STATISTICAL DOWNSCALING OF PROJECTING RAINFALL AMOUNT BASED ON SVC-RVM MODEL

NURUL AININA FILZA BINTI SULAIMAN

SULTAN IDRIS EDUCATION UNIVERSITY

2022

STATISTICAL DOWNSCALING OF PROJECTING RAINFALL AMOUNT BASED ON SVC-RVM MODEL

NURUL AININA FILZA BINTI SULAIMAN

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

DISSERTATION PRESENTED TO QUALIFY FOR A MASTER'S DEGREE IN SCIENCE (APPLIED STATISTICS)
(RESEARCH MODE)

FACULTY OF SCIENCE AND MATHEMATICS
SULTAN IDRIS EDUCATION UNIVERSITY

2022

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

UPSI/IPS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Please tick (√)
Project Paper
Masters by Research
Master by Mixed Mode
PhD

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the 08 day of 02 20 22

i. Student's Declaration:

I, Nurul Ainina Filza Sulaiman, M20181001097, Fakulti Sains dan Matematik (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled STATISTICAL DOWNSCALING OF PROJECTING RAINFALL AMOUNT BASED ON SVC-RVM MODEL is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

Signature of the student

ii. Supervisor's Declaration:

I Dr. Shazlyn Milleana Shaharudin (SUPERVISOR'S NAME) hereby certifies that the work entitled STATISTICAL DOWNSCALING OF PROJECTING RAINFALL AMOUNT BASED ON SVC-RVM MODEL (TITLE) was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a * partial/full fulfillment for the conferment of Master in Science (Applied Statistics) (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

08/02/2022
Date

Signature of the Supervisor

UPSI/IPS-3/BO 31
Pind.: 01 m/s:1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: STATISTICAL DOWNSCALING OF PROJECTING RAINFALL AMOUNT BASED ON SVC-RVM MODEL

No. Matrik /*Matric's No.*: M20181001097

Saya / *I* : Nurul Ainina Filza binti Sulaiman

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (~~Kedoktoran~~/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah / *Please tick (√) for category below:-*

☐	**SULIT/*CONFIDENTIAL***	Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*
☐	**TERHAD/*RESTRICTED***	Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restricted information as specified by the organization where research was done.*
☑	**TIDAK TERHAD / *OPEN ACCESS***	

(Tandatangan Pelajar/ Signature)

Tarikh: 08/02/22

(Tandatangan Penyelia / *Signature of Supervisor*)
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGMENT

First of all, I am beyond grateful to Allah S.W.T. for the strength that I've blessed with throughout this journey full of challenges. My deepest appreciation and millions of thanks to my supervisor, Dr. Shazlyn Milleana binti Shaharudin for all the knowledge shared, time spent and energy contributed to help me complete this study. All her support and sacrifices are surely unrequited. I would also like to thank you my parents, Norsita binti Idris and Sulaiman bin Jaseh for always faithfully listening to the ups and downs in finishing this journey. Not forgetting also for my siblings and friends who are endlessly giving me words of encouragement and praying for success. I believe I could not reach this point without the prayer and support from all of them. I would also like to thank the Fundamental Research Grants Scheme (FRGS/2019/STG06/UPSI/02/4) provided by the Ministry of Education of Malaysia for supporting my study. A lot of thanks to Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris for the chance and facilities provided and also a special thanks to the Jabatan Pengaliran dan Saliran (JPS) for the data and information provided for this study.

ABSTRACT

The objective of this study is to evaluate and compare the proposed statistical downscaling model in Kelantan and Terengganu states. The study also investigates themost accurate imputation methods in handling the missing atmospheric data and the important predictors for a statistical downscaling method by reducing the dimensionality data. The data used in this study include atmospheric data (predictors) and daily rainfall data (predictand) from 1998 until 2007. As part of its methodology, this study had used an imputation method for handling missing data. Then, Principal Component Analysis (PCA) was applied to rectify the issue of high-dimensional data and select predictors for a two-phase model. The two-phase machine learning techniques were introduced as a precise statistical downscaling method in Kelantan and Terengganu states. The first phase is a classification using the Support Vector Classification (SVC) that determines dry and wet days. Subsequently, a regression estimates the amount of rainfall based on the frequency of wet days using the Support Vector Regression (SVR), Artificial Neural Network (ANN), and Relevant Vector Machine (RVM). The proposed model was analysed by using the performance measures that are Root Mean Square Error (RMSE) and Nash-Sutcliffe Efficiency (NSE). The result of imputation methods shows Random Forest (RF) is having the lowest RMSE value and the highest NSE value. The analysis of PCA results indicates two selected Principal Component's cut-off eigenvalues at 1.6 and 70.29% cumulative percentage of the total variance. In the conclusion of this study, the comparison of results from the SVC and RVM hybridizations reveals that the hybrid reproduces the most reasonable daily rainfall projection and supports the high rainfall extremes, making it a perfect candidate for rainfall prediction research. The implication of this study is to establish the relationship between predictand variables and predictors in order to improve predicting accuracy in climate change projections by using a hybridization model.

