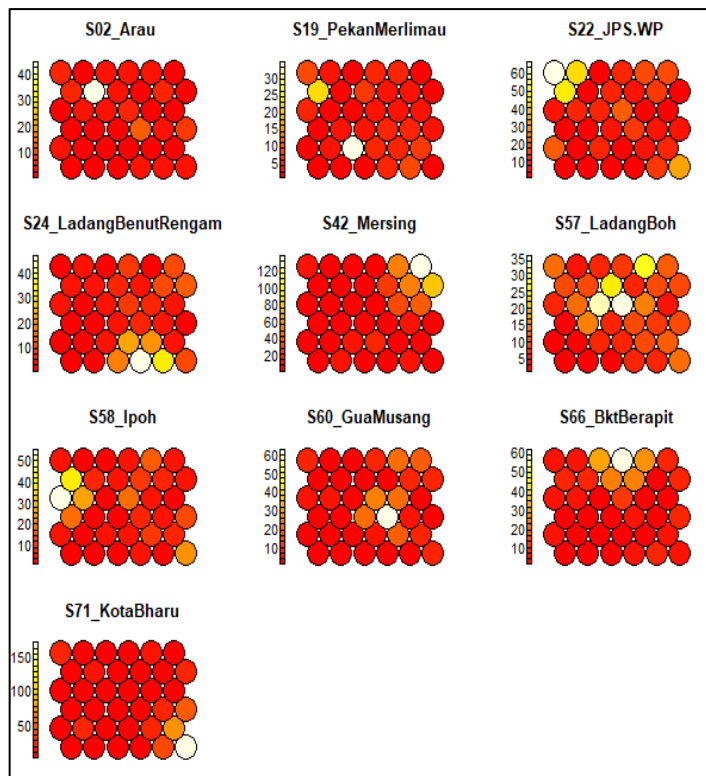


Figure 4.07 Heat Maps of Rainfall Stations (30% Missing Data)



#### 4.4 Performance of the SOFM

The missing daily rainfall observations were estimated via the trained SOFM. The ME which represents the mean difference between the actual observations and the estimated observations were computed for each of the proportion of missing daily rainfall data shown in Table 4.04. The corresponding RMSE were also computed to assess the performance of the SOFM in estimating the missing daily rainfall data.

Table 4.04 Performance Measure of the SOFM

<b>Proportion of Missing Daily Rainfall Data (%)</b>	<b>ME</b>	<b>RMSE</b>
10	-2.2969	16.2989
20	-3.3400	15.7823
30	-4.8293	22.4558

The ME of the estimated missing observations were improved as the proportion of missing daily rainfall data reduces from 30% to 10%. Although the computed ME were reasonably low in the context of daily rainfall value estimation, the negative values of ME suggested that the trained SOFM model for each of the three cases have provided an underestimate for the missing daily rainfall observations on average. Satisfactory results are obtained as the computed RMSE improves when the missing proportion of daily rainfall data was reduced. However, the slight inconsistency when the missing proportion is at 20% could be explained by the complex topological difference, associated with distinct elevations and Euclidean distances among the rainfall stations.

Based on the results, the accuracy of the SOFM in estimating the missing daily rainfall data is the highest when the missing proportion is at 10% as expected. The relative precision is high but is slightly lower than when the missing proportion is at 20%. These were illustrated in the comparison plots between the actual and estimated daily rainfall observations in Figure 4.08, Figure 4.09, and Figure 4.10. The comparison plots agreed with the underestimation of the SOFM model, exhibited by

the red lines (estimated observations) that falls mostly below the blue lines (actual observations).

Most estimated observations matched the actual observations well but carries several highly fluctuated ones for the missing proportion of 10% seen in Figure 4.08. Figure 4.09 shown the estimated observations were also close to the actual observations but was comparatively more consistent with less variation for the missing proportion of 20%. There were slightly more evidences of underestimated rainfall observations with larger variation for which the SOFM have not captured for the missing proportion of 30% illustrated in Figure 4.10. Overall, the missing daily rainfall observations that were highly underestimated were those observations of high rainfall value. In other words, the high peaks seen in the comparison plots were usually the ones that were not very accurately captured by the SOFM model. The primary cause being that the daily rainfall observations that is missing were those of rare and high rainfall values, not being accounted during the training of the SOFM model.

