INTEGRATING A HYBRID STATISTICAL DOWNSCALING-BASED HMM-RF MODEL FOR ENHANCED RAINFALL PREDICTION IN SELANGOR

NOOR HAMIZAH BINTI MOHAMAD SANI

UNIVERSITI PENDIDIKAN SULTAN IDRIS

2025

INTEGRATING A HYBRID STATISTICAL DOWNSCALING-BASED HMM-RF MODEL FOR ENHANCED RAINFALL PREDICTION IN SELANGOR

NOOR HAMIZAH BINTI MOHAMAD SANI

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

FACULTY OF SCIENCE AND MATHEMATICS
UNIVERSITI PENDIDIKAN SULTAN IDRIS

2025

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 ....9th.....day of.....September.....2025.....

i. Student's Declaration:

I, NOOR HAMIZAH BINTI MOHAMAD SANI, M20221001855, FACULTY OF SCIENCE AND MATHEMATICS (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled INTEGRATING A HYBRID STATISTICAL DOWNSCALING-BASED HMM-RF MODEL FOR ENHANCED RAINFALL PREDICTION IN SELANGOR 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 PROF. MADYA DR. SHAZLYN MILLEANA BINTI SHAHARUDIN (SUPERVISOR'S NAME) hereby certifies that the work entitled INTEGRATING A HYBRID STATISTICAL DOWNSCALING-BASED HMM-RF MODEL FOR ENHANCED RAINFALL PREDICTION IN SELANGOR (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 OF SCIENCE (STATISTICS) (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

09/09/2025
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*: INTEGRATING A HYBRID STATISTICAL DOWNSCALING-BASED HMM-RF MODEL FOR ENHANCED RAINFALL PREDICTION IN SELANGOR

No. Matrik /*Matric's No.*: M20221001855

Saya / *I* : NOOR HAMIZAH BINTI MOHAMAD SANI

(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 restircted information as specified by the organization where research was done.*

☑ **TIDAK TERHAD / *OPEN ACCESS***

Hamizah

(Tandatangan Pelajar/ Signature)

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

Tarikh: 09/09/2025

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 and foremost, I am deeply thankful to Allah for the strength and blessings I have received throughout this journey. I would like to extend my heartfelt thanks to my parents and siblings, whose unwavering support has been the cornerstone of my journey. Their endless encouragement, both through their steadfast prayers and generous financial assistance, has provided me with the strength and stability needed to pursue this research. Without their love and belief in me, this accomplishment would not have been possible. My deepest gratitude also goes to my supervisor, Prof. Madya Ts. Dr. Shazlyn Milleana binti Shaharudin, whose dedication and expertise have been truly inspiring. The knowledge shared, the time invested, and the energy contributed to guiding me through this study have been invaluable. Despite my challenges as a slow learner, her patience and commitment have been a beacon of support, allowing me to navigate the complexities of this research and reach this significant milestone. I am also profoundly grateful to my friends, both online and real-life friends, who have stood by me through every phase of this endeavour. Their unwavering support, encouragement, and belief in my abilities have provided me with the motivation needed to push through even the toughest moments. Their companionship and encouragement have made this journey not just bearable, but also memorable. Finally, I want to acknowledge myself for the perseverance and determination I have demonstrated throughout this study. The path was not always smooth, but my commitment to seeing this research through to completion has been a testament to my resolve. The sense of accomplishment that comes from finishing this study is as much a reflection of my own strength and dedication as it is of the support I have received from others.

ABSTRACT

Floods in 2022 caused significant economic losses in Malaysia, totaling RM 6.1 billion ($1.46 billion). Selangor, a densely populated and industrialized state, was the hardest hit, with the manufacturing sector suffering RM 900 million in losses. Accurate rainfall prediction is essential for effective weather forecasting and climate modeling in the region. This study evaluates the effectiveness of a hybrid Statistical Downscaling-based Hidden Markov Model-Machine Learning Model (SD-based HMM-ML) for rainfall prediction. It also explores optimal imputation methods for handling missing data, selects predictors using dimensionality reduction, and addresses uncertainties in zero-bounded rainfall data. Local rainfall (predictand) and atmospheric data (predictor) from 33 stations in Selangor (2008–2018) were analyzed. Seven imputation methods were tested: Mean, Median, Expectation-Maximization (EM), Markov Chain Monte Carlo (MCMC), k-Nearest Neighbor (kNN), Non-iterative Partial Least Square (NIPALS), and Random Forest (RF). Principal Component Analysis (PCA) reduced high-dimensional data, selecting five principal components with a high cumulative variance. HMM addressed zero-bounded rainfall uncertainties, identifying three hidden states with the lowest Bayesian Information Criterion (BIC). Five hybrid models – HMM-RF, HMM-SVM, HMM-DT, HMM-KNN, and HMM-ANN – were evaluated using RMSE, MAE, MFE, and NSE. Median Imputation had the lowest values of RMSE and MAE, and highest value of NSE across all stations. HMM-RF outperformed other models, demonstrating superior accuracy. By leveraging machine learning, this novel hybrid model enhances rainfall prediction accuracy, improving early warning systems, infrastructure planning, and flood mitigation efforts. The study contributes to building a more resilient urban environment, mitigating economic losses from future floods in Selangor.

