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

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

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

NOOR HAMIZAH BINTI MOHAMAD SANI

UNIVERSITI PENDIDIKAN SULTAN IDRIS

2025

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

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

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

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

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 ....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

05-45068 ptbupsi

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

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

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

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:-*

05-45068 ptbupsi

☐ **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.

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

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

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.

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

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

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.

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

CHAPTER 5 CONCLUSION

5.1 Overview 185

5.2 Summary 185

5.3 Future Research 189

REFERENCES 191

APPENDICES 206

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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.,

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

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

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

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

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

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.

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

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.

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

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

1.3 Study Area

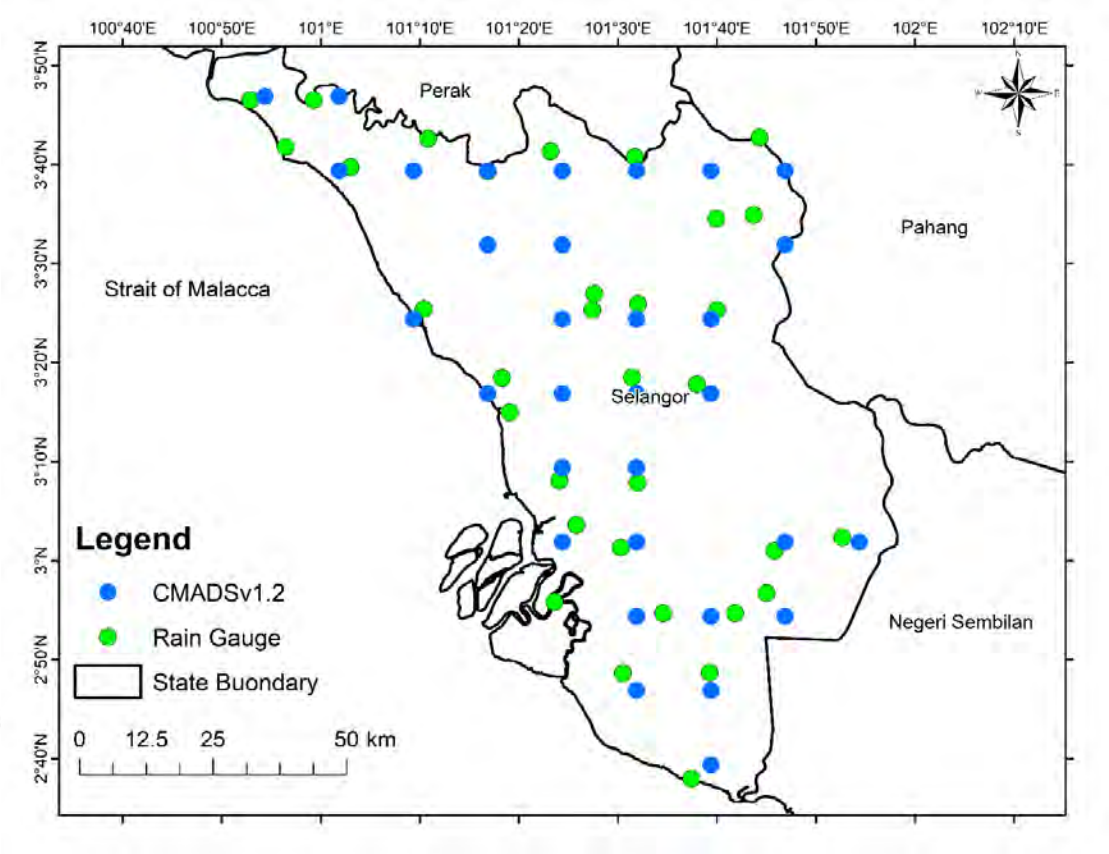
The study focuses on Peninsular Malaysia's Selangor state, located on the south coast of Peninsular Malaysia. Selangor's economy is outpacing that of Malaysia's other 13 states, making it one of the most sought-after places to live due to its strong economy, employment prospects, good facilities, and 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.

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

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

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).

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

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

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

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

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

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.

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

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

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

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

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.

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

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

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

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

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

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

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

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

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.

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

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

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.

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

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

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.

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

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.

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

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

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.

05-4506832 Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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.

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

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

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

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

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

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.

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

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:

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

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

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

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

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

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

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

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.

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

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

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

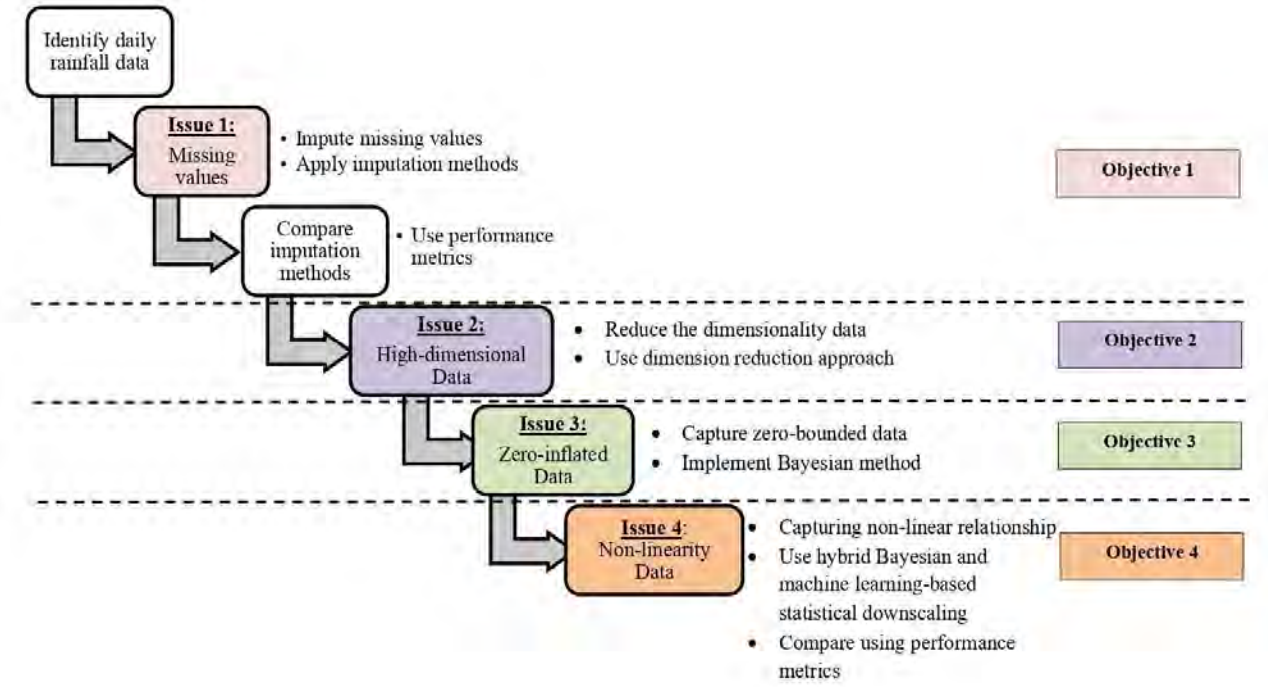
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

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

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

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

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.

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

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

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

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

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.

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

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

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter will delve into an extensive literature review centered on the application of hybrid machine learning-based statistical downscaling models specifically for rainfall prediction. The discussion will thoroughly examine existing research and methodologies in this area, highlighting the evolution and effectiveness of these advanced techniques in predicting rainfall patterns. In addition to exploring hybrid machine learning models, the chapter will also address critical topics related to data challenges that are frequently encountered in such studies. This includes a comprehensive review of literature on handling missing values, where various imputation strategies are discussed in terms of their applicability and effectiveness in ensuring data integrity. The chapter will also explore the complexities associated with high-dimensional data, focusing on techniques used to manage and reduce dimensionality while preserving essential information for accurate model predictions. Further, it will cover literature on zero-inflated data, a common issue in rainfall datasets, where methods to address the disproportionate number of zero values in the data will be discussed. The chapter will also examine non-linearity in data, reviewing

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

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

approaches to capture and model non-linear relationships that are crucial for accurate rainfall projection. Lastly, the discussion will include an evaluation of model performance metrics, considering the various metrics used in the literature to assess and compare the effectiveness of different models in predicting rainfall. Through this comprehensive literature review, the chapter aims to provide a solid theoretical foundation for the study, situating it within the broader context of existing research and highlighting the key challenges and solutions relevant to the field.

2.2 Introduction to Statistical Downscaling-based Machine Learning Model

Downscaling methods have been applied to overcome the scaling gap between coarser-resolution GCM's outputs and regional-scale climatic variables (Wilby et al., 2002; Shi et al., 2020; Maqsood et al., 2023). Traditional statistical downscaling methods rely on large-scale observational data, often reanalysis datasets, as well as local-scale observations (Vandal, Kodra & Ganguly, 2018). Meanwhile, machine learning is the ability of systems to learn from training data that is relevant to a given problem in order to automate the creation of analytical models and complete related activities (Janiesch, Zschech & Heinrich, 2021). Machine learning is also often viewed as a universal toolbox, ready to be utilised for regression tasks, calculating the relationship among provided data samples, and classification tasks, determining a suitable category for a new collection of observations (Rafique & Velasco, 2018). Machine learning models can identify intricate patterns and dependencies within the data (Yang et al., 2019). This capability allows them to capture the fine-scale spatial and temporal variations in climate variables, including rainfall (Pour et al., 2018).

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

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

Mao (2021) stated that machine learning models require a large volume of input data; therefore, data preprocessing, distribution, and spin up procedures are limited by the challenging data collection, where GCM outputs and observed local-scale climate data are collected and cleaned. Relevant predictors are selected from the GCM outputs, and the machine learning algorithms are characterized by a unique ability to learn system behaviour from historical data (Rafique & Velasco, 2018) and learning the relationship between the predictors and the observed local-scale climate variables such as rainfall (Barrera-Animas et al., 2022).

Statistical downscaling-based machine learning offers several advantages over traditional downscaling methods. Machine learning algorithms have the ability to overcome complex and non-linear relationships in the data, leading to more accurate downscaling results (Putri et al., 2021). They can handle high-dimensional data and allowing machine learning to offer more flexible alternatives for analysing high-dimensional, right-censored and heterogeneous data (Spooner et al., 2020). In addition, machine learning can also handle complex spatial and temporal proximities to further improve the estimation performance of spatiotemporal non-stationarity (Wu et al., 2021). In the meantime, some machine learning algorithms have high budget accuracy, and the prediction results are relatively robust to unbalanced and missing data (Fan et al., 2022). The main concept of statistical downscaling is to make a relationship between predictors and predictand by means of statistical method such as linear regression models, nonlinear regression models and weather generators (Nourani, Baghanam & Gokcekus, 2018). Moreover, bias correction is used to fine-tune downscaling findings based on SDSM so that the model can produce outputs that are more in line with the observed data (Fan, Jiang & Gou, 2021). Thus, statistical

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

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

downscaling-based machine learning is efficient, which directly incorporates the relationship between observation and GCM's output and has been actively developed by the scientific community to produce climate projections at the local scale necessary for impact studies (Wood et al., 2002, 2004; Gudmundsson et al., 2012; Hundecha et al., 2015; Sunyer et al., 2015; Su et al., 2016; Liu et al., 2017; Ali et al., 2019).

Al-Mukhtar and Qasim (2019) utilized dynamical or statistical downscaling methods to forecast future precipitation and temperature in Iraq. They predicted the future maximum and minimum temperatures, as well as precipitation levels, for 12 stations across Iraq, focusing on three time periods: the 2020s, 2050s, and 2080s. The projections were based on the Canadian GCM model (CanESM2) under different scenarios (RCP2.5, RCP4.5, and RCP8.5), using the statistical downscaling model (SDSM). The model was initially calibrated and validated with daily data, using 70% of the data for calibration and the remaining 30% for validation. Once calibrated, the model was applied to downscale future CanESM2 scenario predictors. The study demonstrated that SDSM performed satisfactorily in simulating future maximum and minimum temperatures and precipitation. All stations and scenarios consistently indicated an increasing trend in maximum and minimum temperatures and a decreasing trend in precipitation, with the RCP8.5 scenario showing the most severe trends.

Xu et al. (2021) introduced a multi-factor geographically weighted machine learning (MFGWML) algorithm to tackle the challenges of non-linearity and non-stationarity in the relationship between land surface temperature (LST) and various land surface parameters in Beijing. The model integrates three advanced machine learning algorithms—extreme gradient boosting (XGBoost), multivariate adaptive regression

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

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

splines (MARS), and Bayesian ridge regression (BRR)—as base learners to effectively capture the non-linear relationships. MFGWML then employs geographically weighted regression (GWR) to account for spatial non-stationarity by combining the predictions from these base learners. The results indicate that MFGWML surpasses two traditional algorithms, thermal image sharpening (TsHARP) and the high-resolution urban thermal sharpener (HUTS). The study concluded that MFGWML successfully merges the strengths of multiple regression, machine learning, and GWR, effectively capturing local heterogeneity and producing reliable and robust downscaled LST data.

Gonzalez-Abad et al. (2023) proposed utilizing deep learning (DL) models for the statistical downscaling of climate projections, specifically focusing on temperature downscaling within the CORDEX North America domain. They evaluated three Convolutional Neural Network (CNN) models—CNN-UNET, CNN-DeepESD, and CNN-PAN—which differ in complexity and architecture. Additionally, they introduced two innovative eXplainable Artificial Intelligence (XAI)-based diagnostics: Aggregated Saliency Maps (ASM) and Saliency Dispersion Maps (SDM), to assess the models. The motivation behind their approach was to overcome the limitations of Global Climate Models (GCMs) in delivering high-resolution climate data at regional-to-local scales. While DL models have shown potential in downscaling climate projections, concerns about their complexity, lack of transparency, and ability to generalize to future climate scenarios remain. The introduction of XAI techniques aims to enhance the interpretability of DL models' internal mechanisms and identify issues that standard evaluation methods might overlook. The study found that the CNN-UNET model consistently performed well across various regions and future climate scenarios, whereas the CNN-DeepESD and CNN-PAN models exhibited discrepancies,

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

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

particularly in southern regions. XAI diagnostics revealed that the CNN-UNET model effectively captured local relationships, while the other models displayed scattered regions of influence, indicating possible overfitting or structural challenges.

Fan et al. (2022) identified that the scarcity of meteorological stations in the high mountains of northwest China hinders the assessment of non-stationary climatic runoff processes influenced by climate change. To tackle this challenge, they developed an integrated method combining climate downscaling based on reanalysis products, improved complete ensemble empirical mode decomposition with adaptive noise, and four machine learning algorithms—radial basis function artificial neural network, random forests, support vector regression, and eXtreme gradient boosting tree—to simulate streamflow. This method was tested in seven mountainous basins originating from the high Tienshan Mountains. The observed runoff data from hydrological stations showed that the method performed well, with a Nash-Sutcliffe efficiency (NSE) above 0.7 and mean absolute error (MAE) and root mean square error (RMSE) below 3. The study highlights the importance of multi-scale modeling for enhancing the accuracy of runoff simulations. It also found that runoff exhibits a more significant nonlinear relationship on an inter-decadal scale compared to inter-annual and seasonal scales. Additionally, the study indicates that in high mountain regions, temperature primarily controls runoff changes, whereas precipitation has a greater impact on runoff changes in lowland areas.

Alipour et al. (2023) conducted a study on the statistical downscaling of global climate models to analyze temperature trends in Calgary, Canada. They utilized two models: a shallow-layered Artificial Neural Network (ANN) and a deep-based Long

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

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

Short-Term Memory (LSTM) network, to extract historical temperature trends. To ensure the quality of input variables, Mutual Information (MI) was used for screening. The results demonstrated that the LSTM model, when combined with MI for data screening, outperformed the ANN model in extracting historical temperature trends. The LSTM model achieved higher accuracy, with better determination coefficient (DC), correlation coefficient (CC), and lower root mean square error (RMSE) and bias compared to the ANN model. This indicates that the LSTM-based network holds promise as a tool for simulation and impact assessment in the Calgary region.

Sun and Lan (2021) conducted a study on statistical downscaling of daily temperature and precipitation across China using Deep Learning Neural Models. They employed convolutional neural network (CNN) models with various configurations to downscale daily temperature and precipitation, utilizing large-scale atmospheric variables from the ECMWF Interim reanalysis (ERI) as predictors and high-resolution gridded observations as predictands. The study spanned a 21-year period from 1979 to 1999 for calibration, with a warmer period from 2000 to 2017 used for validation to assess the models' extrapolation capabilities. The results revealed that model performance differed across configurations, and for effective multi-site downscaling across China, the convolutional process was essential, requiring more spatial features to accurately parameterize temperature characteristics compared to precipitation. The CNN model demonstrated added value over ERI in reproducing the geographic distributions of seasonal mean climate and seasonal cycles, and in reducing biases in mean and extreme percentiles for both temperature and precipitation. However, ERI outperformed CNN in terms of temporal correlations. The CNN model was further compared with the Generalized Linear Regression Model (GLM) and two quantile

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

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

mapping-based techniques: Bias Correction and Spatial Disaggregation (BCSD) and Bias Correction and Climate Imprint (BCCI). The study found that bias correction methods excelled in reducing biases and representing temporal correlations, especially for precipitation, while the CNN model performed better than GLM in precipitation downscaling. These findings suggest that the CNN model holds significant potential for downscaling applications over large regions.

Kouadri et al. (2021) explored the effectiveness of various artificial intelligence methods in predicting water quality parameters in a dry desert environment. They applied several AI algorithms, including multilinear regression (MLR), random forest (RF), M5P tree (M5P), random subspace (RSS), additive regression (AR), artificial neural network (ANN), support vector regression (SVR), and locally weighted linear regression (LWLR), to generate Water Quality Index (WQI) predictions in the Illizi region of southeast Algeria. Using best subset regression, they developed 12 different input combinations and structured their analysis around two scenarios. The first scenario aimed to reduce the time required for WQI computation by using all parameters as inputs, while the second scenario focused on assessing water quality variations in critical cases where necessary analyses were unavailable, reducing inputs based on sensitivity analysis. The results showed that the MLR model performed exceptionally well in the first scenario, as the linear nature of WQI calculations allowed MLR to effectively handle all parameters as inputs. However, reducing the number of inputs directly impacted model performance, and in the second scenario, the focus was on identifying which model performed best under such conditions. RF models emerged as the top performer in predicting WQI based on Total Hardness (TH) and Total Dissolved Solids (TDS) in the study area.

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

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

Oh et al. (2022) introduced a machine learning-based statistical downscaling model incorporating new topography-derived variables to better interpret the physical characteristics of wind. Focusing primarily on input data, the study utilized the Light Gradient Boosting Machine (LGBM) to assess how different input variables impacted results. By including new derived variables, the model could account for topography shapes that were previously difficult to interpret, leading to a notable performance improvement. The analysis used 1 km Weather Research and Forecasting (WRF) data and ERA5 reanalysis data from South Korea. Two Weibull distribution parameter maps were computed and employed as both input and output data. Three sets of derived variables were developed and compared, with the multi-resolution topography data showing the most significant improvement, reducing the root mean square error (RMSE) by approximately 15% for both linear regression and machine learning models. The land area, in particular, exhibited a 20% reduction in RMSE. The best-performing models achieved RMSE values of 7% and 8% for the two Weibull parameters.

Farikha et al. (2021) investigated the use of projection pursuit regression (PPR) in a statistical downscaling model utilizing an artificial neural network (ANN) for rainfall prediction in Jember, Indonesia. While the General Circulation Model (GCM) provides dynamic rainfall predictions on a monthly scale, its low resolution limits its effectiveness for small-scale predictions, necessitating the use of a statistical downscaling (SD) model. PPR, a non-parametric and non-linear approach, is employed in this SD model to process high-dimensional data by reducing it to smaller dimensions through a projection process. The study further incorporated a neural network-based approach, with ANN serving as the statistical downscaling model to analyze rainfall-

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

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

related events. The study concluded that PPR is effective in reducing dimensions, especially for non-linear data, and that combining PPR with ANN yields smoother performance compared to PPR alone.

Based on previous research into statistical downscaling using machine learning models, such approaches have demonstrated considerable promise for improving rainfall prediction accuracy. These studies highlight the effectiveness of machine learning techniques in refining climate data to better reflect local conditions, which is essential for precise rainfall forecasting. Machine learning models, particularly those leveraging statistical downscaling, can process vast amounts of climate data and discern intricate patterns that traditional methods might overlook. By enhancing the resolution and detail of predictive models, these approaches address the limitations of coarse global climate models, which often lack the granularity needed for accurate local predictions.

Moreover, the integration of machine learning with statistical downscaling techniques provides a robust framework for capturing non-linear relationships and complex interactions within climate data. This is crucial for rainfall prediction, where factors influencing precipitation are often intricate and interdependent. The ability of machine learning models to adapt and improve over time with more data further enhances their suitability for rainfall forecasting. As these models are trained on high-resolution historical data and validated against observed conditions, they become increasingly adept at making reliable predictions. Therefore, applying statistical downscaling-based machine learning models to rainfall prediction is not only a viable

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

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

option but also a potentially transformative approach for enhancing the accuracy and reliability of climate forecasts.

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

Rainfall is a significant meteorological component because it influences human activities in so many different ways, including farming, power generation using water, and managing water supplies (Hsu & Sharma, 2023). Machine learning activities investigate how computers can use information to practice and can be applied to test rainfall projection (Usman et al., 2023). Machine learning has made significant improvements and is a core sub-area of artificial intelligence, and also enables computers to discover themselves without being explicitly programmed (Sani et al., 2020).

Kanani et al. (2023) conducted a study about an AI-Enabled ensemble method for rainfall forecasting using Long-Short term memory (LSTM). The goal of the paper was to provide a customized set of machine learning and deep learning models to provide most accurate results for rainfall prediction. Training classifiers such as eXtreme Gradient Boosting (XGBoost), Random Forest, Kernel Support Vector Machine (SVM), and LSTM were used for the classification task, while models such as Multiple Linear Regressor, XGBoost, Polynomial Regressor, Random Forest Regressor, and LSTM were used for the regression task. The experiment results showed that the proposed approach outperforms several state-of-the-art approaches.

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

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

Patel et al. (2023) focused on how to create long-term data-based models that anticipate the behaviour of hydrologic systems and their components. The main objective was to use a combination of machine learning and forecasting methods to predict rainfall. The methods used in the research were Random Forest, Gradient Boosting and Decision Tree. Those methods were needed to predict the total amount of rain that would fall each day. The researchers claimed that the Gradient Boosting was found to work best among those tested, with features related to and affecting rainfall predictability giving accuracy of 93%, while Random Forest and Decision tree method having 90% and 78.5% of accuracy was achieved respectively.

Next, Praveena et al. (2022) performed a study about the prediction of rainfall analysis using Logistic Regression and Support Vector Machine (SVM). The researchers predicted the rainfall based on nine natural phenomenon features and had to correlate these features by using correlation matrix and this correspondence to be tabulated. Over the entire range of these results, the accuracy of Logistic Regression is 88% while, the accuracy of SVM classifier is 82%. After hyper-tuning, the accuracy of Logistic Regression and SVM classifier were improved by 89% and 91% respectively.

Baljon and Sharma (2023) conducted a study to test the efficacy of several machine learning (ML) approaches for forecasting rainfall. The researchers proposed an improved machine learning technique named Function Fitting Artificial Neural Network (FFANN) for rainfall prediction rate. The findings of the proposed model illustrated the essential factors that played a crucial role during heavy rainfall and all of the extreme rainfall events that occurred during the assessment years could have been precisely predicted by the proposed model, demonstrating its efficacy as a predictor.

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

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

Prottasha et al. (2023) carried out a study that present a machine learning and data-driven that can accurately predict short-term rainfall. The main goal of the study was to develop a short-term rain prediction that can effectively and accurately predict rainfall. The rain-forecasting models were implemented in the study using twenty-nine different machine learning classifiers namely, Neighbors classifiers, Ensemble classifiers, Naïve Bayes classifiers, Semi-supervised classifiers, Discriminant analysis classifiers, Support Vector Machine classifiers, Linear model, Gaussian process classifiers and Neural network. Based on the study, the result showed that hist gradient boosting classifier from Ensemble classifiers has given the highest accuracy and suitable to be a universal and easy rain prediction tool.

Research by Ehteram et al. (2023) introduced a new hybrid deep learning model named convolutional neural network- support vector machine- Gaussian regression process (CONN-SVM-GRP) to predict daily and monthly rainfall data. The new proposed model's primary advantage is that it incorporates uncertainty values into the modelling process. The proposed model also could extract important features without applying preprocessing methods, further improving its efficiency. In conclusion, the CONN-SVM-GRP model could help researchers predict rainfall, which is essential for monitoring water resources and mitigating the impacts of droughts, floods, and other natural disasters.

A study by Adnan et al. (2023) investigated the feasibility of a relevance vector machine tuned with improved Manta-Ray foraging optimization (RVM-IMRFO) in predicting monthly pan evaporation. Evaporation is an important component in hydrological cycle that plays a crucial role in cloud’s moisture budgets (Wooden et al.,

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

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

2007; Adnan et al., 2023) and the magnitude of total evaporation is often equivalent to total rainfall. The accuracy of the proposed model was evaluated by comparing it with RVM tuned with grey wolf optimization (RVM-GWO) and RVM tuned with whale optimization algorithm (RVM-WOA). The researched observed that the new proposed model considerably improved the accuracy of a single RVM in monthly pan evaporation prediction and by importing extraterrestrial radiation into the model, the inputs generally improved the prediction accuracy.

Di Nunno et al. (2022) developed a hybrid machine learning model specifically designed for reliable precipitation prediction, focusing on the tropical monsoon climate of northern Bangladesh, particularly in the Rangpur and Sylhet divisions. The study employed two distinct machine learning algorithms, M5P and support vector regression (SVR), to create a novel hybrid model that combines the strengths of both techniques. The effectiveness of these prediction models was evaluated through a range of metrics and graphical analyses, ensuring a comprehensive assessment of their performance. Additionally, a sensitivity analysis was conducted to examine how the prediction accuracy was influenced by varying the number of exogenous inputs and increasing lag times. This analysis was crucial in understanding the model's robustness under different conditions. Ultimately, the hybrid M5P-SVR model outperformed the individual algorithms, demonstrating superior accuracy in precipitation prediction for the region studied, and thus, was identified as the most reliable model among those tested in the research.

A study by Parviz et al. (2023) aimed at enhancing hybrid models by incorporating ensemble techniques for both linear and non-linear models. The study

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

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

utilized preprocessing configurations and traditional hybrid models based on Gene Expression Programming (GEP), Support Vector Regression (SVR), and the Group Method of Data Handling (GMDH). These models were then compared to a proposed hybrid model that combined all three. Various methods were employed to improve the performance of the hybrid models, and the models were tested using data from two weather stations in Iran—Tabriz and Rasht—using annual and monthly time steps, respectively. The results demonstrated significant improvements in the performance of the proposed hybrid models. Specifically, Theil's coefficient, which measures the degree of disparity between forecasts and observations, improved by 9% and 15% for SVR and GMDH compared to GEP at the Tabriz station. The error metrics further indicated that the proposed hybrid models more accurately represented observations than the traditional hybrid models. At the Rasht station, the mean square error decreased by 67%, and the Nash Sutcliffe efficiency increased by 5% when the three machine learning models were combined using a genetic algorithm instead of relying solely on SVR. Overall, the inclusion of nonlinear models in the improved hybrid models resulted in better performance compared to the traditional hybrid approaches.

The main purpose of a study by Farooq et al. (2023) was focused on predicting seasonal summer rainfall at two stations in the Northern Territory (NT) of Australia using a hybrid machine learning algorithm. The research incorporated observed climate anomalies such as the El Niño Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and Madden-Julian Oscillation (MJO) as input variables for the prediction models. The study first applied two well-known machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF), as base models. To enhance prediction accuracy, the researchers then used wavelet transform (WT) as a

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

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

preprocessing method, combining it with SVM and RF to create the hybrid algorithms W-SVM and W-RF. These hybrid models integrated the wavelet components with the base algorithms to improve performance. The predictive accuracies of these algorithms were evaluated and validated using three statistical metrics. The results revealed that the W-SVM and W-RF hybrid algorithms achieved the highest predictive accuracies for seasonal rainfall, outperforming the individual base algorithms. Moreover, the type of mother wavelet used in the W-RF algorithm influenced its prediction accuracy, whereas the W-SVM consistently delivered high predictive accuracy across all selected wavelets, making it the most reliable model for seasonal rainfall prediction in this study.

Citakoglu and Coşkun (2022) conducted a comprehensive study comparing various hybrid machine learning methods for predicting short-term meteorological droughts at the Sakarya Meteorological Station in Turkey. The research employed several stand-alone machines learning algorithms, including Artificial Neural Networks (ANNs), Adaptive Neuro-Fuzzy Inference System (ANFIS), Gaussian Process Regression (GPR), Support Vector Machine Regression (SVMR), and k-Nearest Neighbors (KNN). To enhance the predictive capabilities of these models, the study also incorporated advanced preprocessing techniques such as Variation Mode Decomposition (VMD), Discrete Wavelet Transforms (DWT), and Empirical Mode Decomposition (EMD) to create hybrid models. The performance of these hybrid models was evaluated against their stand-alone counterparts. The findings revealed that the hybrid models, which combined machine learning algorithms with preprocessing techniques, significantly outperformed the stand-alone models in predicting short-term drought conditions. Specifically, the VMD-GPR hybrid model excelled at 1- and 3-month time scales, while the DWT-ANN hybrid model provided the most accurate

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

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

predictions for the 6-month time scale when forecasting the Standardized Precipitation Index (SPI) drought indices. These results underscore the effectiveness of integrating preprocessing methods with machine learning algorithms to improve the accuracy and reliability of meteorological drought predictions.

Sahoo et al. (2021) conducted a study focused on applying hybrid machine learning models for flood prediction in India. Their research aimed to assess the relative accuracy of models combining Radial Basis Function Neural Network (RBFNN) and Support Vector Machine (SVM) with the Firefly Algorithm (FA) to predict river flood discharge. These hybrid models were compared with traditional models including standalone ANN, RBFNN, and SVM. The evaluation of the results demonstrated that both RBFNN combined with FA (RBF-FA) and SVM combined with FA (SVM-FA) provided more accurate forecasting results compared to the standalone RBFNN, Feed Forward Back Propagation Neural Network (FFBPNN), and SVM models. Among the hybrid approaches, the SVM-FA model achieved superior performance, with an R^2 value of 0.9818, outperforming the RBF-FA model in terms of forecasting accuracy. Additionally, the study found that the simple SVM model offered slightly better performance than the traditional ANN model. These findings highlight the effectiveness of integrating the Firefly Algorithm with machine learning techniques to enhance the precision of flood prediction models, particularly in the context of Indian river systems.

Anaraki et al. (2020) introduced a novel framework for flood frequency analysis under climate change conditions by integrating metaheuristic algorithms, decomposition techniques, and machine learning models. This framework utilizes global climate models (GCMs), including HadCM3 with A2 and B2 scenarios, CGCM3

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

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

with A2 and A1B scenarios, and CanESM2 with RCP2.6, RCP4.5, and RCP8.5 scenarios. The proposed approach involves several components: Multivariate Adaptive Regression Splines (MARS) and M5 Model Tree for classifying precipitation into wet and dry days, Whale Optimization Algorithm (WOA) for training Least Squares Support Vector Machine (LSSVM), and Wavelet Transform (WT) for decomposing precipitation and temperature data. For downscaling precipitation and temperature, the framework employs LSSVM-WOA, LSSVM, K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN). The model simulates discharge for different periods: the present (1972–2000), near future (2020–2040), and far future (2070–2100). Flood frequency analysis is conducted using the log-normal distribution, while Analysis of Variance (ANOVA) and fuzzy methods are used for uncertainty analysis. The Karun Basin in southwest Iran serves as the case study. The results indicate that MARS outperformed the M5 Model Tree. Among the downscaling methods, ANN and LSSVM-WOA slightly exceeded other machine learning algorithms in performance. The discharge simulation revealed that the LSSVM-WOA-WT algorithm provided the best results, with a Nash-Sutcliffe Efficiency (NSE) of 0.911. The flood frequency analysis showed a reduction in the 200-year discharge for all scenarios except for the CanESM2 RCP2.6 scenario in the near future. Uncertainty analysis through ANOVA highlighted that hydrological models are significant sources of uncertainty, with the HadCM3 model exhibiting lower uncertainty in higher return periods, up to 60% lower than other models for the 1000-year return period, according to fuzzy uncertainty analysis.

Lee and Chen (2021) introduced an innovative approach that integrates a Markov chain random field conditioning method with a hybrid machine learning

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

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

technique to enhance long-range precipitation predictions amidst increasingly extreme weather conditions. Traditional precipitation models often struggle with limitations in their time-span capabilities, making it challenging to predict rainfall distribution accurately for specific years in long-range simulations. To address this issue, the researchers proposed a hybrid (ensemble) learning method that operates on a multi-scaled, conditioned functional time series within a sparse l1 space. This advanced methodology enables more accurate long-range forecasting by effectively combining various machine learning techniques and conditioning methods. The resulting algorithm is designed to be applicable in fields such as agriculture and construction, where precise long-term precipitation forecasts are crucial. The study's results demonstrated that the application of the conditioning method and multi-scale decomposition within the sparse l1 space significantly mitigates the impact of statistical variation caused by extreme weather conditions. This advancement offers a promising solution for improving the reliability and accuracy of long-term precipitation predictions, thereby benefiting sectors that rely heavily on weather forecasts for planning and decision-making.