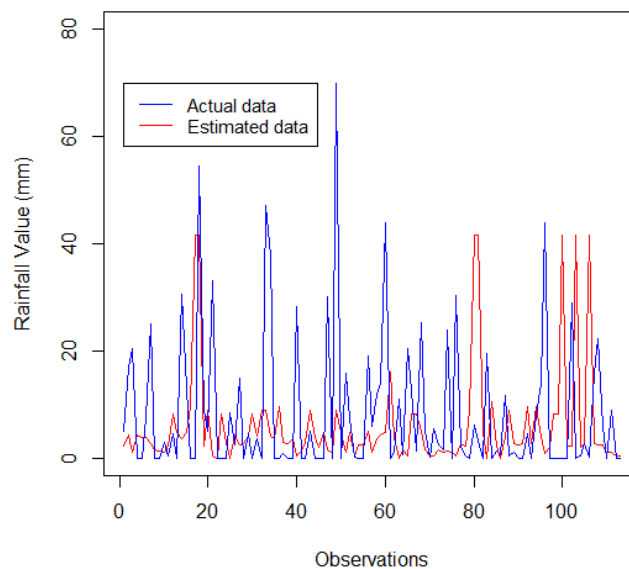


Figure 4.08 Comparison Plot Between Actual and Estimated Rainfall Observations (10% missing data)

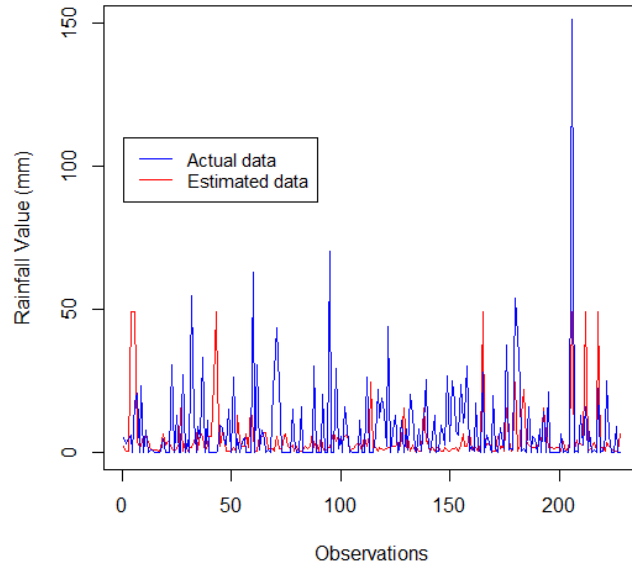


Figure 4.09 Comparison Plot Between Actual and Estimated Rainfall Observations (20% missing data)

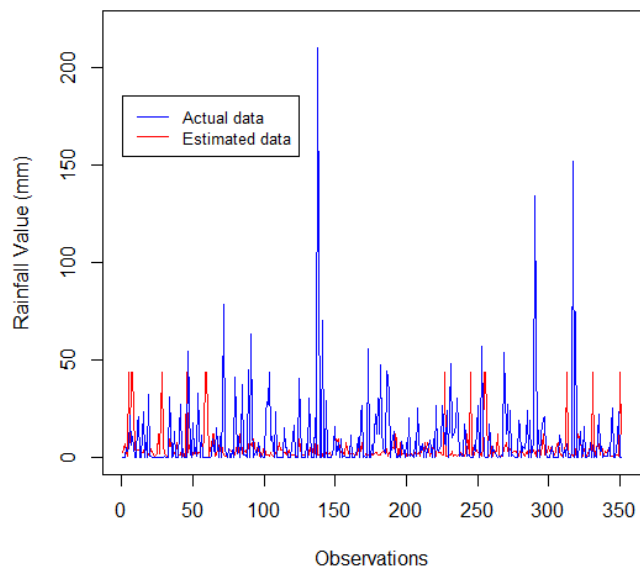


Figure 4.10 Comparison Plot Between Actual and Estimated Rainfall Observations (30% missing data)