MENGINTEGRASIKAN MODEL HIBRID BERASASKAN PENGKECILAN SKALA STATISTIK HMM-RF UNTUK PENINGKATAN RAMALAN HUJAN DI SELANGOR

ABSTRAK

Banjir pada tahun 2022 telah menyebabkan kerugian ekonomi yang besar di Malaysia, berjumlah RM 6.1 bilion ($1.46 bilion). Selangor, sebuah negeri yang padat dengan penduduk serta perindustrian yang pesat, merupakan kawasan yang paling teruk terjejas, dengan sektor pembuatan mengalami kerugian sebanyak RM 900 juta. Ramalan hujan yang tepat adalah penting untuk peramalan cuaca yang berkesan serta pemodelan iklim di rantau ini. Kajian ini menilai keberkesanan model hibrid berasaskan Penskalaan Bawah Statistik dan Model Markov Tersembunyi-Pembelajaran Mesin (SD-based HMM-ML) bagi ramalan hujan. Ia juga meneroka kaedah pengisian yang optimum bagi mengendalikan data yang hilang, memilih peramal menggunakan pengurangan dimensi, dan menangani ketidakpastian dalam data hujan yang terikat sifar. Data hujan tempatan (prediktan) dan data atmosfera (prediktor) dari 33 stesen di Selangor (2008–2018) dianalisis. Tujuh kaedah pengisian diuji: Min, Median, Jangkaan-Maksimum (EM), Rantaian Markov Monte Carlo (MCMC), k-Tetangga Terdekat (kNN), Kuadrat Terkecil Separa Tidak Berulang (NIPALS), dan Hutan Rawak (RF). Analisis Komponen Utama (PCA) digunakan untuk mengurangkan dimensi data yang tinggi, memilih lima komponen utama dengan jumlah varians terkumpul adalah yang tertinggi. HMM menangani ketidakpastian dalam data hujan terhad sifar, mengenal pasti tiga keadaan tersembunyi dengan nilai Kriteria Maklumat Bayesian (BIC) terendah. Lima model hibrid – HMM-RF, HMM-SVM, HMM-DT, HMM-KNN, dan HMM-ANN – dinilai menggunakan RMSE, MAE, MFE, dan NSE. Pengisian Median mempunyai nilai RMSE dan MAE yang paling rendah serta nilai NSE yang paling tinggi di semua stesen. HMM-RF mengatasi model lain, menunjukkan ketepatan yang lebih tinggi. Dengan memanfaatkan pembelajaran mesin, model hibrid baharu ini meningkatkan ketepatan ramalan hujan, memperbaiki sistem amaran awal, perancangan infrastruktur, dan usaha mitigasi banjir. Kajian ini menyumbang kepada pembinaan persekitaran bandar yang lebih berdaya tahan, mengurangkan kerugian ekonomi akibat banjir pada masa hadapan di Selangor.

CONTENTS

		Page
DECLARATION OF ORIGINAL WORK		ii
DECLARATION OF DISSERTATION		iii
ACKNOWLEDGMENT		iv
ABSTRACT		v
ABSTRAK		vi
CONTENTS		vii
LIST OF TABLES		xi
LIST OF FIGURES		xii
LIST OF ABBREVIATIONS		xiv
CHAPTER 1 INTRODUCTION		
1.1	Overview	1
1.2	Background of Study	2
1.3	Study Area	6
1.3.1	Descriptive Statistics of Rainfall Data	10
1.4	Problem Statement	16
1.5	Research Objectives	18
1.6	Contribution of Study	19
1.7	Significance of Study	20

1.8 Framework of Study 21

1.9 Limitation of Study 22

1.10 Summary 23

CHAPTER 2 LITERATURE REVIEW

2.1 Overview 25

2.2 Introduction to Statistical Downscaling-based Machine Learning Model 26

2.2.1 Application of Rainfall Prediction using Statistical Downscaling-based Hybrid Machine Learning Model 35

2.3 Missing Values 45

2.4 High-Dimensionality Data 56

2.5 Zero-Inflated Data 65

2.6 Non-Linearity Data 72

2.7 Model Performance Measure 75

2.8 Conclusions 77

CHAPTER 3 RESEARCH METHODOLOGY

3.1 Introduction 98

3.2 Imputation Methods 99

3.2.1 Mean Imputation and Single Value Imputation 99

3.2.2 *k*-Nearest Neighbour 100

3.2.3 Random Forest 102

3.2.4 EM Algorithm 103

3.2.5 Markov Chain Monte Carlo (MCMC) 107

3.2.6 NIPALS 112

3.3 Principal Component Analysis 115

3.4 The Homogeneous Markov Model (HMM) 117

3.5 Machine Learning 123

3.5.1 Random Forest 123

3.5.2 Support Vector Machine 128

3.5.3 Artificial Neural Networks 132

3.5.4 Decision Tree 135

3.5.5 K-Nearest Neighbors 137

3.6 Disadvantages of Using Single Bayesian Approach and Machine Learning Model for Prediction Modeling 138

3.7 Prediction Modeling of Hybrid Bayesian Approach and Statistical Downscaling-based Machine Learning 140

3.8 Model Performance Metrics 143

3.9 Research Methodology 144

CHAPTER 4 RESULT AND DISCUSSION

4.1 Overview 145

4.2 Handling Missing Value 146

4.3 Reduction of Dimensional Data 149

4.4 The Homogeneous Hidden Markov Model (HMM) 152

4.5 A Hybrid of HMM-Statistical Downscaling-based Machine Learning Model 159

4.5.1 Prediction Modeling using an HMM-RF model 160

4.5.2 Prediction Modeling using an HMM-SVM model 163

4.5.3 Prediction Modeling using an HMM-DT model 167

4.5.4 Prediction Modeling using an HMM-KNN model 172

4.5.5 Prediction Modeling using an HMM-ANN model 176

4.6 Evaluating Performance of HMM – Statistical Downscaling-based Machine Learning Model 179

4.7 Summary 183

CHAPTER 5 CONCLUSION

5.1 Overview 185

5.2 Summary 185

5.3 Future Research 189

REFERENCES 191

APPENDICES 206

LIST OF TABLES

Table No.		Page
1.1	Rainfall stations in Selangor with geographical coordinates	8
1.2	List of predictors	10
1.3	Summary statistics of daily rainfall data amount for each station	11
2.1	Summary of methodology and contributions of previous studies	79
4.1	Performance values for each imputation methods	148
4.2	Result of Principal Components (PCs)	149
4.3	Result of factor loadings for each predictor	151
4.4	Comparison number of iteration for each hidden state layer based on BIC	153
4.5	The performance values of SD-based HMM-RF model	162
4.6	SVM Parameters	164
4.7	The performance values of SD-based HMM-SVM model	165
4.8	The performance values of SD-based HMM-DT model	171
4.9	Number of neighbors used in KNN	173
4.10	The performance values of SD-based HMM-KNN model	175
4.11	The performance values of SD-based HMM-ANN model	178
4.12	The overall results of the performance of the five hybrid models	180

LIST OF FIGURES

No.	Figures	Page
1.1	The location of rainfall and atmospheric stations in Selangor state	7
1.2	Framework of the study	23
2.1	Missing data pattern	46
2.2	Algorithm for handling missing data	48
3.1	Steps involve in kNN imputation method	101
3.2	Steps involve in RF imputation method	104
3.3	Steps involve in EM Algorithm	107
3.4	Steps involve in MCMC	111
3.5	Steps involve in NIPALS	114
3.6	Procedure of PCA model	117
3.7	Steps involve in HMM	122
3.8	A workflow of an ML framework	124
3.9	Steps involve in RF	128
3.10	A schematic representation of SVM	129
3.11	Steps involve in SVM	131
3.12	ANN's layer	133
3.13	The steps involve in ANN	135
3.14	Steps involve in DT	136
3.15	Steps involve in KNN	138
3.16	The flowchart of the study	142