In a nutshell, hybrid machine learning models have made a great development in weather forecasting to advance and enhance decision-making tool (Patel et al., 2023). The hybrid machine learning models is suitable for rainfall prediction (Sani et al., 2020) due to their ability to identify highly non-linear and irregular patterns in rainfall data (Poornima et al., 2023). Building a hybrid machine learning-based model is extremely desirable for the sustainable development of environmental systems due to the importance of the issue and the dynamic character of the atmosphere, which makes

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

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

machine learning methods preferred to other approaches (Tuysuzoglu, Birant & Birant, 2023).

Building on existing research into hybrid statistical downscaling methods and machine learning for rainfall prediction, it is evident that employing a hybrid model represents a highly effective strategy for enhancing forecasting accuracy. Previous studies have demonstrated that combining various machine learning techniques with statistical downscaling approaches significantly improves the precision of rainfall predictions. These hybrid models leverage the strengths of different methodologies, integrating them to address the limitations inherent in individual models. By combining machine learning algorithms with statistical techniques, such as regression analysis or data assimilation, these models can better capture the complex patterns and nuances of rainfall data, leading to more reliable and accurate forecasts.

Furthermore, the integration of hybrid models in rainfall prediction has shown substantial advantages in various research contexts. For instance, the fusion of machine learning algorithms like artificial neural networks or support vector machines with statistical downscaling methods allows for a more comprehensive analysis of meteorological data. This approach enhances the model's ability to process and interpret large datasets, effectively managing the variability and uncertainty associated with rainfall forecasting. As a result, hybrid models offer a robust solution for predicting rainfall, particularly in regions with complex climatic conditions or limited observational data. The continued development and application of these hybrid models hold significant promise for advancing the field of meteorological forecasting and improving decision-making processes in sectors reliant on accurate rainfall predictions.

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

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

2.3 Missing Values

Missing data is when an observation has no value assigned to it (Mirzaei et al., 2021). The issue of missing data is inevitable in data collection. It may result from equipment failure, reluctance to respond, the change of investigation plans, data availability or collection errors (Nikfalazar et al., 2019). Missing data are usually attributed to human error when processing data, machine error due to the malfunctioning of equipment, respondents' refusal to answer certain questions and merging unrelated data (Suthar, Patel & Goswami, 2012; Houari, Bounceur, Tari & Kecha, 2014; Emmanuel et al., 2021). The missing data problem is usually common in all domains that deal with data and causes different issues like performance degradation, data analysis problems and biased outcomes lead by the differences in missing and complete data (Ayilara et al., 2019; Emmanuel et al., 2021). Moreover, the seriousness of missing data depends in part on how much data is missing, the pattern of missing data, and the mechanism underlying the missingness of the data (Kang, 2013; Emmanuel et al., 2021).

Missing data can be classified based on the pattern of missing values and the mechanism of missing values. Missing data patterns describe which values are missing and observed in a data set (Emmanuel et al., 2021). In Figure 2.1, there are three missing data patterns: univariate, monotone and no-monotone. Univariate happened when there is only one variable with missing data. This pattern is rare in most disciplines and arises in experimental studies. Missing data pattern is said to be monotone if the variables in the data can be arranged, this pattern is usually associated with longitudinal studies where members drop out and never return. The monotone data pattern is easier to deal with since the patterns among the missing values are easily observable. Non-monotone

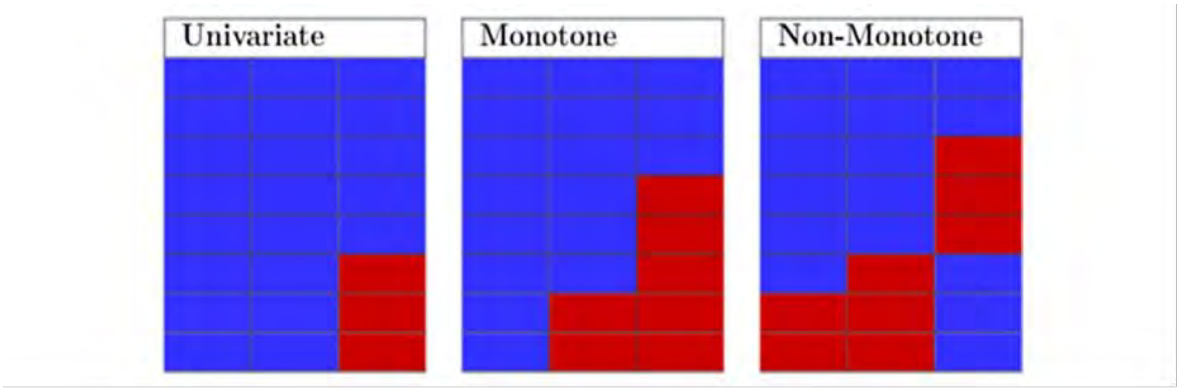
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

is a missing data pattern whereby the missingness of one variable does not affect the missingness of any variables.

Figure 2.1

Missing data pattern. Blue represents observed values; red represents missing values



Adapted from Emmanuel et al. (2021)

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

Mostly mechanisms that lead to the missing values on data effect some assumptions supporting most missing data handling methods. According to Little and Rubin (1987), there are three types of missingness mechanisms that can cause an incomplete dataset (Lin & Tsai, 2019). These mechanisms can be categorized as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Rubin 2004; Psychogyios et al., 2023). MCAR can be defined as when a missing value is not related to any other value in the dataset. According to Mirzaei et al. (2021), data that are MCAR are not usually attributed to a question in the survey or other phenomenon, whether observable or unobservable. Nevertheless, MAR is present if the reason for dropout depends on the observed outcomes and possibly on the observed variables at any point (Ayilara et al., 2019). In other words, when the distribution of an instance having missing values for an attribute depends on the

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

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

observed data, but does not depend on the missing data (Lin & Tsai, 2019). Lastly, NMAR contains non-ignorable missingness, are data that do not meet the criteria of either MCAR or MAR (Mirzaie et al., 2021). Emmanuel et al. (2021) stated that in this mechanism, handling the missing values is usually impossible, as it depends on the unseen data. Thus, research by Graham (2009) and Dantan et al. (2008), it is mostly impossible to unambiguously categorise missing data into these three mechanisms since imagining that missing data is completely not related to other non-missing variables (Emmanuel et al., 2021). Figure 2 shows an algorithm for handling missing data by Mirzaie et al. (2021).

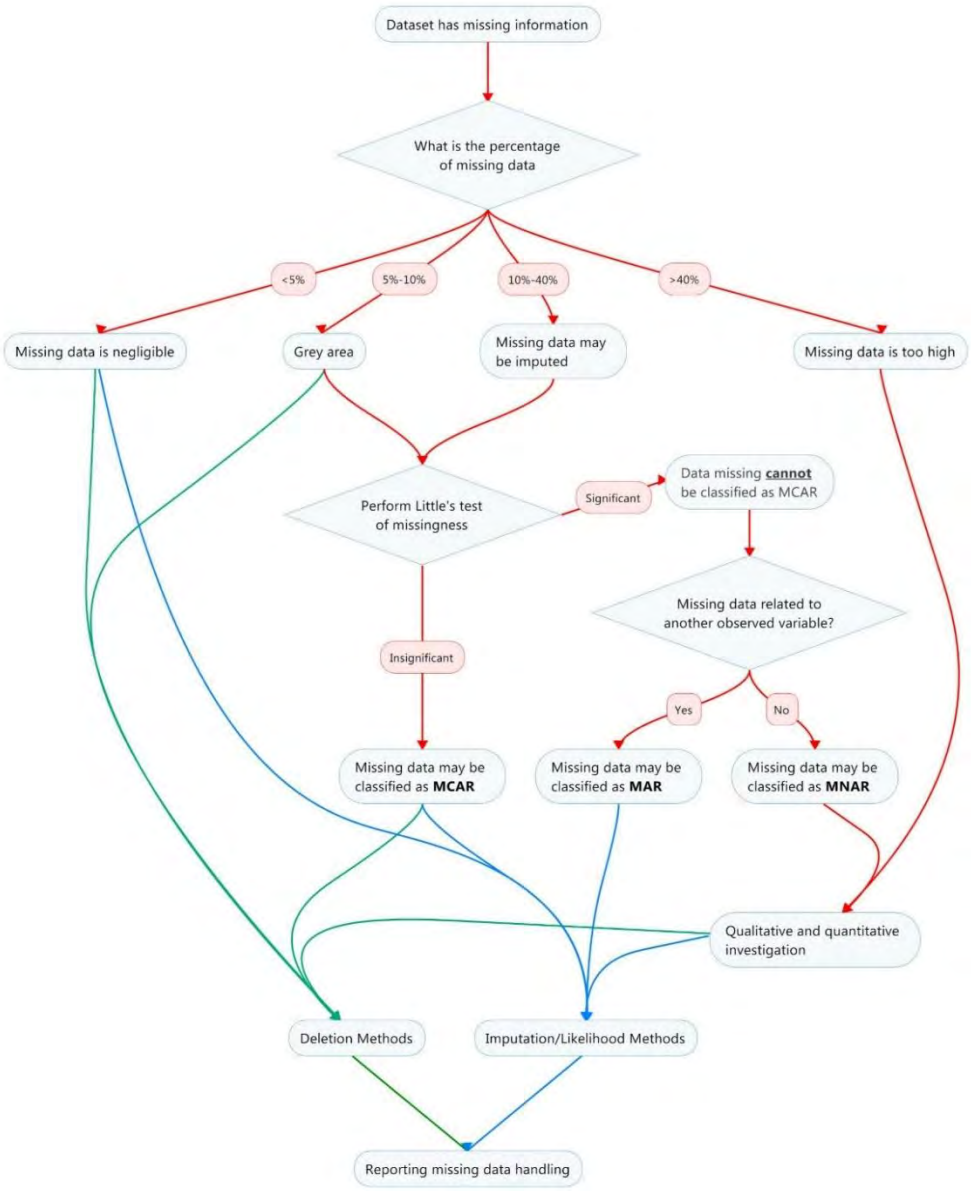
Imputation method is a process that involves replacing missing values by some predicted values where the non-missing values data set is normally used to predict the values to replace the missing values (Donders et al., 2006; Emmanuel et al., 2021). Imputation methods can be categorized into five types namely deletion methods, single imputation, multiple imputation, likelihood methods and indicator methods (Mirzaei et al., 2021).

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

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

Figure 2.2

Algorithm for handling missing data



Adapted from Mirzaie et al. (2021).

Research by Saeipourdizaj, Sarbakhsh and Gholampour (2021) explained that deletion method can be done in two ways which are pairwise deletion and listwise deletion. In the pairwise deletion, the number of observations needed for analysis varies from variable to variable since only the missing data are eliminated while, in the listwise

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

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

deletion, sometimes referred to as case deletion or comprehensive analysis, is the process of deleting all observations that have missing data for one or more variables. The entire data series can be retrieved by eliminating the missing values if the MCAR missing data technique is used and the missingness rate is less than 5%.

Single imputation refers to imputing one plausible value for each missing value of a particular variable in the dataset and then performing analysis as if all data were originally observed (Jadhav, Pramod & Ramanathan, 2019). According to Uusitalo et al. (2015) in Emmanuel et al. (2021), multiple imputation is where the distribution of the observed data is utilized to approximate numerous values that reflect the uncertainty around the true value, and this method was mostly implemented to solve the limitations of single imputation. For each missing observation, many possible values are substituted in multiple imputation rather than just one to express uncertainty about the correct values to impute. Hence, multiple imputation generates "m" different complete datasets with observed and imputed values (Jadhav, Pramod & Ramanathan, 2019). Thus, the results of the "m" will differ because the m imputations differ (Nengsih et al., 2019).

Likelihood methods may also be used when handling missing data and the most widely used are Expectation-Maximisation (EM) and Full Information Maximum likelihood (FIML) (Lee & Chiu, 1990; Enders, 2001; Mirzaie et al., 2021). These are model-based techniques that do not rely on prior imputation for parameter estimation when there are missing data. By maximising the likely value based on all other observed parameters in the model, necessary but missing parameters are approximated (Mirzaie et al., 2021). Lastly, indicator method is an alternative technique involves the addition

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

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

of a "missing" category to partially observed categories before incorporating confounding into a propensity score model, including both the confounding factor and a missingness indicator; however, this approach often results in biased outcomes, and a more generalized missing pattern analysis suffers when the instances with missing patterns are fewer than the observed covariates, typically arising in data with numerous missing patterns (Yang et al., 2023).

A study by Aieb et al. (2019) was about a new approach for processing climate missing databases applied to daily rainfall data. The aim of the study was to propose a new imputation algorithm based on the optimization of some regression methods, which are Hot Deck, k-Nearest Neighbors Imputation, Weighed k-Nearest Neighbors Imputation, Multiple Imputation, Linear Regression and Simple Average Method. The study showed two results; first, explaining the best choice between several methods, according to the results of comparison using statistical tests in different cases of missing data amount, it showed that the Hot-Deck, k-Nearest Neighbors Imputation and Weight k-Nearest Neighbors Imputation methods had given the best filling. Second, showing the influence of altitude, latitude and longitude on the correlation index and the residual average between neighboring stations and reference station – this variation affects the imputation results obtained by multiple imputation, linear regression and simple average method. Aieb et al. (2019) also stated that the proposed filling methodology can be also applied to estimate datasets for another case studies when the missing data distribution is random.

Moreover, Nor et al. (2020) focused on a comparative study of different imputation methods for daily rainfall which the main goal was to performed data

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

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

imputation using Replace by Mean, Nearest Neighbor, Markov Chain Monte Carlo (MCMC), Nonlinear Iterative Partial Least Squares (NIPALS) and Random Forest (RF) methods for daily rainfall data. Then, evaluated the performance of imputation methods combined with Multiple Linear Regression (MLR) model in predicting the future daily rainfall data. Based on the experiment, it showed that Replace by Mean is the best result for single imputation but RF had shown its supremacy by giving the best result and more fitting for satisfying the missing data when combined with MLR.

Next, Addi et al. (2022) experimented the evaluation of 12 imputation methods for infilling missing data in daily rainfall data. The imputation methods that were applied namely Mean, Regression, Multiple Imputation by Chain Equation (MICE), k-Nearest Neighbor (kNN), Probabilistic Principal Component Analysis (PPCA), Random Forest (RF) – executed using missForest algorithm, Linear Interpolation (LI), Hot-Deck (HD), Expectation Maximization (EM), Gaussian Copula (GC), Inverse Distance Weighting (IDW) and Kriging. However, Regression, PPCA and RF were showing a positive and excellence results for estimating the numbers of dry and wet periods and moderate to extreme rainfall values.

Hamzah et al. (2021) conducted a comprehensive study to evaluate the effectiveness of three distinct imputation methods—robust random regression imputation (RRRI), k-nearest neighbors (k-NN), and classification and regression tree (CART)—in predicting recurrences within streamflow datasets. The primary objective was to determine which method most accurately addresses the issue of missing data in streamflow records, a common challenge in hydrological studies. To achieve this, the researchers utilized the complete historical daily streamflow data from 2012 to 2014 as

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

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

a training dataset. This dataset served as a benchmark for assessing and validating the performance of the imputation methods. Following the initial analysis, each of these imputation techniques was coupled with multiple linear regression (MLR) and applied to restore missing streamflow rates in the Langat River Basin in Malaysia, covering the period from 1978 to 2016. The effectiveness of the estimation techniques was rigorously evaluated using various metrics, including the Nash-Sutcliffe efficiency coefficient (CE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE). The findings of the study revealed that the combination of RRRI with MLR (RRRI-MLR) significantly outperformed the other imputation methods in terms of accuracy and reliability. Specifically, the RRRI-MLR approach achieved the lowest RMSE and MAPE values, indicating superior performance in filling missing data within daily streamflow datasets. This suggests that RRRI-MLR is particularly well-suited for addressing the complexities and variabilities inherent in hydrological data, making it the most effective method among those tested. The study's results underscore the importance of selecting appropriate imputation techniques for hydrological modeling, as the choice of method can have a substantial impact on the accuracy of streamflow predictions. Given the critical role of streamflow data in water resource management and planning, the study's conclusions provide valuable insights for researchers and practitioners working in the field of hydrology. By demonstrating the efficacy of RRRI-MLR, the researchers offer a robust solution for improving the quality and reliability of streamflow datasets, particularly in regions where missing data is a significant concern.

Hırca and Türkkan (2024) presented a comprehensive methodology for tackling the challenge of missing data by thoroughly examining the structure and techniques

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

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

associated with missing data. Recognizing that missing data is often overlooked in hydrological studies, they created simulated datasets that carefully considered the number of missing data points, the patterns in which data were missing, and the underlying mechanisms that cause data to be absent. These simulations were designed to reflect real-world datasets that contain gaps, enabling a more accurate assessment of various missing data techniques. The study applied the most widely used methods for handling missing data, not only in hydrology but also in other scientific fields, to these simulated datasets. The researchers employed a range of techniques for data imputation. Simple imputation methods and expectation maximization (EM) were executed using SPSS software, while more advanced machine learning techniques like k-nearest neighbor (kNN) and the hot-deck method were implemented using Python. To determine the effectiveness of each method, the performance was evaluated based on error metrics. The results of this analysis indicated that the EM method outperformed the other techniques, emerging as the most suitable approach for completing missing data. The study highlights the importance of selecting the right method for data imputation, as it can significantly influence the accuracy of hydrological analyses. By providing a detailed comparison of various techniques and their performance, the researchers offer valuable insights for researchers and practitioners dealing with missing data in hydrology and other fields.

Aguilera et al. (2020) conducted a study that compared three different techniques for filling in large amounts of missing daily precipitation data: spatio-temporal kriging (STK), multiple imputation by chained equations using predictive mean matching (PMM), and the random forest (RF) machine learning algorithm. The findings indicated that both STK and RF were capable of handling extreme levels of

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

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

missing data, whereas PMM required a larger sample size of observed data to function effectively. Among the methods, STK was identified as the most robust, particularly well-suited for dealing with chronological patterns of missing data, while RF was found to be more efficient when faced with random patterns of missing data. Although model evaluation typically focuses on performance and error measures, this study emphasized the importance of not solely relying on these metrics without also checking for consistency. Notably, the RF algorithm was found to have overestimated daily precipitation outside of the validation period in certain instances, largely due to an over-detection of rainy days when time-dependent missing patterns were present.

Chuan et al. (2020) evaluated the effectiveness of a Probabilistic Principal Component Analysis (PPCA) Model and Expectation Maximisation Algorithm (EM Algorithm) in treating missing daily rainfall data. The main goal of this research is to introduce a fresh multiple-imputation technique for managing gaps in rainfall data. This novel approach involves utilizing the M-component probabilistic principal component analysis model in conjunction with the expectation-maximization algorithm (MPPCA-EM).The analysis outcomes indicate that for handling absent historical daily rainfall data from coastal monitoring stations, the 2-component probabilistic principal component analysis model combined with the expectation-maximization algorithm (2PPCA-EM) demonstrated better performance compared to the single and multiple imputation methods suggested in earlier research. Conversely, the single imputation techniques from prior studies outperformed the MPPCA-EM imputation methods for addressing missing historical daily rainfall data collected from inland monitoring stations.

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

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

To conduct a meaningful analysis of rainfall data, it is crucial to predict the missing data within the daily rainfall dataset (Jahan et al., 2019). Several methods are available for addressing gaps in hydrological time series, specifically rainfall data, but the suitability of each approach hinges on factors like available information and the extent of gaps within the time series that needs to be completed (Sahoo & Ghose, 2022). Hence, when managing with missing data, approach is always crucial as mishandling missing data could lead in deriving inaccurate inferences (Emmanuel et al., 2021).

Building upon previous research on missing data, this study will implement a variety of imputation methods to effectively address and manage missing values in the dataset. Recognizing the critical importance of accurate imputation for maintaining the integrity and reliability of the data, the study will apply a comprehensive suite of imputation techniques. These techniques include Mean Imputation, which substitutes missing values with the average of available data, and Median Imputation, which uses the median as a more robust measure, particularly in the presence of outliers. The k-Nearest Neighbor (kNN) method will also be employed, leveraging the proximity of data points in the dataset to estimate missing values based on similar neighboring observations.

In addition to these methods, the study will incorporate more advanced techniques such as the Markov Chain Monte Carlo (MCMC) method, which uses a probabilistic model to generate multiple imputations based on the distribution of the data. The Non-linear Iterative Partial Least Squares (NIPALS) algorithm will be used, particularly useful for handling missing data in high-dimensional datasets. The Expectation-Maximization (EM) Algorithm, a well-established statistical method, will

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

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

be applied to estimate missing values by iteratively refining parameter estimates until they converge to the most likely values.

Finally, the study will utilize Random Forest, a powerful ensemble learning method, to predict and impute missing values by building multiple decision trees and averaging their predictions. By applying this diverse array of imputation methods, the study aims to ensure that the missing data is handled with the highest possible accuracy, thereby enhancing the overall quality and robustness of the research findings.

2.4 High-Dimensionality Data

According to Tahvili and Hatvani (2022), high-dimensional data involves datasets where the number of features or covariates exceeds the quantity of observations, whereas low-dimensional data exhibits the reverse, with a surplus of observations compared to features. In other words, high-dimensional data can occur when there are so many variables per subject (Mihalik et al., 2022). Data that were collected are often high-dimensional, censored, heterogenous and contain missing information, presenting challenges to traditional statistical methods (Spooner et al., 2020). The complexity of rainfall prediction studies stems from the interdependency of various hydrological cycle components and their close connection to multiple causal factors, making the dimensionality of the issue a persistent concern (Das, Sachindra & Chanda, 2022). High-dimensional inputs must be translated to inputs with lower dimensions in order to reduce the dimensionality, making neighbouring points on the manifold correspond to similar points in the input space (Reddy et al., 2020). To retain the essence of the data,

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

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

dimensionality reduction methods transform the initial high-dimensional dataset into a condensed low-dimensional form, thereby mitigating the challenges posed by the curse of dimensionality (Zebari et al., 2020).

For decades, researchers used dimensionality reduction approaches for different reasons such as to reduce computational processing overhead, reduce noise in the data, and for the better data visualization and interpretation (Fodor, 2002; Sorzano, Vargas & Montano, 2014; Abdulhammed et al., 2019). Research by Zebari et al. (2020) stated that dimensionality reduction is the operation of transforming the high dimensional representation of data in low dimensional representations. Meanwhile, Hasan & Abdulazeez (2021) described that dimensionality reduction approach is an approach that aims to reduce the distance in a latent space between distributions of different data sets to allow efficient transfer learning.

Dimensionality reduction approach usually implemented in preprocessing phase before going to the training phase (Prottasha et al., 2023). In preprocessing phase, the high-dimensional data will transform to low-dimensional data and also to fix missing data and eliminate noise (Hsu & Sharma, 2023). Therefore, implementing dimensionality reduction approach could give a lot of benefits which include: (i) reducing dimension count and conserving storage space, (ii) expediting computation, (iii) eliminating irrelevant, noisy, and redundant data, (iv) potentially enhancing data quality, (v) promoting algorithmic efficiency and accuracy, (vi) enabling data visualization, and (vii) simplifying classification processes while bolstering overall performance (Liu & Zsu, 2009; Juvonen, Sipola & Hämäläinen, 2015; Hasan & Abdulazeez, 2021).

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

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

According to Reddy et al. (2020), Principal Component Analysis (PCA) is a statistical procedure which uses as orthogonal transformation and converts a group of correlated variables to a group of uncorrelated variables. The main purpose of PCA is to reduce the dimension of a dataset including a large number of information available and this can be implemented by transforming to a new set of variables which is called principal component (Hassan, Abdulazeez & Tiryaki, 2018). Salo, Nassif & Essex (2018) proposed a novel hybrid dimensionality reduction, information gain and principal component analysis (IG-PCA) to cope with large-scale and heterogeneous data and it exhibited better performance regarding classification accuracy, detection rate and false alarm rate.

Subsequently, Abdulhammed et al. (2019) implemented PCA as dimensionality reduction approach for machine learning based network intrusion detection and the result confirmed that the proposed framework with the reduced feature dimensionality achieved a maximum precision value, confirming the efficiency and effectiveness of the intrusion detection process. Sulaiman et al. (2022) also applied PCA to reduce the number of predictors by extracting the number of principal components without losing any significant information for rainfall datasets, resulting in the same number of components as the initially chosen variables. Thus, the effectiveness of the PCA algorithm in dimension reduction for large and complex datasets, along with its adaptability through various modifications, is evident in its accurate data compression and enhanced classification capabilities across diverse fields (Hasan & Abdulazeez, 2021).

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

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

Bharadiya (2023) delved into the classification challenge of anomaly detection in network data using machine learning techniques, focusing on how to effectively manage and analyze large datasets. The study specifically used the KDD99 dataset, a well-known benchmark for network intrusion detection systems (IDS), to explore the impacts of dimensionality reduction and classification methods. In the context of network data applications, the study employed Principal Component Analysis (PCA) for dimensionality reduction and Support Vector Machine (SVM) for classification, critically assessing the outcomes of these methods. The results of Bharadiya's analysis highlighted significant improvements in the efficiency and accuracy of the classification process when PCA was used to reduce the dimensionality of the input data. Specifically, the reduction in data dimensions led to a decrease in execution time, making the classification process faster and more efficient. Moreover, the precision and recall metrics for the SVM classification algorithm showed enhanced accuracy, with a notable reduction in the rate of misclassification when paired with PCA. This demonstrates that the SVM combined with PCA not only optimizes computational efficiency but also improves the reliability of anomaly detection in network data. Bharadiya also pointed out the broader implications of this approach in other fields, such as health research, where large datasets are increasingly common but challenging to interpret. By applying PCA, researchers can reduce the complexity of these datasets, making them easier to analyze while retaining most of the critical information, ultimately enhancing the interpretability and value of data-driven studies.

Costa et al. (2020) conducted a study to assess the effectiveness of various dimensionality reduction methods within the framework of an additive-dominant model, comparing these methods against the genomic best linear unbiased prediction

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

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

(G-BLUP) method. The study specifically evaluated three dimensionality reduction techniques: Principal Components Regression (PCR), Partial Least Squares (PLS), and Independent Components Regression (ICR). To conduct their analysis, the researchers used a simulated dataset comprising 1,000 individuals and 2,000 single-nucleotide polymorphisms (SNPs). The dataset was examined across four different scenarios, which varied based on two levels of heritability and two types of genetic architecture. In their analysis, Costa et al. explored how well these dimensionality reduction methods could capture additive, dominant, and total genomic information. They found that PCR generally yielded higher accuracy values compared to PLS and ICR, suggesting it may be more effective in certain contexts. However, despite PCR's relatively better performance, the study revealed that none of the evaluated methods, including G-BLUP, were capable of accurately recovering true genomic heritabilities. Each method produced biased estimates, either underestimating or overestimating the genomic genetic values, which highlights the challenges and limitations associated with using these dimensionality reduction techniques in genomic prediction models. The findings underscore the need for further refinement and development of methodologies that can more accurately predict genetic values without introducing significant biases.

Portnova-Fahreeva et al. (2020) conducted a study with the objective of identifying an efficient way to represent hand kinematics data that could improve the control of prosthetic hands. The study focused on comparing two different methods for capturing the essential characteristics of a wide range of hand gestures and actions: Principal Component Analysis (PCA) and a non-linear Autoencoder Network (nAEN). The researchers evaluated the performance of these methods based on several criteria, including their ability to accurately reconstruct hand kinematic data from a latent

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

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

manifold with reduced dimensions, the distribution of variance across these latent dimensions, and the separability of hand movements in both compressed and reconstructed representations when analyzed using a linear classifier. The results of the study revealed that the nAEN outperformed PCA in reconstructing hand kinematic data from a reduced-dimensional latent manifold. Specifically, while PCA could account for 78% of the input data variance within two dimensions of the latent manifold, the nAEN managed to capture 94% of the variance, demonstrating its superior ability to preserve the essential information in the data. Moreover, the nAEN's latent manifold exhibited a more uniform distribution of signal variance across its dimensions compared to PCA, suggesting that it provided a more balanced representation of the data. Additionally, when the reconstructed hand movements were classified using a linear classifier, such as SoftMax regression, the nAEN was able to produce a manifold where different tasks were more distinguishable than those produced by PCA. The study concludes that non-linear dimensionality reduction techniques, like the nAEN, may offer a more effective platform for controlling prosthetic hands compared to traditional linear methods such as PCA. This is particularly important for applications where precise and reliable control of prosthetic devices is crucial, as the ability to accurately and distinctly represent complex hand movements could significantly enhance the functionality and user experience of prosthetic hands.

Shirmard et al. (2020) conducted an in-depth study focusing on the integration of selective dimensionality reduction techniques for mineral exploration, particularly using ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) satellite data. The study aimed to evaluate and compare the effectiveness of three widely used dimensionality reduction methods: Principal Component Analysis (PCA),

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

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

Independent Component Analysis (ICA), and Minimum Noise Fraction (MNF). These techniques were not only assessed individually but also integrated to explore how they could be jointly applied to enhance remote sensing data analysis. The specific goal was to map hydrothermal alteration zones associated with epithermal copper-gold (Cu–Au) deposits in the Toroud-Chahshirin range, located in Central Iran. The researchers applied these dimensionality reduction techniques to carefully selected subsets of ASTER spectral bands. These bands were particularly useful for identifying gossans and various hydrothermal alteration zones, including argillic, propylitic, and phyllic zones, which are key indicators of potential mineral deposits. To integrate the most effective thematic layers derived from the transformation techniques, the study employed a fuzzy logic model. This integration process resulted in a highly efficient remote sensing evidential layer that significantly enhanced mineral prospectivity mapping. The findings of the study highlighted ICA as the most robust technique among the three for generating hydrothermal alteration thematic layers. ICA's superior ability to separate source signals from noise proved crucial in the enhanced detection and mapping of geological features, particularly the specific alteration zones associated with mineral deposits. The study demonstrated that by integrating these dimensionality reduction techniques, particularly emphasizing the strength of ICA, the remote sensing analysis could be substantially improved, leading to more accurate and effective mineral exploration efforts.

Mazher (2020) conducted an in-depth study focusing on the development of a visualization framework for high-dimensional spatio-temporal hydrological gridded datasets, utilizing machine learning techniques to enhance the understanding and analysis of such complex data. The study introduced a comprehensive and generic

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

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

workflow designed to achieve three main objectives: first, to project high-dimensional spatio-temporal data onto a two-dimensional (2D) plane with accuracy; second, to compare various dimensionality reduction techniques (DRTs) in terms of resolution and computational efficiency; and third, to spatially represent the 2D projection using a perceptually uniform background color map, which aids in visual interpretation. To accomplish these objectives, the study compared several machine learning-based DRTs commonly used for data visualization. These included Principal Component Analysis (PCA), Generative Topographic Mapping (GTM), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). Each technique was evaluated for its accuracy, resolution, and computational efficiency in handling massive datasets, which are typical in hydrological studies. The accuracy of the visualizations was assessed using a quality metric derived from a co-ranking framework, ensuring that the results were both reliable and informative. The workflow was applied to a specific case study involving an output from the Australian Water Resource Assessment (AWRA) model, focusing on the region of Tasmania, Australia. The dataset used in this study consisted of daily time series data for nine components of the water balance, recorded at a 5 km grid cell resolution over the entire year of 2017. The case study revealed several key insights: PCA was found to be highly effective for rapidly visualizing global data structures, making it suitable for getting an overall sense of the data. However, for more detailed and accurate representations of local trends within the data, t-SNE and UMAP were superior. Among these, UMAP stood out as being computationally more efficient than t-SNE, and it was also less affected by outliers compared to GTM, making it a robust choice for visualizing complex hydrological datasets. The study by Mazher underscores the importance of selecting appropriate DRTs based on the specific requirements of the dataset and the desired

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

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

outcomes of the visualization. By providing a clear comparison of these techniques, the research offers valuable guidance for scientists and engineers working with high-dimensional spatio-temporal data, particularly in the field of hydrology, where such data is increasingly common and critical for resource management and environmental studies.

Based on previous research about high-dimensional data, this study will implement a dimensionality reduction approach to effectively manage and simplify the complexity inherent in datasets with a large number of variables. High-dimensional data can pose significant challenges in data analysis, including issues such as overfitting, increased computational costs, and difficulties in visualizing and interpreting relationships within the data. To address these challenges, this study will apply Principal Component Analysis (PCA), a widely recognized and powerful technique for dimensionality reduction.

PCA works by transforming the original high-dimensional dataset into a new set of variables called principal components. These components are linear combinations of the original variables and are designed to capture the maximum variance within the data, while also reducing the number of dimensions. The first principal component captures the most significant variance, and each subsequent component captures the remaining variance in decreasing order of importance. By focusing on the principal components that account for the majority of the data's variance, PCA allows the study to retain the most critical information while discarding the less important details, thus reducing the dimensionality of the dataset.