## CHAPTER 5

### CONCLUSIONS

#### 5.1 Conclusions

The application of SOFM model in estimating the missing daily rainfall data has shown to be appropriate based on the encouraging results that were established. SOFM is indeed a reliable ANN model that is recommended, where it could also be implemented in various hydrological-related researches, in particularly, in estimating missing daily rainfall data. With the capability of the SOFM model in determining the existence of spatial relationship of rainfall events between the studied rainfall stations, results suggested that each of the stations possessed distinct characteristics in terms of rainfall patterns and distributions due to the effects caused by different climate factors, geographical, and regional variability. The performance measures were evaluated for all three proportions of missing daily rainfall data. The estimation for all levels of missing proportions resulted in consistently low and negative values of the ME, indicating an underestimation by the SOFM model associated with sensible values of the RMSE produced.

#### 5.2 Recommendations

The results of the estimation of missing daily rainfall data via the SOFM model may be improved by incorporating more relevant climate factors such as humidity, temperature, and atmospheric pressure into the SOFM model to help analyse the rainfall patterns and distributions. It is also recommended to further extend the study by exploring with more homogenous rainfall stations based on cluster regions, no limited to Euclidean distance-based criterion . Consequently, only rainfall stations of the same cluster region will be studied and the individual SOFM will be trained separately for each of these clusters. The corresponding rainfall stations from the same cluster region defined are then known to be homogenous. The study could also extend or shorten the data period or include the SEM rainfall data so that more relevant and constructive data will be accounted by the SOFM model to enhance the result of estimations for the missing daily rainfall data.

## REFERENCES

- [1] Chuah, H.L., 2016. *Statistical models for daily rainfall data: A case study in Selangor, Malaysia*. Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman.
- [2] Fauziana, A., Tomoki, U. and Takahiro, S., 2017. *Determination of Z-R relationship and inundation analysis for Kuantan river basin*. Malaysian Meteorological Department. Ministry of Science, Technology, and Innovation.
- [3] Mahmood, R., Jia, S.F. and Zhu, W.B., 2019. *Analysis of climate variability, trends, and prediction in the most active parts of the Lake Chad basin, Africa*. Scientific Reports. 9:6317, 1-18. <https://doi.org/10.1038/s41598-019-42811-9>.
- [4] Kalteh, A.M., 2007. *Rainfall-runoff modelling using artificial neural networks (ANNs)*. Department of Water Resources Engineering, Lund Institute of Technology, Lund University.
- [5] Ho, M.K. and Yusof, F., 2012. *Application of Self-Organizing Map (SOM) in Missing Daily Rainfall Data in Malaysia*. International Journal of Computer Applications. 0975 -888:48,5.
- [6] Dawson, C.W. and Robert, W., 2001. *Hydrological Modelling Using Artificial Neural Networks*. Progress in Physical Geography. 25:1, 80-108. DOI: [10.1177/030913330102500104](https://doi.org/10.1177/030913330102500104).
- [7] Govindaraju, R.S., 2000, *Artificial neural network in hydrology. I: Preliminary concepts*. Journal of Hydrologic Engineering. 5:2, 115-123.
- [8] Adib, P., 2019. *Malaysia 2 Dead, 15,000 Displaced as Floods Worsen in Kelantan and Terengganu*. <http://floodlist.com/asia/malaysia-floods-kelantan-terengganu-december-2019>.
- [9] Joshua, B. and Isaac, Y., 2020. *66 people now killed by flooding in Jakarta and more rain appear to be on the way*. CNN. <https://cnnphilippines.com/world/2020/1/6/jakarta-monsoon-rain-flooding-deaths.html>.
- [10] Das, S.K., Gupta R.K. and Varma, H.K., 2007. *Flood and drought management through water resources development in India*. WMO Bulletin. 3:56, 179-188.
- [11] Googhari, et al., 2010. *Neural networks for forecasting daily reservoir inflows*. Pertanika Journal of Science and Technology. 0128-7680. 18:1, 33-41.
- [12] Nastar, et al., 2019. *Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks*. 2352-3409. <https://doi.org/10.1016/j.dib.2019.104517>.