4.1	The visualization of missing data	147
4.2	Comparison number of iterations with the value of BIC	155
4.3	Trace and Density plot	156
4.4	The distribution of rainfall values for both the original dataset and SD-based HMM-RF model prediction	162
4.5	The performance of SD-based HMM-RF model in predicting rainfall amounts in calibration period	163
4.6	The distribution of rainfall values for both the original dataset and SD-based HMM-SVM model prediction	165
4.7	The performance of SD-based HMM-SVM model in predicting rainfall amounts in calibration period	166
4.8	The Residuals vs. Fitted Plot	168
4.9	The Q-Q Residuals	168
4.10	The Scale-Location plot	169
4.11	The Residuals vs. Leverage plot	169
4.12	The distribution of rainfall values for both the original dataset and SD-based HMM-DT model prediction	171
4.13	The performance of SD-based HMM-DT model in predicting rainfall amounts in calibration period	172
4.14	RMSE vs. Number of neighbors plot	173
4.15	The distribution of rainfall values for both the original dataset and SD-based HMM-KNN model prediction	175
4.16	The performance of SD-based HMM-KNN model in predicting rainfall amounts in calibration period	176
4.17	Neural Network Diagram	177
4.18	The distribution of rainfall values for both the original dataset and SD-based HMM-ANN model prediction	178
4.19	The performance of SD-based HMM-ANN model in predicting rainfall amounts in calibration period	179
4.20	The short forecast for two weeks of five hybrid models in predicting daily rainfall amount	182

LIST OF ABBREVIATIONS

2PPCA-EM	2-Component Probabilistic Principal Component Analysis-Expectation Maximization
AI	Artificial Intelligent
ANN	Artificial Neural Network
CanESM2	Canadian GCM model
CNN	Convolutional Neural Network
CNN-DeepESD	Convolutional Neural Network - Deep Earthquake Signal Detection
CNN-PAN	Convolutional Neural Network - Parallel Attention Network
CNN-UNET	Convolutional Neural Network - U-Net
CONN-SVM-GRP	Convolutional Neural Network-Support Vector Machine-Gaussian Regression Process
COVID-19	Coronavirus Disease 2019
DT	Decision Tree
ECMWF	European Centre for Medium-Range Weather Forecasts
EM	Expectation-Maximization
FFANN	Function Fitting Artificial Neural Network
FIML	Full Information Maximum Likelihood
GC	Gaussian Copula
GCM	General Circulation Method
HadCM3	Hadley Centre Coupled Model, version 3

HD	Hot-Deck
HMM	Homogeneous Markov Model
IDW	Inverse Distance Weighting
IG-PCA	Information Gain-Principal Component Analysis
KDD99	Knowledge Discovery in Databases 1999
*k*NN	*k*-Nearest Neighbour
LI	Linear Interpolation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MFE	Mean Forecast Error
MFGWML	Multi Factor Weighted Machine Learning
MICE	Multiple Imputation by Chain Equation
ML	Machine Learning
MLR	Multiple Linear Regression
MPPCA-EM	M-Component Probabilistic Principal Component Analysis-Expectation Maximization
NCEP-CFSR	National Centers of Environment Prediction – Climate Forest System Reanalysis
NIPALS	Non-iterative Partial Least Square
NMAR	Not Missing at Random
NSE	Nash–Sutcliffe model efficiency
PC	Principal Component
PCA	Principal Component Analysis

PPCA	Probabilistic Component Analysis
RF	Random Forest
RMSE	Root Mean Square Error
RVM	Relevance Vector Machine
RVM-GWO	Relevance Vector Machine-Grey Wolf Optimization
RVM-IMRFO	Relevance Vector Machine-Improved Manta-Ray Foraging Optimization
RVM-WOA	Relevance Vector Machine-Whale Optimization Algorithm
SD-based HMM-ANN	Statistical Downscaling-based Homogeneous Markov Model – Artificial Neural Network
SD-based HMM-DT	Statistical Downscaling-based Homogeneous Markov Model – Decision Tree
SD-based HMM-KNN	Statistical Downscaling-based Homogeneous Markov Model – K-Nearest Neighbors
SD-based HMM-ML	Statistical Downscaling-based Hidden Markov Model-Machine Learning
SD-based HMM-RF	Statistical Downscaling-based Homogeneous Markov Model – Random Forest
SD-based HMM-SVM	Statistical Downscaling-based Homogeneous Markov Model – Support Vector Machine
SDSM	Statistical Downscaling Method
SVM	Support Vector Machine
TsHARP	Thermal Image Sharpening
XGBoost	eXtreme Gradient Boosting

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter begins by laying a comprehensive foundation for the study, starting with an exploration of the background and context that underscores the importance of the research, followed by an overview of the specific study area and its relevance. We will then delve into the problem statement, clearly defining the key challenges this study seeks to address, which naturally leads into a discussion of the research objectives that guide the investigation. The chapter also highlights the contributions of the study, detailing how it advances existing knowledge and offers new insights, along with its broader significance and potential impact on both the field and practical applications. Additionally, the methodological framework of the study is outlined, providing a structured approach to achieving the research goals. Finally, we acknowledge the limitations of the study, discussing the constraints and challenges that may influence the findings, thereby offering a balanced perspective on the scope and potential outcomes of the research. This comprehensive introduction sets the stage for the detailed analysis and discussions that will follow in subsequent chapters.

1.2 Background of Study

Rainfall is one of the most impactful meteorological factors influencing various aspects of human daily lives (Barrera-Animas et al., 2022). This is because it has been crucial in the development and sustenance of human civilizations (Kanani et al., 2023). All water consumers will be impacted by altered patterns of rainfall and can trigger a range of natural disasters such as floods and that are sure to harm human (Usman et al., 2023).

Extreme weather events like floods and tsunamis will occur due to climate change's effects on the hydrological cycle. Sulaiman et al. (2022) stated that the most devastating climate change impact experienced in Malaysia is flooding. The main cause of flooding is heavy rainfall for days nonstop. The effect of floods on people is it can disrupt daily activities, and this effect can last for weeks. Thus, an accurate rainfall prediction model is essential to reduce the risk of human life (Prottasha et al., 2023), and enhance disaster preparedness by providing early warnings for extreme weather events. Due to the increasing variability and unpredictability of weather patterns caused by climate change, the demand for accurate and reliable rainfall predictions has grown, driving advancements in predictive modeling techniques (Sani et al., 2020).