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

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

The application of PCA in this study is particularly crucial for enhancing the efficiency and effectiveness of data analysis. By reducing the number of variables, PCA helps to mitigate the risk of overfitting, where models become too tailored to the training data and perform poorly on new, unseen data. Additionally, PCA reduces the computational burden associated with analyzing high-dimensional data, making it more feasible to conduct complex analyses within a reasonable timeframe. Furthermore, the reduced dimensionality makes it easier to visualize and interpret the relationships and patterns within the data, facilitating more insightful and accurate conclusions.

Overall, the application of PCA as a dimensionality reduction technique in this study is expected to significantly enhance the quality and robustness of the research findings, ensuring that the analysis remains focused on the most relevant and informative aspects of the data. This approach not only strengthens the study's outcomes but also contributes to advancing the broader field by demonstrating the value of dimensionality reduction in managing and interpreting complex data sets.

2.5 Zero-Inflated Data

According to Gramosa, Nascimento and Morales (2020), in numerous practical scenarios, it is common to encounter with excessive zero values in the datasets, posing challenges for precise statistical analysis as standard models are not designed to accommodate this case. In big data analysis, Jun (2023) also stated that the datasets frequently contain a significant number of zero values, which profoundly affect the analysis; failure to adequately account for excessive zero values can result in an

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

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

unsuitable fit for both zero and nonzero counts as stated by Feng (2021). Thus, to capture zero-inflated data, the Bayesian approach is recommended due to its ability to provide mechanisms that facilitate estimation and yield efficient results (Choi, Chapkin & Ni, 2020). Workie and Azene (2021) described that Bayesian approach allows for parameter estimation through Markov model simulation that can generate random values and thus offering greater flexibility in estimation.

Kang et al. (2023) studied about analysis of zero-and-one inflated bounded count time series with applications to climate and crime data. They proposed a novel variant of the first-order binomial autoregressive (BAR(1)) process that incorporates zero-and-one inflated binomial marginals using the concept of hidden Markov models. This new model encompasses the BAR(1) and other existing processes as particular instances. The stochastic properties of the model are explored, and its parameters are estimated using probability-based methods, quasi-maximum likelihood, maximum likelihood, and Bayesian techniques. The findings of their study demonstrated that the proposed model effectively captures the zero-inflation, one-inflation, and overdispersion characteristics of the data. The predictive distributions are utilized to detect anomalies and subsequently develop an early warning system for risk.

Research by Hughes, Guttorp and Charles (1999) proposed a non-homogeneous hidden Markov model (NHMM) that aims to link precipitation events across various rain-gauge stations with large-scale atmospheric circulation patterns. They analyzed a 15-year sequence of winter data from 30 rain stations in southwestern Australia, utilizing the initial 10 years for model development and the subsequent 5 years for model evaluation. The result showed that despite a shift in atmospheric circulation and

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

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

consequent rainfall patterns between the two periods, the fitted model effectively replicates observed rainfall statistics in the reserved data and offers valuable insights into the underlying processes influencing rainfall in this area.

Long et al. (2019) developed an operational scheme for forecasting precipitation, harnessing the Bayesian Non-Homogeneous Hidden Markov Chain Model (NHMM), which affords the modeling of non-stationary transition probabilities and emission matrices. Their model's efficacy was exemplified through its application in analyzing precipitation across multiple stations in NingXiang County, an area situated inland in southern China and characterized by dense population. The findings of their research suggest that the developed model adeptly captures the varied precipitation patterns observed at different stations. Furthermore, the proposed model exhibits versatility, with potential applications extending beyond the specific case study, enabling the simulation and generation of stochastic monthly precipitation data across diverse geographical regions.

Similarly, Stoner and Economou (2020) undertook the task of modifying a hidden Markov model to craft a sophisticated framework for simulating sub-daily rainfall patterns. Their adapted model exhibits a remarkable capacity to encapsulate key features of sub-daily rainfall dynamics, encompassing extended dry spells or droughts, as well as the nuanced seasonal and temporal variations in both occurrence and intensity, along with the predisposition towards extreme values. These modifications encompass enhancements such as the incorporation of clone states and temporally non-uniform state persistence probabilities. The outcomes derived from this model not only yield valuable insights but also facilitate a comprehensive examination of the diurnal,

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

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

seasonal, and annual fluctuations inherent in sub-daily rainfall occurrences and intensities, thereby offering a deeper understanding of the intricate dynamics governing rainfall patterns at finer temporal scales.

In a study by Wang, Asefa and Sarkar (2020), they proposed a novel semiparametric Bayesian non-homogeneous hidden Markov model, uniquely designed to explicitly integrate the influences of observed covariates, such as climate indices, on both the state transition probabilities and the distribution of observation emissions. This novel model underwent rigorous testing across three rainfall monitoring stations situated within the service jurisdiction of Tampa Bay Water, a prominent regional water supply agency in the Southeastern United States. The findings unveiled the model's adeptness in accurately capturing the historical properties of rainfall patterns at these locations, thus positioning it as a promising alternative for furnishing operational rainfall forecasts that hold significant practical utility.

Nyongesa, Zeng and Ongoma (2019) also employed a non-homogeneous Markov model to analyze daily rainfall patterns across 16 synoptic stations distributed throughout Kenya. This analytical framework operates under the assumption that the occurrence of diurnal rainfall events at these monitoring sites is influenced by latent, or unobserved, “weather states”. These states' progression is modeled according to a first-order Markov criterion, wherein the probabilities of transitioning from one state to another are contingent upon certain atmospheric variable indices. Through the application of the Bayes information criterion (BIC), five distinct states were identified for inclusion in the model. To refine the analysis and provide detailed insights into daily rainfall occurrences across the network of 16 stations, the non-homogeneous Markov

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

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

model utilized outputs from global circulation models (GCMs), specifically projections of daily precipitation and sea surface temperatures pertinent to the study period. The outcomes derived from this study underscore the suitability of the modeling tool for statistically downscaling daily rainfall occurrences across a multisite station network, thereby presenting valuable implications for various domains such as drought and flood preparedness, water resource management, and the refinement of inputs for crop models.

Furthermore, Guo et al. (2022) utilized a nonhomogeneous Markov model (NHMM) to simulate daily summer rainfall patterns, incorporating future projections. Through this investigation, they underscored the efficacy of the NHMM algorithm, showcasing its ability to accurately capture the coherent relationships between regional rainfall dynamics within the middle and lower reaches of the Yangtze River basin and the prevailing large-scale atmospheric circulation features surrounding the region. The findings highlighted the NHMM's exceptional capacity to elucidate the intricate interplay between rainfall patterns and the corresponding regional atmospheric circulation regimes. Similarly, Jiang et al. (2023) devised an innovative NHMM variant aimed at enhancing the accuracy of reproducing spatiotemporal precipitation characteristics at specific geographic locations, particularly focusing on extreme precipitation events. The outcomes of their research, conducted within the same study area, underscored the model's high efficiency and superior performance in simulating precipitation variability and extreme events. Concurrently, Cao et al. (2023) employed a Bayesian-NHMM approach to forecast daily rainfall during the rainy season across the same study area. Their findings demonstrated the favorable performance of

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

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

Bayesian-NHMM in capturing the seasonality and interannual variability of precipitation across diverse climatic conditions.

In a related study, Ramadhan and Devianto (2020) utilized a hidden Markov model to forecast rainfall data availability at a weather station in West Sumatra. The study aimed to predict hidden states of rainfall data availability using the Baum-Welch Algorithm within the hidden Markov model framework. The results demonstrated that the hidden Markov model is effective for forecasting the probability of rainfall data being available. Specifically, the predictions indicated that for the upcoming one-day period, the likelihood of rainfall data availability from three weather stations could be accurately assessed using the Viterbi algorithm.

Tan, Yusof, and Yusop (2013) highlighted the significance of the nonhomogeneous Markov model (NHMM) in generating rainfall patterns based on a limited number of weather states, which act as intermediaries linking the large-scale atmospheric conditions. Motivated by this, they employed NHMM to analyze daily rainfall data collected from 20 stations across Peninsular Malaysia during the northeast monsoon season. Their findings revealed that NHMM exhibited a commendable ability to accurately replicate various aspects of observed daily mean rainfall, including correlations between rainfall amounts recorded at different stations and the quantile-quantile plot. Consequently, it was deduced that NHMM emerges as a valuable tool for simulating daily rainfall amounts, offering insights crucial for devising strategies and plans to mitigate unforeseen disasters such as floods and droughts. However, in a subsequent study conducted in 2017, Tan, Yusof, and Yusop sought to further explore the modeling capabilities, this time employing a homogeneous Markov model to

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

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

examine the variability of rainfall patterns and atmospheric circulation across 40 rainfall monitoring stations in Peninsular Malaysia. The outcomes of this investigation demonstrated the model's efficacy in depicting the spatiotemporal variability inherent in monsoon rainfall patterns. Thus, it can be inferred that the Bayesian approach holds promise in capturing the intricate temporal dependencies and transitions evident in daily rainfall patterns, thereby contributing to enhanced understanding and predictive capabilities in hydrological and climate modeling endeavors.

Drawing from previous research on zero-inflated data, this study applies a Hidden Markov Model (HMM) within a Bayesian framework to effectively manage and account for the uncertainties associated with zero-bounded datasets. Zero-inflated data, characterized by an excess of zero values, presents unique challenges for statistical modeling, particularly in accurately capturing the underlying distribution and variability of the non-zero observations. By employing an HMM in a Bayesian context, the study aims to address these challenges by providing a robust approach to quantify parameter uncertainties. The Bayesian approach facilitates a comprehensive assessment of the uncertainties related to the model parameters by incorporating prior distributions and updating them with observed data to produce posterior distributions. This methodology allows for a more nuanced understanding of the data's structure and variability, improving the model's ability to handle the peculiarities of zero-bounded data. Through this approach, the study seeks to enhance the accuracy and reliability of predictions and inferences derived from zero-inflated datasets, thereby advancing the overall effectiveness of statistical analyses in contexts where zero values are prevalent.

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

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

2.6 Non-Linearity Data

In reality, conceptually based methods that rely on historical input data can generate non-linear correlations between the simulated parameters and the estimators without the necessity for prior knowledge of the flow's nature (Hadi et al., 2019). Capturing non-linearity in data using machine learning-based statistical downscaling involves developing models that can accurately predict relationships between atmospheric variables and local-scale observations (Ian et al., 2023). Loecher et al. (2023) also stated that machine learning models can find trend and predict outcomes by exploiting and learning from large, complex and non-linear data sets.

Bowden, Foster and Parkes (2023) discovered that the connections between weather and agricultural productivity are intricate and nonlinear; they observed that the machine learning model can uncover intricate non-linear relationships and interactions between climate and variations in rice production, which depend on factors like extreme weather timing and magnitude during the growth period. Meanwhile, according to Shakya (2019), machine learning methods are utilized within optical fiber cables to diminish non-linearity, offering a more favourable and economically efficient approach for monitoring various impairments in optical networks. Peng and Nagata (2020) applied machine learning model for an empirical overview of non-linearity and overfitting using COVID-19 data and the result showed the model that incorporates the highest degree of non-linearity had the best performance.

In hydrology, Hadi et al. (2019) applied machine learning models to encounter several difficulties in predicting streamflow, stemming from the intricate dynamics of

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

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

various natural factors like non-stationarity, non-linearity, and unpredictability, posing considerable challenges. Kumar et al. (2021) introduced a deep learning-based downscaling of summer monsoon rainfall data as rainfall variability is dependent on the complex spatio-temporal process leading to non-linear or chaotic spatio-temporal variation and it was found that the proposed model provided better information on rainfall intensity. He et al. (2022) proposed two machine learning models for predicting runoff, which took into account the non-linear and real-time characteristics of hourly rainfall and runoff data; their findings demonstrated that these models effectively enhanced the model's generalization capacity and provided improved guidance in identifying optimal non-linear relationships.

Furthermore, Li et al. (2021) introduced a regression-based ensemble method called the Variational Bayesian Long Short-Term Memory network (VB-LSTM) to tackle model structural uncertainty across various fields. This approach utilizes an advanced variational inference (VI) algorithm, which offers faster and more scalable performance compared to the Bayesian Markov Chain Monte Carlo (MCMC) method, to approximate the posterior distributions of numerous parameters within LSTM networks. Their research showed that the VB-LSTM model effectively reduces the impact of member selection and combines different model components to create a more robust model than those achievable through Bayesian Model Averaging (BMA). The results indicate that VB-LSTM, with its sophisticated architecture and extensive parameter set, is capable of capturing non-linear relationships and temporal dependencies effectively. The extent of the improvement in model performance correlates with the strength of the non-linearities and time dependencies in the data.

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

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

Next, Jose and Yasala (2024) introduced a hybrid model that combines the Hidden Markov Model (HMM) with Artificial Neural Networks (ANN) to enhance predictions of groundwater quality in the Kanyakumari District, India. This integrated approach leverages the strengths of both models: HMM is used to capture sequential patterns and state transitions, while ANN excels at identifying complex relationships within individual features. By merging these models, the study achieves a more nuanced understanding of groundwater quality dynamics and improves the accuracy of quality assessments. The HMM-ANN hybrid model demonstrated a high prediction accuracy of 97.41%, illustrating its effectiveness. This innovative methodology advances groundwater quality evaluation and supports better decision-making for water resource management and environmental conservation. Thus, it is confirmed that machine learning's strength is that it can develop non-linear correlation in data and interactions between predictors (Sumayli, 2023).

Building on prior research, it has been observed that there has been a notable absence of studies exploring the integration of Bayesian approaches with machine learning-based statistical downscaling for rainfall prediction. To address this gap, this study proposes a novel hybrid model name Statistical Downscaling-based Homogeneous Markov Model-Machine Learning model, referred to as SD-based HMM-ML, to enhance the accuracy of rainfall forecasts.

The Homogeneous Markov Model is adept at managing parameter uncertainties related to zero-inflated data, which is crucial for handling datasets where zero values are prevalent. By incorporating HMM, the model can effectively account for the variability and uncertainties inherent in such data. On the other hand, SD-based ML

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

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

techniques are designed to capture complex, non-linear relationships between predictors and the variable being predicted (predictand). This capability is essential for accurately modeling the intricate interactions that influence rainfall patterns.

By combining these two advanced methodologies, the SD-based HMM-ML model aims to leverage the strengths of both approaches. The HMM component provides robust handling of data uncertainties and sequence transitions, while the SD-based ML component enhances the model's ability to understand and predict non-linear relationships. This hybrid approach is expected to significantly improve the precision of rainfall predictions, offering a more comprehensive and reliable forecasting tool.

Overall, the integration of SD-based HMM-ML represents a significant advancement in the field of rainfall prediction, addressing the limitations of existing models and paving the way for more accurate and insightful weather forecasts. This innovative approach not only fills a critical research gap but also holds the potential to contribute to more effective water resource management and climate adaptation strategies.

2.7 Model Performance Metrics

Performance metrics are essential parts of the frameworks used for evaluation in numerous fields. Botchkarev (2019) defined performance metrics as a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted. In machine learning regression experiments,

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

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

performance metrics are employed to assess the alignment between model predictions and actual observed data from the test dataset, a factor that directly influences the selection of suitable machine learning algorithms (Makridakis, Spiliotis, & Kopoulus, 2018).

To effectively analyze and evaluate the predicted results by different models, model performance metrics is applied to evaluate the anomalies of data prediction accuracy and measurement of performance (Ian et al., 2023). RMSE is applied for evaluating the efficiency and performance of different prediction models (Chai et al., 2013; Savage et al., 2013; Chai & Draxler, 2014; Sahoo & Ghose, 2022). Consequently, a smaller RMSE value indicates superior performance of a method (Addi et al., 2022). The RMSE quantifies the magnitude of average squared errors, resulting in a heightened sensitivity towards larger discrepancies (Duarte et al., 2022).

Rafik et al. (2023) summarised that The Nash-Sutcliffe efficiency coefficient (NSE) is an extensively utilized statistical metric designed to assess the performance of hydrological models. Introduced by Nash and Sutcliffe in 1970, this coefficient provides a quantitative evaluation of how well a model's simulated streamflow matches the observed data. The NSE value ranges from negative infinity to 1, where a value close to 1 signifies an excellent agreement between the observed and simulated stream flows, indicating that the model performs very well (Agarwal et al., 2023). Conversely, an NSE value close to 0 suggests poor model performance, implying that the model simulations do not closely replicate the observed data.

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

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

Next, Agarwal et al. (2023) indicated that they calculated errors in predictions at each point by determining the variance between observed values and predicted outcomes. The average of these calculated errors provided the MAE for each predicted dataset. The MAE, which ranges from 0 to positive infinity, ideally approaches 0, signifying perfect accuracy where predicted values precisely match the target values (Van et al., 2022).

Lastly, Agustian et al. (2020) described Mean Forecast Error (MFE) as the average error obtained by dividing the total errors by the number of forecasted periods, serves as an effective means to determine if forecasted values over a given period are excessively high or low, computed by summing all forecast errors across forecasting periods and dividing by the number of periods. MFE typical ranges from negative to positive infinity, with zero representing perfect accuracy (Satrio et al., 2021).

In a nutshell, to analyze the performance of the proposed model in this study, RMSE, NSE, MAE and MFE will be applied. These models can measure the accuracy of the proposed model and the stability of the model proposed in this study.

2.8 Conclusion

In this chapter, an in-depth discussion is provided on the literature review, drawing on previous research that addresses several critical aspects of data analysis and modeling relevant to this study. The chapter begins by exploring the issue of missing data, a common challenge in many datasets, particularly in meteorological studies. It delves

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

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

into various imputation methods that have been employed in previous research to address these gaps, evaluating their effectiveness and suitability for different types of data. Additionally, the chapter examines the complexities of zero-inflated data, where a significant proportion of observations are zero, which poses unique challenges for analysis and modeling. To tackle these challenges, the chapter reviews dimension reduction approaches that are essential for managing high-dimensional data, ensuring that the most relevant information is retained while reducing the computational burden. Beyond these topics, the chapter also discusses previous studies that have utilized Bayesian approaches, highlighting their application in capturing uncertainties and improving model predictions. Furthermore, the chapter provides a comprehensive overview of statistical downscaling techniques, particularly those that integrate machine learning, offering insights into how these methods have been used to refine predictions and improve model accuracy. The discussion is rounded out by a review of model performance metrics, which are crucial for evaluating the effectiveness of various models in different research contexts. By synthesizing these diverse strands of literature, the chapter not only underscores the challenges and methodologies pertinent to this study but also situates the research within a broader academic context, providing a robust foundation for the subsequent analysis and discussion. Finally, the following is a summary of the methodology and contributions of previous studies, organized by sub-topics and publication year.

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

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

Table 2.1

Summary of methodology and contributions of previous studies

No.	Year Published	Author	Title	Methods	Contribution
2.2 Statistical Downscaling-based Machine Learning Model					
1.	2019	Al-Mukhtar & Qasim	Future predictions of precipitation and temperature in Iraq using the statistical downscaling model	SDSM	The study demonstrated that SDSM performed satisfactorily in simulating future maximum and minimum temperatures and precipitation.
2.	2021	Farikha et al.	Projection pursuit regression in statistical downscaling model using artificial neural network for rainfall prediction	PPR, ANN	The study concluded that PPR is effective in reducing dimensions, especially for non-linear data, and that combining PPR with ANN yields smoother performance compared to PPR alone.
3.	2021	Kouadri et al.	Performance of machine learning models in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)	MLR, RF, M5P, RSS, AR, ANN, SVR & LWLR	RF model emerged as the top performer in predicting WQI on total hardness and total dissolved solids in the study area.

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

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

No.	Year Published	Author	Title	Methods	Contribution
4.	2021	Sun & Lan	Statistical downscaling of daily temperature and precipitation over China using deep learning neural models: localization and comparison with other methods	CNN, GLM, BCSD & BCCI	The study found that bias correction methods excelled in reducing biases and representing temporal correlations, especially precipitation, while the CNN model holds significant potential for downscaling.
5.	2021	Xu et al.	Spatial downscaling of land surface temperature based on a multi-factor geographically weighted machine learning model	MFGWML algorithm – XGBoost, MARS & BRR	The study concluded that MFGWML successfully merges the strength of multiple regression, machine learning and GWR, effectively capturing local heterogeneity and producing reliable and robust downscaled LST data.
6.	2022	Fan et al.	Simulating the climate driven runoff in data-scarce mountains by machine learning and downscaling reanalysis data	RBF-ANN, RF, SVR & XGBoost	The study indicates that in high mountain regions, temperature primarily control runoff changes, whereas precipitation has a greater impact on runoff changes in lowland areas.

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

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

No.	Year Published	Author	Title	Methods	Contribution
7.	2022	Oh et al.	Machine learning-based statistical downscaling of wind resource maps using multi-resolution topographical data	LGBM	The model achieved RMSE values of 7% and 8% for the two Weibull parameters.
8.	2023	Alipour et al.	Statistical downscaling of global climate models for temperature trend analysis in Calgary, Canada	ANN & LSTM	The results demonstrated that LSTM model, when combined MI for data screening, outperformed the ANN model in extracting historical temperature trends. This indicates that the LSTM-based network holds promise as a tool for simulation and impact assessment in Calgary, Canada.
9.	2023	Gonzalez-Abad et al.	Using explainability to inform statistic downscaling based on deep learning beyond standard validation approaches	CNN-UNET, CNN-DeepESD & CNN-PAN	The study found that the CNN-UNET model consistently performed well across various regions and future climate scenarios, whereas the CNN-DeepESD and CNN-PAN models exhibited discrepancies.

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

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

No.	Year Published	Author	Title	Methods	Contribution
2.2.1 Application of Rainfall Prediction using Statistical Downscaling-based Hybrid Machine Learning Model					
1.	2020	Anaraki et al.	Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods	MARS, M5, WOA, LSSVM & WT (decomposing) LSSVM-WOA, LSSVM, KNN & ANN (downscaling)	The results indicate that MARS outperformed the M5 model. Among the downscaling methods, ANN and LSSVM-WOA slightly exceeded other machine learning algorithms in performance.
2.	2021	Lee & Chen	Precipitation modeling for extreme weather based on sparse hybrid machine learning and Markov chain rain field in a multi-scale subspace	MCRF	The study's results demonstrated that the application of the conditioning method and multi-scale decomposition within the sparse l_1 space significantly mitigates the impact of statistical variation caused by extreme weather conditions.
3.	2021	Sahoo et al.	Prediction of flood in Barak River using hybrid machine learning approaches: a case study	RBFNN & SVM combine with FA, ANN	SVM-FA model achieved superior performance, outperforming the RBF-FA model in terms of forecasting accuracy.

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

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

No.	Year Published	Author	Title	Methods	Contribution
4.	2022	Citakoglu & Coşkun	Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts in Sakarya Meteorological Station in Turkey	ANN, ANFIS, GPR, SVMR, KNN, VMD, DWT, & EMD	The findings revealed that VMD-GPR hybrid model excelled at 1- and 3-month time scales, while the DWT-ANN hybrid model provided the most accurate predictions for the 6-month time scale when forecasting the SPI drought indices.
5.	2022	Di Nunno et al.	Precipitation forecasting in Northen Bangladesh using a hybrid machine learning model	M5P & SVR	The hybrid M5P-SVR model outperformed the individual algorithms, demonstrating superior accuracy in precipitation prediction for the region studied, and thus, was identified as the most reliable model among those tested in the research.

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

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

No.	Year Published	Author	Title	Methods	Contribution
6.	2023	Adnan et al.	Pan evaporation estimation by relevance vector machine tuned with new metaheuristic algorithms using limited climatic data.	RVM-IMRFO	The new proposed model considerably prediction and by importing extraterrestrial radiation into the model, the inputs generally improved the prediction accuracy.
7.	2023	Baljon & Sharma	Rainfall prediction rate in Saudi Arabia using improved machine learning techniques	Improved FFANN	The findings of the proposed model illustrated the essential factors that played a crucial role during heavy rainfall and all of the extreme rainfall events that occurred during the assessment years could have been precisely predicted by the proposed model, demonstrating its efficacy as a predictor.
8.	2023	Ehteram et al.	Convolutional neural network-support vector machine-gaussian process regression: a new machine model for predicting monthly and daily rainfall	CONN-SVM-GRP	The proposed model could help researchers predict rainfall, which is essential for monitoring water resources and mitigating the impacts of droughts, floods, and other natural disasters.

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

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

No.	Year Published	Author	Title	Methods	Contribution
9.	2023	Farooq et al.	Prediction of seasonal rainfall using hybrid machine learning algorithms	W-SVM & W-RF	W-SVM consistently delivered high predictive accuracy across all selected wavelets, making it the most reliable model for seasonal rainfall prediction in this study.
10.	2023	Kanani et al.	An AI-enabled ensemble method for rainfall forecasting using long-short term memory	LSTM	The experiment results showed that the proposed approach outperforms several state-of-the-art approaches.
11.	2023	Parviz et al.	Improving hybrid models for precipitation forecasting by combining non-linear machine learning methods	GEP, SVR, & GMDH	The inclusion of nonlinear models in the improved hybrid models resulted in better performance compared to the traditional hybrid approaches.
12.	2023	Patel et al.	Rainfall prediction using machine learning techniques for Sabarmati River Basin, Gujarat, India	RF, GB & DT	The researchers claimed that the GB was found to work best among RF and DT, with features related to and affecting rainfall predictability.

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

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

No.	Year Published	Author	Title	Methods	Contribution
13.	2023	Praveena et al.	Prediction of rainfall analysis using logistic regression and support vector machine	LR & SVM	The result showed that the accuracy of LR is 88% while, the accuracy of SVM is 82%. After hyper-tuning, the accuracy of LR and SVM classifier were improved by 89% and 91% respectively.
14.	2023	Prottasha et al.	Short-term rainfall prediction using supervised machine learning	Twenty-nine different ML classifiers	The result showed that hist gradient boosting classifier from Ensemble classifiers has given the highest accuracy and suitable to be a universal and easy rain prediction tool.
			2.3 Missing Values		
1.	2019	Aieb et al.	A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria	HD, KNN, WKNN, MI, LR, & SAM	The HD, KNNI, and WKNNI consistently yield the best imputation results across varying missing data levels. Additionally, altitude, latitude, and longitude significantly impact correlation and residual averages, influencing the performance of MI, LR, and SAM methods.

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

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

No.	Year Published	Author	Title	Methods	Contribution
2.	2020	Aguilera et al.	Estimating extremely large amounts of missing precipitation data	STK, PMM & RF	Among the methods, STK was identified as the most robust, particularly well-suited for dealing with chronological patterns of missing data, while RF was found to be more efficient when faced with random patterns of missing data.
3.	2020	Chuan et al.	The effectiveness of a probabilistic principal component analysis model and expectation maximisation algorithm in treating missing daily rainfall data	2PPCA-EM & MPPCA-EM	The analysis outcomes indicate that for handling absent historical daily rainfall data from coastal monitoring stations, the 2PPCA-EM demonstrated better performance compared to the single and multiple methods suggested in earlier research.

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

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

No.	Year Published	Author	Title	Methods	Contribution
4.	2020	Nor et al.	A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia	RBM, NN, MCMC, NIPALS, & RF combined with MLR	Based on the experiment, it showed that RBM is the best result for single imputation but RF had shown its supremacy by giving the best result and more fitting for satisfying the missing data when combined with MLR.
5.	2021	Hamzah et al.	A comparison of multiple imputation methods for recovering missing data in hydrological studies	RRRI, KNN & CART combined with MLR	The findings of the study revealed that the combination of RRRI-MLR significantly outperformed the other imputation methods in terms of accuracy and reliability.
6.	2022	Addi et al.	Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana	12 imputation methods	Regression, PPCA and RF were showing a positive and excellence results for estimating the numbers of dry and wet periods and moderate to extreme rainfall values.

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

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

No.	Year Published	Author	Title	Methods	Contribution
7.	2024	Hırca & Türkkan	Assessment of different methods for estimation of missing rainfall data	Simple imputation methods & EM	The analysis indicated that the EM method outperformed the other techniques, emerging as the most suitable approach for completing missing data.
		2.4	**High-Dimensionality Data**		
1.	2020	Costa et al.	Genomic prediction with the additive-dominant model by dimensionality reduction methods	PCR, PLS & ICR	To assist in selecting the number of components, the results were analyzed based on additive, dominant, and total genomic information. Overall, PCR demonstrated greater accuracy compared to the other methods.
2.	2020	Mazher	Visualization Framework for High-Dimensional Spatio-Temporal Hydrological Gridded Datasets using Machine-Learning Techniques	PCA, GTM, t-SNE & UMAP	The study demonstrates that PCA enables quick visualization of global data patterns, while t-SNE and UMAP provide more precise depictions of local trends. Additionally, UMAP is more computationally efficient than t-SNE and less influenced by outliers compared to GTM.

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

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

No.	Year Published	Author	Title	Methods	Contribution
3.	2020	Portnova-Fahreeva et al.	Linear and non-linear dimensionality-reduction techniques on full hand kinematics	PCA & nAEN	The results of the study revealed that the nAEN outperformed PCA in reconstructing hand kinematic data from a reduced-dimensional latent manifold.
4.	2020	Shirmard et al.	Integration of selective dimensionality reduction techniques for mineral exploration using ASTER satellite data	PCA, ICA & MNF	The findings of the study highlighted ICA as the most robust technique among the three for generating hydrothermal alteration thematic layers.
5.	2023	Bharadiya	A tutorial on principal component analysis for dimensionality reduction in machine learning	PCA, SVM	By applying PCA, researchers can reduce the complexity of these datasets, making them easier to analyze while retaining most of the critical information, ultimately enhancing the interpretability and value of data-driven studies.

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

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

No.	Year Published	Author	Title	Methods	Contribution
		2.5	**Zero-Inflated Data**		
1.	1999	Hughes, Guttorp & Charles	A non-homogeneous hidden Markov model for precipitation occurrence	NHMM	The result showed that despite a shift in an atmospheric circulation and consequent rainfall patterns between the two periods, the fitted model effectively replicates observed rainfall statistics in the reserved data and offers valuable insights into the underlying processes influencing rainfall in this area.
2.	2013	Tan, Yusof & Yusop	Non-Homogeneous Hidden Markov Model for Daily Rainfall Amount in Peninsular Malaysia	NHMM	It was deduced that NHMM emerges as a valuable tool for simulating daily rainfall amounts, offering insights crucial for devising strategies and plans to mitigate unforeseen disasters such as floods and droughts.

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

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

No.	Year Published	Author	Title	Methods	Contribution
3.	2017	Tan, Yusof & Yusop	Subseasonal to multidecadal variability of northeast monsoon daily rainfall over Peninsular Malaysia using a hidden Markov model	HMM	The outcomes of this investigation demonstrated the model's efficacy in depicting the spatiotemporal variability inherent in monsoon rainfall patterns.
4.	2019	Long et al.	Monthly precipitation modeling using Bayesian non-homogenous hidden Markov chain	NHMM	It shows that the model effectively captures precipitation patterns across stations and demonstrates versatility. It can be applied beyond the case study to simulate and generate stochastic monthly precipitation data in various regions.
5.	2019	Nyongesa, Zeng & Ongoma	Non-homogenous hidden Markov model for downscaling of short rains occurrence in Kenya	GCM-NHMM	It indicates that the modeling tool is well-suited for statistical downscaling of daily rainfall occurrences across a multisite station network. The model's statistical insights are valuable for drought and flood preparedness, water resource management, and as inputs for crop models.

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

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

No.	Year Published	Author	Title	Methods	Contribution
6.	2020	Ramadhan & Devianto	A Hidden Markov Model for Forecasting Rainfall Data Availability at the Weather Station in West Sumatra	HMM	The results demonstrated that the HMM is effective for forecasting the probability of rainfall data being available.
7.	2020	Stoner & Economou	Multivariate hierarchical frameworks for modeling delayed reporting in count data	HMM	The model provides valuable insights into diurnal, seasonal, and annual rainfall variations, enabling a deeper understanding of sub-daily rainfall dynamics.
8.	2020	Wang, Asefa & Sarkar	A novel non-homogeneous hidden Markov model for simulating and predicting monthly rainfall	NHMM	The findings unveiled the model's adeptness in accurately capturing the historical properties of rainfall patterns at these locations, thus positioning it as a promising alternative for furnishing operational rainfall forecasts that hold significant practical utility.

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

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

No.	Year Published	Author	Title	Methods	Contribution
9.	2022	Guo et al.	Increase of future summer rainfall in the middle and lower reach of the Yangtze River basin projected with a nonhomogeneous hidden Markov model	NHMM	The findings highlighted the NHMM's exceptional capacity to elucidate the intricate interplay between rainfall patterns and the corresponding regional atmospheric circulation regimes.
10.	2023	Cao et al.	The predictability of daily rainfall during rainy season over East Asia by a Bayesian nonhomogeneous hidden Markov model	Bayesian-NHMM	Their findings demonstrated the favorable performance of Bayesian-NHMM in capturing the seasonality and interannual variability of precipitation across diverse climatic conditions.
11.	2023	Jiang et al.	A stacked ensemble learning and non-homogeneous hidden Markov model for daily precipitation downscaling and projection	NHMM	The outcomes of their research, conducted within the same study area, underscored the model's high efficiency and superior performance in simulating precipitation variability and extreme events.