- [13] Madhuri, D., and Hardaha, M.K., 2019. *Application of Standard Models and Artificial Neural Network for Missing Rainfall Estimation*. International Journal of Current Microbiology and Applied Science. 8:1, 1564-1572. <https://doi.org/10.20546/ijcmas.2019.801.164>.
- [14] Kim, J.W., and Pachepsky, Y.A., 2010. *Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT stream flow simulation*. ISSN:0022-1694. DOI: 10.1016/j.jhydrol.2010.09.005.
- [15] Moradkhani, et al., 2004. *Improved streamflow forecasting using self-organizing radial basis function artificial neural networks*. Journal of Hydrology. 295, 246-262.
- [16] Aieb, et al., 2019. *A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria*. Heliyon, e01247. DOI: 10.1016/j.heliyon.2019. e01247.
- [17] Annalisa, D.P., et al., 2011. *Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy*. ISSN:0303-2434. DOI:10.1016/j.jag.2011.01.005.
- [18] Nkuna, T.R., and Odiyo, J.O., 2011. *Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks*. Journal of Physics and Chemistry of the Earth. 830-835.
- [19] Barrios, A., Trincado, G., and Garreaud, R., 2018. *Alternative approaches for estimating missing climate data: application to monthly precipitation records in South Central Chile*. Forest Ecosystems. 5:28 <https://doi.org/10.1186/s40663-018-0147-x>.
- [20] Amar, G.A., Al-Darwish, A.Q., and Slieman, A.A., 2018. *Infilling daily data using feedforward back-propagation artificial neural network (ANN), Hama, Syria*. American journal of innovative research and applied sciences. 7:2, 109-117.
- [21] Radan, H., and Luci, P., 2005. *A simultaneous analysis of climatic trends in multiple variables: An example of application of multivariate statistical methods*. International Journal of Climatology. 25:4,469-484. <https://doi.org/10.1002/joc.1146>.
- [22] Kramer, M.A., 1991, *Non-linear principal component analysis using auto-associative neural networks*. AIChE Journal. 37:2, 233-243.
- [23] Licciardi, G. and Chanussot, J., 2018, *Spectral transformation based on non-linear principal component analysis for dimensionality reduction of hyperspectral images*. European Journal of Remote Sensing. 51:1, 375-390. DOI: 10.1080/22797254.2018.1441670.

- [24] Sliva, D.R.P., Dayawansa, N.D.K., and Ratnasiri, M.D., 2007. *A comparison of methods used in estimating missing rainfall data*. The Journal of Agricultural Sciences. 3:2.
- [25] Ramesh, T., 2018. *Trends and changes in hydroclimatic variables: Links to climate variability and change*. Elsevier. ISBN:9780128109861, 416.
- [26] Sun, H., et al., 2018. *Optimizing kNN for mapping vegetation cover of arid and semi-arid areas using Landsat images*. Remote Sens. 10, 1248.
- [27] Andras, B., and Geoffrey, P., 2014. *Infilling missing precipitation records: A comparison of a new copula-based method with other techniques*. Journal of Hydrology. 519, 1162-1170. DOI: [10.1016/j.jhydrol.2014.08.025](https://doi.org/10.1016/j.jhydrol.2014.08.025).
- [28] Kagoda, P.A., and Ndiritu, J., 2009. *Application of radial basis function neural networks to short-term streamflow forecasting*. Journal of Physics and Chemistry of the Earth. 35, 571-581.
- [29] Muhammad, A.A., Radziah, O., and Che, F.I, 2017. *Soils of Malaysia*, CRC Press. 214.
- [30] Gupta, H.V. and Sorooshian, S.,1998. *Toward improved calibration of hydrologic models: Multiple and non-commensurable measures of information*. Water Resources Research. 34:4, 751-763.
- [31] Murao, H. et al., 1996. *A hybrid neural network system for the rainfall estimation using satellite imagery*, Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), Nagoya, Japan,1993. 2,1211-1214. DOI: [10.1109/IJCNN.1993.716761](https://doi.org/10.1109/IJCNN.1993.716761).
- [32] Natita, W., Wiboonsak, W. and Dusadee S, 2016. *Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand*. International Journal of Modeling and Optimization. 6:1. DOI: [10.7763/IJMO.2016.V6.504](https://doi.org/10.7763/IJMO.2016.V6.504).