PENURUNAN STATISTIK DENGAN MODEL SVC-RVM BAGI MENGUNJURKAN JUMLAH HUJAN

ABSTRAK

Objektif kajian ini adalah untuk menilai dan membandingkan model penskalaan statistik yang dicadangkan di negeri Kelantan dan Terengganu. Kajian ini juga menyiasat kaedah imputasi yang paling tepat dalam mengendalikan data atmosfera yang hilang dan memilih peramal yang penting untuk kaedah penurunan skala statistik dengan mengurangkan data dimensi. Data yang digunakan dalam kajian ini ialah data atmosfera (peramal) dan data hujan harian (ramalan) dari tahun 1998 hingga 2007. Sebagai sebahagian daripada metodologi, kajian ini telah menggunakan kaedah imputasi untuk mengendalikan data yang hilang. Kemudian, Analisis Komponen Prinsipal (PCA) digunakan untuk menyelesaikan isu data dimensi tinggi dan memilih peramal untuk model dua fasa. Teknik pembelajaran mesin dua fasa diperkenalkan sebagai kaedah penurunan skala statistik yang tepat di negeri Kelantan dan Terengganu. Fasa pertama ialah pengelasan menggunakan Klasifikasi Vektor Sokongan (SVC) yang menentukan hari kering dan basah. Selepas itu, regresi yang menganggarkan jumlah hujan berdasarkan kekerapan hari basah menggunakan Regresi Vektor Sokongan (SVR), Rangkaian Neural Buatan (ANN), dan Mesin Vektor Berkaitan (RVM). Model yang digunakan telah dianalisis dengan menggunakan ukuran prestasi iaitu Ralat Purata Punca Kuasa Dua (RMSE) dan Kecekapan Nash-Sutcliffe (NSE). Hasil kaedah imputasi menunjukkan Hutan Rawak (RF) mempunyai nilai RMSE yang terendah dan nilai NSE yang tertinggi. Analisis keputusan PCA menunjukkan dua Komponen Prinsipal (PC) yang terpilih dipotong pada nilai eigen 1.6 dan peratusan kumulatif 70.29% dari jumlah varians. Kesimpulannya, hasil perbandingan daripada hibridisasi SVC dan RVM mendedahkan bahawa hibrid ini telah mengeluarkan semula unjuran hujan harian yang paling wajar dan mengambil kira hujan lebat yang tinggi, justeru menjadikannya calon yang tepat bagi penyelidikan ramalan hujan. Implikasi kajian ini adalah untuk mewujudkan hubungan antara pembolehubah ramalan dan peramal bagi meningkatkan ketepatan ramalan dalam unjuran perubahan iklim dengan menggunakan model hibridisasi.

CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF ABBREVIATIONS

ANFIS	Adaptive Neuro-Fuzzy Interferences
ANN	Artificial Neural Network
ARD	Automatic Relevance Determination
BMA	Bayesian Model Average
CART	Classification And Regression Tree
CV	Cross Validation
DID	Department of Irrigation and Drainage
DR	Dimension Reduction
DRT	Dimension Reduction Techniques
ELM	Extreme Learning Machine
EM	Expectation Maximization
FA	Factor Analysis
GCM	Global Climate Model
GP	General Programming
GPR	Gaussian Process Regression
GP-SPCA	Generalized Power method for Sparse Principal Component Analysis
GRNNSK	Generalized Regression Neural Network
KCV	K-fold Cross Validation
KNN	K-Nearest Neighbor
LSSVM	Least Squared Support Vector Machine
MAE	Mean Absolute Error
MAR	Missing at Random
MCA	Maximum Covariance Analysis
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo

LIST OF ABBREVIATIONS

MI	Multiple Imputation
MICE	Multivariate Imputation by Chained Equations
ML	Machine Learning
MLR	Multiple Linear Regression
MNAR	Missing Not at Random
NCEP-CFSR	National Centers of Environment Prediction (NCEP) Climate Forecast System Reanalysis
NIPALS	Non-linear Interactive Partial Least Squared
NN	Nearest Neighbor
NSE	Nash-Sutcliffe Error
PC	Principal Component
PCA	Principal Component Analysis
PCC	Pearson's Correlation Coefficient
R^2	Determination Coefficient
Rbias	Relative Bias
RCM	Regional Climate Model
RF	Random Forest
RMSE	Root Mean Square Error
RSM	Response Surface Methodology
RVM	Relevance Vector Machine
SDSM	Statistical Downscaling Model
SVC	Support Vector Classification
SVR	Support Vector Regression
TM-PCA	Tree Structure Multi linear Principal Component Analysis
WT	Wavelength Transform

APPENDIX LIST

CHAPTER 1

INTRODUCTION

1.1 Overview

Statistical downscaling is widely applied in many regions and over a range of climate change impacts. This chapter will discuss the background of the study, problem statement, research objectives, the significance of the study, notation, research methodology, thesis organization, and the chapter will close with the limitation of study.

1.2 Background of study

Climate change analyses the behaviour of weather like forecasting rainfall where specific features such as humidity and wind are used to predict rainfall in specific locations (Zainudin et al., 2016). Changing rainfall and river flow patterns such as the intensification of droughts, floods, typhoons, and monsoons will affect all water users. The study of climate change is widely expanding worldwide by focusing on various aspects like impacts of climate change, patterns of data, and extreme rainfalls. Flood is

the most significant natural hazard in Malaysia in terms of population affected, frequency, area extent, flood duration, and socio-economic damage (Weng Chan, 2012). According to previous studies (D/iya et al., 2014), 29 flooding events were recorded from 1980 to 2010. Hence, due to the volatility in extreme cases of unforeseen climate change events in the region over the recent years, the importance of predicting future climate at regional and local levels has become even direr. For that reason, prediction rainfall in Malaysia is analysed by meteorologists of hydrologists with a focus on heavy rainfall events by using machine learning techniques.