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

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

No.	Year Published	Author	Title	Methods	Contribution
12.	2023	Kang et al.	Analysis of zero-and-one inflated bounded count time with applications to climate and crime data	BAR(1)	The findings of their study demonstrated that the proposed model effectively captures the zero-inflation, one-inflation, and overdispersion characteristics of the data.
			2.6 Non-Linearity Data		
1.	2019	Hadi et al.	Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation	XGB, ELM	The XGBELM model outperformed the stand-alone schema of both XGB and ELM models and the high-lagged schema of the XGB. It is important to indicate that the XGBELM model found to improve the prediction ability with minimum variables number.
2.	2019	Shakya	Machine learning based nonlinearity determination for optical fiber communication-review	ML	The ML models offering a more favorable and economically efficient approach efficient for monitoring various impairments in optical networks.

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

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

No.	Year Published	Author	Title	Methods	Contribution
3.	2020	Peng & Nagata	An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data	ML	The result showed the model that incorporates the highest degree of non-linearity had the best performance.
4.	2021	Kumar et al.	Deep learning–based downscaling of summer monsoon rainfall data over Indian region	SRCNN, stacked SRCNN, & DeepSD	It was found that the DeepSD provided better information on rainfall intensity.
5.	2021	Li et al.	Application of a hybrid algorithm of LSTM and Transformer based on random search optimization for improving rainfall-runoff simulation	VB-LSTM	The results indicate that the proposed model, with its sophisticated architecture and extensive parameter set, is capable of capturing non-linear relationships and temporal dependencies effectively.
6.	2022	He et al.	Shortm-term runoff prediction optimization method based on BGRU-BP and BLSTM-BP neural networks	BGRU-BP & BLSTM-BP	The findings show that these models improve generalization and enhance the identification of optimal non-linear relationships.

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

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

No.	Year Published	Author	Title	Methods	Contribution
7.	2023	Bowden, Foster & Parkes	Identifying links between monsoon variability and rice production in India through machine learning	ML	ML model can uncover intricate non-linear relationships and interactions between climate and variations in rice production, which depend on factors like extreme weather timing and magnitude during the growth period.
8.	2024	Jose & Yasala	Machine learning-based ensemble model for groundwater quality prediction: A case study	HMM-ANN	The proposed hybrid model demonstrated a high prediction accuracy of 97.41%, illustrating its effectiveness. This innovative methodology advances groundwater quality evaluation and supports better decision-making for water resource management and environmental conservation.

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

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

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the methodology employed in this research. The process starts with seven imputation methods namely Mean and Median Imputation, k-Nearest Neighbor (kNN), Random Forest (RF), Expectation-Maximization (EM) Algorithm, Markov Chain Monte Carlo (MCMC), and Nonlinear Iterative Partial Least Squares (NIPALS), followed by dimensional reduction techniques such as Principal Component Analysis (PCA), a Bayesian approach known as the Homogeneous Markov Model and Statistical Downscaling-based Machine Learning models namely Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT) and k-Nearest Neighbor (KNN). Additionally, this chapter also discusses about disadvantages of using single Bayesian Approach and Statistical Downscaling-based Machine Learning model for prediction modeling, prediction modeling of hybrid Statistical Downscaling-based Bayesian-Machine Learning model and end with model performance metrics used in this study.

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

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

3.2 Imputation Methods

According to Kwak and Kim (2017), imputation involves replacing missing values with substituted values obtained from a statistical analysis to produce a complete data set without missing for analysis. It is important to the logical relationship between data features first when filling in missing data, then establish mathematical model analysis, and lastly generating models to ensure that the data filling program is simple and reliable (Huang, 2021). Donders et al. (2006) also stated in Emmanuel et al. (2021) that the process of imputation involves replacing missing values by some predicted values, where the non-missing values data set is normally used to predict the values used to replace the missing values.

05-4506832 Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah ptbupsi

Single imputation method generates specific value for a missing real value in a data set (Khan & Hoque, 2020). Less computational cost is required for this method because the value can be obtained by mean, median or mode of the available values of that variable. In the meantime, multiple imputation is a technique that was primarily used to address the shortcomings of single imputation (Uusitalo et al. 2015; Emmanuel et al., 2021). It involves using the distribution of the observed data to approximatively represent a number of values that reflect the uncertainty surrounding the true value.

3.2.1 Mean and Median Imputation

Mean imputation is easy to implement and simple because it replaces missing value with sample mean or median or mode depending or distribution of the data (Jadhav, Pramod & Ramanathan, 2019). The formula of mean imputation can be written as:

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

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

$$P_x = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.1}$$

where P_x is the imputed mean rainfall value, X_i represents the observed rainfall values and n is the number of observed rainfall data.

With median, a constant is used to replace the missing data (Jadhav, Pramod & Ramanathan, 2019). In the event of a categorical variable, the value "1" could be used to replace all missing values.

3.2.2 *k*-Nearest Neighbor

The closest neighbours of missing values are classified by the *k*NN algorithm, and those neighbours are then used for imputation using a distance metric between instances (Maillo et al., 2017; Emmanuel et al., 2021). Below, an explanation of the *k*NN imputation is provided using the Euclidean distance metric:

$$dist_{xy} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \tag{3.2}$$

where $dist_{xy}$ is the Euclidian distance between two data *x* and *y*, *m* is the number of data attributes, X_{ik} is the value of the *k*-th attribute of the data point with the missing value and X_{jk} is the value of the *k*-th attribute of the complete data point.

Once the distances are calculated, the *k* nearest data points with the smallest Euclidean distances to the data point with the missing value are selected. These nearest neighbors will be used for the imputation. Assign weights to these *k* nearest neighbors

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

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

based on their distances. Closer neighbors receive higher weights, reflecting their greater relevance to the missing value. The weight w_j for the j-th neighbor is given by:

$$w_j = \frac{1}{dist_{xy}^2} \tag{3.3}$$

where $dist_{(x,y)}$ is Euclidian distance between the data point with the missing value and the j-th neighbor.

Finally, compute the weighted mean of the k nearest neighbors to estimate the missing rainfall value. The imputed rainfall value $\hat{R}_{missing}$ is calculated using the formula:

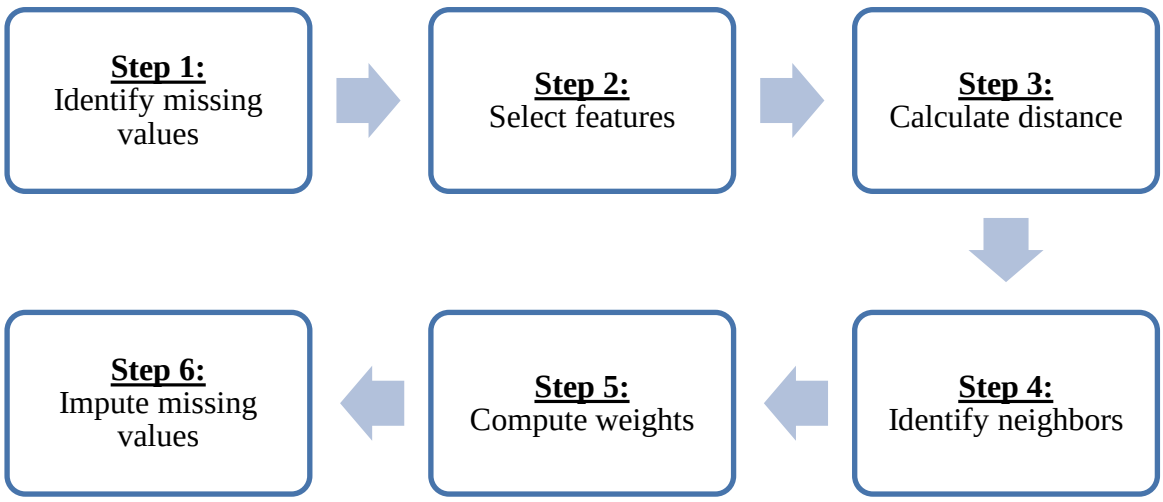
$$\hat{R}_{missing} = \frac{\sum_{j=1}^{k} w_j R_j}{\sum_{j=1}^{k} w_j} \tag{3.4}$$

where R_j are the rainfall values of the k nearest neighbors and w_j are their respective weights.

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

Figure 3.1

Steps involve in kNN imputation method



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

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

3.2.3 Random Forest

Each classification tree in the Random Forest model is built using a collection of random subsets of input data that include input and output variables (Mishra & Datta-Gupta, 2017). As expected, the remaining "out-of-bag" samples can be used to verify the effectiveness of the trained model since each tree only consumes a portion of the whole data set (Bhattacharya & Mishra, 2018). Several network characteristics are necessary for RF, including the number of maximum depths, the predictor variables at each node, and the overall number of trees.

On the observed data sets, RF was taught to anticipate the missing values. Suppose that $X = (X_1, X_2, \dots, X_p)$ is an $n \times p$ dimensional data matrix. For an arbitrary variable, X_s including missing values at entries $i_{mis}^{(s)} \subseteq \{1, \dots\dots, n\}$, we may divide the dataset into four steps (Stekhoven & Bühlmann, 2012). First of all, the observed values of variable X_s are denoted by $y_{obs}^{(s)}$. Then, the missing values of variables X are denoted by $y_{mis}^{(s)}$. Next, variables other than X_s with missing values $i_{mis}^{(s)} \subseteq \{1, \dots\dots, n\} \backslash i_{mis}^{(s)}$ are denoted by $x_{mis}^{(s)}$. It should be noted that $x_{obs}^{(s)}$ is rarely completely observed because the index $i_{obs}^{(s)}$ corresponds to the observed values of the variable x_x. Likewise, $x_{mis}^{(s)}$ is not always completely missing.

The variables $X_s, s = 1, \dots\dots, p$ by the number of missing values were arranged beginning with the fewest. The missing values X_s were imputed for each variable by first fitting an RF with response $y_{obs}^{(s)}$ and predictors $x_{obs}^{(s)}$, then predicting the missing values $y_{mis}^{(s)}$ by applying the trained RF to $x_{mis}^{(s)}$. When the disparity between the freshly

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

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

imputed data matrix and the prior one grows for the first time with regard to both variable types, the stopping creation, γ is met. For the set of continuous variables N, the difference is defined as:

$$\Delta_N = \frac{\sum_{j \in N} (X_{new,j}^{imp} - X_{old,j}^{imp})^2}{\sum_{j \in N} (X_{new,j}^{imp})^2} \tag{3.5}$$

where X_{new}^{imp} is the imputed matrix updated using predicted $y_{mis}^{(s)}$ and X_{old}^{imp} is the previously imputed matrix. The imputation procedure is repeated until criterion for stopping is met. The RF method is summarized as follows:

3.2.4 EM Algorithm

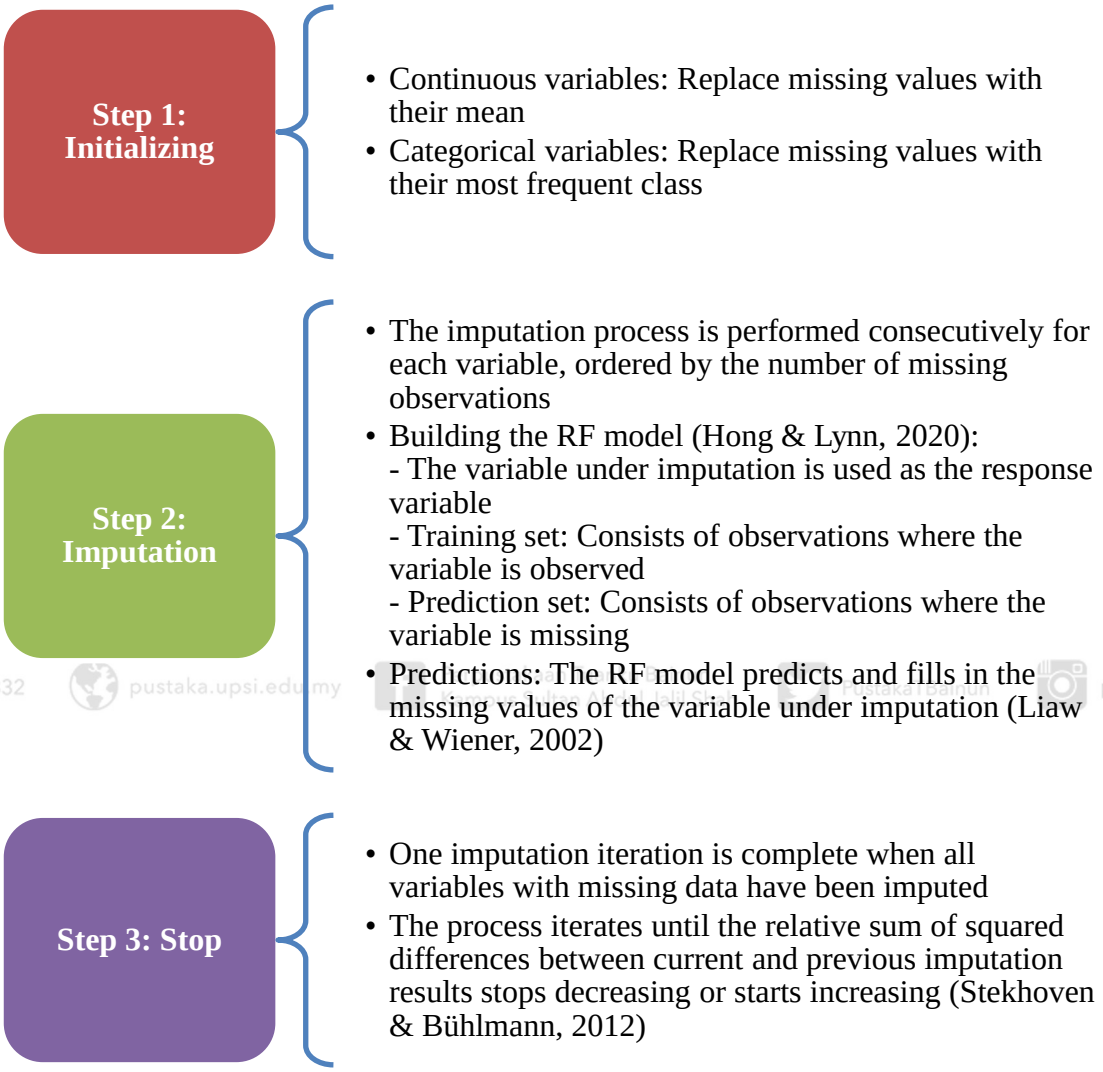
The Expectation-Maximization (EM) algorithm iteratively uses equations to estimate the mean vector and covariance matrix of a multivariate distribution from the exponential family, maximizing the log likelihood to determine parameters, especially in cases with missing values (Rumaling et al. 2020). The EM algorithm stands out among current multiple imputation methods due to its simplicity, smooth operation, and faster performance, rendering it one of the most favored choices for imputation. According to Avtar et al. (2019), it consists of two stages: the expectation step (E-step) and the maximization step (M-step).

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

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

Figure 3.2

Steps involve in RF imputation method



The process begins with the initialization step, where an initial guess for the missing values is made. This initial guess can be derived from simple statistics of the observed data, such as the mean or median for continuous variables, or it can be a random sample from the observed data. Let X_{obs} represent the observed data and X_{miss} represent the missing data. The initial guess for the missing values is denoted as $X_{miss}^{(0)}$.

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

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

In the expectation step (E-step), the expected value of the missing data is calculated given the observed data and the current estimates of the parameters. This step involves computing the expected log-likelihood of the complete data (both observed and missing) based on the current parameter estimates $\theta^{(t)}$.

$$Q\left(\theta\middle|\theta^{(t)}\right) = \mathbb{E}_{missing|observed,\theta^{(t)}}[\log p\,(\text{complete data}|\theta)] \tag{3.6}$$

Here, $Q\left(\theta\middle|\theta^{(t)}\right)$ represents the expected value of the complete-data log-likelihood. θ represents the parameters of the model, and $\theta^{(t)}$ denotes the parameter estimates at the t-th iteration. $\mathbb{E}_{missing|observed,\theta^{(t)}}$ denotes the expectation with respect to the distribution of the missing data, conditioned on the observed data and the current parameter estimates $\theta^{(t)}$. $[\log p\,(\text{complete data}|\theta)]$ is the log-likelihood of the complete data (both observed and missing) given the parameter θ. The complete data refers to the full dataset, including both the observed data and the missing data that we are trying to estimate.

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

Following the E-step, the maximization step (M-step) involves maximizing the expected log-likelihood obtained in the E-step with respect to the parameters to update the parameter estimates. This step finds the parameter values that maximize the expected log-likelihood:

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta\middle|\theta^{(t)}\right) \tag{3.7}$$

where $\theta^{(t+1)}$ represents the updated parameter estimates at the $(t+1)$-th iteration. $\arg\max_{\theta} Q\left(\theta\middle|\theta^{(t)}\right)$ means "the argument of the maximum". It indicates that we are looking for the value of θ that maximizes the function following the $\arg\max$ operator. $Q\left(\theta\middle|\theta^{(t)}\right)$ is the expected log-likelihood function calculated in the E-step.

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

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

Next, using the updated parameters from the M-step, the estimates of the missing values are updated. This step computes the conditional expectation of the missing data given the observed data and the updated parameters:

$$X_{miss}^{(t+1)} = \mathbb{E}[X_{miss}|X_{obs}, \theta^{(t+1)}] \tag{3.8}$$

where $X_{miss}^{(t+1)}$ represents the updated estimate of the missing data at the $(t+1)$-th iteration. $\mathbb{E}$ denotes the expectation (expected value) operator. It calculates the average value based on a probability distribution. X_{miss} is the missing data that we aim to estimate and $X_{obs}, \theta^{(t+1)}$ indicates that the expectation is conditional on the observed data, X_{obs} and the updated parameter estimates $\theta^{(t+1)}$.

Lastly, the algorithm iterates between the E-step and the M-step until convergence is achieved. Convergence is typically defined as the point where the changes in the parameter estimates or the imputed values between iterations are below a predefined threshold, ϵ:

$$\left|\theta^{(t+1)} - \theta^{(t)}\right| < \epsilon \tag{3.9}$$

The purpose of this convergence criterion is to determine when the algorithm should stop iterating. It provides a formal way to decide when the parameter estimates have stabilized, indicating that further iterations are unlikely to result in significant changes. The absolute difference measures how much the parameter estimates have changed between iterations; if this difference is very small, it suggests that the estimates are close to their optimal values.

In practical implementation, the algorithm updates the parameter estimates at each step and then computes the absolute difference between the current and previous

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

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

estimates. This difference is compared to the tolerance level ϵ. If the difference is less than ϵ, the algorithm stops iterating, signifying that the parameter estimates have converged. If the difference is greater than or equal to ϵ, the algorithm continues to the next iteration.

This convergence criterion ensures efficient use of computational resources by stopping the algorithm once satisfactory convergence is achieved. It also ensures that the parameter estimates are accurate to within the specified tolerance, providing reliable results for the model. Therefore, it is crucial for ensuring that the iterative process terminates appropriately, balancing efficiency and precision.

Figure 3.3

Steps involve in EM Algorithm



3.2.5 Markov Chain Monte Carlo (MCMC)

Shaharudin et al. (2020) stated that the Markov Chain Monte Carlo (MCMC) algorithm is another valuable approach for imputing missing data. This technique creates simulated random samples from probability distributions using Markov chains, where each variable's distribution depends on the preceding one within the sequence (Susianto

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

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

et al., 2017). MCMC relies on Bayesian inference, where information regarding unknown parameters is conveyed through a posterior probability distribution (Nor et al., 2020). This process involves several systematic steps, each crucial for ensuring that the imputed values are plausible and statistically sound.

The first step in MCMC imputation is to initialize the missing data with some initial values. These values can be set to the mean, median, or another statistic derived from the observed data, or they can be imputed using a simple method. Mathematically, if Y_{obs} represents observed data and Y_{mis} represents the missing data, the initial step can be represented as:

$$Y_{mis}^{(0)} = \text{initial values} \tag{3.10}$$

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun PustakaTBainun ptbupsi

Next, the Bayesian model is defined by specifying the joint probability distribution of the data, which includes the likelihood function and the prior distributions for the parameters. Let $Y = (Y_{obs}, Y_{mis})$ be the complete data and θ be the parameters. The goal is to sample from the posterior distribution $P(Y_{mis}, \theta | Y_{obs})$. Assuming Y follows a normal distribution, the likelihood function is:

$$P(Y|\theta) = \prod_{i=1}^{n} \mathcal{N}(Y_i|\mu, \sigma^2) \tag{3.11}$$

where $P(Y|\theta)$ is the likelihood function, which represents the probability of observing the data Y given the parameters θ. $\Pi_{i=1}^{n}$ denotes the product over all data points from $i = 1$ to $i = n$. It indicates that we are considering the joint probability of all n data points. $\mathcal{N}(Y_i|\mu, \sigma^2)$ is the probability density function of the normal (Gaussian) distribution for the i-th data point, given the mean μ and variance σ^2. For the parameters $\theta = (\mu, \sigma^2)$, we specify the prior distributions:

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

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

$$P(\mu) = \mathcal{N}(\mu_0, \tau_0^2) \tag{3.12}$$

$$P(\sigma^2) = \text{inverse} - \text{gamma}(\alpha_0, \beta_0) \tag{3.13}$$

Gibbs sampling, a popular MCMC technique, is then used to iteratively sample from the conditional distributions of each parameter given the current values of the other parameters and the data. There are three steps involve in Gibbs sampling. The first step is we update the parameter μ by drawing a sample from its conditional distribution given the observed data Y and the current value of σ^2. The conditional distribution of μ given Y and σ^2 is a normal distribution. Mathematically, this can be expressed as:

$$\mu^{(t+1)} \sim P(\mu|\mathrm{Y}, \sigma^2) \tag{3.14}$$

For a normal distribution, the conditional distribution of the mean μ is derived from the data. The updated sample for μ is drawn from a normal distribution with mean and variance:

$$\mu^{(t+1)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n}\frac{Y_i}{\sigma^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}\right) \tag{3.15}$$

where Y_i are the observed data points, σ^2 is the current estimate of the variance, μ_0 and τ_0^2 are the prior mean and variance of μ, respectively and n is the number of observed data. Next, we update the parameter σ^2 by drawing a sample from its conditional distribution given the observed data Y and the newly updated μ. The conditional distribution of σ^2 is an inverse-gamma distribution, which is typical for the variance of a normal distribution when using a conjugate prior. This can be written as:

$$\sigma^{2(t+1)} \sim P(\sigma^2|Y, \mu) \tag{3.16}$$

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

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

For an inverse-gamma distribution, the parameters are updated based on the data and the prior as follows:

$$\sigma^{2(t+1)} \sim \text{Inverse} - \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \mu^{(t+1)}\right)^2\right) \quad (3.17)$$

where α_0 and β_0 are the shape and scale parameters of the inverse-gamma prior for σ^2, n is the number of data points, Y_i are the observed data points and $\mu^{(t+1)}$ is the newly updated mean. The term $\alpha_0 + \frac{n}{2}$ represents the updated shape parameter, and $\beta_0 + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \mu^{(t+1)}\right)^2$ represents the updated scale parameter, incorporating the deviation of the data from the mean. In the final step, we impute the missing data Y_{mis} by drawing samples from its conditional distribution given the observed data Y_{obs}, the updated mean μ, and the updated variance σ^2. The conditional distribution of the missing data Y_{mis} is a normal distribution:

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} | Y_{obs}, \mu^{(t+1)}, \sigma^{2(t+1)}\right) \quad (3.18)$$

where $\mu^{(t+1)}$ is the updated mean and $\sigma^{2(t+1)}$ is the updated variance. This step ensures that the imputed values are consistent with the distribution of the observed data, taking into account the current estimates of the parameters.

These Gibbs sampling steps are repeated for a large number of iterations until the Markov chain reaches its stationary distribution. This indicates that the samples are being drawn from the true posterior distribution.

After a burn-in period, where initial samples are discarded to allow the Markov chain to converge, the sampled values of the missing data from subsequent iterations are used as imputations. These imputed values can be averaged or used to create multiple imputed datasets for further analysis:

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

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

$$\hat{Y}_{mis} = \frac{1}{T} \sum_{t=\text{burn}-\text{in}}^{T} Y_{mis}^{(t)} \tag{3.19}$$

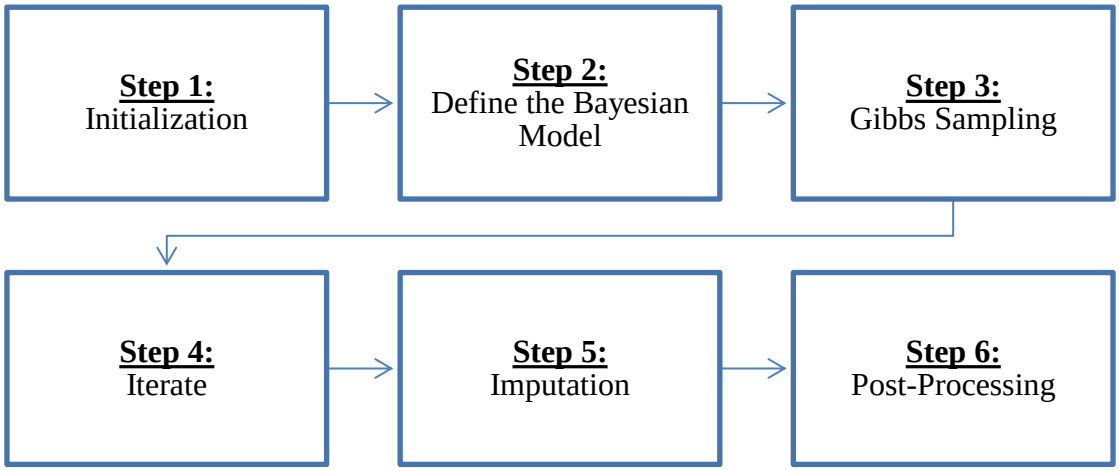
where $\hat{Y}_{mis}$ denotes the estimated value of the missing data, Y_{mis} after the MCMC process. T is the total number of iterations of the MCMC sampling process. $\text{burn} - \text{in}$ refers to the initial set of iterations that are discarded to allow the Markov chain to reach its stationary distribution. It helps in eliminating the effect of the initial values on the final estimates. $Y_{mis}^{(t)}$ represents the value of the missing data, Y_{mis} at iteration, t of the MCMC process.

Finally, the imputed dataset is analyzed or multiple imputed datasets are combined if multiple chains were used. This can involve further statistical analysis, such as regression or other modeling techniques, using the imputed data.

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

Figure 3.4

Steps involve in MCMC



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

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

3.2.6 NIPALS

Data imputation is not the sole remedy for handling missing data challenges (Ochoa-Muñoz & Contreras-Reyes, 2023). Another approach is employing the nonlinear iterative partial least squares (NIPALS) algorithm introduced by Wold (1966). Unlike direct data imputation, NIPALS operates based on the principle of utilizing available data, stepping in when imputation is not viable or produces nonsensical results. Essentially, it decomposes the data matrix into singular values through iterative sequences of orthogonal projections, achieved by point products, akin to the geometric concept of regression (Ochoa-Muñoz, González-Rojas & Pardo, 2019).

To begin, we start with an initial guess for the missing values in the data matrix, $\boldsymbol{X}$. A common approach is to replace the missing values with the mean of the observed values in the corresponding column. Let $\boldsymbol{X}$ be an $n \times p$ data matrix with missing values. The initial imputation for the missing values x_{ij} can be done as follows:

$$x_{ij} = \frac{1}{n_i} \sum_{i \in obs(j)} x_{ij}, \qquad \forall (i,j) \in \text{miss} \tag{3.20}$$

where n_i is the number of observed values in column j, and $obs(j)$ denotes the set of indices with observed values in column j. After the initial imputation, the data matrix $\boldsymbol{X}$ is centered by subtracting the column means.

The next step involves estimating the first principal component. We initialize the score vector, t_1 with one of the columns of $\boldsymbol{X}$ that has no missing values. An iterative process is then used to refine t_1 and compute the loading vector p_1.

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

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

Initially, we set:

$$t_1 = \mathbf{X}_{:,1} \tag{3.21}$$

The iterative process involves the following steps:

$$p_1 = \frac{\mathbf{X}^T t_1}{t_1^T t_1} \tag{3.22}$$

This step computes the loading vector p_1. To ensure p_1 is a unit vector, we normalize it:

$$p_1 = \frac{p_1}{\|p_1\|} \tag{3.23}$$

Next, we update the score vector, t_1:

$$t_1 = \frac{\mathbf{X}p_1}{p_1^T p_1} \tag{3.24}$$

This process is repeated until convergence.

Based on the equations above, $\mathbf{X}^T$ is the transpose of the matrix $\mathbf{X}$. The purpose of transposing $\mathbf{X}$ is to align it for the operation with the vector t_1. Next, t_1^T represents the transpose of the vector t_1, turning t_1 into a row vector. The notation $\|p_1\|$ represents the Euclidean norm of the vector p_1. Lastly, the operation $\mathbf{X}p_1$ involves multiplying the data matrix $\mathbf{X}$ by the loading vector p_1 and P_1^T represents the transpose of the vector p_1.

Once the first principal component (score vector t_1 and loading vector p_1) is estimated, we deflate the data matrix $\mathbf{X}$ to remove the information captured by this component. The deflated matrix $\mathbf{X}_1$ is computed as follows:

$$\mathbf{X}_1 = \mathbf{X} - t_1 p_1^T \tag{3.25}$$

This step ensures that the subsequent principal components will capture new variations in the data that were not explained by the first principal component.

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

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

With the principal components estimated, we can impute the missing values. For each missing value x_{ij} in **X**, the imputed value, $\hat{x}_{ij}$ can be calculated using the relationship:

$$\hat{x}_{ij} = t_i p_j \tag{3.26}$$

where t_i is the score vector and p_j is the loading vector corresponding to the missing entry. The original matrix **X** is then updated with these imputed values.

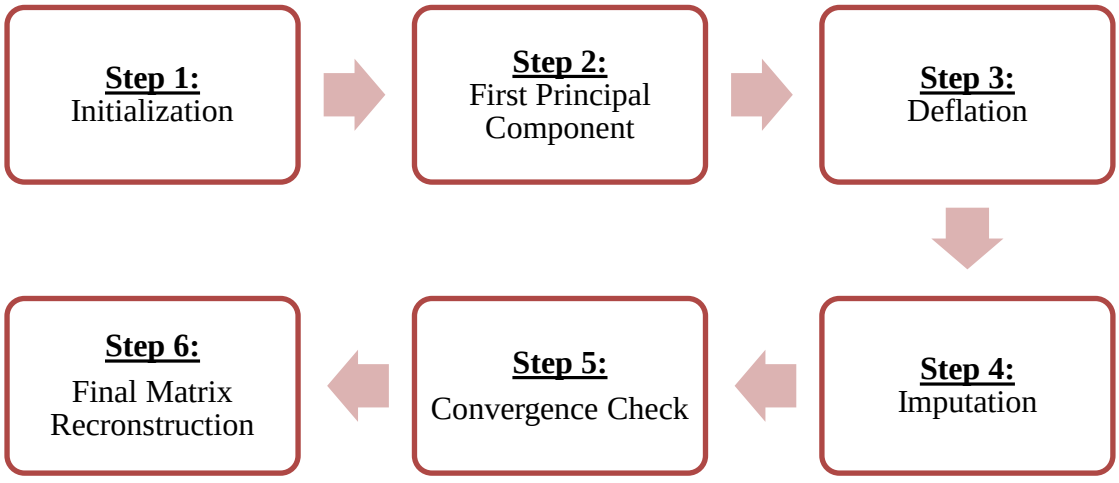
After imputing the missing values, we check for convergence by comparing the previous and current imputed values. If the imputed values have not stabilized, we repeat the process of estimating principal components and imputing missing values.

Once convergence is achieved, the imputed matrix **X** is finalized. If the data was centered initially, we add back the column means that were subtracted during the centering step to obtain the final imputed data matrix.

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

Figure 3.5

Steps involve in NIPALS



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

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

3.3 Principal Component Analysis

Lu et al. (2019) describes that Principal Component Analysis (PCA) is a fundamental approach for data analysis, wherein it exploits low dimensional structure in high dimensional data, which commonly exists in different types of data, such as image, text, video and bioinformatics. The main objective of PCA is to reduce the dimension of a dataset including a large number of information available and this can be implemented by transforming to a new set of variables which is called principal component (PC).

The first step in PCA is to standardize the data. Standardization ensures that each feature has a mean of zero and a standard deviation of one, which is essential because PCA is sensitive to the scale of the data. For a data matrix X, where each column represents a feature and each row represents an observation, the standardized data Z is computed as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{3.27}$$

where X is the original data matrix, μ is the mean of each feature, σ is the standard deviation of each feature and Z is the standardized data matrix.

After standardizing the data, the next step is to compute the covariance matrix. The covariance matrix captures the pairwise covariances between the features, providing insight into how features vary with respect to each other. The covariance matrix C is calculated using the standardized data matrix Z:

$$C = \frac{1}{n-1} Z^T Z \tag{3.28}$$

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

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

where C is the covariance matrix, Z^T is the transpose of the standardized data matrix, Z and n is the number of observations.