One of the critical challenges in rainfall prediction is dealing with missing data (Chiu et al., 2021). Missing values in a dataset can greatly increase computational costs and distort the results (Khan & Hoque, 2020). According to Addi et al. (2021), missing daily rainfall data can lead to substantial errors and biases in analysis, resulting in heightened uncertainty in water resources assessments. Therefore, employing an

imputation method is crucial for addressing the issue of missing rainfall data. Without proper imputation techniques, the integrity of the dataset is compromised, which can lead to inaccurate analyses and unreliable predictions. Imputation methods help fill in the gaps in the data, allowing for more robust and consistent results in rainfall prediction models (Aieb et al., 2019). These methods are vital for maintaining the quality and completeness of the dataset, thereby ensuring that subsequent analyses and forecasts are based on accurate and comprehensive information.

Rainfall prediction models must process big amounts of atmospheric data, including temperature, humidity, wind speed, and precipitation, which are often high-dimensional. High-dimensional data can cause the difficulty of knowledge discovery and pattern classification due to the presence of numerous redundant and irrelevant features (Zebari et al., 2020). Dimensionality reduction is a method that can reduce the high-dimensional data to low-dimensional data (Hasan & Abdulazeez, 2021). Over the past few decades, a variety of dimensionality reduction methods have been employed to filter data samples within the dataset being considered (Reddy et al., 2020). Principal Component Analysis (PCA) is a method that is widely employed as a dimensionality reduction approach because it is relatively cost-efficient in terms of computation and capable in handling big data (Abdulhammed et al., 2019). Hence, applying a dimensional reduction approach can help in reducing computational complexity without losing critical predictive features.

Rainfall is crucial for fitting models based on probability distribution functions, which estimate variability, but in many practical situations, especially in low rainfall regions, large datasets often contain numerous zero rainfall values (Gramosa et al.,

2019; Zamani & Bazrafshan, 2020). Perumean-Chaney et al. (2012) concluded that ignoring zeros in the data led to poor estimation and the omission of statistically significant findings. Therefore, a specialized model that accounts for zero-bounded data is essential for generating realistic predictions, thereby helping to manage and reduce parameter uncertainties. The Bayesian method can accurately capture parameter uncertainties, making them particularly valuable in complex modeling scenarios such as rainfall prediction (Cao et al., 2023). This enables more robust decision-making by taking into account the range of potential outcomes and their associated probabilities (Bharadiya, 2023).

Non-linearity in rainfall data poses significant challenges for statistical models, which often assume linear relationships between variables (Zhang et al., 2023). This limitation is particularly evident in statistical downscaling models used for climate projections, where the complex, non-linear interactions between large-scale atmospheric variables and local rainfall patterns need to be accurately captured. Statistical downscaling methods have been applied to overcome the scaling gap between coarser-resolution General Circulation Models (GCM) outputs and regional-scale climatic variables (Maqsood et al., 2023). Machine learning-based statistical downscaling has the ability to overcome complex and non-linear relationships in the data, leading to more accurate downscaling results (Putri et al., 2021). By integrating hybrid Bayesian approaches with machine learning techniques, these models can better handle non-linear relationships. Bayesian methods provide a robust framework for incorporating prior knowledge and quantifying uncertainty, while machine learning techniques excel at modeling complex, non-linear interactions. Together, they enhance the predictive power and accuracy of statistical downscaling models, making them

more adept at forecasting rainfall under the influence of non-linear dynamics. This integrated approach addresses the inadequacies of the models, offering a more sophisticated solution for rainfall prediction in the face of climate variability.

Despite significant advancements, there remains a gap in integrating Bayesian methods with machine learning models for rainfall prediction. Current research has primarily focused on either improving Bayesian techniques or enhancing machine learning algorithms independently. Few studies have explored the synergy of combining these approaches to address the multifaceted challenges of rainfall prediction. This study aims to fill this gap by developing a new framework that integrates hybrid Bayesian and machine learning models based statistical downscaling to enhance the accuracy and reliability of rainfall predictions.

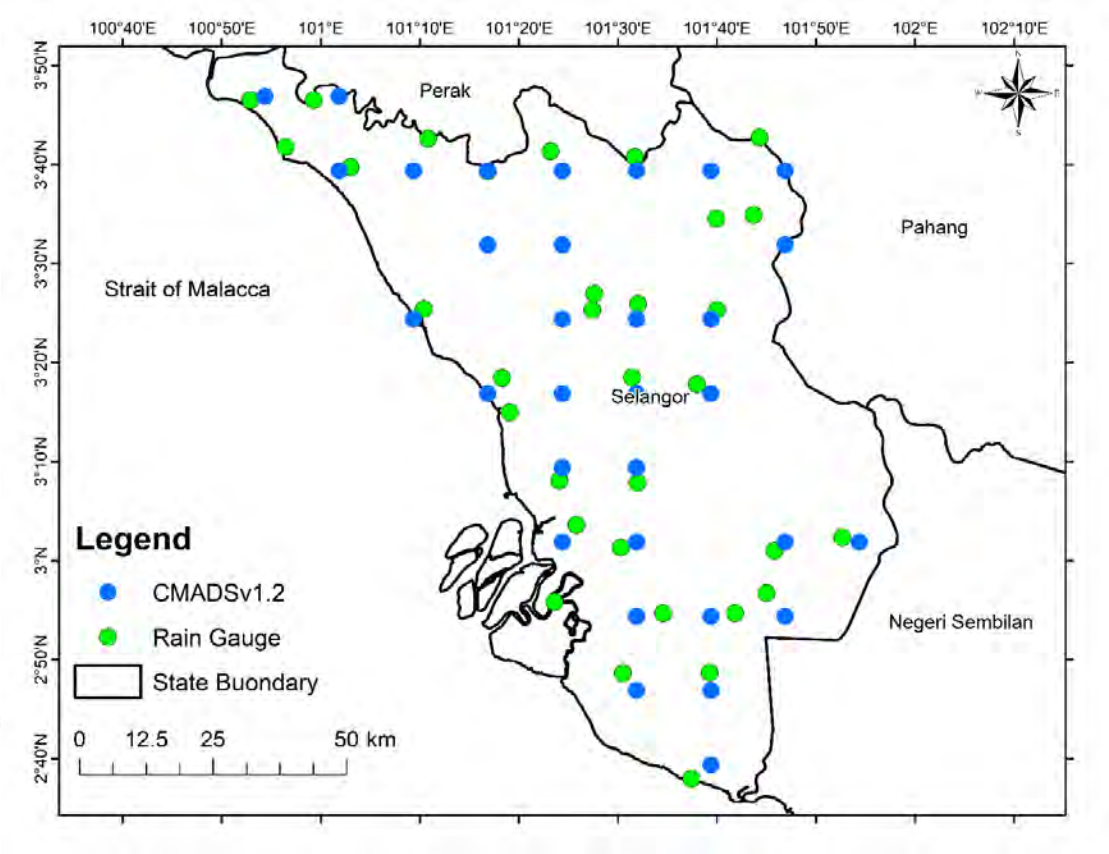
The development of an integrated framework combining Bayesian and machine learning models with statistical downscaling represents a significant advancement in rainfall prediction. This approach has the potential to address the critical challenges of missing data, high-dimensional inputs, zero-bounded constraints, and non-linear relationships in a unified manner. Improved rainfall predictions can lead to better-informed decision-making in agriculture, water management, and disaster preparedness, ultimately contributing to societal resilience in the face of climate variability and change.