Hydrologists are constantly dealing with large datasets because they frequently use time series data. The time series data is the data collected at a specific time like daily, hourly, weekly, monthly, and annually (Aftab et al., 2018). The collected data will be in high dimensional or big data with multiple variables or attributes. Essentially, there is no consensus in the literature on the minimum number of dimensions required to make a data set high dimensional (Assent, 2012). There are some issues and challenges that must be resolved or managed when dealing with this type of data set. The high dimensionality of the data increases the time to create the training data set and algorithms required to solve the function prediction problem (Loyola R et al., 2016). Other than that, it is harder to extract values and interpret information from large data sets during the capturing and analysis processes (Katal et al., 2013), resulting in poor analysis model performance.

Aside from that, when dealing with time series data, the presence of missing data is unavoidable. Missing data in hydrological time series data is a common but serious problem since it can lead to biased results and, in the worst-case scenario,

prevent important analyses of the variables from being performed. The biased result occurs when the data set is reduced and, in certain cases, the row of data is removed and ignored (Kaiser, 2014). Normally, shortages of data problem is due to rainfall station relocation, erroneous sampling, insufficient sample size, or recording issues (Kamaruzaman et al., 2017). The studies by Barnard and Meng (1999) introduced the problems associated with missing values, which are efficiency loss and complications in handling and interpreting data. These issues arose because methods and algorithms were incapable of dealing with missing data that must be analysed during the pre-processing phase, which is a time-consuming process. Hence, determining the best method for missing data handling is imperative to resolve issues that arise from missing values.

Nowadays, the application of downscaling in climate change data has become popular due to its capability to extract the relationship between local climate and atmospheric variables. The downscaling of global climate change projections has been developed to serve the needs of decision-makers who require local climate information for impact assessments. Global Climate Model (GCM) provide information at scales on the order of 100–500 kilometres for studies that focus on large geographic regions and direction of change, e.g., increase or decrease in temperature. Hence, downscaling to 10–50 kilometres is necessary for the assessment of regional and station-scale climate information. GCM output can be downscaled via dynamical and statistical means that vary in sophistication and applicability. Based on research, more studies are commonly using statistical downscaling compared to dynamical statistic (Schoof, 2013) because statistical downscaling projections are less computationally intensive than dynamical statistic.

In statistical downscaling, relationships are established to link large- and regional-scale climate data. However, statistical downscaling presents some challenges in terms of selecting a statistical method. The nature of the predictand (local scale meteorological variables such as rainfall) influences the statistical method selection to some extent. For example, a local climate variable that approaches normal distribution, such as monthly mean temperature, will only require linear regression analysis, such as Multiple Linear Regression (MLR), for analysis. Such large-scale climate predictors tend to be normally distributed by assuming the linearity of the relationship variable (Wilby et al., 2004). However, a highly diversified and discontinuous in space and time local or non-normality variable like daily precipitation will almost certainly necessitate a more complex and difficult non-linear approach or transformation of the raw data (Bürger, 1996). Large amounts of observational data are frequently required to fit such complex models.

There are also challenges in selecting predictors in statistical downscaling. Sometimes the best predictors defined through statistical analysis of observations are insufficient for climate change applications. Geopotential heights, for example, can influence daily rainfall in extratropical areas. However, changes in geopotential heights caused by global warming will contain a non-dynamical signal that will erroneously affect rainfall estimation. This non-dynamical component should be corrected by subtracting the average changes in geopotential height over a sufficiently large area or using geopotential thickness as predictors rather than geopotential heights (Wilby et al., 2004). On the other hand, excluding key predictors change, possibly due to a high degree of covariance with another variable under current climates, could result in a

critical loss of information about various prediction responses to changes in large-scale forcing.

Because of the aforementioned downscaling issues, this research is designed to handle missing data as part of the pre-processing process. The dimensions are then reduced, and the predictors are chosen as the steps before introducing the statistical downscaling method and approaches. Kelantan and Terengganu in east coast Peninsular Malaysia is the focus of the study area because it is one of the states most affected by monsoon rains, particularly from the end of the year to the beginning of the following year. This research could help resolve problems for affected states if they recur in the future.

1.3 Study Area

Peninsular Malaysia is the country region that lies in the Equatorial zones of Northern latitude between 1 and 7° N and Eastern longitude from 100 to 103° E. The weather in Peninsular Malaysia is hot and humid all year. Generally, the climate in Malaysia is influenced by winds blowing from the Indian Ocean known as the Southwest Monsoon Wind that occur from May to September, and winds blowing from the South China Sea known as the Northeast Monsoon Wind, occurring from November to March. The intermonsoon periods, which occur in March to April and September to October, are recognized as the transition period between the two monsoons and introducing extreme convective rain to many areas of the peninsula.