The next step is to find the eigenvalues and eigenvectors of the covariance matrix. Eigenvectors represent the directions of maximum variance in the data, and the corresponding eigenvalues indicate the magnitude of these directions. For the covariance matrix C, the eigenvalue equation is:

$$Cv_i = \lambda_i v_i \tag{3.29}$$

where C is the covariance matrix, v_i is the *i*-th eigenvector and λ_i is the *i*-th eigenvalue.

Once the eigenvalues and eigenvectors are computed, sort them in descending order of eigenvalues. The eigenvectors corresponding to the largest eigenvalues represent the principal components with the most variance. Sorting these eigenvalues and their associated eigenvectors helps in selecting the top principal components.

The feature vector is a matrix formed from the eigenvectors corresponding to the top k eigenvalues, where k is the desired number of dimensions. This matrix is used to transform the original data into a lower-dimensional space:

$$W = [v_1, v_2, \dots, v_k] \tag{3.30}$$

where W is the matrix of the top k eigenvectors (principal components) and v_i represents the eigenvectors corresponding to the largest eigenvalues.

Finally, transform the original standardized data into the new lower-dimensional space defined by the principal components. This is done by multiplying the standardized data matrix Z by the feature vector W:

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

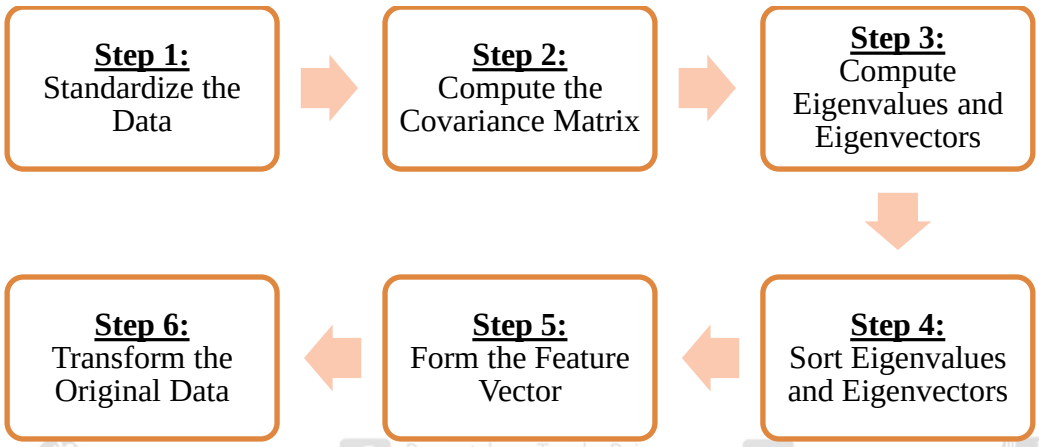
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

$$Y = ZW \tag{3.31}$$

where Y is the transformed data matrix in the reduced feature space, Z is the standardized data matrix and W is the matrix of top k eigenvectors.

Figure 3.6

Procedure of PCA model



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

3.4 The Homogeneous Markov Model (HMM)

Using similar notation to Hughes et al. (1999) and Robertson et al. (2011), let $R_t = (R_t^1, \dots, R_t^M)$ represent a multivariate random vector of rainfall occurrences for a network of M rain stations. The observed value $r_t^i = 1$ indicates rain at station i on day t, while $r_t^i = 0$ indicates no rain. Let S_t be the hidden rainfall state for day t, which can take any of the K values from 1 to K. We denote the daily sequences of precipitation occurrences and hidden rainfall states by $R_{1:T}$ and $S_{1:T}$ respectively. A hidden Markov model (HMM) for rainfall data makes two conditional independence assumptions. HMM for rainfall data assumes that the multivariate precipitation observations R_t at

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

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

time t are independent of all other variables in the model up to time t, given the weather state S_t at time t,

$$P(R_t|S_t, R_{1:t-1}, S_{1:t-1}) = P(R_t|S_t) \tag{3.32}$$

The second assumption is that the hidden state process, $S_{1:T}$, follows a first-order Markov process,

$$P(S_t|S_{1:t-1}) = P(S_t|S_{t-1}) \tag{3.33}$$

Additionally, the process is time-homogeneous, which means the transition probabilities between states remain constant over time. In other words, the probability of transitioning from one state to another is the same at all time steps, represented by a $K \times K$ transition probability matrix, where K is the number of possible states. This assumption simplifies the modeling of the hidden state sequence, as the transition probabilities do not change over time (Hughes & Guttorp 1994).

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

The probability models for individual stations, $P(R_t|S_t)$, are defined using a delta function for zero precipitation and a mixture gamma distribution for modeling the distribution of non-zero precipitation amounts. (Cioffi et al., 2020). In $P(R_t|S_t)$, the rainfall observation at each station at time t is assumed to be independent of observations at other stations at the same time, given the hidden state:

$$P(S_t = s) = \prod_{m=1}^{M} P(R_t^m = r|S_t = s) = \frac{1}{M}\prod_{m=1}^{M} p_{smr} \tag{3.34}$$

where $r \in \{0,1\}$, each $p_{smr} \in [0,1]$, and $p_{sm0} + p_{sm1} = 1$. In the equation, m represents the specific rain station in the network while M represents the total number of rain stations in the network. $P(S_t = s)$ represents the probability that the state at

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

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

time t is s. This is the overall probability distribution of the state variable S_t being s. $P(R_t^m = r|S_t = s)$ represents the conditional probability of observing rainfall R_t^m at station m given that the state at time t is s. p_{smr} represents the probability of rainfall amount r at station m given the state s. Although we assume conditional independence given the states, the rainfall process is not spatially independent (which would be unrealistic). In this study, a single distribution provided the best performance based on standard model fitting metrics like the Bayesian Information Criterion (BIC).

A comprehensive discussion on parameter fitting and prediction using homogeneous HMMs can be found in previous studies (Robertson, 2004). The hidden process z follows a Markov model with an inhomogeneous $K \times K$ transition matrix Q_t, with components q_{ijt}, where $i, j \in \{1, \dots, K\}$. In an HMM, the probability of hidden state transitions varies over time as a function of the exogenous variable X.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun PustakaTBainun ptbupsi

The log-likelihood of the data for HMM becomes:

$$l = log\, P(R|X) = log \left[\sum_s \left[P(S_1|X_1) \prod_{t=2}^{T} P(X_t)\right] \prod_{t=1}^{T} P(R_t|S_t)\right] \quad (3.35)$$

In Equation 3.35, the hidden state transition probabilities at time t are modeled using multinomial logistic regression based on X_t (Hughes et al., 1999):

$$q_{ijt} = P(z_{t-1} = i|x_t, \zeta) = \frac{exp\,(\xi_{ij} + x_t'\rho_j)}{\sum_{m=1}^{K} exp\,(\xi_{im} + x_t'\rho_m)} \quad (3.36)$$

where, x_t is a B-dimensional exogenous covariate time series for $t = 1, \dots, T$, and ρ_j is a B-dimensional vector of coefficients corresponding to the B components of $x_t = (x_{1t}, \dots, x_{Bt})$. For notation convenience, let $\zeta = \zeta_{ij} = (\rho_j, \xi_{ij})$ for all $i, j \in \{1, \dots, K\}$.

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

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

Additionally, HMMs include a set of state-dependent emission distributions $f(y_t|z_t = k, \theta)$, where $k = 1, \ldots, K$ and θ represents the set of all parameters of the emission distribution. These emission distributions vary depending on the state k and station s combinations. In a Bayesian analysis of HMM, the posterior full conditional distributions for z, ζ, and θ can be computed independently using the MCMC algorithm. This model can incorporate two latent variables: one for sampling the coefficients of the hidden states' transition probabilities (Polson, Scott, & Windle, 2013) and another associated with the emission distributions (Albert and Chib, 1993).

Given the observed rainfall data y, x and w, and conditioned on the sampled values of the parameters ζ and θ, the posterior full conditional distribution of the hidden state z_t^n at the nth sampling iteration is as follows:

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

$$z_t^n|\zeta, \theta, \ldots \sim \text{Multi}\left(\frac{q_{z_{t-1}^n,1} q_{1,z_{t+1}^{(n-1)}} f_1(\cdot)}{\sum_{k=1}^K q_{z_{t-1}^n,k} q_{k,z_{t+1}^{n-1}} f_k(\cdot)}, \ldots, \frac{q_{z_{t-1}^n,K} q_{K,z_{t+1}^{n-1}} f_K(\cdot)}{\sum_{k=1}^K q_{z_{t-1}^n,k} q_{k,z_{t+1}^{n-1}} f_k(\cdot)}\right) \quad (3.37)$$

where $f_{zt}(\cdot) = f(z_t = k, \theta)$ is the emission distribution for state $k = 1, \ldots, K$. Each of the z_t are sampled in succession for all $t = 2, \ldots, T$ at each of the $n = 1, .., N$ iterations of the larger MCMC algorithm. As part of this study, state one is assigned to the first hidden state associated with the first day of the time series as describe as: $Pr(z_{t=1} = 1) = 1$.

This study model daily rainfall amounts using a zero-inflated mixture of two exponential distributions, following previous literature (Tracy, 2017). This distribution has a physical interpretation with three components corresponding to no rain, light rain, and heavy rain. The delta function at zero (δ_0) accounts for additional dry days, while the light rain and heavy rain are each represented by an exponential distribution, where:

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

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

$$y_{ts}|z_t, \theta \sim p_{0_{ts}}\delta_0 + p_{1_{ts}}Exp\left(\lambda_{1_{z_{ts}}}\right) + p_{2_{ts}}Exp\left(\lambda_{2_{z_{ts}}}\right) \tag{3.38}$$

where $z_t = k$ and θ represents the mixing probability parameters and rate parameters of the emission distributions. Additionally, the probabilities $p = \left(p_{0_{ts}}, p_{1_{ts}}, p_{2_{ts}}\right)$ are assumed to depend on the exogenous variables.

In this study, many rainfall stations in Selangor have missing data points, with less than 10% of the values missing. These missing data points can be treated as random variables, and their posterior distributions are inferred along with the model's other variables using Bayesian methods. For each missing data point, the posterior conditional distribution is computed as (y_t° at time t and station s) $y_t^\circ|z_t^\circ, \ldots \sim f(w_t)$.

This paper simulates these forecast chains using exogenous variables for held-out years of input. A new time step r can be expressed as follows:

$$\begin{gathered} q_{ijr}^*|X_r, \zeta^\circ = g^{-1}(X_r'\varsigma^\circ) \\ z_r^*|q_{z_{r-1}^* jr}^* \sim Multi\left(q_{z_{r-1}^* 1r}^*, \ldots, q_{z_{r-1}^* Kr}^*\right) \end{gathered} \tag{3.39}$$

where $z_r^* = k$ and q_{ijr}^*, z_r^* and y_r^* are new predictive draws at time r.

The study utilized data from the first eight years, spanning from 2008 to 2015, to generate predictive conditional chains for the years 2016 to 2018. These predictive chains were then compared with the observed y data from the three years that were held out, enabling a thorough assessment of the model's performance through model selection and distributional checks.

The implementation of this model was carried out using the HMM R Package available on the Comprehensive R Archive Network (CRAN). This proposed model

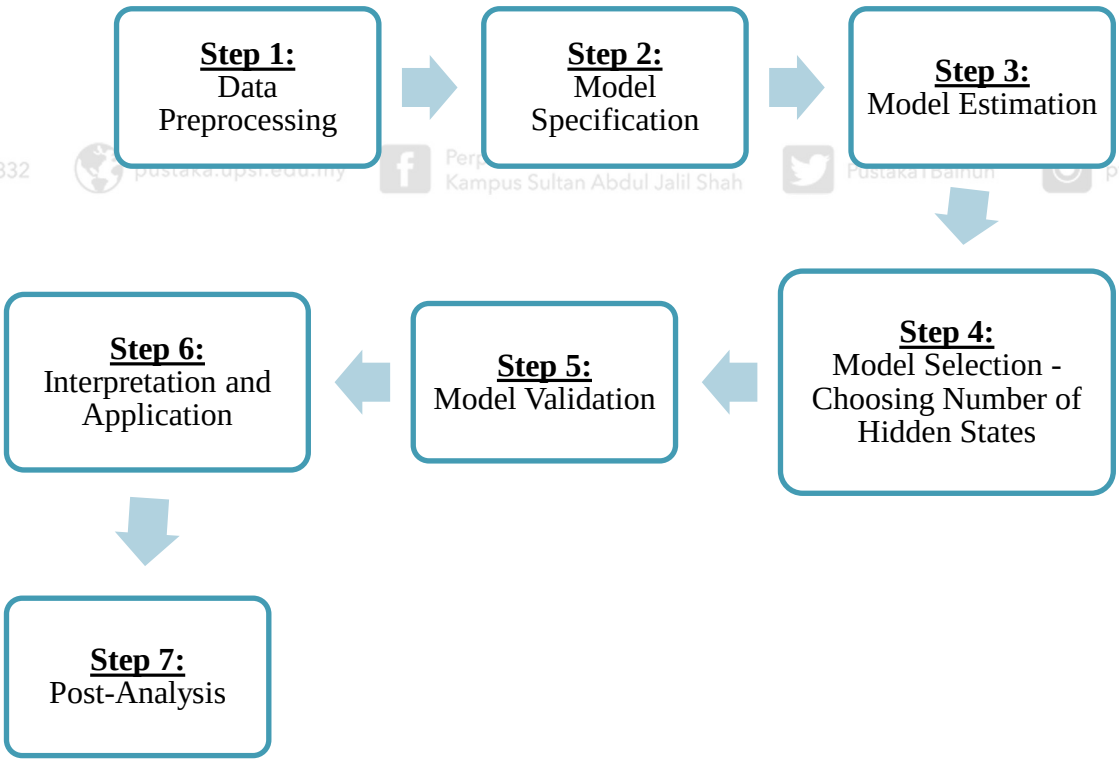
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

differs from the classical HMM in that it introduces inhomogeneity in both the transition probabilities and the state-dependent emission distributions. Furthermore, the model advances by developing a fully Bayesian estimation scheme, which enhances its capability to handle high-dimensional data sets effectively. This comprehensive approach allows for more accurate modeling of complex data, providing a robust framework for analyzing and predicting rainfall patterns (Holsclaw, 2017).

Figure 3.7

Steps involve in HMM



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

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

3.5 Machine Learning

Machine Learning is typically thought of as a universal toolbox, ready to be used for classification problems, identifying a suitable category for a new set of observations, and regression tasks, estimating the relationship among given data samples (Rafique & Velasco, 2018). In its most direct form, ML refers to the computational representation of a phenomena intended for task execution, given a specific performance and based on a specified environment. Data is obtained from historical databases and pre-processed to handle missing data, remove or normalise outliers, filter and aggregate parameters during the initial construction phase of ML models. The following stage deals with data transformation, choosing pertinent data features, reducing redundancy, adapting data formats to the assigned task, etc. This is an essential phase in the ML workflow since, due to the dimensionality curse, not all available data is necessarily helpful in terms of algorithm performance and complexity. Model selection and behaviour learning's most crucial stages are frequently completed concurrently. High-dimensional parameter optimisations and associated assessments are used in the process to trade off performance, complexity, and computational effort. Lastly, test data is fed into the models, results are translated to representative knowledge (such as a specific pattern), and insights are supplied to a dashboard or other relevant software components. Figure 3.8 shows a typical workflow of an ML framework.

3.5.1 Random Forest

Random Forest (RF) by Breiman (2001) is a versatile machine learning technique applicable to both regression and classification tasks. In the RF approach, each decision

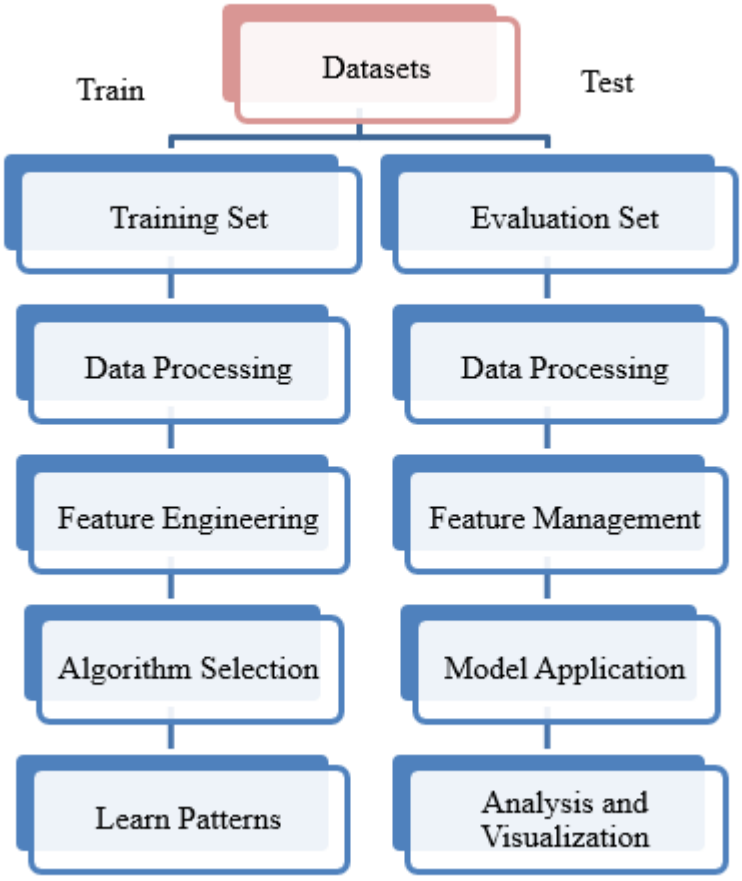
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

tree is built using a bootstrap sample from the training set, meaning a random subset of the training data is used for growing each tree (Benmakhlouf et al., 2023). To further prevent overfitting and enhance stability, only a subset of the features is considered at each node when determining the optimal split. This randomization process helps to decrease the correlation among individual trees, leading to a more diverse and robust ensemble.

Figure 3.8

A workflow of an ML framework



Adapted from Rafique and Velasco (2018)

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

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

According to Agarwal et al. (2023), RF combines multiple decision trees to predict the class of a dataset, which means that while some trees may individually predict the correct output and others may not, the ensemble of all trees tends to produce the correct prediction collectively. In constructing these trees, a number m (referred to as $m$try) is randomly selected from the total input variables at each node. The optimal split is then determined based on these m variables, and this split is used to partition the node. Importantly, the value of m remains constant throughout the growth of the forest, ensuring consistency.

Every tree in the forest is grown to its full potential without pruning. Pruning, a technique often used in machine learning and search algorithms to reduce the size of decision trees by eliminating non-critical and redundant parts, is not employed in RF. Consequently, each tree is allowed to expand fully, capturing as much detail from the data as possible.

The parameter n (ntree) represents the number of decision trees grown in the algorithm. This parameter is set before the maximum voting or prediction averages are calculated. Increasing the number of trees generally enhances the model's efficiency, as a larger ensemble reduces variance and improves the accuracy of the final predictions. The robustness of RF lies in this ability to aggregate the predictions from multiple trees, balancing out errors from individual trees and leading to more reliable and stable outcomes.

Formally, a random forest is a predictor composed of a collection of randomized base regression trees $\{r_n(x, \theta_m, D), m \geq 1\}$, where $\theta_1, \theta_2, \ldots$ are

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

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

independent and identically distributed outputs of a randomizing variable θ. These individual random trees are aggregated to form the overall regression estimate

$$\underline{r}_n(X, D_n) = \mathbb{E}_\theta[r_n(X, \theta, D_n)] \tag{3.40}$$

where E_θ denotes the expectation with respect to the random parameter, conditional on X and the data set D_n. To simplify notation, the dependency of the estimates on the sample will be omitted in the following text, and for instance, $\underline{r}_n(X)$ will be written instead of $\underline{r}_n(X, D_n)$.

The randomizing variable θ determines how the successive cuts are performed when building individual trees, including the selection of the coordinate to split and the position of the split. In the model under consideration, θ is assumed to be independent of X and the training sample D_n, thereby excluding any bootstrapping or resampling steps in the training set and ruling out any data-dependent strategies for tree construction, such as optimizing splits based on actual observations. However, θ can be based on a second sample, independent of but similarly distributed as D_n. With these considerations, each individual random tree is constructed by associating all nodes with rectangular cells so that, at each step of the tree's construction, the collection of cells associated with the tree's leaves forms a partition of $[0, 1]^d$. The root of the tree is $[0, 1]^d$ itself. The procedure is then repeated $\lceil k_n \rceil$ times, where log_2 is the base-2 logarithm, $\lceil . \rceil$ is the ceiling function, and $k_n \geq 2$ is a deterministic parameter set beforehand by the user, potentially depending on n.

1. At each node, a coordinate from $X = (X^{(1)}, \ldots, X^{(d)})$ is selected, with the j-th feature having a selection probability of $p_{nj} \in (0,1)$.

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

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

2. Once the coordinate is selected at each node, the split occurs at the midpoint of the chosen dimension.

Each randomized tree $r_n(X,\theta)$ computes the average of all Y_i for which the corresponding vectors X_i fall within the same cell of the random partition as X. In essence, if $A_n(X,\theta)$ represents the rectangular cell of the random partition containing X,

$$r_n(X,\theta) = \frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,\theta)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]}} 1_{E_n(X,\theta)} \tag{3.41}$$

where $E_n(X,\theta)$ is defined as

$$E_n(X,\theta) = \left[\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]} \neq 0\right] \tag{3.42}$$

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

As per convention, the estimate is assigned a value of 0 for empty cells. Ultimately, after considering the expectation with respect to the parameter θ, the random forests regression estimate can be expressed as:

$$\underline{r}_n(X) = E_\theta[r_n(X,\theta)] = E_\theta\left[\frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,\theta)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]}} 1_{E_n(X,\theta)}\right] \tag{3.43}$$

Ultimately, the necessity for splits to consistently occur at the midpoint of the cell sides is primarily technical in nature, and theoretically, it could be substituted with a more intricate random mechanism, possibly derived from the second sample. However, this would come at the expense of significantly complicating the analysis process.

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

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

Figure 3.9

Steps involve in RF



3.5.2 Support Vector Machine

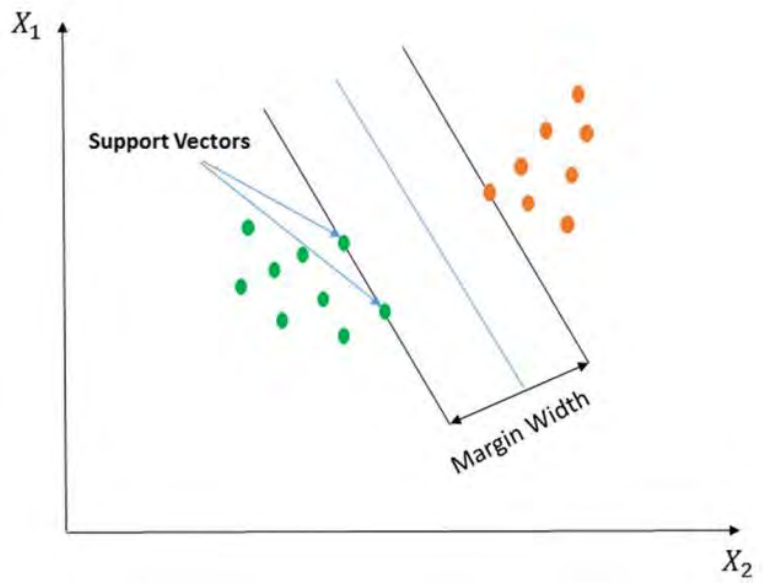
Support Vector Machine model is ideal for the prediction due to its excellent learning ability in processing small sample data (Li et al., 2019). The main objective of SVM is to solve a convex quadratic optimization problem for obtaining a globally optimal solution in theory and hence, overcoming the local extremum dilemma of other machine learning techniques (Sheykhmousa et al., 2020). SVM is a cutting-edge classifier and can speculate direct characterization limits in a multidimensional space (Cortes & Vapnik, 1995; Corter & Vapnik, 1998; Vapnik, 1998; Meshram, 2020). SVM is configured by hyperplane (Cortes & Vapnik, 1995; Meshram, 2020) that implies the choice limits named “support vectors” as shown below.

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

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

Figure 3.10

A schematic representation of SVM



Adapted from Meshram (2020)

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

Consider that the datasets from PCs were split into two sets, training data and testing data. The training data with two classes $[(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_i, y_i)]$ and the input vector is x_i, with the output of y_i. The output can be identified as $y_i \in \{+1, -1\}$. The classifier for the problem of binary classification is

$$f(x) = sign[w^T \cdot \phi(x) + b] \tag{3.44}$$

where the input vector (x) was mapped with a feature space by non-linear function $\phi(x)$. Then, the *w* and *b* are the classifier parameters. Solving the optimization problems is equivalent to determining the SVC classifier from the theory,

$$\begin{aligned} \text{min} \quad & \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i[w^T \cdot \phi(x) + b] \geq 1 - \xi_i \\ \xi_i \geq 0, \quad & i = 1, \ldots., l \end{aligned} \tag{3.45}$$

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

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

where ξ_i is a non-negative slack variable that influences objective function when data are misclassified. Basically, C is a penalty parameter with a positive value. The optimization problems of C will be solved by using Lagrange multiple, α_i where $0 \leq \alpha_i \leq C$.

$$f(x) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right) \tag{3.46}$$

The kernel function, $K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j)$, was introduced to calculate the inner products. Various kernels might be used in SVM but the standard kernels used are linear, polynomial, radial basis function, and sigmoid. The most popular and capable kernel is the radial basis function (RBF) with parameter γ.

$$K(x, x') = \begin{cases} x^T \cdot x', & \text{linear} \\ (x^T \cdot x' + 1)^d, & \text{polynomial} \\ \exp(-\gamma \|x - x'\|^2), & \text{RBF} \\ \tanh(\gamma x^T \cdot x' + C), & \text{sigmoid} \end{cases} \tag{3.47}$$

In the final classifier, only nonzero Lagrange multiple will take part. The data with nonzero corresponding Lagrange multiple will be named as support vector. The classifier will be written as

$$f(x) = sign(\sum_{k=1}^{l} \alpha_k y_k K(x_k, x) + b \tag{3.48}$$

where x_k is the support vector and m is the number of support vector. In the support vector classifier, there are two parameters to calibrate, which are C (penalty term) and type of kernel. The Lagrange multiple and parameter b in the equation above can be found using the SVM algorithm by determining these parameters. Then, the accuracy of the classification model by SVM was measured as below;

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

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

$$Accuracy: \frac{\hat{A} + \hat{B}}{A + B} \times 100 \tag{3.49}$$

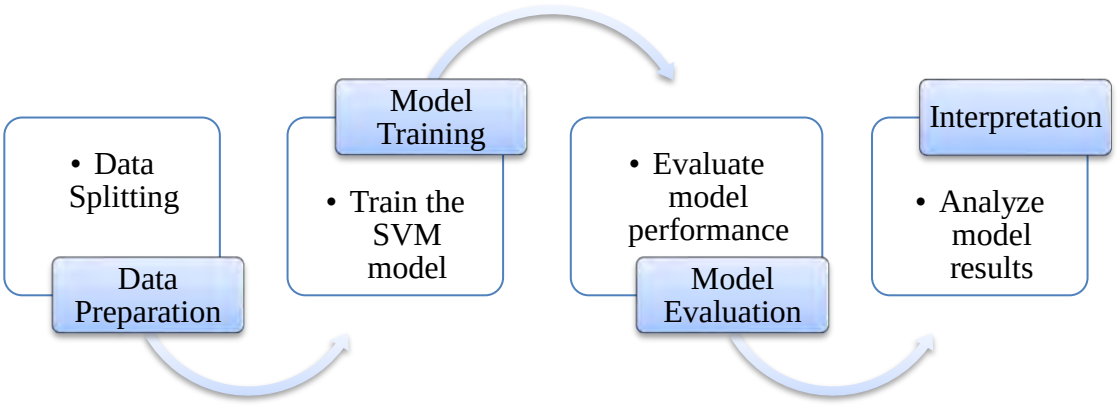
where A is the number of dry days while B is the number of wet days. $\hat{A}$ denotes as the number of dry days that are correctly classified as dry days, and $\hat{B}$ as the number of wet days that are correctly classified as wet days. The difference between actual and predicted values resulted in misclassification error (Kassambara, 2018), which represents that individuals classified as positive values were incorrectly assigned to negative values class and vice versa. Subsequently, the misclassification error by SVM can be calculated as follows;

$$misclassification\ error = \frac{FP + FN}{A + B} \tag{3.50}$$

where FP is the False Positive that, by the predicted value, were falsely predicted (the actual values are negative, but the predicted values are positive). The FN is the False Negative that represents the type of error when the actual values are positive, meanwhile the model predicted a negative value.

Figure 3.11

Steps involve in SVM



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

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

3.5.3 Artificial Neural Network

An artificial neural network (ANN) is a computational model that is based on the biological neural networks in the human brain that process information (Asteris & Mokos, 2019). Large datasets gathered from experiments or simulations can be "learned" and corroborated by neural networks. The trained neural network functions as an analytical tool for accurate predictions of the final outcomes. They can produce excellent prediction accuracy ratings due to effective approaches for training and validation. Ali and Shahbaz (2020) described that an artificial neuron in ANN performs two major functions. In order to get the net input of the neuron the weighted total of all of the inputs to the neuron it first applies a summation function. After that, it uses the net input as the input for an activation or transfer function to create an output.

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

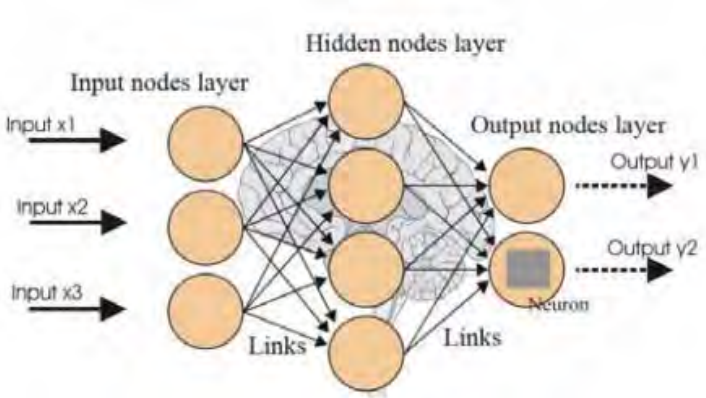
Mijwil (2018) stated that a neural network can be categorized into three layers which is an input layer, an intermediate layer and an output layer, as shown in Figure 3.12 The neurons are represented by the grey box in the figure below, while the connection points are represented by the arrows. The input layer prepares a data set for training, which is presented to the network. The network gives the connection points in the intermediate layer the weights of the events it learns. Some points can be zero and not all of them need to have a value. In order to prevent the zero values at the connection points from also becoming zero, a threshold value is inserted between the layers.

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

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

Figure 3.12

The ANN's layers



By Wijmil (2018)

Finding a set of weights that produces a mapping that closely matches input and output is the goal of the learning process. An arbitrary nonlinear function like the one below could be considered:

$$F(x, w) = y \tag{3.51}$$

where x represents the input vector (predictors) presented to the network, w represents the weight of the network and y represents the corresponding output, which is approximated or predicted by the network. The number of neurons in the hidden layer can be computed as below:

$$l = \frac{1}{2}(i + o) + \sqrt{N} \tag{3.52}$$

where l denotes the hidden layer's number of nodes, i the number of input neurons, and o the number of output neurons. The training set's record count, N, is given. The ANN algorithm is structured as follows.

$$\hat{y}_k = f_o\left[\sum_{j=1}^{l} \omega_{k\,j} \cdot f_h\left(\sum_{i=1}^{N_n} \omega_{ji} x_i + \omega_{j_o}\right) + \omega_{k0}\right] \tag{3.53}$$

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

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

where ω_{ji} denotes the weight of the hidden layer, which is connecting to the *i*th neuron in the input layer to the *j*th neuron in the hidden layer, and $\omega_{k\,j}$ is the *j*th neuron bias. The f_h in the equation above represents *j*th hidden neuron activation function. Moreover, based on the equation above, ω_{j_o} denotes the parameter of the output layer that connects to the *j*th neuron in the hidden layer to the *k*th neuron in the output layer, and ω_{k0} represents the bias for the *k*th output neuron. Furthermore, the f_o is the activation function for the output neuron, while x_i is the *i*th input data. Lastly, N_n denoted as the number of neurons in input layers, meanwhile $\hat{y}_k$ denoted as the calculated output.