1.3 Study Area

The study focuses on Peninsular Malaysia's Selangor state, located on the south coast of Peninsular Malaysia. Selangor's economy is outpacing that of Malaysia's other 13 states, making it one of the most sought-after places to live due to its strong economy, employment prospects, good facilities, and superb infrastructure (Husin et. al., 2020). However, floods have been occurring in Selangor, particularly in the metropolitan area, resulting in financial losses and property destruction (Muhammad & Shaidin, 2022). The selection of this state is because of the extreme floods (Muhammad & Shaidin, 2022) and continuous rain (Zahari, Zainol & Ismail, 2022) that occurred rapidly and simultaneously at numerous places in Selangor. Based on Figure 1.1, the ground-based rainfall observation stations (green dots) serve as predictands, while the atmospheric stations (blue dots) act as predictors. The predictand is the variable or outcome being predicted, whereas the predictors are the variables used for making that prediction. It is important to note that for data synchronization, the positions of the atmospheric and ground stations must align.

Figure 1.1

The location of rainfall and atmospheric stations in Selangor state



From Department of Irrigation and Drainage (DID)

Daily rainfall data (predictand) and atmospheric data (predictors) are the data used in this research. Daily rainfall data series were obtained from 2008 to 2018 from the Department of Irrigation and Drainage (DID) involving 33 stations in Selangor. These stations from each region were chosen based on the data's completeness and the duration of the records. The specifics of rainfall stations in Selangor are outlined in Table 1.1. Meanwhile, large-scale atmospheric data was obtained from the National Centers of Environment Prediction (NCEP) – Climate Forecast System Reanalysis (CFSR).

Table 1.1

Rainfall stations in Selangor with geographical coordinates

Code of Rainfall Station	Name of Rainfall Station	Longitude (°)	Latitude (°)
22-334	P/A Sg. Pelek, Sepang	02 37 58.9	02 37 58.9
23-333	P/A Pekan Banting	02 48 38	02 48 38
23-334	Rtb Bukit Changgang	02 48 41	02 48 41
24-332	P/Kwln P/S Telok Gong	02 55 50	02 55 50
24-333	Ldg. Bkt. Cheeding	02 54 40	02 54 40
24-334	Puncak Niaga	02 54 40	02 54 40
24-335	Klm Takungan S. Merab	02 56 43	02 56 43
25-332	Sg. Udang	03 03 37	01 35 77
25-333	Kg. Jawa	03 01 18	03 01 18
25-335	Sg. Balak H. Langat	03 00 59	03 00 59
25-336	Pejabat Jps. Klang	03 02 20	03 02 20
26-332	Ldg. Harpenden	03 08 05	03 08 05
26-333	Rumah Pam R. Panjang	03 07 51	01 86 94
27-331	Ldg. Kuala Selangor	03 18 28	03 18 28
27-332	Ldg. Braunston	03 14 58.5	03 14 58.5
27-333	Taman Desa Kundang	03 18 30	01 10 35
27-334	Taman Templer	03 17 49	01 79 74
28-330	Stor Jps. Tg. Karang	03 25 25	03 25 25
28-332	Pengorekan Bijih Berju	03 25 20	03 25 20
28-333	Kg. Sungai Buaya	03 25 56	03 25 56
28-334	Taman Desa Kelisa	03 25 20	04 39 15
29-331	Sg. Burung	03 41 46.6	03 41 46.6
29-332	Ldg. Hopeful	03 26 59	02 16 35
29-335	Kampung Pertak	03 34 54	03 34 54
30-329	S.R.K Parit 4 Sg. Haji Dorani	03 39 45	03 39 45
30-330	Kg. Belia Sg. Panjang	03 42 37	03 42 37

Code of Rainfall Station	Name of Rainfall Station	Longitude (°)	Latitude (°)
30-331	Ldg. Pkps Sg. Panjang	03 39 19	03 39 19
30-332	Felda Soeharto	03 40 81	03 40 81
30-333	Kom. P'hulu Ulu Bernam	03 40 46	03 40 46
30-334	Loji Air Kuala Kubu Bahru	03 34 33	03 34 33
30-335	Bukit Fraser	03 42 43	03 42 43
31-328	Bagan Nakhoda Omar	03 46 30	03 46 30
31-329	Pekan Sabak Bernam	03 46 30	03 46 30

Table 1.2 shows a list of the predictors, and due to the diverse units of variables, a standardization task was undertaken. Data standardization is a process of converting data from various sources or units into a consistent format that complies with the standard (Egnyte, 2023). The predictor data employed in this study gathered a total of 132,594 days (from 2008 to 2018) with 2.13% of missing values. According to Mirzaie et al. (2021), employing the multiple imputation method might not yield significant advantages if the amount of missing data is less than 5%. However, the primary emphasis in this study was on the northeast monsoon season, known for its frequent floods in the Selangor region. To explore factors contributing to these flood occurrences, a thorough filtering process was implemented on the dataset, focusing on specific months within the northeast monsoon season, namely November to March. As a result, the dataset was meticulously curated, comprising a total of 1,664 observations across 33 stations in Selangor.

Table 1.2

List of the predictors

Variables	Unit
Precipitation	mm
Minimum Temperature	℃
Maximum Temperature	℃
Relative Humidity	%
Solar Radiation	W/m^2
Wind	knot

1.3.1 Descriptive Statistics of Rainfall Data

Descriptive statistics are the first step in data analysis, providing the groundwork for further statistical analysis. It helps in making complex data understandable and accessible, allowing for better interpretation and communication of data insights. Descriptive statistics for the rainfall data, including mean, standard deviation, skewness, and coefficient of variation, as presented in Table 1.3, are calculated to offer a concise summary of the rainfall data for the study area.