The study focuses on Peninsular Malaysia's Kelantan and Terengganu states, located on the east coast of Peninsular Malaysia. The selection of both states due to their received more than the average rainfall yearly and frequency of flood events. Figure 1.1 shows the ground-based rainfall observation stations (green dot) and the stations of atmospheric (black dot). It should be noted that the positions of ground stations and atmospheric stations must match for the data collected to be synchronized. For the case study, there are ten rainfall observation stations located on the Kelantan and Terengganu states. However, nine stations were chosen based on their proximity to the nearest atmospheric station. For example, the two nearest ground stations sharing the same atmospheric station are code 8 and 59, but code 59 was chosen since it is the closest ground station to the atmospheric station. The details of the selected stations are shown in Table 1.1.

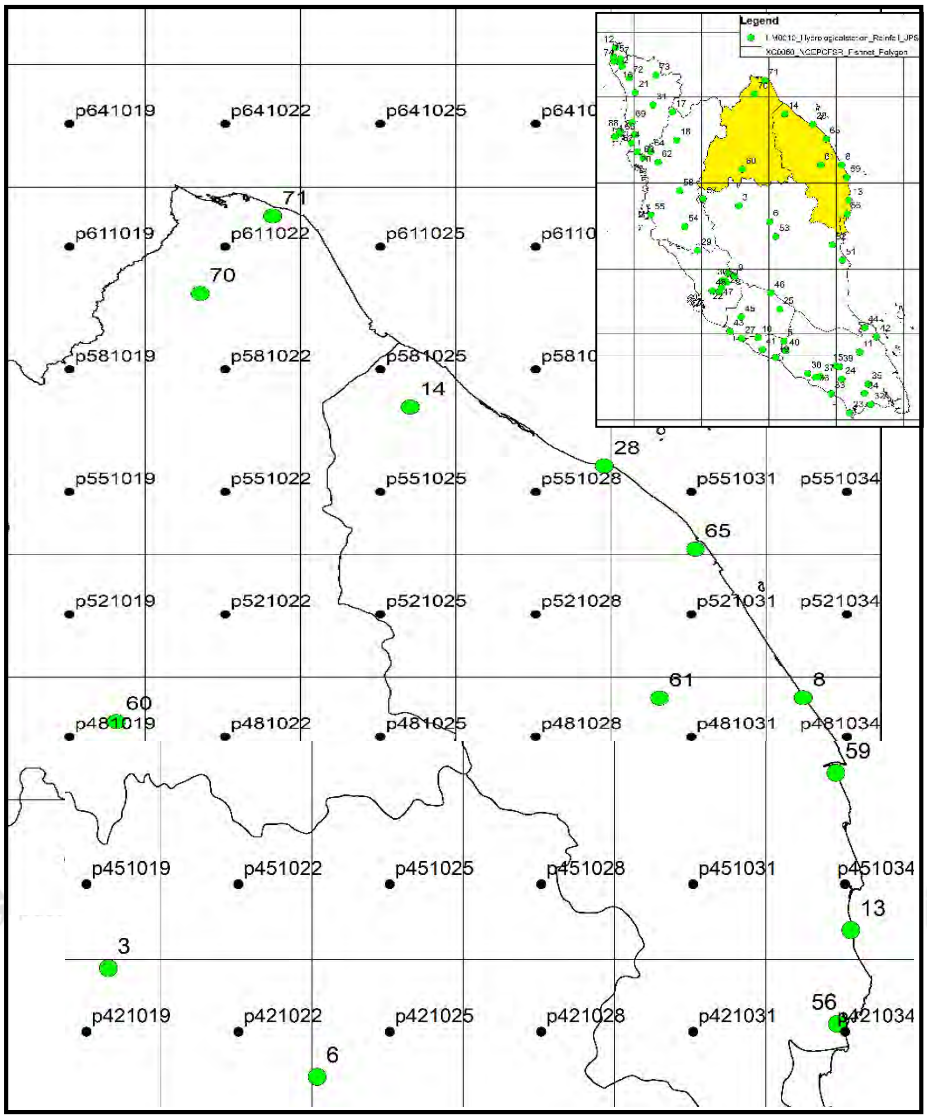

Figure 1.1. The location of rainfall and atmospheric stations in Kelantan and Terengganu states.

Table 1.1

Geographical coordinates of nine match stations in Kelantan and Terengganu states between rainfall and atmospheric stations

State	Code of atmospheric station	Code of ground station	Name of ground station	Longitude (°E)	Latitude (°N)
Kelantan	P611022	71	Kota Bharu	102.28	6.17
	P481019	60	Gua Musang	101.97	4.88

(continued)

Table 1.1 *(continued)*

Geographical coordinates of nine match stations in Kelantan and Terengganu states between rainfall and atmospheric stations

State	Code of atmospheric station	Code of ground station	Name of ground station	Longitude (°E)	Latitude (°N)
Terengganu	P581025	14	Sek. Keb, Kg Jabi	102.56	5.68
	P551031	65	Stor JPS Kuala Terengganu	103.13	5.32
	P551028	28	Kg Merang, Setiu	102.95	5.68
	P481031	61	Kg. Menerong	103.06	4.94
	P481034	59	Sek. Men. Sultan Omar, Dungun	103.42	4.76
	P421034	56	JPS Kemaman	103.42	4.23
	P451034	13	Sek. Keb. Kemasek	103.45	4.43

1.3.1 Database of predictor and predictand

The data used in this study include atmospheric data (predictors) and daily rainfall data (predictand) at selected stations stated in Table 1.1. Daily rainfall data series from 1998 to 2007 were obtained from the Department of Irrigation and Drainage (DID). The daily rainfall data was measured by using a bucket rain gauge. The bucket rain gauge is made up of a funnel that collects rainfall and channels it into a seesaw-like container (Department of Irrigation and Drainage Malaysia, 2018). Every six months, the rain gauges are calibrated, and any malfunction is repaired within seven days. The complete

rainfall data without missing values are a total of 32,869 daily measurements. A huge amount of time series data is needed to obtain an appropriate overview analysis of rainfall (Abdul Aziz, 2015).