In this study, like in most others, a feed-forward MLP network with a backpropagation (BP) error-correction learning algorithm is used as the model. MLP is widely used ANN technique in modelling hydrological activities (Traore et. al., 2010) and it consists of an input layer, one or more hidden layer(s), and an output layer (Ali & Shahbaz, 2020). The three-layer formula can be seen as below:

$$y(t) = f_o\left[\sum_{i=1}^{m} w_{kj} f_h\left[\sum_{i=1}^{n} w_{ij} x_i(t) + \theta_j\right] + \theta_0\right] \tag{3.54}$$

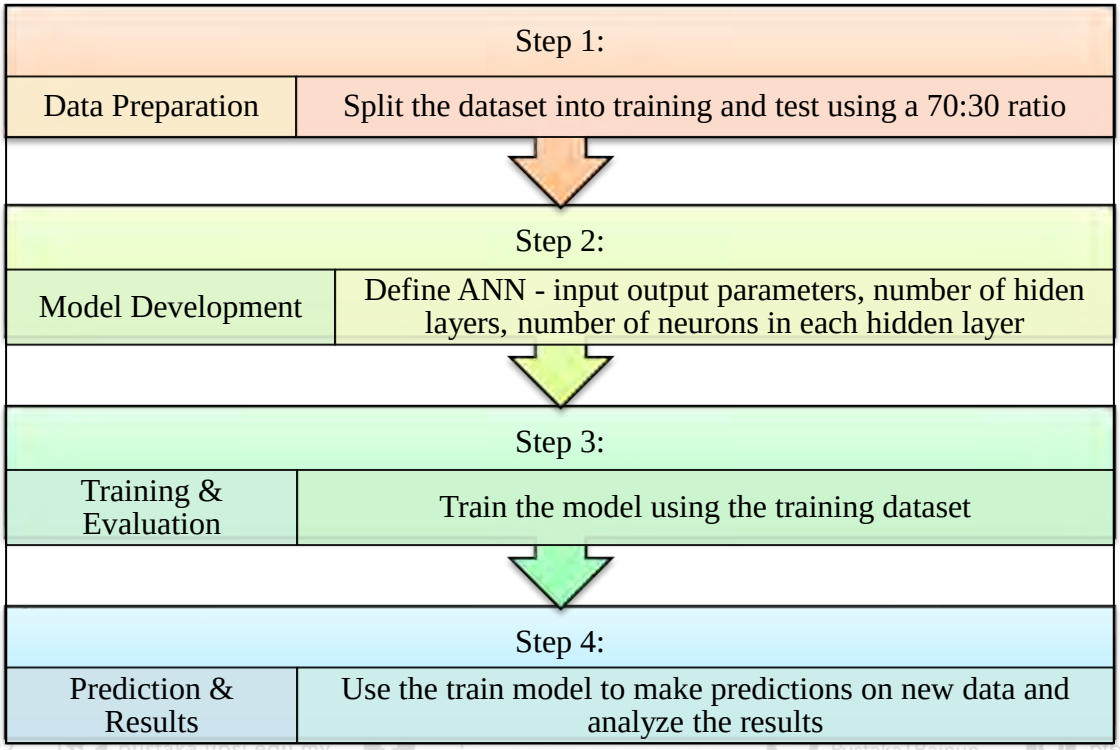
where $y(t)$ represents the output of a predictand at time *t*, meanwhile f_o and f_h is the transfer function of the output and hidden layers. $x_i(t)$ denotes as the input of the predictors *i* at time *t*. Next, w_{kj} can be described as the weight between the hidden and output layers, while w_{ij} as the weight between input and hidden layers. Then, *n* is the number of predictors, while *m* is the number of hidden nodes. Subsequently, Figure 3.13 below shows the steps involve in ANN.

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

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

Figure 3.13

The steps involve in ANN



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

3.5.4 Decision Tree

A decision tree is a structure where each node represents an attribute, each branch represents a decision rule, and each leaf denotes an outcome, either categorical or continuous. By mimicking human reasoning, decision trees clearly display all possible alternatives and trace each to its conclusion, facilitating easy comparison and interpretation of the data (Kokilavani et al., 2022). Ulkir et al. (2024) stated that a decision tree is a directed structure with a root node that has no input and internal nodes that each receive one input, where nodes whose outputs serve as inputs to other nodes are called internal or test nodes, and those whose outputs do not are called leaf nodes.

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

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

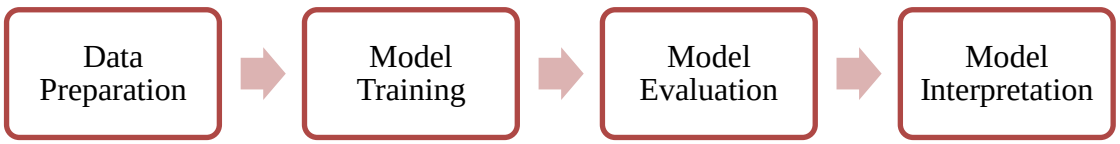
Banadkooki et al. (2019) applied a 'divide-and-conquer' approach, a model is created where N data points are either assigned to a leaf or subjected to a test criterion that splits them into subsets based on the test outcome. The criterion for splitting relies on evaluating the standard deviation of class values and calculating the reduction in error. Specifically, the standard deviation reduction values are determined using the following equation:

$$SDR = sd(T) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(T_i - \bar{T})^2} \tag{3.55}$$

In this context, T represents a series of samples that reach a particular node, T_i is the i-th output of the model, i denotes the number of available data records, and sd stands for standard deviation. Once the number of split branches is maximized, the tree model selects the split that optimizes the expected reduction in error. However, the extensive number of divisions and branches can lead to overfitting. To address this, a pruning stage is implemented, which involves reducing the complexity of the tree by removing parts that provide little predictive power. Pruning effectively divides the input model into smaller, more manageable areas. Additionally, based on a greedy algorithm, the model eliminates variables that are minimally relevant to the overall prediction, further refining the model to enhance its performance and generalizability.

Figure 3.14

The steps involve in DT



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

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

3.5.5 *k*-Nearest Neighbor (KNN)

In the realm of classification and regression, the K-Nearest Neighbors (KNN) algorithm stands out as one of the most renowned and effective machine learning techniques (Anaraki et al., 2023). KNN is a nonparametric algorithm, which means it does not make any assumptions about the distribution of the training data, allowing it to address both regression and classification problems flexibly (Nurwatik et al., 2022). Its simplicity and effectiveness make KNN a prominent and widely used machine learning algorithm. Due to its versatility and straightforward implementation, KNN is particularly useful in handling complex data and can be applied across a diverse range of applications (Hadi et al., 2024). Its capacity to adapt to various data structures and problems underscores its importance and popularity in the field of machine learning.

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

According to Anaraki et al. (2023), in the KNN algorithm, the training data points are treated as neighboring points to the testing data. To determine the significance of each neighbor, the algorithm calculates the inverse of the Euclidean distance between the testing data and the neighboring points, using this inverse distance as the weight for each neighbor. This means that points closer to the testing data have higher weights, reflecting their greater relevance. The algorithm then sorts these neighboring points based on their weights, identifying the top K points with the highest weights. These selected K neighbors are pivotal in computing the final output. For each input dataset, KNN calculates the output by taking a weighted average of these K neighbors, ensuring that closer points have a larger influence on the result. This approach leverages the spatial proximity of the data points to deliver accurate predictions for both classification and regression tasks:

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

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

$$R_{Output,i} = \frac{\sum_{j=1}^{K} W_j R_j}{\sum_{j=1}^{K} W_j} \tag{3.56}$$

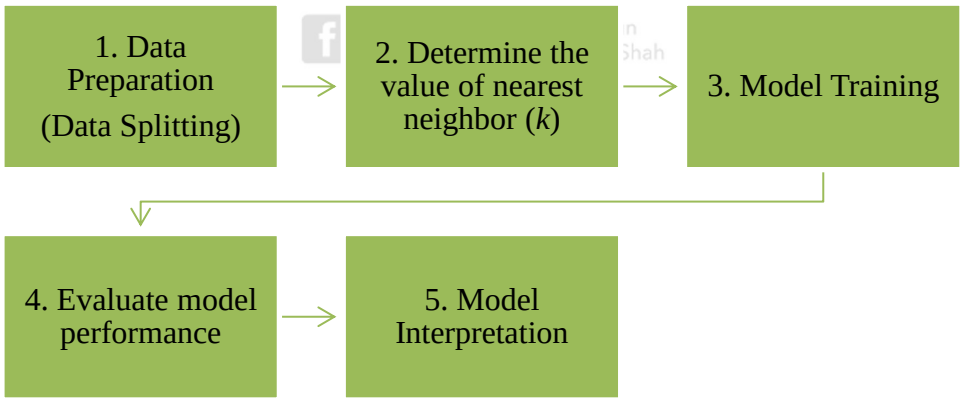
where R_j is the j-th observed runoff during the training period, $R_{Output,i}$ is the i-th estimated runoff, and W_j is the j-th weight of the neighbor, which can be calculated as follows:

$$W_j = \frac{1}{\|X - X_j\|} \tag{3.57}$$

where X represents as testing input data and X_j represents as training input data.

Figure 3.15

The steps involve in kNN



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

Relying solely on either a Bayesian approach or a single machine learning model for rainfall prediction presents significant limitations, particularly when dealing with issues like zero-bounded rainfall data and non-linearity. The Bayesian approach, while

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

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

powerful in probabilistic modeling, often struggles with zero-bounded data where a significant portion of the dataset consists of zeros, such as no rainfall events. Standard Bayesian methods may not adequately capture the probability mass around these zeros, leading to biased or inaccurate predictions. Additionally, Bayesian models typically assume a specific parametric form, which can fail to accurately model complex, non-linear relationships in rainfall data, resulting in poor predictive performance. Furthermore, Bayesian methods can be computationally intensive, making them less practical for real-time or large-scale rainfall prediction.

On the other hand, single machine learning models, especially non-probabilistic ones like support vector machines or neural networks, often do not provide a natural way to quantify uncertainties in predictions. This is a significant drawback when dealing with rainfall data, where understanding uncertainty is crucial due to the stochastic nature of weather events. These models are also prone to overfitting, particularly with complex or non-linear data, leading to unreliable predictions on unseen data. Moreover, many machine learning models lack interpretability, making it difficult to understand the underlying patterns and relationships in the data, which can hinder model refinement and decision-making.

To overcome these challenges, a hybrid approach that integrates Bayesian methods with machine learning models is necessary. Such an approach leverages the strengths of both: Bayesian methods provide a probabilistic framework to capture uncertainties and incorporate prior knowledge, while machine learning models excel at handling complex, non-linear relationships and large datasets. By combining these methodologies, a hybrid model can better accommodate zero-bounded rainfall data,

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

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

ensuring that the probability mass around zero is correctly modeled and adapting to non-linear patterns. This integration enhances predictive accuracy, reduces overfitting, and improves the model's ability to generalize to new data. Moreover, a hybrid approach allows for better uncertainty quantification, offering a more comprehensive understanding of prediction confidence, which is critical for making informed decisions in weather and climate-related contexts.

3.7 Prediction Modeling of Hybrid Statistical Downscaling-based Bayesian-Machine Learning model

In this study, the primary objective of the hybrid model is to achieve accurate rainfall prediction in Selangor, Malaysia. The research is conducted in three distinct stages: preprocessing, prediction modeling, and evaluation. In the Preprocessing stage, the study begins with two primary datasets: local surface rainfall data (y-variable) and atmospheric data (x-variable). The local surface rainfall data, which serves as the target variable, contains missing values. These missing values are handled using various imputation techniques, including Mean Imputation, Median Imputation, Nonlinear Iterative Partial Least Squares (NIPALS), the Expectation-Maximization (EM) Algorithm, Markov Chain Monte Carlo (MCMC), K-Nearest Neighbors (KNN), and Random Forest. Each imputation method is evaluated using performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash-Sutcliffe Efficiency (NSE) to ensure the accuracy of the imputed data. Additionally, for datasets with an excess of zero values, a Hidden Markov Model (HMM) is employed to address the zero-inflation issue, enabling a more accurate representation of the data.

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

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

Concurrently, the atmospheric data, which serves as the predictor variable, is standardized to ensure that all variables contribute equally to the analysis. Atmospheric data require standardization due to different units in the data. This process brings the data to a uniform scale, ensuring that all features have an equal impact on the analysis, thereby improving the performance of the proposed prediction model. Following this, dimensionality reduction is performed on the standardized atmospheric data using Principal Component Analysis (PCA). This step reduces the complexity of the data by identifying and retaining only the most significant components, making the subsequent modeling more efficient.

During the Prediction Modeling phase, the research utilizes a sophisticated hybrid model that combines a Statistical Downscaling-based Bayesian Approach with a Machine Learning technique, resulting in what is referred to as the Statistical Downscaling-based Hidden Markov Model-Machine Learning model (SD-based HMM-ML). This hybrid model is specifically designed to predict rainfall by leveraging preprocessed atmospheric data, which has been meticulously prepared to serve as input for the prediction process. The fusion of these advanced methodologies—Statistical Downscaling and Machine Learning, underpinned by a Bayesian framework—aims to significantly improve the precision and reliability of rainfall forecasts. By integrating these approaches, the model seeks to capture the complex patterns and relationships within the atmospheric data, thereby offering more accurate predictions compared to traditional methods. This enhancement in prediction accuracy is crucial for applications such as weather forecasting, climate research, and hydrological studies, where precise rainfall predictions can have substantial impacts.

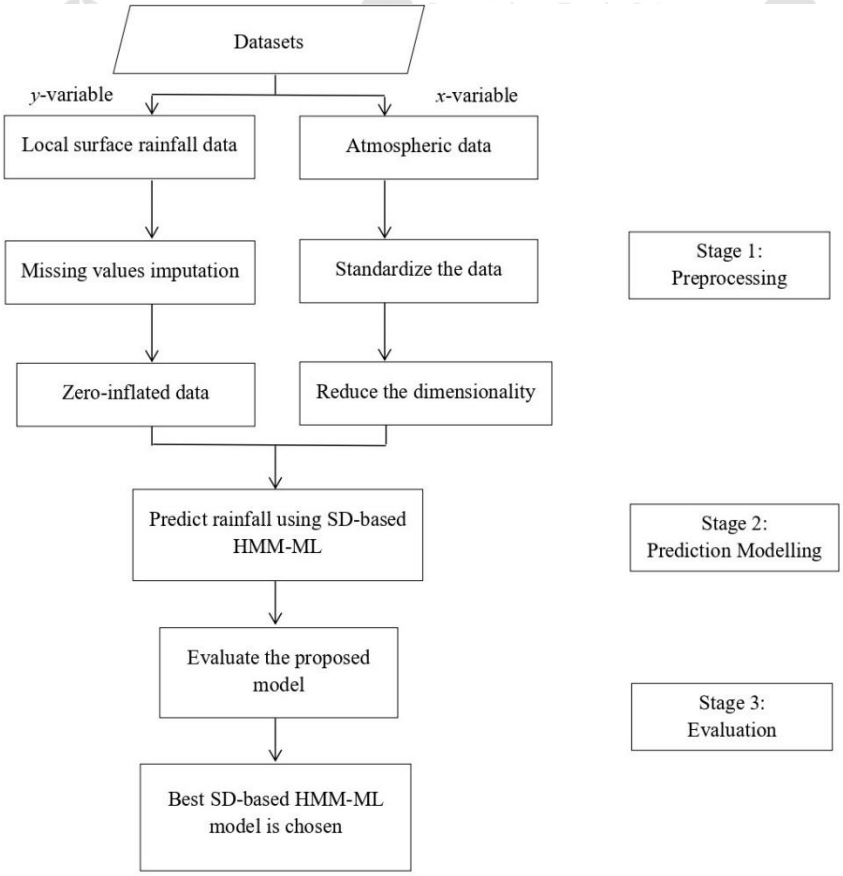
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

Finally, in the Evaluation stage, the performance of the proposed model is assessed using several metrics, including RMSE, MAE, NSE, and Mean Forecast Error (MFE). These metrics provide a comprehensive evaluation of the model's accuracy, bias, and predictive power. Based on the evaluations, the study selects the best-performing SD-based HMM-ML model. This comprehensive evaluation ensures that the chosen model is well-suited for practical application in rainfall prediction for Selangor, Malaysia, offering a reliable novel tool for weather forecasting and related decision-making processes. The flowchart of prediction rainfall data using SD-based HMM-ML model is illustrated in Figure 3.16.

Figure 3.16

The flowchart of the study



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

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

3.8 Model Performance Metrics

The performance of the proposed model will be evaluated with the performance measures namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash-Sutcliffe Efficiency (NSE), and Mean Forecast Error (MFE). In addition, the imputation methods that used in this study (which are mean imputation, single value imputation, *k*-nearest neighbour, EM Algorithm, NIPALS, MCMC and random forest) were compared by using RMSE. The forecasting model that yields the lowest values of MAE, MFE, and RMSE is designated as the best model for forecasting (Masngut et al., 2020). Meanwhile, the NSE ranges from $-\infty$ to 1, where NSE = 1 indicates a perfect match between the downscaled model and the predictand, NSE = 0 means the model's accuracy is equivalent to the mean of the predictand, and NSE < 0 signifies that the observed mean predicts better than the model (Nourani et al., 2018). The following gives the mathematical expression of RMSE, MAE, NSE and MFE;

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - y_t)^2} \tag{3.58}$$

$$MAE = \frac{\sum_{t=1}^{n}|x_t - y_t|}{n} \tag{3.59}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(x_t - y_t)^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{3.60}$$

$$MFE = \frac{\sum_{t=1}^{n}(x_t - y_t)}{n} \tag{3.61}$$

where x_t represents the measured value of rainfall, while y_t represents the predicted value of rainfall, $\overline{x}$ is mean of measured value and *n* denoted as the number of data sets.

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

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

3.9 Conclusion

This chapter provides a comprehensive overview of the methodology employed in this study, detailing each of the key stages that form the foundation of the research process. The methodology is divided into three primary stages: the preprocessing stage, the prediction modeling stage, and the evaluation stage. In the preprocessing stage, the focus is on preparing the data for analysis by addressing missing values, standardizing variables, and reducing dimensionality, among other critical tasks. The prediction modeling stage then builds on this groundwork, utilizing advanced machine learning techniques to develop models that forecast daily rainfall based on the preprocessed data. This stage is crucial as it determines the effectiveness of the models in capturing the complex patterns within the dataset. Finally, the evaluation stage involves a thorough assessment of the models' performance using a range of metrics, ensuring that the models not only predict accurately but also do so with reliability and minimal bias. The methodologies outlined in this chapter are instrumental in achieving the study's objectives, providing a clear and systematic approach to tackling the challenges of rainfall prediction in Selangor, Malaysia. The results derived from applying these methodologies, along with an in-depth discussion of their implications, will be presented in the subsequent chapter, where the study's findings will be analyzed and interpreted in detail.

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

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

CHAPTER 4

RESULT AND DISCUSSION

4.1 Overview

In this chapter, the focus will be on presenting and discussing the results obtained from each step of the study, providing a detailed analysis of how the various methods were applied and their outcomes. The chapter begins by exploring the visualization of missing data within the dataset, followed by a thorough examination of the imputation techniques used to address these gaps. Subsequently, the chapter delves into the process of determining the optimal number of components for Principal Component Analysis (PCA), a critical step in reducing the dimensionality of the atmospheric data. This reduction facilitates a more efficient analysis and enhances the predictive capabilities of the models. The chapter also discusses the application of a Hidden Markov Model (HMM) to effectively capture and handle zero-bounded data, which is a common characteristic in rainfall datasets. Following this, the chapter details the projection of rainfall using a SD-based HMM-ML model, encompassing a range of algorithms including Random Forest (SD-based HMM-RF), Support Vector Machine (SD-based HMM-SVM), Artificial Neural Network (SD-based HMM-ANN), Decision Tree (SD-based HMM-DT), and K-Nearest Neighbors (SD-based HMM-KNN). The

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

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

performance of each model is meticulously evaluated using a set of performance metrics, allowing for a comprehensive comparison of their predictive accuracy, reliability, and overall effectiveness in modeling daily rainfall in Selangor, Malaysia. Through this comparative analysis, the chapter aims to identify the most suitable model for accurate and reliable rainfall forecasting, contributing to the broader goal of improving climate-related predictions and decision-making.

4.2 Handling Missing Data

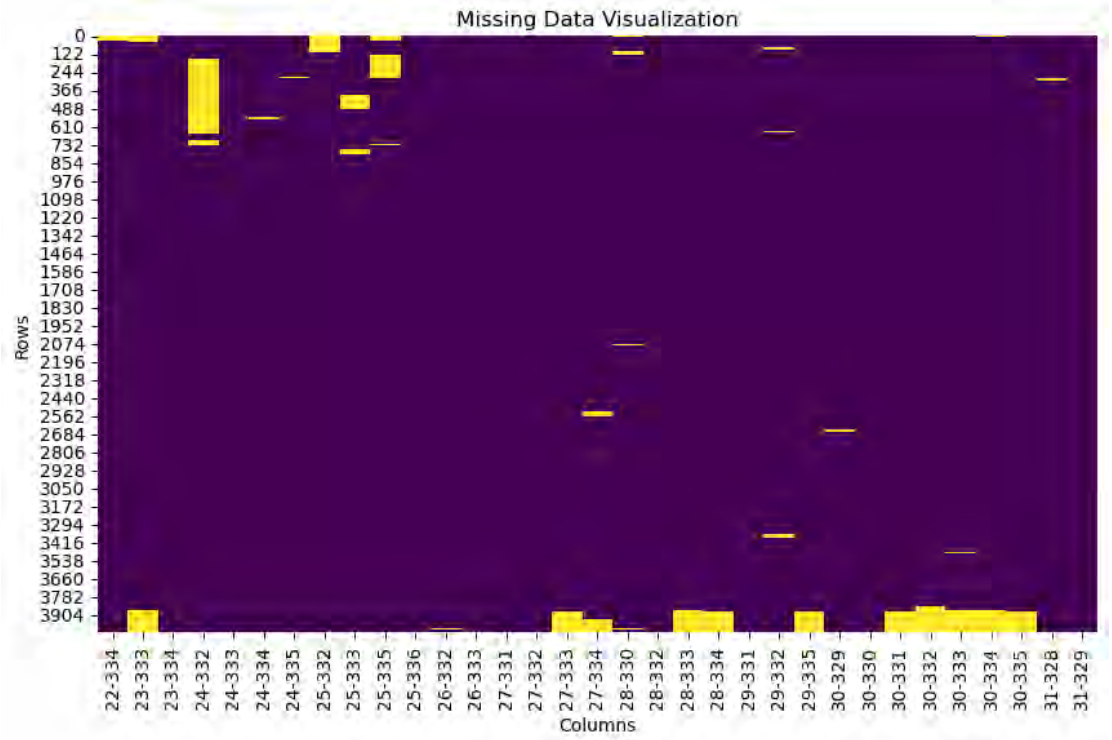
Based on the figure below, the daily rainfall data used in this study spans 132,594 days (from 2008 to 2018) with 2.1% missing values. According to Mirzaie et al. (2021), when missing data is less than 5%, the multiple imputation method may not offer significant benefits. The visualization of missing data was generated using the naniar package in R (Ghazali et. al., 2020). The yellow colour indicates missing data, while the purple colour represents data that is present. The highest percentage of missing data was from station 24-332 (Pusat Kawalan JPS Telok Gong) with 13%, followed by station 23-333 (Pekan Banting) and station 25-335 (Sg. Balak H. Langat) with 5%. As for the remaining stations, there were missing data percentages of 4%, 3%, 1%, or less.

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

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

Figure 4.1

The visualization of missing data



Seven different imputation methods were employed to address the missing data: mean imputation, median imputation, NIPALS, MCMC, the EM algorithm, kNN, and RF. Each of these methods was assessed based on performance metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Nash-Sutcliffe efficiency (NSE). The performance metrics measure the differences between values predicted by the model and the values actually observed. A lower RMSE and MAE indicate a better fit, as it signifies smaller discrepancies between the estimated and observed values. Meanwhile, NSE values range from negative infinity to 1, where 1 indicates a perfect model fit to observed data.

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

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

These imputation methods were compared because the performance of each method can vary depending on the type of variables and the characteristics of the dataset. By evaluating the performance for each method, the study aimed to identify which method provided the most accurate imputation for the missing data. The method with the lowest RMSE and MAE and the highest NSE was deemed the most effective, as it resulted in the smallest estimation errors and was the best approach for imputing missing values in the dataset. This careful comparison ensured that the chosen imputation method was robust and provided reliable estimates for the missing data.

Table 4.1

Performance values for each imputation methods

Method	RMSE	MAE	NSE
Mean Imputation	0.982	0.141	0.996
Median Imputation	0.146*	0.021*	1.000*
NIPALS	1.271	0.165	0.986
MCMC	2.097	0.254	0.980
EM Algorithm	3.318	0.280	0.950
***k*-Nearest Neighbor**	2.197	0.138	0.978
Random Forest	2.116	0.231	0.980

* indicates the best result

In this study, the imputation methods for estimating missing rainfall data were implemented through XLSTAT 2021 Software. The experimental results for each imputation method are presented in Table 4.1, revealing a clear hierarchy in their effectiveness. The median proved to be the most effective, with the lowest RMSE and

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

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

MAE values of 0.146 and 0.021, respectively, and the highest NSE value of 1, demonstrating its superior accuracy in estimating missing data. The findings underscore that median is the most suitable method for handling missing daily rainfall data in the Selangor state, providing the smallest estimation errors. Conversely, the EM Algorithm, with its highest RMSE, MAE and the lowest NSE value, is deemed the least effective method for this study, as it produced the largest discrepancies between the estimated and observed values.

4.3 Reduction of Dimensional Data

In this study, PCA was used as a dimension reduction approach to reduce the dimension by extracting the number of principal components (PC) without losing any significant information. The result of the analysis includes the eigenvalue, variation and cumulative percentage of variation is shown below in Table 4.2.

Table 4.2

Result of Principal Components (PCs)

	Eigenvalue	**Variability (%)**	**Cumulative (%)**
PC1	2.138	35.640	35.640
PC2	1.056	17.604	53.244
PC3	1.003	16.720	69.964
PC4	0.863	14.379	84.344
PC5	0.535	8.910	93.254
PC6	0.405	6.746	100.000

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

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

Table 4.2 presents a detailed analysis of six components derived from eigenvalues and total variation. Typically, in Principal Component Analysis (PCA), the number of components corresponds to the number of selected variables, with their respective percentages of variation and eigenvalues listed in descending order. Following the Kaiser criterion (Ocampo-Marulanda, 2022), components with eigenvalues exceeding 1.00 are considered significant for interpretation. Consequently, components PC1, PC2, and PC3, with eigenvalues of 2.138, 1.056, and 1.003 respectively, were identified as meaningful. Furthermore, it is recommended that the cumulative percentage of variation for hydrological data exceed 70% to be considered optimal (Shaharudin et al., 2018; Sulaiman et al., 2022). Therefore, components that achieve cumulative percentages of 83.344% and 93.254% were also selected as they surpassed this threshold.

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

Next, Yamamoto et al. (2014) defined factor loading as the correlation coefficients between principal component (PC) scores and the variables, which is essential for conducting statistical hypothesis testing and objectively selecting significant metabolites based on statistical criteria. Building on this, Kuznietsov and Biedunkova (2024) classified PC loadings with correlation coefficients as follows: ‘weak’ for 0.30-0.49, ‘moderate’ for 0.50-0.74, and ‘strong’ for coefficients greater than 0.75. This classification aids in distinguishing the impact of variables within the PCA framework, ensuring a rigorous selection process for significant components.

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

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

Table 4.3

Result of factor loadings for each predictor

	F1	F2	F3	F4	F5
Precipitation	0.440	-0.506	0.080	0.721*	-0.058
Relative Humidity	0.831*	0.101	-0.014	0.082	0.326
Solar	-0.744	-0.151	-0.254	0.168	0.573*
Wind	-0.142	0.822*	-0.204	0.505	-0.080
Maximum Temperature	-0.801*	-0.228	0.066	0.215	-0.251
Minimum Temperature	-0.200	0.198	0.941*	0.079	0.164

* indicates the best result

According to Table 4.3, the selection of predictor variables was based on the highest or strongest PC loadings for each variable, as outlined by Blackburn et. al. (2024). The analysis of factor loadings reveals distinct associations between the variables and their corresponding factors. Factor 1 is primarily characterized by a strong positive loading for relative humidity (0.831) and a strong negative loading for maximum temperature (-0.801), suggesting that this factor represents a contrast between high humidity and low maximum temperatures, where an increase in relative humidity tends to coincide with a decrease in maximum temperature, and vice versa. Factor 2 is dominated by wind, with a strong positive loading (0.822), indicating that this factor is closely associated with wind speed, making wind the most significant variable within this factor. Factor 3 is heavily influenced by minimum temperature, with a strong positive loading (0.941), reflecting the importance of variations in nighttime or early morning temperatures. Factor 4 is defined by precipitation, which has the highest positive loading (0.721), highlighting its significance in this factor. Factor 5 is

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

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

characterized by solar radiation, with a strong positive loading (0.573), suggesting that this factor represents the influence of solar radiation on environmental conditions. Due to their high correlation coefficients in their respective factors, these variables were selected as key predictors for further analysis. This selection process ensures that the most influential variables within each factor are considered, enhancing the accuracy and robustness of subsequent modeling or analytical efforts by capturing different aspects of environmental conditions. This led to the creation of a new dataset formatted as a matrix with 4,018 rows representing the number of days per station and five columns corresponding to the extracted factors (4,018 × 5).

4.4 The Homogeneous Hidden Markov Model (HMM)

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

To generate the posterior sample, the HMM package by Holsclaw (2020) was utilized in R software. Generating posterior samples from an HMM allows for the quantification of uncertainty in the predictions. A 10-year historical rainfall dataset from Selangor was used, marking "NA" for days without rain. The MCMC algorithm was run for varying iterations, ranging from 500 to 2000, with an additional 10% burn-in, and tested across hidden state layers from 0 to 10. At 2000 iterations, the parameter dataset can overcome redundancy of initial information while achieving stationarity (Yao et al., 2024). After the initial burn-in phase, 2000 iterations were deemed sufficient to obtain realizations from the posterior distribution.

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

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

The number of hidden states for the HMM model must be determined. By running the HMM with various values of K (number of hidden states) and applying a model selection criterion, the optimal number of hidden states can be selected from large datasets. In this study, the Bayesian Information Criterion (BIC) is used for model selection. The optimal number of hidden states corresponds to the BIC score that reaches its minimum turning point. The HMM model was tested with K values ranging from 0 to 10 (Cao et al., 2022), without incorporating climate variables such as temperature, relative humidity, or sea level pressure, ensuring a uniform distribution of rainfall throughout the year.

Table 4.4 presents the metrics for determining the number of hidden states (K) in the HMM model. It includes the BIC values and the number of parameters (p), with the minimum value marked by an asterisk (*). The model with K = 1 corresponds to a single constant state, equivalent to having no states (K = 0). The table compares the BIC values across different iterations for daily rainfall data in Selangor, showing that the BIC achieves a local minimum at three hidden states for each iteration from 500 to 2000. Additionally, it was observed that the BIC score reached a turning point at three hidden states at 2000 iterations.

Table 4.4

Comparison number of iterations for each hidden state layer based on BIC

k	Parameter Count	BIC			
		Iterations: 500	**Iterations: 1000**	**Iterations: 1500**	**Iterations: 2000**
0-1	297*	280,789.12	280,770.36	280,772.79	280,773.36
2	596	263,386.03	263,330.71	263,353.77	263,334.60
3	897	260,099.07*	260,054.62*	260,039.33*	260,018.30*

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

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

k	Parameter Count	BIC			
		Iterations: 500	Iterations: 1000	Iterations: 1500	Iterations: 2000
4	1,200	260,271.78	260,243.24	260,245.79	260,210.88
5	1,505	260,598.30	260,545.23	260,589.40	260,329.93
6	1,812	262,696.25	261,638.82	261,516.00	261,709.51
7	2,121	262,984.36	262,482.05	262,983.75	263,163.52
8	2,432	264,446.87	264,412.07	264,222.40	264,408.67
9	2,745	266,828.56	266,598.41	266,744.14	266,205.23
10	3,060	267,385.67	268,405.90	267,369.37	268,143.18

* indicates the lowest value

Figure 4.2 depicts the x-axis representing the number of hidden state and the y-axis representing a specific metric, namely the Bayesian Information Criterion (BIC). The graph includes four lines: orange, grey, yellow and blue, representing different iterations of the HMM. The orange line corresponds to iteration 500, the grey line corresponds to iteration 1000, the yellow line corresponds to iteration 2000 and the blue line corresponds to iteration 2000. When comparing the iteration counts, the BIC values for 500 iterations are generally higher than those for 1000, 1500, and 2000, indicating that more iterations lead to a better model fit. The lines for 1000, 1500, and 2000 iterations are relatively close, reflecting diminishing returns in BIC improvement with increased iterations beyond 1000. Ultimately, the lowest BIC is observed for 2000 iterations, suggesting it is the most optimal model among the tested iterations.