Table 1.3

Summary statistics of daily rainfall data amount for each station

Code of Rainfall Station	Name of Rainfall Station	Values					
		Mean	**SD**	**Min**	**Max**	**Skew**	**CV**
22-334	P/A Sg. Pelek, Sepang	4.98	12.61	0	130	4.22	2.53
23-333	P/A Pekan Banting	5.29	12.66	0	178	4.84	2.39
23-334	Rtb Bukit Changgang	5.73	14.11	0	420.5	8.82	2.46
24-332	P/Kwln P/S Telok Gong	5.40	12.57	0	141	3.96	2.33
24-333	Ladang Bukit Cheeding	5.62	12.21	0	166	3.63	2.17
24-334	Puncak Niaga	6.21	13.68	0	121.5	3.53	2.20
24-335	Klm Takungan S. Merab	7.99	36.60	0	1873	36.36	4.58
25-332	Sg. Udang	5.81	12.94	0	179.5	3.88	2.23
25-333	Kg. Jawa	6.57	13.27	0	113	3.09	2.02
25-335	Sg. Balak H. Langat	7.27	14.83	0	146.5	3.23	2.04
25-336	Pejabat Jps. Klang	6.13	13.46	0	141.5	3.56	2.20
26-332	Ldg. Harpenden	5.58	13.16	0	141.5	4.20	2.36
26-333	Rumah Pam R. Panjang	6.14	13.88	0	166	3.79	2.26
27-331	Ladang Kuala Selangor	4.81	11.95	0	127	4.22	2.49
27-332	Ladang Braunston	4.80	12.18	0	150	4.68	2.54
27-333	Taman Desa Kundang	6.98	13.76	0	122	2.96	1.97

Code of Rainfall Station	Name of Rainfall Station	Values					
		Mean	SD	Min	Max	Skew	CV
27-334	Taman Templer	8.73	16.95	0	140.5	2.92	1.94
28-330	Stor Jps. Tanjung Karang	4.48	11.20	0	138	4.17	2.50
28-332	Pengorekan Bijih Berju	4.72	11.31	0	148	4.22	2.40
28-333	Kampung Sungai Buaya	7.54	14.96	0	167	3.14	1.98
28-334	Taman Desa Kelisa	8.33	15.95	0	136.5	2.89	1.91
29-331	Sungai Burung	4.96	12.44	0	163	4.80	2.51
29-332	Ladang Hopeful	6.89	15.19	0	287.5	4.62	2.21
29-335	Kampung Pertak	7.11	14.53	0	154	3.54	2.04
30-329	S.R.K Parit 4 Sg. Haji Dorani	4.24	14.01	0	357.5	10.67	3.30
30-330	Kg. Belia Sg. Panjang	5.17	12.03	0	130	3.96	2.33
30-331	Ldg. Pkps Sg. Panjang	6.57	13.64	0	111.5	3.23	2.07
30-332	Felda Soeharto	7.52	15.16	0	150.5	3.33	2.02
30-333	Kom. P'hulu Ulu Bernam	7.51	14.91	0	214.1	3.73	1.98
30-334	Loji Air Kuala Kubu Bahru	7.28	15.48	0	193.5	3.77	2.13
30-335	Bukit Fraser	6.61	12.26	0	120	3.11	1.86
31-328	Bagan Nakhoda Omar	4.64	11.82	0	115.5	3.97	2.55
31-329	Pekan Sabak Bernam	4.72	14.88	0	426	10.44	3.15

The table provides a detailed summary of rainfall statistics across different stations, focusing on the mean, standard deviation (SD), minimum and maximum values, skewness, and coefficient of variation (CV). The mean rainfall ranges from 4.80 mm (Ladang Braunston) to 7.99 mm (Kolam Takungan Sungai Merab), where higher means suggest more consistent rainfall, but they could also indicate a higher risk of flooding. Stations with lower mean rainfall might experience more dry periods, while those with higher means could be prone to more frequent and intense rainfall events.

The SD varies from 11.20 (Stor Jps. Tanjung Karang) to 36.60 (Kolam Takungan Sungai Merab), reflecting the variability in rainfall at different stations. A lower SD, such as 11.20, indicates that the rainfall amounts are more consistent and closer to the mean, signifying relatively stable weather patterns. In contrast, a higher SD, like 36.60, suggests greater variability, meaning the station experiences a broader range of rainfall amounts, including both very dry and very wet periods. This higher variability can complicate predictions and water management efforts, as it indicates a higher likelihood of extreme rainfall fluctuations.

The minimum values across all stations are uniformly 0 mm, indicating that there are instances where no rainfall occurs at all. This is a common occurrence, but it could become problematic if these dry spells happen frequently during critical periods, such as the growing season for crops or when water resources are already stressed. Prolonged periods of no rainfall can lead to drought conditions, affecting agriculture, water supply, and overall ecosystem health.

On the other hand, the maximum values range from 111.5 mm (Ladang Pkps Sungai Panjang) to 1873 mm (Kolam Takungan Sungai Merab), reflecting the highest recorded rainfall at each station. Stations with higher maximum values, like 1873 mm, are more prone to experiencing extreme rainfall events. Such peaks can lead to severe weather conditions, including flash floods, landslides, and significant infrastructure damage. The wide range in maximum values across stations suggests variability in the intensity of rainfall events, where some areas may be more susceptible to extreme weather and flooding than others. This variability underscores the need for localized flood management strategies and preparedness plans to mitigate the potential impacts of these severe weather events.

Skewness values, ranging from 2.89 (Taman Desa Kelisa) to 36.36 (Kolam Takungan Sungai Merab), indicate a positive skew across all stations, meaning that the distribution of rainfall data is asymmetrical, with a longer tail on the right side. In other words, there are more instances of lower rainfall amounts, but the presence of a few extreme high-rainfall events skews the average upward.

Higher skewness is less desirable because it signals an increased likelihood of extreme rainfall events. For example, a station with a skewness of 36.36 is likely to experience infrequent but very intense rainfall, which could lead to sudden and severe weather conditions like flash floods. This positive skew can be particularly challenging for water resource management and disaster preparedness, as it implies that while most rainfall events are moderate, the station is at risk for occasional but extreme rainfall that can have significant impacts. High skewness suggests a need for careful planning to manage the risks associated with these potentially catastrophic weather events.

The coefficient of variation (CV), which represents the relative variability of rainfall, is a crucial metric for understanding the consistency of rainfall patterns at different stations. The CV is calculated as the ratio of the standard deviation to the mean, expressed as a percentage. Lower CV values are generally more favorable because they indicate that the rainfall at a particular station is more consistent and stable, with less fluctuation around the mean.