For this study, the large-scale atmospheric variables from the National Centers of Environment Prediction (NCEP) Climate Forecast System Reanalysis (CFSR) reanalysis data set are candidates for predictors. NCEP-CFSR is the third-generation global coupling seasonal forecast reanalysis data with a spatial resolution of 0.3125° (~38km) (Zhang et al., 2020). The reanalysis data were interpolated to fit the grid size to maintain the compatibility of the downscaling model when used in projection. The list of predictors is shown in Table 1.2. A set of six large-scale atmospheric variables were converted by using Fishnet Polygon as the input format. A task of standardized data was performed because the predictor data involved different units of variables. Data standardization is the process of ensuring that data is internally consistent, meaning that each type of data has the same format (Ferguson, 2019). The predictor data employed in this study gathered a total of 27,470 days and 5.5% of missing values.

Table 1.2

The six large-scale atmospheric reanalyses of NCEP-CFSR variables

Variables	Unit
Minimum temperature	°C
Maximum temperature	°C
Precipitation	mm
Wind	knot
Relative humidity	%
Solar	W/m^2

1.3.2 Database of Machine Learning

The core of this research is a machine learning algorithm that creates a set of mathematical models of a data set with inputs and outputs. This type of machine learning includes classification and regression. The entire data set (predictors and predictands) was split into two by a 70:30 ratio in both types of machine learning, with the first period between 1998 and 2004 (about 23,008 days) for model calibration and the second period between 2005 and 2007 (9,861 days) for model validation. According to studies by Sachindra et al., (2018) and Chen et al., (2010), the split of data set by sequence is more practical when dealing with time series data. The data set was split to identify the behaviour of the machine learning model and reduce the bias in the validation model (Tokuç, 2021). However, the amount of input data in regression analysis was different because this analysis was dependent on the result class of wet days.

1.4 Problem Statement

For a few past decades, climate change and the impact of water resources have emerged as an interesting topic to the hydrologic research community. Climate change vulnerability assessments and adaptation planning necessitate local or regional climate change projection data. Because climate change impact applications are highly sensitive to local climate variation, they require information proportional to the point climate observations (Smid & Costa, 2017). In Malaysia, the impact of climate change on water resources can be seen through flooding events. The flood always happens during extreme rainfalls and is caused by the overflow of water covering land surfaces (Noor et al., 2018). However, it is also difficult to accurately simulate extreme precipitation and many statistical downscaling studies have been conducted in order to find a method for better prediction of extreme precipitation (Castellano & DeGaetano, 2017). Hence, the impact of climate change can be assessed using a statistical downscaling process, which can be useful for hydrologists or climatologists in planning future events.

In hydrological modelling, the use of long and continuous time series is always plagued by the issues of missing values. To reduce the time and effort required to overcome the missing data problem, imputation methods based on statistical models are preferred. Statistical models can fill in the data gaps in the time series measurements of various hydrologic parameters such as rainfall, maximum or minimum temperature, or water levels. In recent years, statisticians have developed imputation techniques such as regression-based imputation, mean imputation, or maximum likelihood techniques

based on the expectation-maximization (EM) algorithm and multiple imputation (MI) (Gao et al., 2018) for missing data issues. In general, the missing value imputation technique is described as replacing missing data with a reasonable estimate value, most commonly a mean value (van der Heijden et al., 2006). Different imputation methods are necessary due to the various characteristics of rainfall in terms of autocorrelation and variance. These methods offer promising solutions, but their effectiveness is dependent on the specific application and a thorough understanding of the theoretical background (Soley-bori, 2013). Hence, the selection of appropriate imputation methods should be considered a significant feature for hydrological data.

Numerous statistical downscaling models and software had been developed in the past literature. They had documented the performance comparison of downscaling approaches based on machine learning techniques (e.g., artificial neural networks (ANN), support vector machines (SVM), genetic programming (GP)), and traditional analytic methods like multiple linear regression (MLR) and autoregression. For example, Coulibaly (2004) discovered that GP-based downscaling models outperform MLR-based downscaling models in simulated daily minimum temperatures in a downscaling experiment. After that, the study by Sachindra et al. (2013) discovered that, even though both techniques over-predicted the most streamflow, Least Square Support Vector Machine (LSSVM) surpassed MLR in capturing the streamflow trend during validation. In a temperature downscaling study, Duhan and Pandey (2014) found that SVM-based downscaling models perform slightly better than ANN and MLR-based models in simulating temperature. According to the above studies, the downscaling models based on machine learning techniques outperform the traditional statistical regression models. However, there is no detailed investigation in the current

literature on the selection of appropriate machine learning approaches for developing downscaling models based on characteristics of the climate data in the study area.