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

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

Figure 4.2

Comparison number of hidden state with the value of BIC

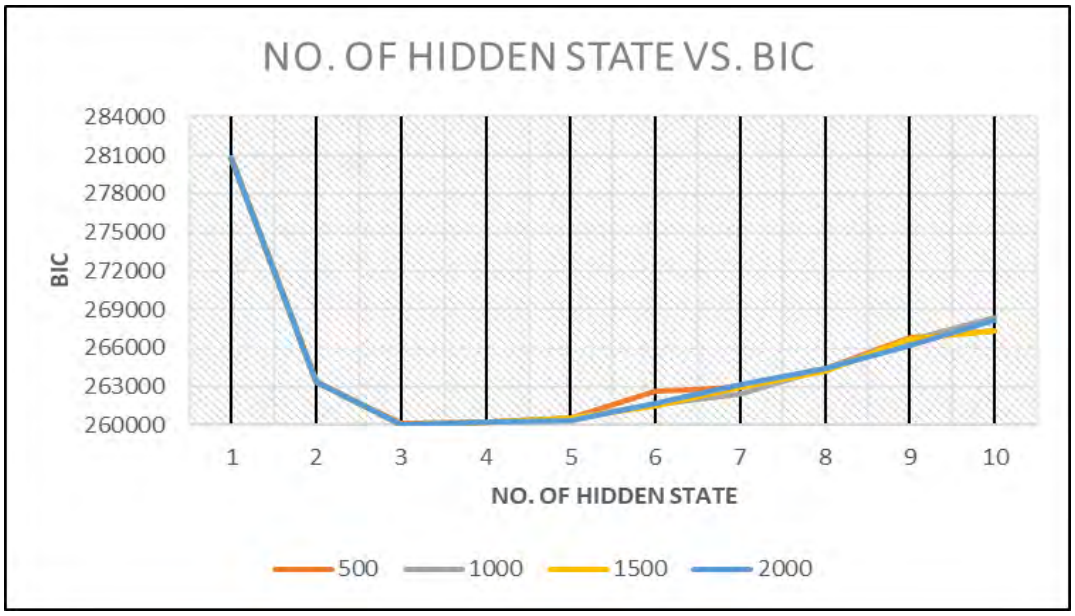


Figure 4.3 illustrates trace and density plots for the three hidden states, used to evaluate the convergence of the posterior distribution over 2000 iterations. The trace plots (the top plot for each state) display the evolution of the estimated posterior probabilities for each state at each iteration. Ideally, as the Markov Chain Monte Carlo (MCMC) sampling progresses, these plots should show a random, meandering pattern, indicating that the chain has sufficiently explored the parameter space. The density plots (the bottom plot for each state) show the distribution of the posterior probabilities for each state across all iterations. A well-defined, symmetric distribution resembling a bell curve is an indicator of convergence and good mixing of the Markov chain. In the context of K=3, the trace and density plots suggest that the posterior distribution has achieved perfect convergence. The trace plots show no discernible pattern or trend, indicating that the MCMC sampler has thoroughly explored the parameter space. The density plots exhibit well-defined, symmetric distributions, confirming the

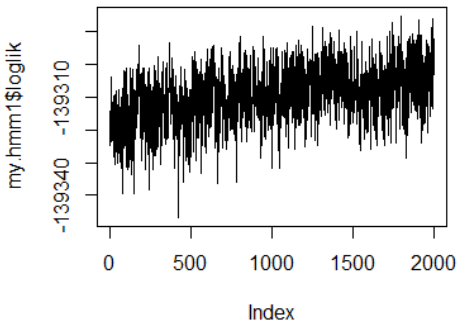
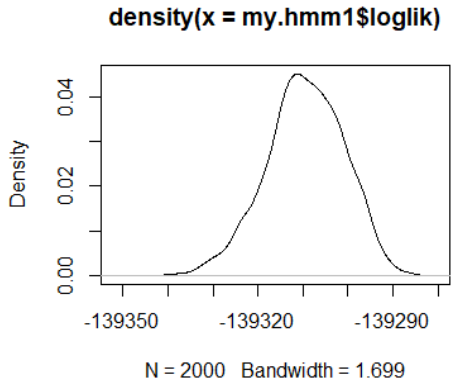
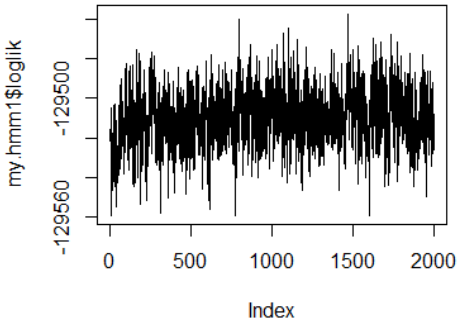
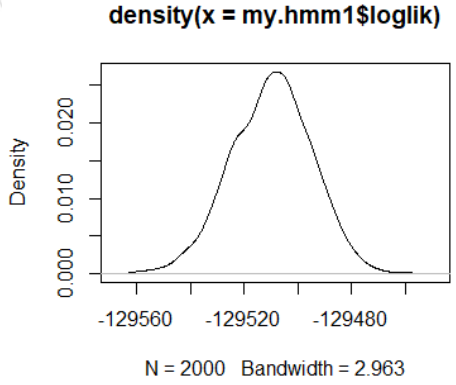
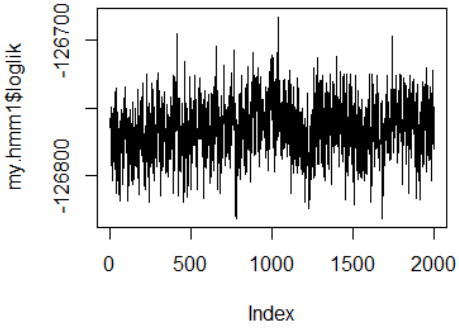
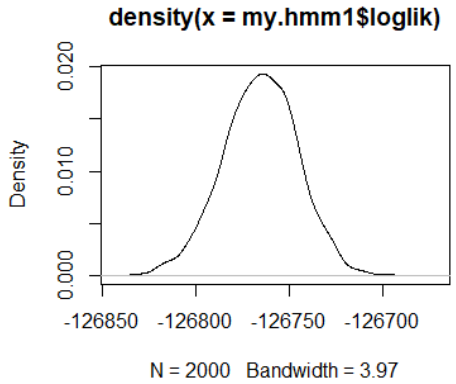
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

convergence. This is a strong indication that the estimated posterior probabilities for each state are reliable and accurate.

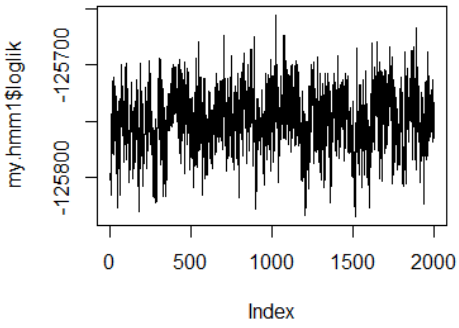
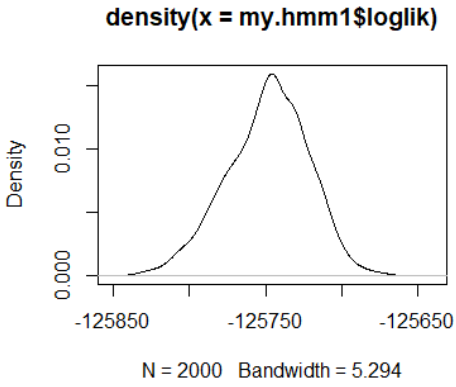
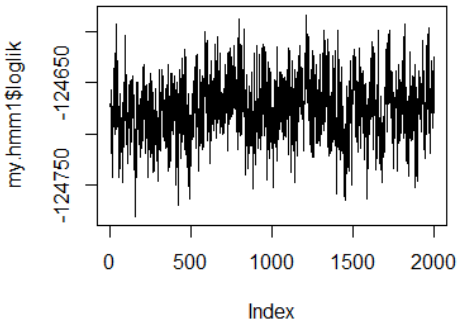
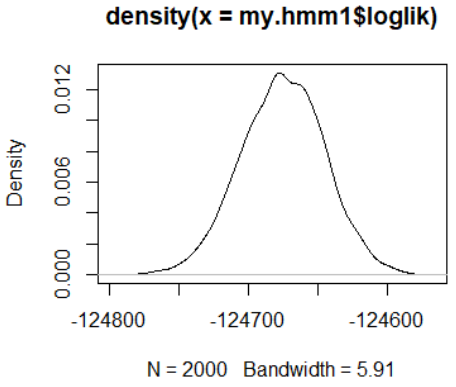
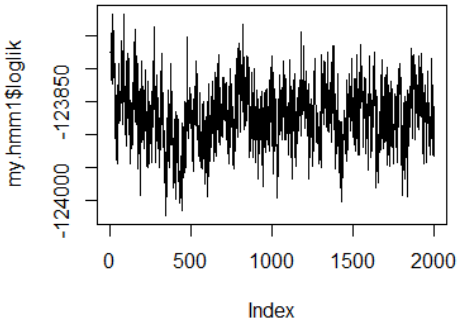
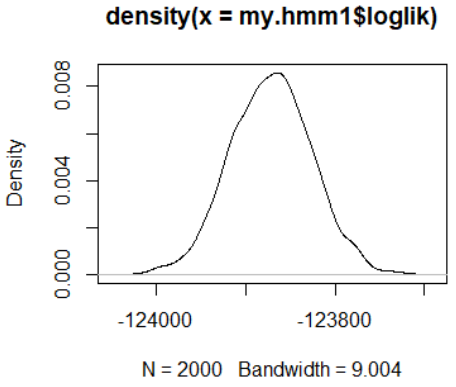
Figure 4.3

Trace and Density plot

K	Trace	Density Plot
1		
2		
3		

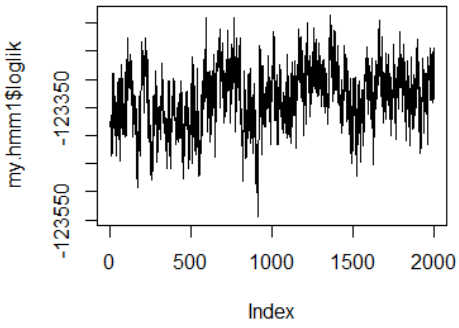
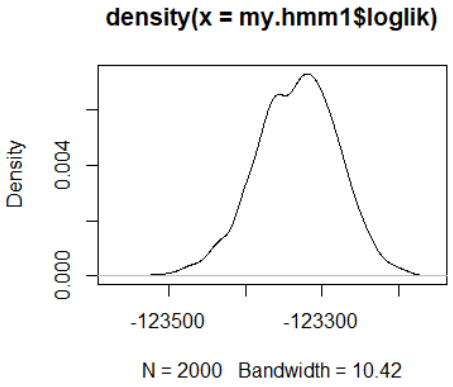
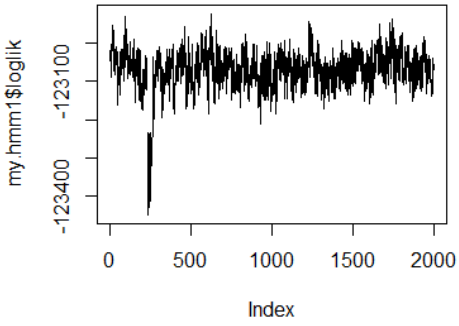
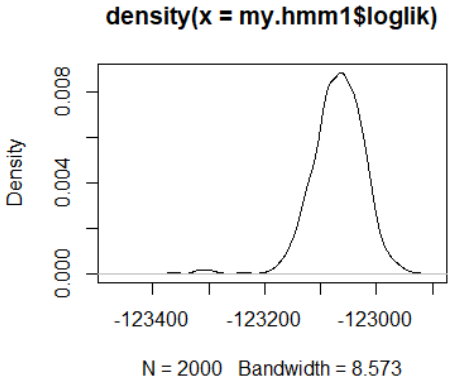
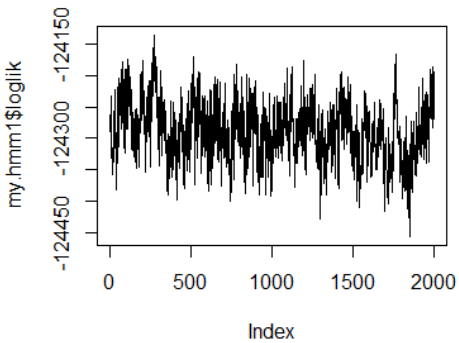
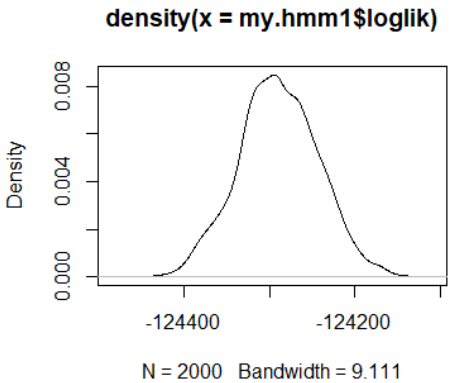
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

K	Trace	Density Plot
4		
5		
7		

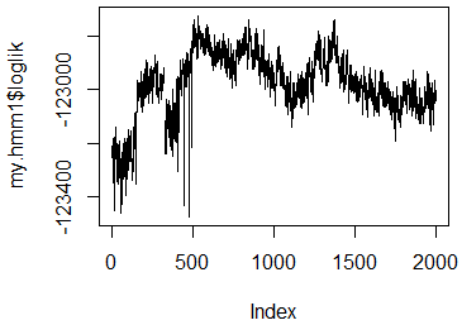
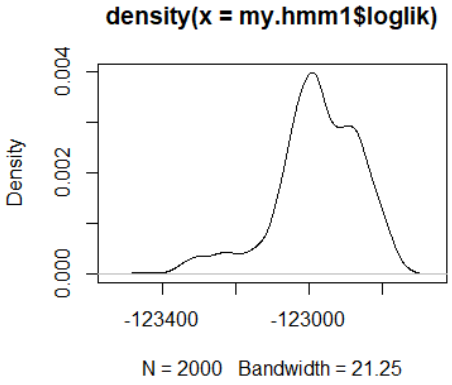
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

K	Trace	Density Plot
8		
9		
6		

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

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

K	Trace	Density Plot
10		

In a zero-inflated dataset, a significant proportion of the observations are zeros, which can be challenging to model using traditional distributions. However, the density plots for K=3 show a clear peak at zero, indicating that the model is able to capture the zero-inflation in the data. The fact that the density plots are well-defined and symmetric, with a clear peak at zero, suggests that the model is able to accurately represent the underlying distribution of the data, including the zero-inflation. This is a good indication that the model is effectively capturing the characteristics of the data, including the excess zeros (Wang et al., 2023). Thus, a new rainfall dataset with no excess zeros is created for further analysis.

4.5 A Hybrid of Statistical Downscaling-based Homogeneous Markov Model-Machine Learning (SD-based HMM-ML) model

Regression analysis is an essential tool for modeling and analyzing data, particularly in predictive analysis, such as forecasting rainfall (Mohammed et al., 2020). A hybrid of HMM-Statistical Downscaling-based Machine Learning model was used for rainfall

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

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

prediction in Selangor. The machine learning models used in this study include Random Forest (SD-based HMM-RF), Support Vector Machine (SD-based HMM-SVM), Artificial Neural Network (SD-based HMM-ANN), Decision Tree (SD-based HMM-DT), and K-Nearest Neighbors (SD-based HMM-KNN). Furthermore, the prediction values obtained from the HMM will be utilized to enhance predictive accuracy through machine learning techniques. Therefore, in this study, the *y*-variable is represented by a new dataset generated from an HMM, while the *x*-variable consists of principal components derived from PCA. Patel et al. (2023) also noted that a 70:30 split between the training and test datasets is optimal, based on empirical studies indicating that this ratio produces the best results.

4.5.1 Prediction modeling using an SD-based HMM-RF model

Random Forest creates several decision trees and combines their results to produce a more precise prediction (Patel et al., 2023). A random forest regression algorithm, consisting of 500 decision trees, was employed to forecast future rainfall trends. The dataset was divided into two sections, with 70% of the data used for the calibration phase and the remaining 30% for validation purposes (Kouadri et al., 2021).

The histogram in Figure 4.4 shows the distribution of rainfall values for both the original dataset (purple) and the SD-based HMM-RF model predictions (cyan). The histogram reveals a high concentration of values near zero, indicating many dry or low-rainfall instances in the dataset. The SD-based HMM-RF model frequently predicts rainfall between 2 and 4, suggesting a tendency to underestimate higher values, while

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

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

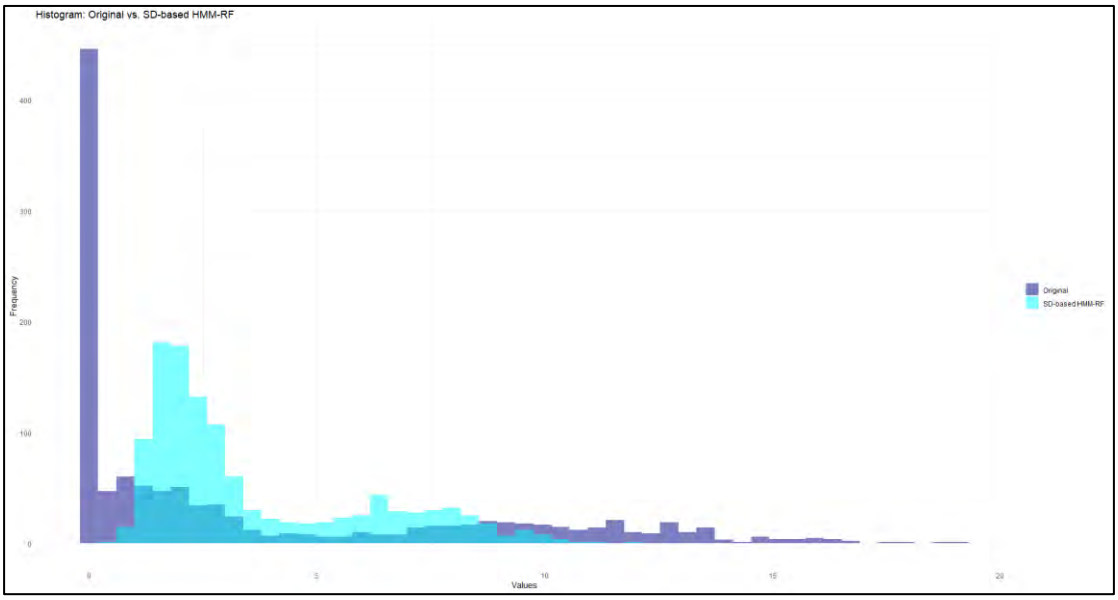
the original data exhibits a broader spread, including mid-range (4 - 6) and higher values. The lack of mid-range predictions suggests the model struggles to fully capture rainfall variability, potentially missing intermediate rainfall events. Although the model does capture some higher rainfall values (> 6), its lower frequency compared to the original data indicates room for improvement in modeling extreme rainfall distributions. According to Table 4.5, the Random Forest model's performance was evaluated using metrics such as RMSE, NSE, MAE, and MFE for both the calibration and validation periods. The model's RMSE of 2.298 and MAE of 1.900 indicate moderate prediction accuracy, with RMSE reflecting the average error magnitude. An NSE of 0.752 demonstrates decent predictive capability, as values close to 1 signify better performance. The MFE of -0.049, being close to zero, suggests minimal bias in the predictions. Figure 4.5 displays the model's performance in predicting rainfall amounts over several days, with blue bars representing actual rainfall and orange bars showing predicted values. While the model captures the overall trend, there are noticeable discrepancies, especially during higher rainfall periods. These results confirm that the model approximates the general rainfall pattern but struggles with precise predictions, particularly underestimating higher rainfall amounts, as indicated by the smoother predicted values in the graph.

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

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

Figure 4.4

The distribution of rainfall values for both the original dataset (purple) and the SD-based HMM-RF model predictions (cyan)



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

Table 4.5

The performance of SD-based HMM-RF model

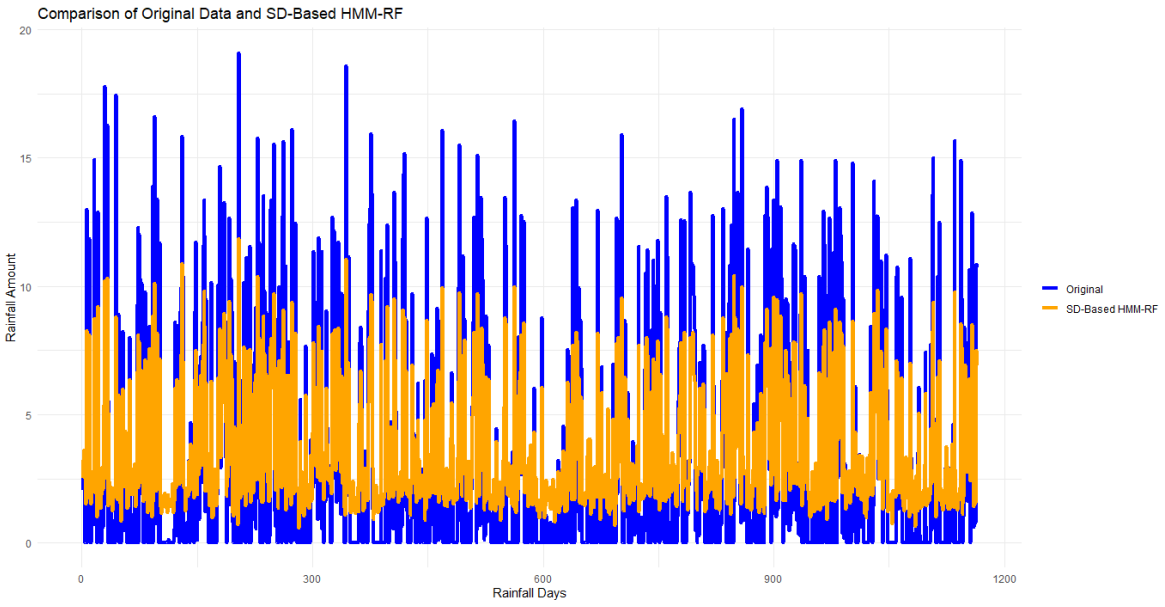
Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
2.298	0.752	1.900	-0.049	4.937	0.006	3.851	0.867

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

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

Figure 4.5

The performance of SD-based HMM-RF model in predicting rainfall amounts in calibration period



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

4.5.2 Prediction modeling using an SD-based HMM-SVM model

Before applying Support Vector Machine, Vapnik (1995) noted that the performance of SVM is influenced by the penalty terms, C and kernel function parameters. Thus, it is important to decide the kernel type and selecting the gamma (γ), penalty terms (C) and epsilon (ε) parameters. Various kernel models can be used to construct different SVMs, such as Sigmoid, Radial Basis Function (RBF), Polynomial, and Linear kernels (Sheykhmousa et. al., 2020). In this study, the RBF kernel is selected because it has consistently proven to be the most accurate and efficient option, particularly for rainfall prediction (Truong et al., 2021; Velasco et al., 2022; Sammen et al., 2023; Saha, 2023; Someetheram et al., 2024). Table 4.6 below shows a parameter used in this study.

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

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

Table 4.6

SVM parameters

	Parameters:
SVM-Type	Epsilon-Regression
SVM-kernel	Radial Basis Function
Cost	5
Gamma	0.5
Epsilon	0.1

The dataset will be divided into 70% for training and the remaining 30% for testing to minimize the discrepancies and understand the model's characteristics better (Rahman et al., 2022). The histogram illustrates the distribution of rainfall values for both the original dataset (red) and the SD-based HMM-SVM model predictions (green). The model's predictions are heavily concentrated at lower values, with a noticeable peak around zero, suggesting it frequently forecasts little to no rainfall. While the model does capture some higher rainfall values, the overall distribution appears more constrained compared to the original data, indicating potential underestimation of extreme events. The spread of predictions at mid-range values suggests that the RBF kernel captures some non-linear relationships, but the noticeable differences in distribution imply that the model may still struggle to fully represent the variability present in the original data.

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

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

Figure 4.6

The distribution of rainfall values for both original datasets (red) and the SD-based HMM-SVM model prediction (green)



Accompanying this visual are several performance metrics in Table 4.7: a Root Mean Square Error (RMSE) of 4.204, Mean Absolute Error (MAE) of 2.543, Nash-Sutcliffe Efficiency (NSE) of 0.170, and Mean Forecast Error (MFE) of 1.468. These metrics suggest that the model has some degree of error in its predictions, with the relatively low NSE indicating that the model's predictive power is not particularly strong. The positive MFE further suggests a tendency for the model to underpredict values.

Table 4.7

The performance values of SD-based HMM-SVM model

Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
4.204	0.170	2.543	1.468	5.687	-0.318	4.453	1.179

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

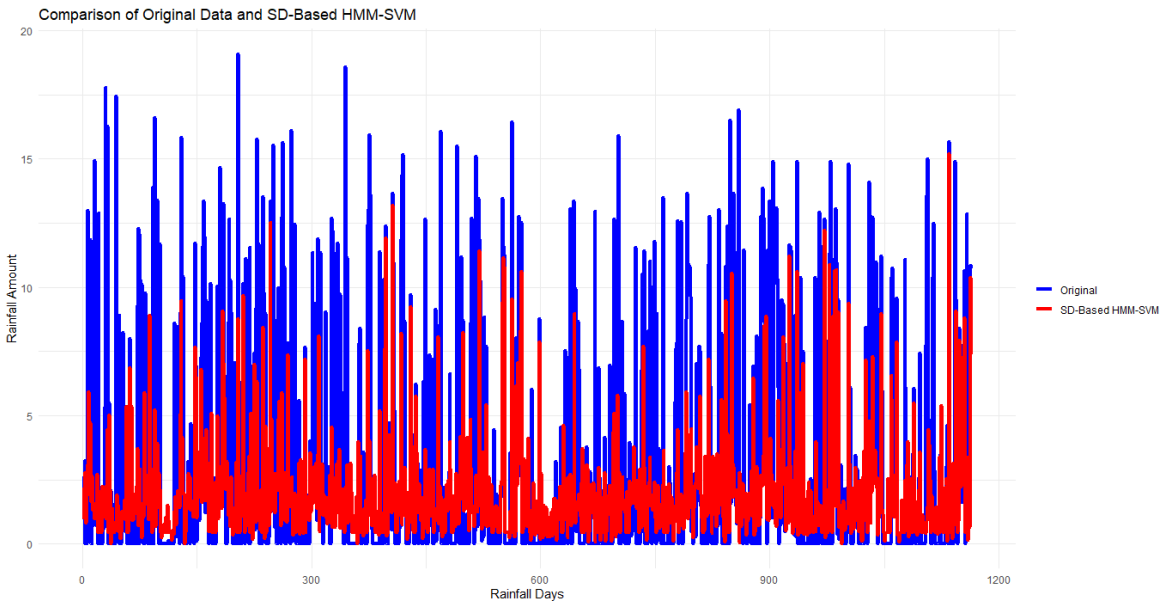
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

The graph in Figure 4.7 compares actual and predicted rainfall amounts over rainfall days, with actual values shown in blue and predicted values in red. The blue bars reveal significant variability in daily rainfall, with some peaks reaching 20-25 units, while the red bars show that the SVM model consistently underestimates these values, particularly failing to capture higher rainfall events. The model's predictions follow the general timing of rainfall but exhibit a much narrower range, indicating a tendency to smooth out predictions and miss the extremes present in the actual data. This underestimation aligns with previous findings, where the model struggled with accuracy, likely due to its current hyperparameters.

Figure 4.7

The performance of SD-based HMM-SVM model in predicting rainfall amounts in calibration period

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



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

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

4.5.3 Prediction modeling using an SD-based HMM-DT model

The combination of HMM and decision tree allows the decision tree to make predictions based on a more refined understanding of the rainfall process. Before applying the Decision Tree, a ratio of 70:30 again was used to split the datasets into training data and testing data respectively. DT builds a tree based on training data to predict continuous numerical values (Zhou et al., 2024).

Figure 4.8 shows the residuals versus fitted plot where the residuals represent errors between observed and predicted values on the y-axis meanwhile, the fitted values (predicted values) on the x-axis. The plot reveals heteroscedasticity and potential outliers, indicating that the variance of errors increases with the predicted values and suggesting issues with the model capturing data patterns. Figure 4.9 displays the Q-Q residuals plot which compares the standardized residuals to the theoretical quantiles of a standard normal distribution. The plot shows that the residuals deviate from normality, particularly in the tails, indicating that extreme values are more frequent than expected under a normal distribution.

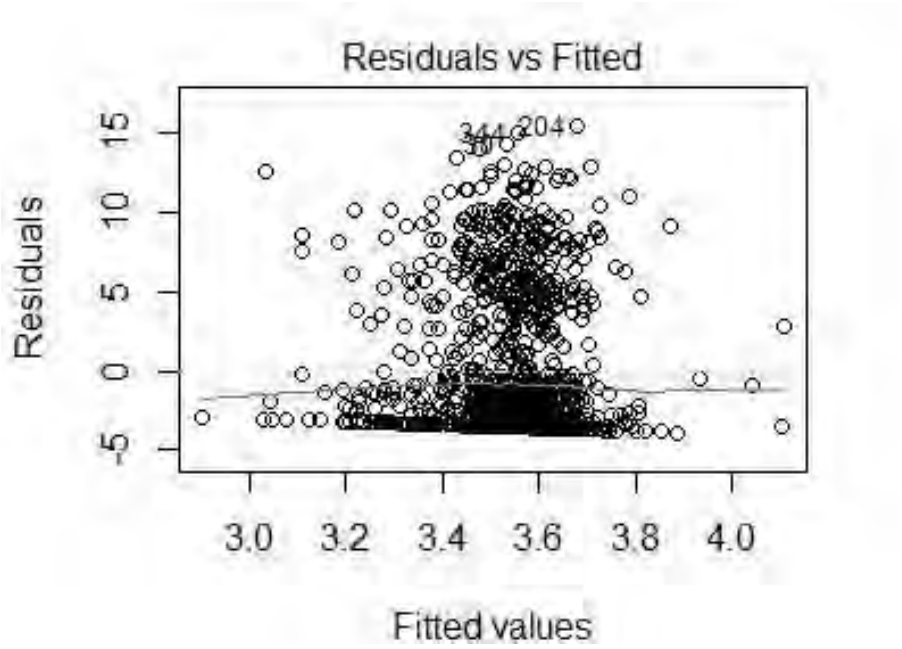
Next, the Scale-Location plot in Figure 4.10 is used to check the assumption of homoscedasticity (constant variance) in the residuals. The plot suggests mild heteroscedasticity, with a slightly increasing spread of residuals as the fitted values increase, indicating that the variance of residuals is not perfectly constant. The Residuals versus Leverage plot in Figure 4.11 is used to identify influential data points. The identifies a few influential data points with high leverage, which may disproportionately affect the model.