In this case, the CV ranges from 1.86 to 4.58 across the stations, with Bukit Fraser showing the lowest variability (CV of 1.86). This suggests that Bukit Fraser experiences relatively stable and predictable rainfall, which is beneficial for activities such as agriculture, water resource management, and infrastructure planning, as the risk of sudden, extreme weather conditions is lower.

On the other hand, Kolam Takungan Sungai Merab has the highest CV (4.58), indicating greater variability in rainfall. A higher CV means that rainfall amounts are less predictable and more erratic, with larger deviations from the average. This could lead to challenges in managing water resources, as the station may experience both periods of heavy rainfall and drought. The higher variability at Kolam Takungan Sungai Merab could also increase the likelihood of extreme weather events, making it crucial for local authorities to implement robust flood management and disaster preparedness strategies.

1.4 Problem Statement

Selangor's robust economy, employment prospects, quality facilities, and excellent infrastructure make it a highly sought-after place to live (Husin et al., 2020). However, the region has been plagued by flash floods, particularly in metropolitan areas, causing financial losses and property damage (Muhammad & Shaidin, 2022). Flooding occurs when intense rainfall overwhelms the drainage capacity (Fitria & Amalia, 2018), leading to significant disruptions. For example, persistent heavy rainfall has necessitated the evacuation of residents due to flooding across Selangor (New Straits Times, 2018), and flash floods in 2017 caused transportation disturbances and road closures despite no casualties or evacuations (Crisis24, 2017). These events highlight the need for accurate rainfall projections to aid hydrologists and climatologists in preparing for future occurrences.

Accurate flood prediction models are crucial for reducing the effects of floods and improving community resilience (Nakhaei et al., 2023). Currently, Malaysia's WRF-MMD (Weather Research Forecast – Malaysian Meteorological Department) model struggles with inconsistencies, deterministic outputs, and limited forecasting lead times (Zaidi et al., 2022; Rosmadi et al., 2023). There is an urgent need for sophisticated predictive models that incorporate various data sources to more precisely identify areas at risk of flooding. Developing these advanced models requires overcoming several challenges, including addressing missing data, managing high-dimensional datasets, handling zero-bound data, and accounting for the inherent non-linearity of climate data. Successfully tackling these issues is essential for improving the model's accuracy and ensuring effective flood risk assessment.

Rainfall data, crucial for accurate climate and weather forecasting, is often plagued by gaps due to various reasons such as equipment failure, human error, or environmental factors (Shaharudin et al., 2020). These missing values pose a significant challenge in developing reliable rainfall projection models, as they can lead to biased predictions or incorrect trend identification. Handling these missing values while preserving the temporal and spatial characteristics of the data is essential (Ginkel et al., 2020). This issue necessitates the development of more sophisticated methods that can effectively handle missing data without compromising the accuracy of the model.

Another challenge is managing the high-dimensional nature of rainfall data, which includes numerous predictor variables like temperature, pressure, humidity, wind speed, and geographical parameters (Abdulhammed et al., 2019). The inclusion of such a large number of variables can lead to complications in data visualization, analysis, and model development (Mazher, 2020). High-dimensional data not only makes it difficult to discern meaningful patterns but also introduces the risk of overfitting, where the model becomes excessively complex and performs well on training data but poorly on unseen data (Bacro et al., 2020; Azeem et al., 2023). Thus, to effectively address this issue, a specialized method is required that can distill this high-dimensional data into a more manageable form without losing critical information.

Rainfall datasets also often contain a significant number of zero values, making them zero-bounded (Khooripan et al., 2023). This characteristic makes it challenging to apply standard regression techniques, which assume normally distributed data. Failure to address zero-inflated data can result in poor fitting for both

zero and non-zero counts (Perumean-Chaney et al., 2013; Feng, 2021). Addressing this issue requires the development of specialized regression techniques that can accommodate zero-bounded data and improve the robustness of rainfall projections.

Moreover, the non-linear relationships between predictor factors and rainfall patterns add another layer of complexity, as these relationships are influenced by intricate atmospheric and geographic factors (Kumar et al., 2021). Statistical downscaling-based machine learning models offer a solution by allowing for the modeling of non-linear relationships, but they too come with challenges such as the risk of overfitting and the need for large datasets. Therefore, there is a pressing need for a hybrid-regression framework that combines the strengths of both statistical downscaling-based machine learning models and Bayesian method to effectively model non-linear rainfall data.

Thus, to address these issues, a prediction modeling a hybrid of Statistical Downscaling-based Bayesian Approach and Machine Learning model. This novel model seeks to substantially enhance rainfall prediction by harnessing the data-driven strengths of machine learning. By providing a probabilistic framework, this model captures a wide range of variability, ultimately improving the accuracy of rainfall projections.

1.5 Research Objectives

The objectives of this research are:

i. To identify the most suitable imputation methods in handling of missing values in the rainfall dataset.

ii. To reduce high-dimensional rainfall data by applying dimensional reduction approach.

iii. To capture uncertainties associated with zero-bounded rainfall data by implementing a Bayesian method.

iv. To investigate and evaluate the effectiveness of hybrid Statistical Downscaling-based Bayesian Approach and Machine Learning model in capturing and modeling non-linear relationships between large-scale atmospheric predictor variables and local rainfall patterns.

1.6 Contribution of Study

The contributions of this study are:

i. This study will contribute by determining the most effective techniques for handling missing values in rainfall datasets, ensuring more complete and accurate data for analysis.

ii. By applying dimensional reduction approaches, the research will provide methods to simplify complex rainfall datasets, making them more manageable and interpretable.

iii. The study will offer insights into capturing uncertainties associated with zero-bounded rainfall data through Bayesian methods, improving the reliability of predictions.

iv. This research will evaluate the performance of hybrid Statistical Downscaling-based Bayesian Approach and Machine Learning

models, contributing to the understanding of non-linear relationships between predictor variables and rainfall patterns, ultimately leading to more precise rainfall predictions.

1.7 Significance of Study

The significances of this research are:

i. Improved Data Handling: By identifying the most suitable imputation methods for handling missing values in rainfall datasets, the study ensures data completeness and quality, which are critical for accurate rainfall predictions.

ii. Enhanced Model Efficiency: The application of dimensional reduction techniques addresses the challenge of high-dimensional rainfall data, making the predictive models more efficient and easier to interpret.

iii. Increased Prediction Reliability: Implementing a Bayesian method to capture uncertainties associated with zero-bounded rainfall data leads to more robust and reliable rainfall predictions under varying conditions.

iv. There Advancement in Predictive Capabilities: By investigating and comparing hybrid Statistical Downscaling-based Bayesian Approach and Machine Learning, the study enhances the ability to model complex, non-linear relationships between atmospheric predictors and rainfall patterns, leading to more precise and accurate rainfall forecasts.

v. Contribution to Flood Management and Climate Adaptation: The findings of this study are expected to improve flood prediction accuracy, which is essential for effective flood management, disaster risk reduction, and adaptation to climate variability.