Machine learning is broadly classified into two types: supervised learning and unsupervised learning. There are two techniques in supervised learning, namely classification, and regression. In recent years, the combination classification and regression methods of machine learning have received attention in various studies field. For example, in an age recognition study, van Heerden et al. (2010) used combining classification and regression methods to improve automatic speakers. The SVR was used to generate the estimation of fine age and SVC design to classify the gender of the speaker. The results showed that combination methods outperform the 7-class direct method. In a food engineering study, Qiu et al. (2014) showed that Extreme Learning Machine (ELM) outperformed other methods in classification and regression for an electronics nose data treatment. Chen et al. (2010) used a two-step statistical downscaling, namely the classification and regression method with SVM, to project daily precipitation. According to the findings, the SVM combination model produced more accurate results and showed better agreement with extreme events than multivariate analysis. Hence, researchers believe that the combination of two techniques could increase the prediction accuracies and overcome the weakness in a single model by applying the combination model as an alternative way to resolve the problem in the predicted area. The proposed predicting tool is deemed to be able to predict the climate data and simultaneously overcome the weakness of the existing single models such as analog method, fuzzy classification, and Monte Carlo methods. An effective statistical downscaling based on machine learning techniques is expected to enhance the accuracy of climate values prediction.

To identify the most suitable machine learning model, the selection of predictors (atmospheric variables) is one of the issues that must be addressed. Suitable predictors should be informative, and the relationship between the predictors and predictands (local climate variable) should be stationary. Informative predictors can be identified using dimension reduction methods, such as maximum covariance analysis (MCA), independent component analysis, and principal component analysis (PCA). Interactive model-fitting approaches are also used in predictor selection. The selection of suitable dimensional reduction methods could save time and effort in selecting and extracting relevant features analysis. Furthermore, using a dimensional reduction approach strategy to identify and extract relevant characteristics (predictors) for analysis allows for faster predictors while requiring less information from the original data. Simultaneously, employing the dimensional reduction strategy will assist in extract the small set of valuable features that describe the large dataset (Saini & Sharma, 2018). This is another advantage of using dimension reduction, as it aids in reducing the dimensions of high-dimensional data. The importance of using the dimension reduction method in high dimensional data is improving the accuracy of the classification model while lowering the cost of computational (Tanwar et al., 2018). Therefore, a precise statistical downscaling method with an inbuilt predictor selection mechanism will be helpful for researchers studying climate change impact.

1.5 Research Objectives

The objectives of this research are:

a) To investigate the most accurate imputation methods in handling the missing atmospheric data in Kelantan and Terengganu states.

b) To identify the important predictors for a statistical downscaling method by reducing the dimensionality data.

c) To propose a new framework of statistical downscaling approaches based on machine learning techniques for daily rainfall projection in Kelantan and Terengganu states.

d) To evaluate and compare the proposed downscaling statistical model in predicting the daily rainfall in Kelantan and Terengganu states.

1.6 Contribution of study

The contribution of this research are:

a) Discovering the most accurate imputation model performance for handling missing atmospheric data in Kelantan and Terengganu states.

b) Identifying important predictors for statistical downscaling methods by reducing the data dimensionality.

c) Building a new framework by combining classification and regression based on machine learning techniques as statistical downscaling approaches for climate projection rainfall in Kelantan and Terengganu states.

d) Assist in further analysis and development of an appropriate statistical downscaling method based on machine learning techniques for the prediction of rainfall over Kelantan and Terengganu states.

1.7 Significance of study

The significance of this research are:

a) By introducing the new framework of machine learning techniques, there will be an expansion of the current understanding statistical downscaling development in Kelantan and Terengganu states.

b) The development of statistical downscaling approaches will benefit the government departments like the Malaysian Meteorological Department to forecast future climate change events.

c) A detailed explanation of the development of statistical downscaling approaches involved in this research may serve as a tool for further studies to innovate the current strategies being employed in the industries related to climate change events like agriculture.

1.8 Notation

To facilitate computation, the database is configured to take the form of large rectangular rows, n by columns, p for matrix, M. In a supervised learning machine, x represents the input data vector and y or $\vec{y}$ denotes the desired output label. We use $\vec{z}$ for hidden variables and sometimes, q for hidden discrete variables. We use upper case letters to denote constants such as C, M, K, etc., and lower case as dummy indexes for appropriate range.

Throughout this thesis, we use the terms ‘rows’ of the data matrix as days, and ‘columns’ of the data matrix represent predictors and predictand variables.

1.9 Research Methodology

Briefly, this thesis focuses on the implementation of a statistical downscaling method based on machine learning techniques for projection rainfall occurrence in Kelantan and Terengganu states. This section comprises two subsections that illustrate the framework of the study by resolving the issues that arise and the flowchart of the statistical downscaling method using in this study.

1.9.1 Framework of study

The framework of the study shows the flow of study regarding achieving those objectives stated in research objectives. Firstly, handling an issue of missing values in predictors data by using the imputation methods such as K-Nearest Neighbor (KNN), Random Forest (RF), and mean substitution. The performance of the imputation model was evaluated by using statistical tests such as root mean square error (RMSE) and Nash Sutcliffe error (NSE) and compared to determine the best imputation approach for predictors data of the Kelantan and Terengganu states. Then, PCA was used as a dimensional reduction approach to assist in the selection of predictors for the statistical downscaling model. Following that, the statistical downscaling approach was introduced by applying the supervised machine learning techniques (classification and regression) to identify the daily climate projections of rainfall. Lastly, the effectiveness of the statistical downscaling approaches based on supervised machine learning models is measured and compared by using a statistical test. The structure of the study framework is shown in Figure 1.2.

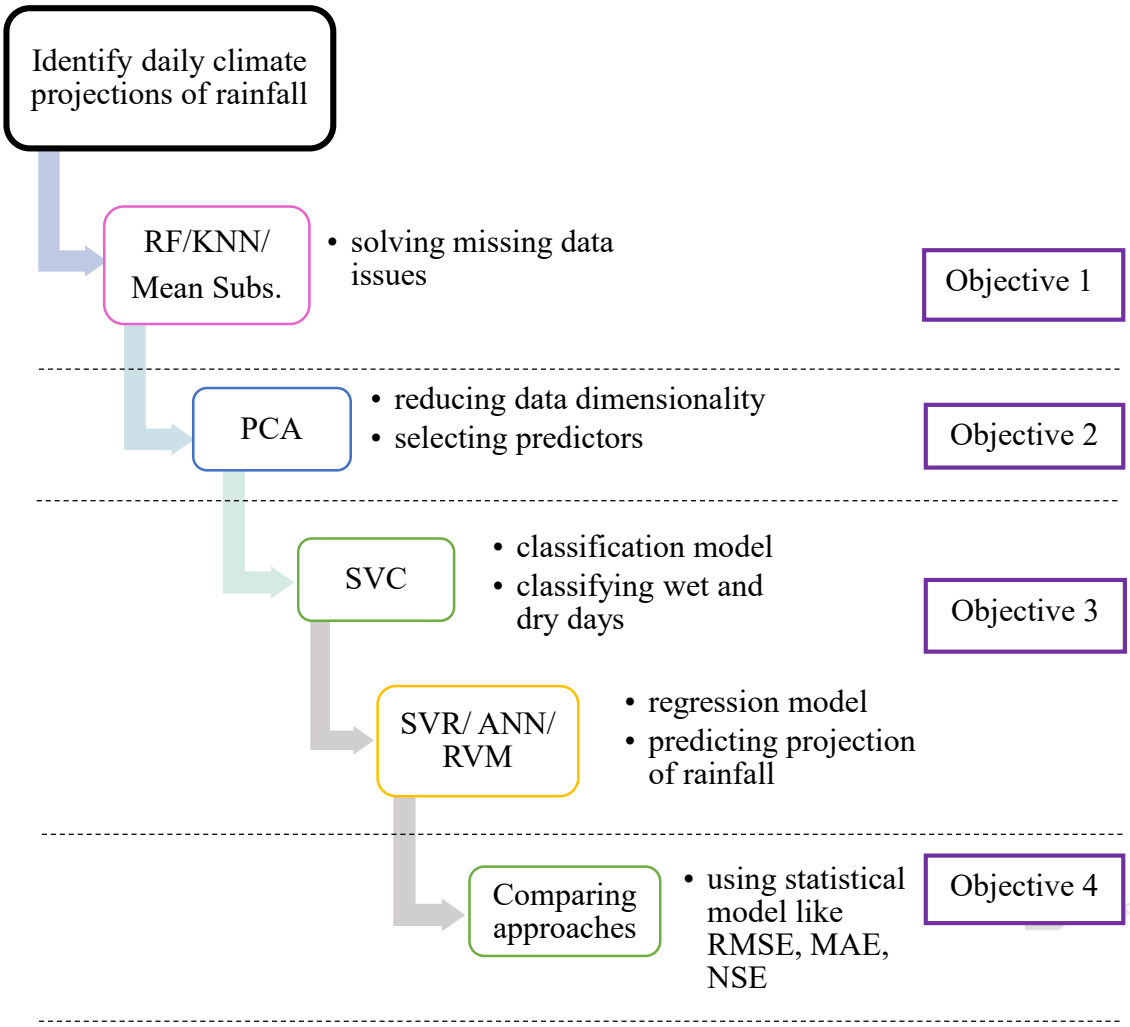

Figure 1.2. Framework of study

1.9.2 Flowchart of statistical downscaling method

The proposed statistical downscaling method for daily rainfall consists of both classification and regression models, which were developed using machine learning techniques described in this section. The proposed model consists of data collection from predictors (atmospheric data) and predictand (rainfall data). As a pre-processing phase, the predictors were standardized and replenished by using the imputation method. Then, PCA was used to select predictor variables while also reducing the data's dimensionality. The new data matrices between the selected predictors and predictand were then created. Following that, a classification model was used to classify the days as dry or wet, in which a day with the amount of rainfall exceeding 1mm is considered a wet day (Shaharudin, et al., 2018). If the day is classified as wet days, suitable regression techniques are used to predict the rainfall values. The proposed statistical downscaling model is shown in Figure 1.3.