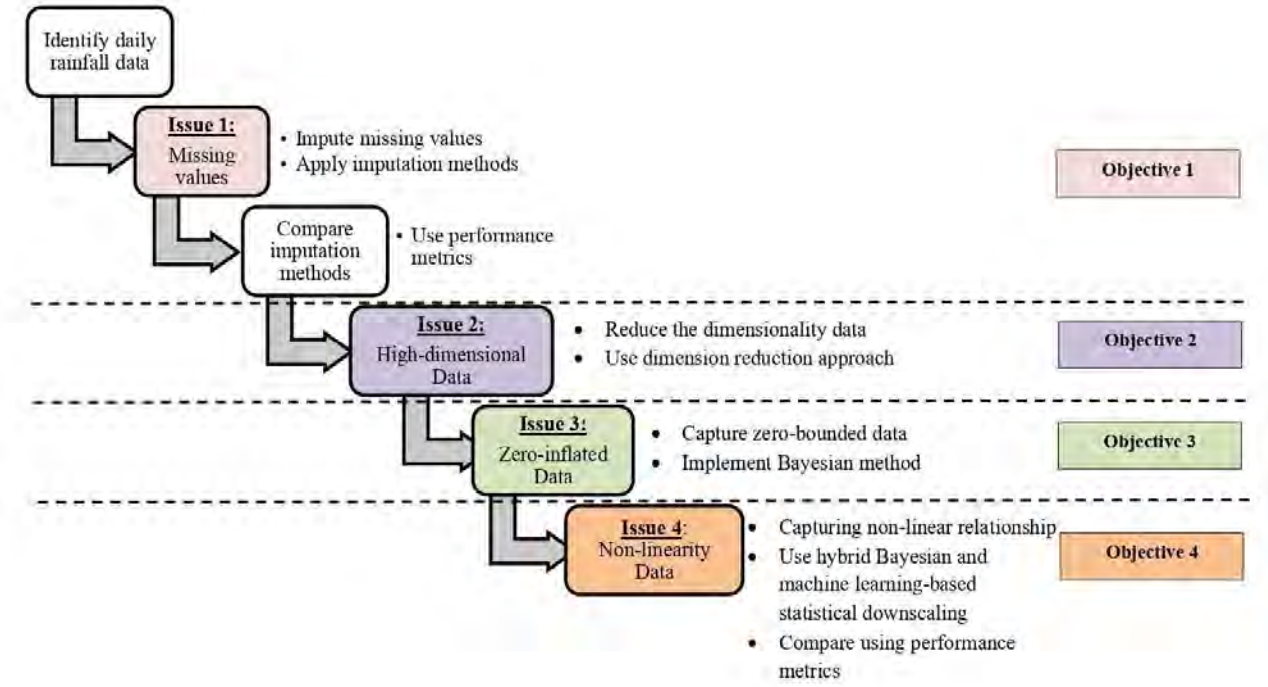
1.8 Framework of Study

The study's framework outlines the systematic process undertaken to achieve the research objectives, providing a clear roadmap from data handling to model evaluation. The first step in this process addresses the issue of missing values in rainfall data, where various imputation methods are applied to fill in the gaps. Each imputation method is then rigorously evaluated using performance metrics to identify the most effective approach for dealing with daily rainfall data. Following this, the study tackles the challenge of high-dimensional data by implementing dimensional reduction techniques, which condense the data into a more manageable, low-dimensional form while retaining essential information. To handle the uncertainties inherent in zero-bounded rainfall data, a Bayesian approach is employed, effectively mitigating the zero-bound issue. This is followed by the development of a hybrid model that combines the Bayesian approach with machine learning-based statistical downscaling techniques, designed to capture the complex, non-linear relationships within the data. Finally, the study employs a set of performance metrics to evaluate and compare the effectiveness of the proposed models, ensuring that the best model is identified based on its ability to accurately predict daily rainfall and handle the unique challenges of the data. This

structured approach ensures that each objective is systematically addressed, leading to robust and reliable results. The structure of the study framework is shown in Figure 1.2.

Figure 1.2

Framework of study



1.9 Limitation of Study

The lack of access to the most recent atmospheric data from ground stations in Selangor has led to a situation where the data from these stations have not been updated. This limitation is significant because it affects the comprehensiveness of the atmospheric variables available for analysis. The challenge is further compounded by the constraints of the server used for processing this high-dimensional data, particularly in the case of rainfall datasets. High-dimensional data, which involves numerous variables and

extensive records, often presents difficulties in terms of both the sheer volume of data and the lengthy processing times required. These issues are common in the analysis of such data and highlight the need for a more powerful and faster server to handle the computational demands effectively. Additionally, machine learning techniques, which are integral to this study, rely on sophisticated multi-algorithmic approaches that demand considerable storage capacity and substantial iteration time, especially when working with large datasets. The iterative nature of these algorithms, combined with the extensive data involved in time series analysis, underscores the necessity for substantial computational resources. In summary, conducting an accurate and reliable analysis of machine learning models applied to time series data, such as in this study, requires significant computing power to manage the complexities and demands of the data processing involved.

1.10 Summary

In conclusion, Chapter 1 provided a comprehensive introduction to the foundational aspects of this study. It began by outlining the background, setting the context for the research, and presenting the problem statements that highlight the key challenges and issues this study seeks to address. The chapter also offered a brief yet informative overview of the study area, focusing on the state of Selangor, which serves as the geographical focus of the research. In addition, it detailed the specific dataset employed in the study, providing clarity on the type of data analyzed and its relevance to the research objectives. Furthermore, the chapter outlined the research objectives, clearly articulating the goals the study aims to achieve. This was followed by a discussion on

the contributions of the study, emphasizing how this research adds to existing knowledge and its potential implications. The significance of the study was also highlighted, underscoring its importance in the broader context of the field. The chapter then presented the framework of the study, offering a structured overview of the methodology and approach taken to address the research objectives. Lastly, it acknowledged the limitations of the study, providing a balanced perspective on the challenges that may influence the outcomes of the research. Collectively, this chapter sets the stage for the subsequent chapters, establishing a solid foundation for understanding the study's purpose, scope, and significance.