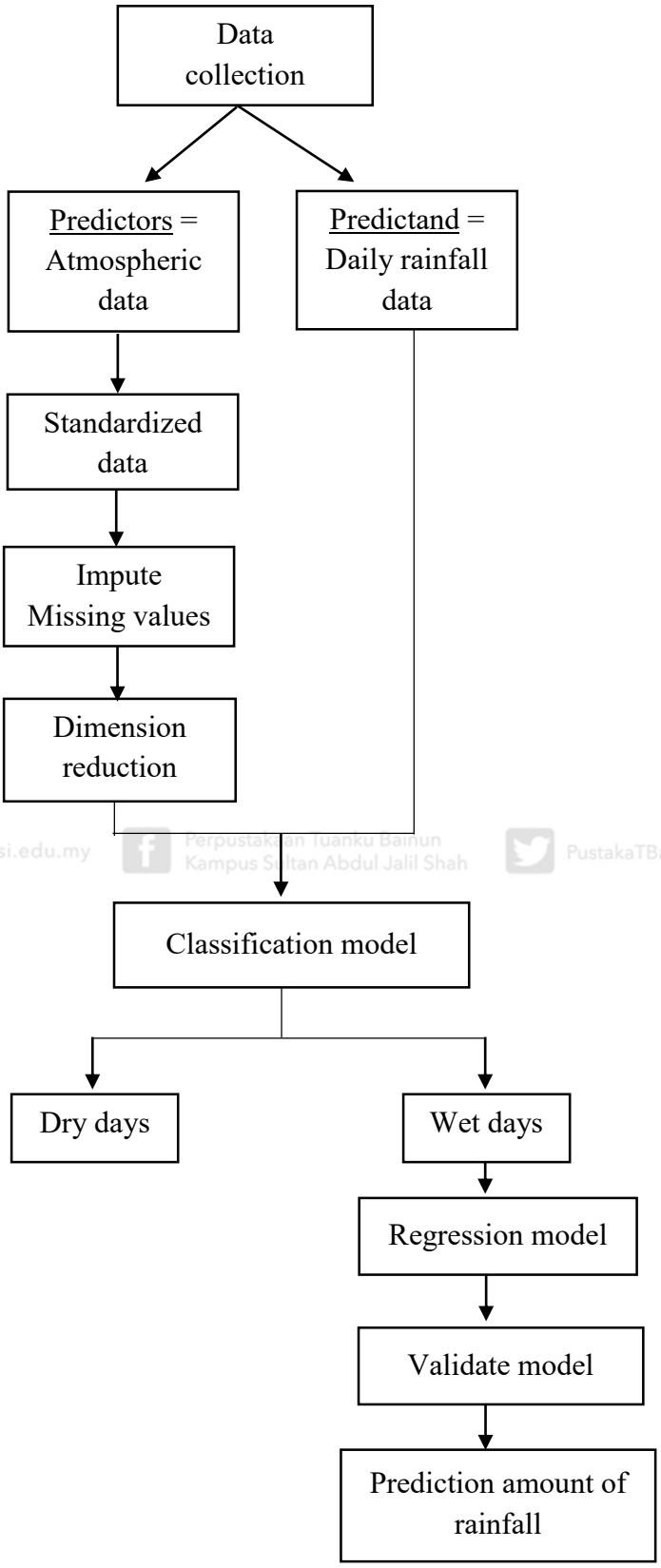

Figure 1.3. Flowchart of statistical downscaling method

1.10 Limitation of Study

In this study, the suggested methodologies are only explained using a series of predictors and predictand variables in Kelantan and Terengganu states. Hence, this study does not focus on all the rainfall stations in Peninsular Malaysia. The involved ground stations in Kelantan and Terengganu states are not updated because they consist of unavailable atmospheric data for the most recent ground stations. As for the available atmospheric data, more variables cannot be ruled out, partly due to the server used for the high-dimensional data processing, especially rainfall data. When dealing with high-dimensional data, the common problems include managing the lengthy data and the time-consuming processing. Hence, a more powerful and higher speed of server would be needed. As widely known, machine learning techniques have specific multi-algorithms for analysis, which requires a large amount of storage space and iteration time for long data. Essentially, studies on machine learning in time series data necessitate a high computationally demand for accurate analysis.

1.11 Thesis Organization

Chapter 1 introduces the background of the problem and related issues that arise from the implementation of statistical downscaling approaches based on machine learning techniques in Kelantan and Terengganu states. It also includes a description of the research area and defines the specific dataset used in this study. Following that, this chapter explains the objectives of the study, the significance of the study, notation, and research methodology. In Chapter 2, the existing works related to the methods or

approaches used in this study are reviewed. The chapter begins with an overview of missing data mechanisms and patterns, as well as related studies using the imputation model. Then, the next section describes important works of literature related to the dimensional reduction approach. Additionally, the advantages, drawbacks, and application of using machine learning techniques as a statistical downscaling method are also described in this section. Essentially, prior to that, there is an overview of the downscaling.

A novel statistical downscaling approach to identify the daily climate rainfall projection in Kelantan and Terengganu states is discussed in Chapter 3. This chapter clearly explains the algorithm and procedure of related statistical models used in this study - starting with imputation models, followed by dimensional reduction model, and supervised machine learning model. Subsequently, Chapter 4 presents and discusses the result of imputation models, dimensional reduction models, and machine learning models. In this chapter, we also compare the performance of different imputation models as well as the different machine learning models' efficiency as a statistical downscaling approach by using a statistical test. Finally, Chapter 5 summarizes the findings and discussions of all the problems investigated in this study. This chapter also discusses possible future research to gain a better understanding of the problems considered.