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

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

Figure 4.8

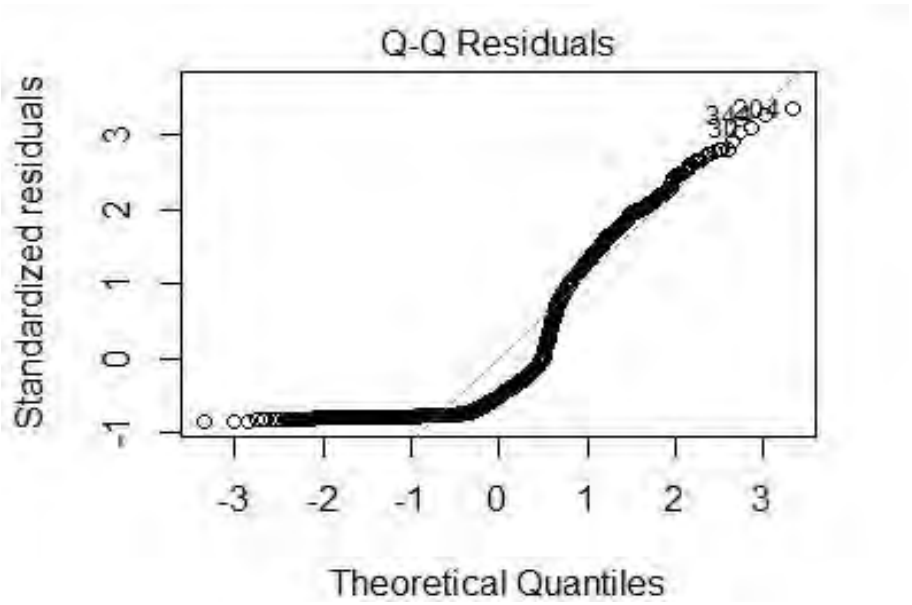
The Residuals vs. Fitted plot



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

Figure 4.9

The Q-Q Residuals plot

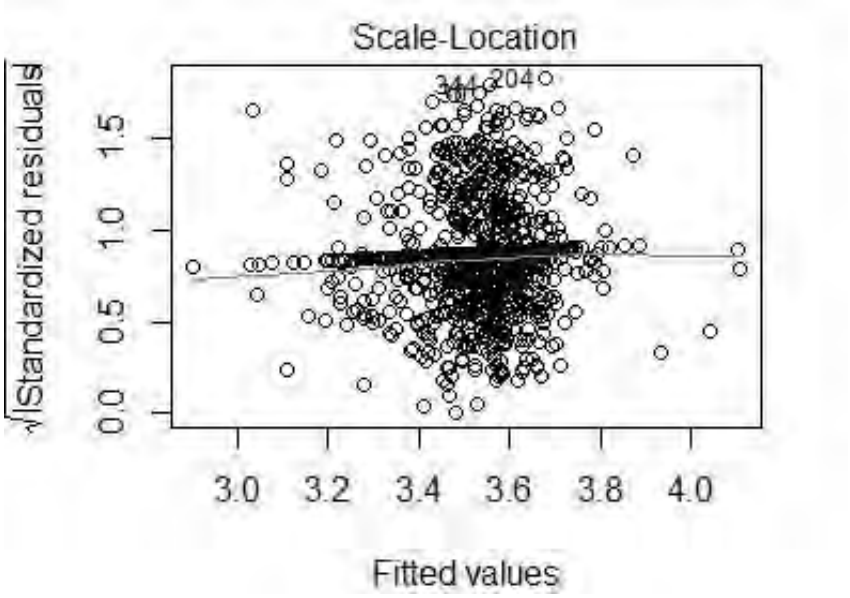


05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

Figure 4.10

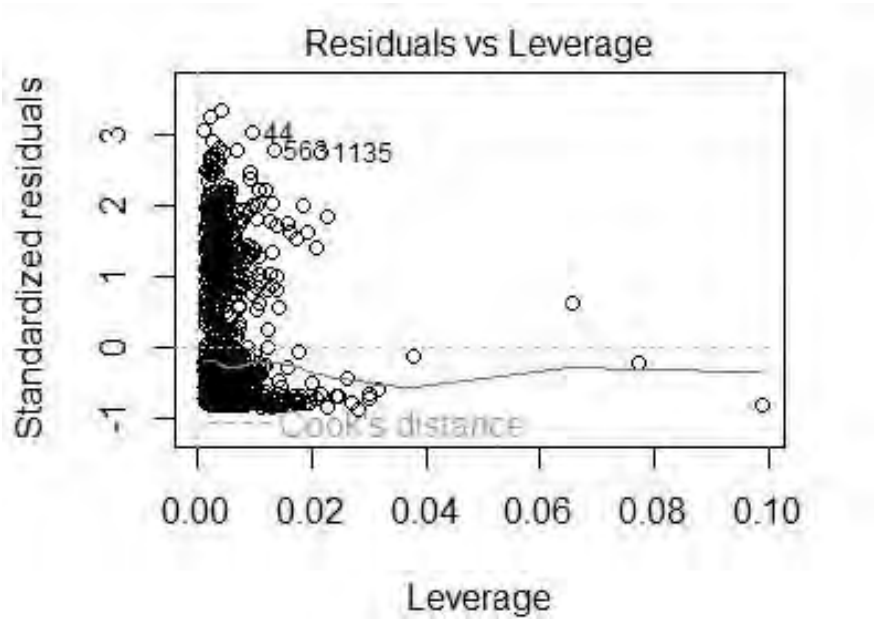
The Scale-Location plot



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

Figure 4.11

The Residuals vs. Leverage plot



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

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

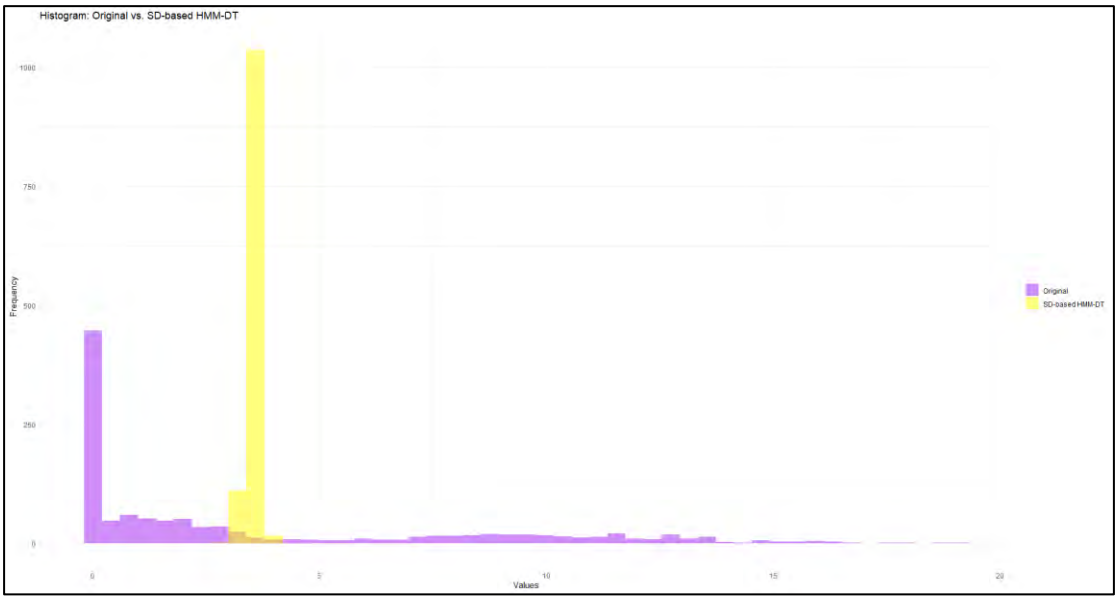
The histogram illustrates the comparison between the original dataset and the SD-based HMM-DT model's predictions. The original data distribution, represented in purple, is more dispersed, indicating a broader range of values. In contrast, the SD-based HMM-DT model, shown in yellow, exhibits a highly concentrated peak at a specific value, suggesting that the model generates predictions that fall within a narrow range. This pattern implies potential underfitting, as the model fails to capture the diversity and complexity present in the original data, resulting in overly similar outputs across different instances. It can be seen clearly in the graph in Figure 4.13 where the blue line represents actual rainfall amount and the orange line represents predicted rainfall amount. The blue line shows significant daily variability in rainfall, while the orange line remains nearly flat, indicating a lack of variation in the predicted values. This disparity highlights the poor performance of the decision tree model, which is further confirmed by its performance metrics in Table 4.8: an RMSE of 4.614, indicating notable prediction errors; an NSE of 0.001, reflecting that the model's predictions are barely better than using the mean of the observed data; and an MAE of 3.866, showing that predictions deviate by an average of 3.866 units from the actual values. Although the mean forecast error is close to zero (1.12e-11), indicating no systematic bias, this does not compensate for the overall inaccuracies.

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

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

Figure 4.12

The distribution of rainfall values for both original datasets (purple) and the SD-based HMM-DT model prediction (yellow)



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

Table 4.8

The performance values of SD-based HMM-DT model

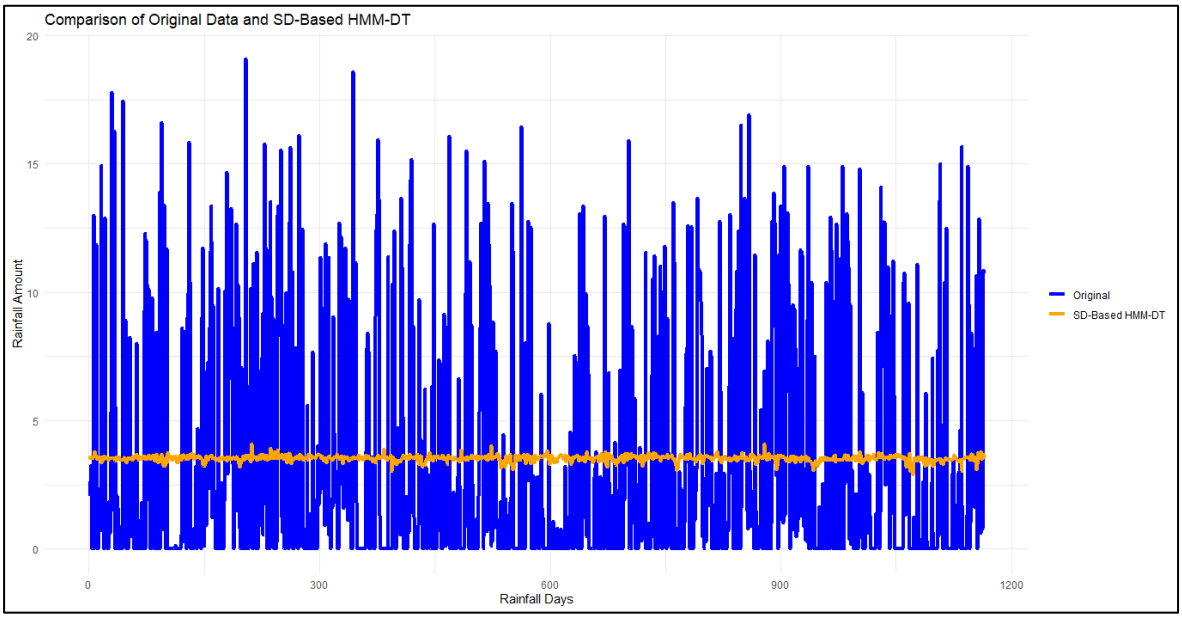
Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
4.614	0.001	3.866	1.12e-11	2.060	-0.044	4.096	1.054

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

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

Figure 4.13

The performance of SD-based HMM-DT model in predicting rainfall amounts in calibration period



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

4.5.4 Prediction modeling using an SD-based HMM-KNN model

HMM introduces variability in transition probabilities based on *x* variables, which could lead to more nuanced and accurate modeling of rainfall data. This dynamically modeled data can enhance the performance of KNN by providing more contextually accurate *y* variable. In applying k-Nearest Neighbour (KNN) model, the datasets were splitted into the ratio of 70:30 for training and test.

Appiah-Badu et al. (2022) stated that the effectiveness of KNN modeling depends on the number of neighbors (K) used. Table 4.9 and Figure 4.14 show the number of neighbors (k) used in the k-nearest neighbors (KNN) algorithm and the root

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

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

mean square error (RMSE) obtained using bootstrap resampling. As the number of neighbors increases from 5 to 9, the RMSE decreases, indicating that the model's prediction accuracy improves with more neighbors.

Table 4.9

Number of neighbors used in KNN

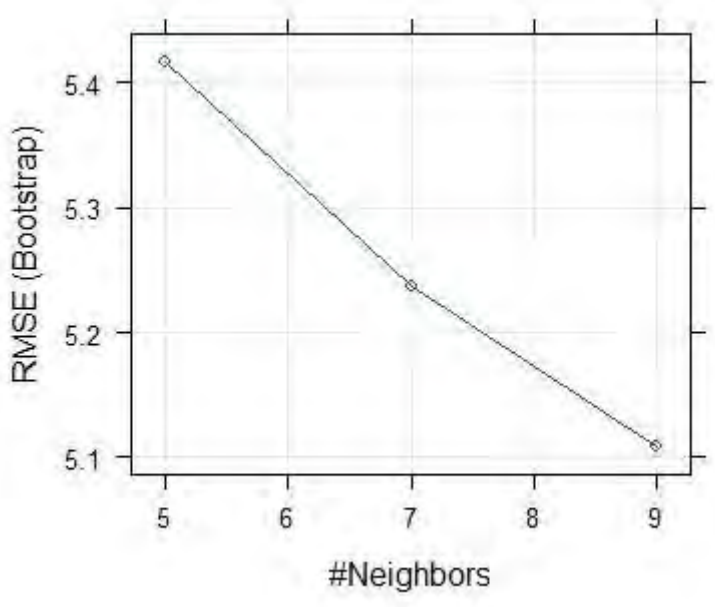
No. of neighbor, k	RMSE	R^2	MAE
5	5.417	0.001624	4.224
7	5.236	0.001664	4.144
9*	5.108*	0.001338*	4.081*

* indicates the best result

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

Figure 4.14

RMSE vs. Number of neighbors plot



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

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

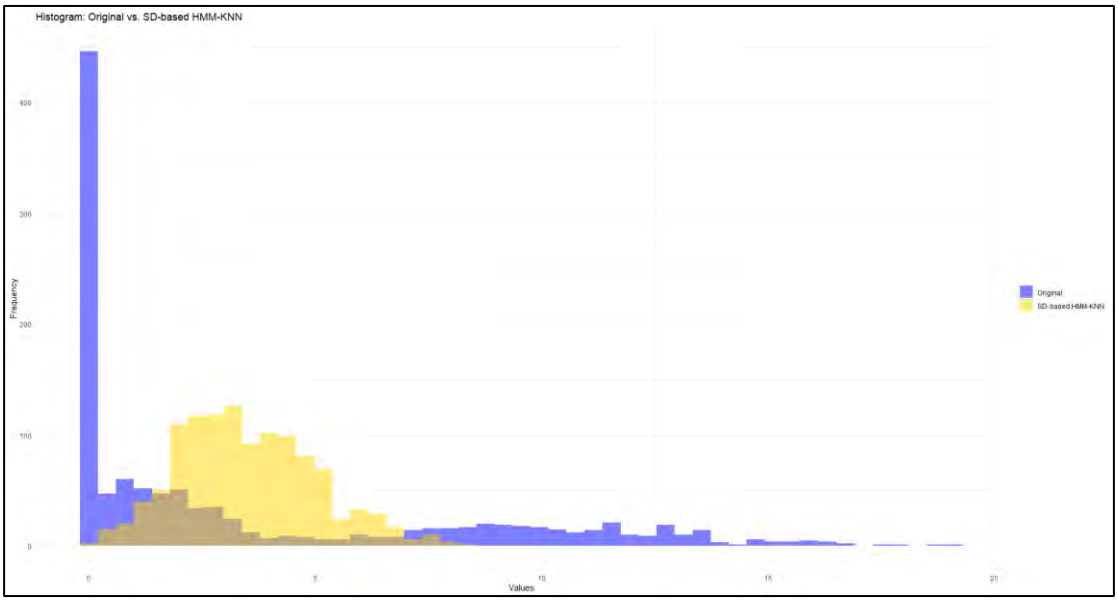
The histogram compares the frequency distribution of the original and SD-based HMM-KNN model data. The overlapping and spread of bars indicate the differences and similarities between the two datasets. The yellow bars representing the SD-based HMM-KNN appear more distributed across higher values, while the original dataset (blue bars) shows a higher frequency of lower values, indicating potential zero-inflated characteristics. This visualization suggests that the SD-based HMM-KNN captures a broader range of values compared to the original dataset. It can be seen clearly as the graph in Figure 4.16 shows a comparison between the original rainfall values and the predicted values. The original values are depicted in blue, while the predicted values are shown in orange. The graph suggests that the model does not perform particularly well in predicting daily rainfall amounts. The predictions tend to smooth out the variability, leading to underestimation of high rainfall days and overestimation of low rainfall days. This discrepancy underscores the poor performance of the KNN model, as evidenced by its performance metrics in Table 4.10: an RMSE of 4.400 suggests that while the model captures some trends, there is still considerable error in the predictions, an NSE of 0.091 suggests the model has limited skill in accurately predicting rainfall amounts, an MAE of 3.553 indicates the average magnitude of the errors in the predictions, without considering their direction and an MFE of 0.029 indicates that the model has no significant bias in overestimating or underestimating the rainfall amounts over the long term.

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

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

Figure 4.15

The distribution of rainfall values for both the original dataset (blue) and SD-based HMM-KNN model (yellow) prediction



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

Table 4.10

The performance values of SD-based HMM-KNN model

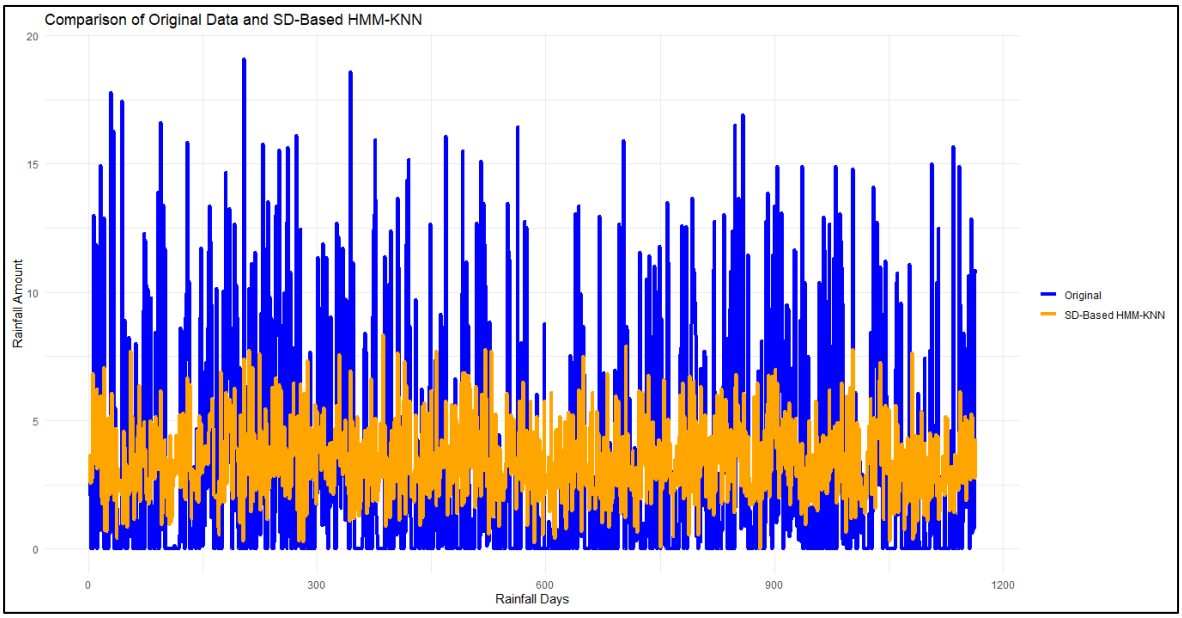
Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
4.400	0.091	3.553	0.029	5.248	-0.123	4.133	1.137

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

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

Figure 4.16

The performance of SD-based HMM-KNN model in predicting rainfall amounts in calibration period



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

4.5.5 Prediction modeling using an SD-based HMM-ANN Model

HMM enhances this model by allowing the transitions to vary based on *y* variable, leading to a more accurate and dynamic representation. This dynamic modeling of rainfall data can significantly enhance the performance of Artificial Neural Network (ANN), leading to more accurate and reliable predictions. In applying ANN, the datasets were divided into two parts, 70% of the data were used for calibrating and the remaining 30% of the data for validation purposes.

Figure 4.17 displays the neural network diagram. Based on the diagram, 5 neurons labeled as PC1 to PC5, representing principal components as input layer and two hidden layers with numerous neurons interconnected by weighted edges. The first

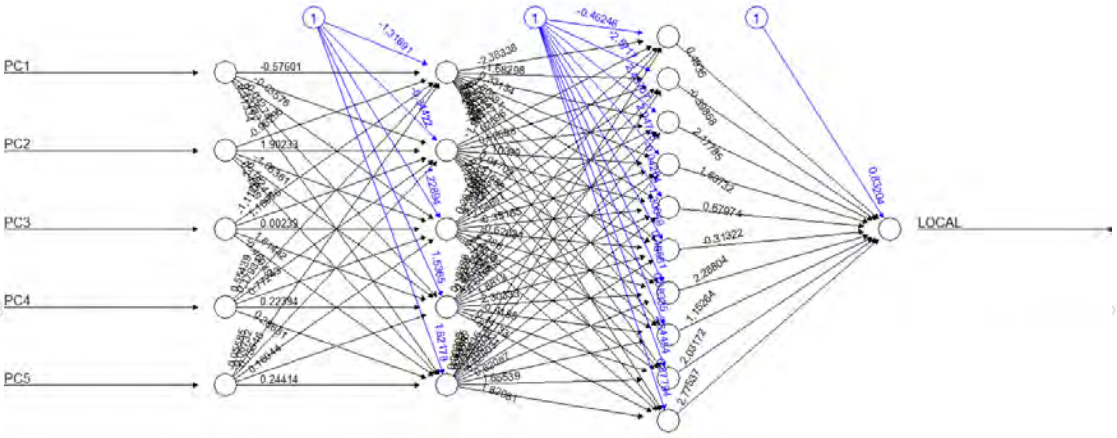
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

hidden layer has a dense connection from the input layer, and the second hidden layer receives inputs from the first hidden layer. Adjusted connection weights were essential in the ANN model as they enabled hidden neurons to receive all inputs from the input neurons, which are then used in the non-linear function (Tangri et al., 2008; Sulaiman et al., 2022).

Figure 4.17

Neural Network Diagram



The histogram illustrates the frequency distribution for the original and SD-based HMM-ANN datasets. The green bars representing the SD-based HMM-ANN data display a strong clustering near zero, indicating limited variability and potentially highlighting overfitting or model rigidity. Conversely, the original dataset (orange bars) demonstrates a broader distribution, capturing more variability across different values. The analysis of ANN performance in Table 4.11 indicates poor results of the ANN in predicting rainfall; a high RMSE of 5.265 indicates that there are significant errors in the model's predictions, a negative NSE (-0.302) highlights significant deficiencies in the model's ability to capture the variability in the data, an MAE of 3.367 confirms that

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

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

the predictions are substantially off from the actual values, and an MFE of 2.535 suggests a consistent overestimation bias in the model. Figure 4.19 illustrates a graph comparing original values (blue line) and predicted values (orange line). The model's predictions are almost a flat line near to zero which indicates that the model fails to capture any of the variability present in the actual rainfall data.

Figure 4.18

The distribution of rainfall values for both the original dataset (orange) and SD-based HMM-ANN model (green) prediction

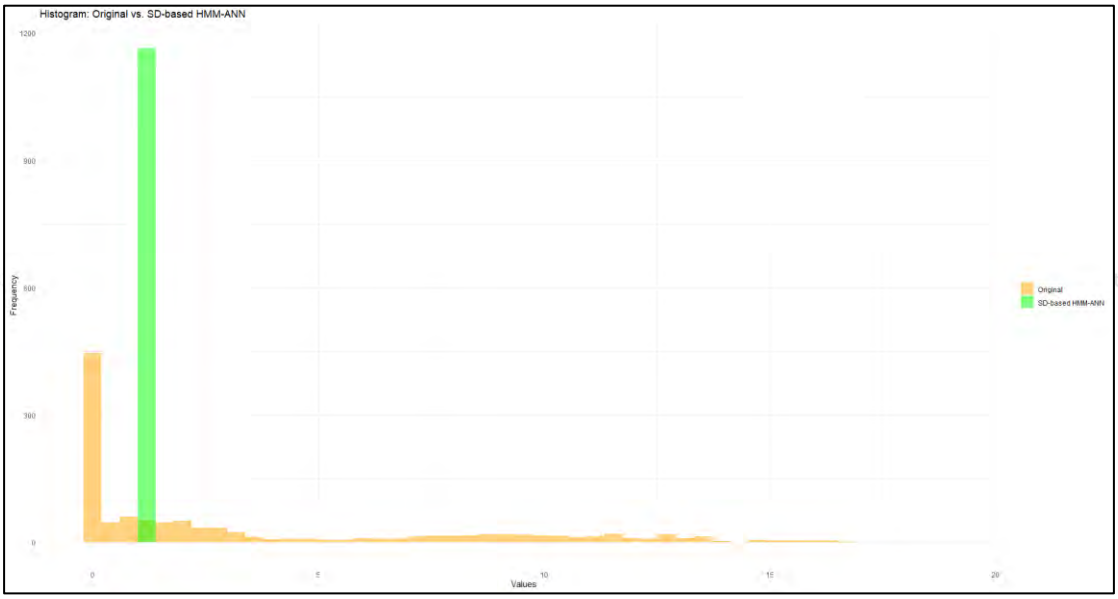


Table 4.11

The performance values of SD-based HMM-ANN model

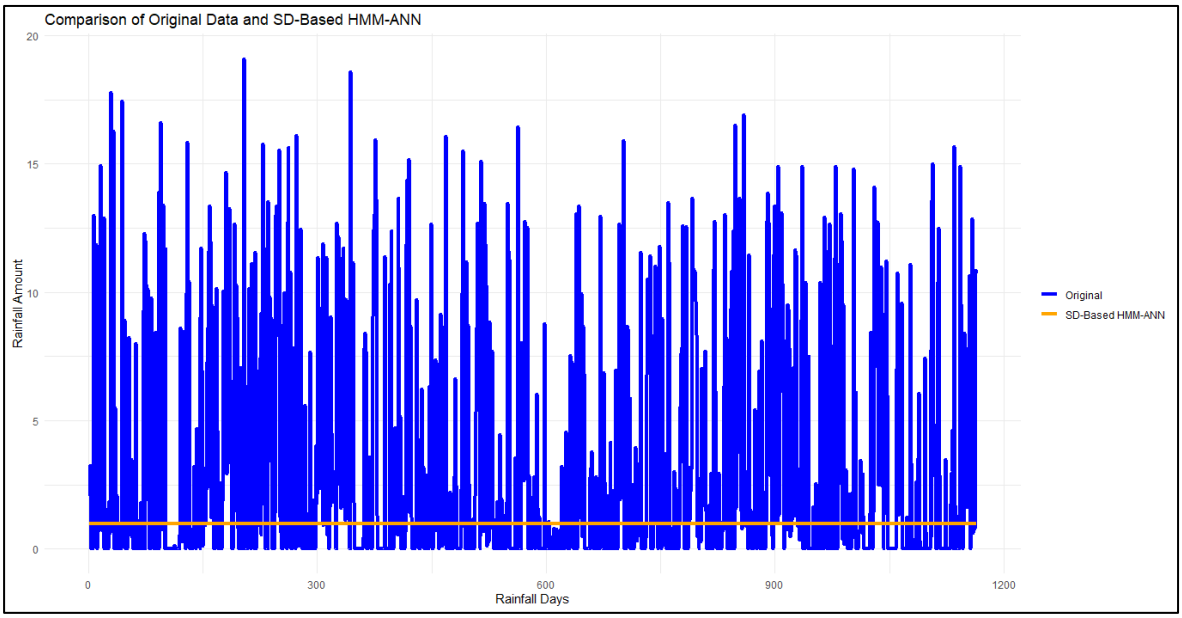
Calibration				Validation			
RMSE	**NSE**	**MAE**	**MFE**	**RMSE**	**NSE**	**MAE**	**MFE**
5.265	-0.302	3.367	2.535	5.991	-0.463	4.062	3.370

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

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

Figure 4.19

The performance of SD-based HMM-ANN model in predicting rainfall amounts in calibration period



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

4.6 Evaluating performance of Statistical Downscaling-based Homogeneous Markov Model-Machine Learning model (SD-based HMM-ML model)

A hybrid model is anticipated to enhance both the accuracy and efficiency compared to a single model (Sulaiman et al., 2022). The hybrid models that applied in this study are SD-based HMM-RF, SD-based HMM-SVM, SD-based HMM-DT, SD-based HMM-KNN and SD-based HMM-ANN. The performance of the proposed models was measured by using the performance metrics of Root Mean Square Error (RMSE), Nash-Sutcliffe Error (NSE), Mean Absolute Error (MAE) and Mean Forecast Error (MFE). Table 4.12 displays the overall results of the performance of the five hybrid models.

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

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

Table 4.12

The overall results of the performance of the five hybrid models

		SD-based HMM-RF	SD-based HMM-SVM	SD-based HMM-DT	SD-based HMM-KNN	SD-based HMM-ANN
RMSE	**70%**	2.298*	4.204	4.614	4.400	5.265
	30%	4.937*	5.687	5.060	5.248	5.991
NSE	**70%**	0.752*	0.170	0.001	0.091	-0.302
	30%	0.006*	-0.318	-0.044	-0.123	-0.463
MAE	**70%**	1.900*	2.543	3.866	3.553	3.367
	30%	3.851*	4.453	4.096	4.133	4.062
MFE	**70%**	-0.049*	1.468	1.12e-11	0.029	2.535
	30%	0.867*	1.179	1.054	1.137	3.370

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

* indicates the best result

In terms of RMSE, the SD-based HMM-RF model demonstrates the lowest values (2.298 and 4.937), indicating the highest accuracy in predictions among the models. Lower RMSE values signify a better fit of the model to the data. SD-based HMM-SVM, SD-based HMM-DT, SD-based HMM-KNN, and SD-based HMM-ANN models have higher RMSE values, with the SD-based HMM-ANN model showing the highest values (5.991), reflecting less accurate predictions.

For NSE, the SD-based HMM-RF model again leads with the highest values (0.752 and 0.006). NSE values close to 1 suggest better model performance, while negative values indicate performance worse than the mean of the observed data.

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

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

Statistical downscaling-based SD-based HMM-ANN exhibits negative NSE value, underscoring its lower efficiency and less reliable predictions.

When evaluating MAE, the SD-based HMM-RF model achieves the lowest values (1.900 and 3.851), indicating smaller average errors between predicted and actual values. This suggests that statistical downscaling-based SD-based HMM-RF model's predictions are closer to the actual values compared to the other models. The other models show higher MAE values, with statistical downscaling-based SD-based HMM-SVM model, SD-based HMM-DT model, SD-based HMM-KNN model, and SD-based HMM-ANN model all demonstrating larger errors.

The MFE values for SD-based HMM-RF model are -0.049 and 0.867, with the first value being the closest to zero, indicating minimal bias in predictions. MFE values close to zero reflect that the model neither consistently underestimates nor overestimates the predictions. In contrast, the other models exhibit higher MFE values, suggesting a more significant bias. Notably, SD-based HMM-ANN models the highest MFE values (2.535 and 3.370 respectively), indicating a pronounced bias in its forecasts.

The graph in Figure 4.20 displays the short forecast for two weeks of observed rainfall values (in blue) against predictions made by various hybrid models, including SD-based HMM-RF (red), SD-based HMM-SVM (green), SD-based HMM-KNN (pink), SD-based HMM-DT (orange), and SD-based HMM-ANN (brown) model. The graph covers a two-week forecast period since the Department of Irrigation and Drainage Selangor prioritizes short-term predictions. This timeframe provides essential

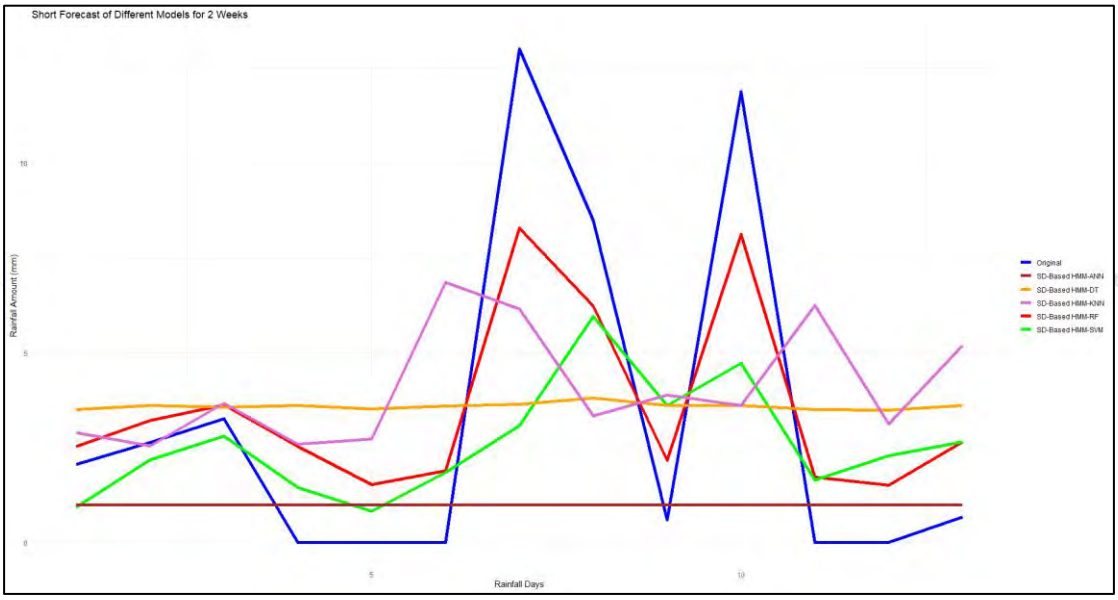
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

insights for flash flood mitigation efforts in specific locations over the upcoming two weeks. Among the models, the SD-based HMM-RF predictions (represented by the red line) show a pattern that is generally more aligned with the peaks and troughs of the original rainfall values (blue line). This suggests that the RF component is effective at capturing the variability and dynamics of the rainfall data when informed by the HMM.

Figure 4.20

The short forecast for two weeks of five hybrid models in predicting daily rainfall amount



Overall, SD-based HMM-RF consistently outperforms the other models across the majority of the metrics. It provides the most accurate and reliable predictions with minimal bias, as indicated by its superior performance in RMSE, NSE, and MAE, and its near-zero MFE. SD-based HMM-ANN tends to perform the worst, with higher errors and lower efficiency. SD-based HMM-SVM, SD-based HMM-DT and SD-based HMM-KNN show intermediate performance but are generally outclassed by SD-based HMM-RF.

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

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

4.7 Summary

The primary objective of this study is to forecast daily rainfall data in Selangor, Malaysia, using a suite of hybrid Statistical Downscaling-based Homogeneous Markov Model-Machine Learning model (SD-based HMM-ML model). The proposed models include SD-based HMM-RF, SD-based HMM-SVM, SD-based HMM-DT, SD-based HMM-KNN, and SD-based HMM-ANN. Among these, the SD-based HMM-RF model stands out as the most effective, consistently delivering superior performance across a variety of evaluation metrics.

Specifically, the SD-based HMM-RF model exhibits the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), reflecting its precision in predicting daily rainfall amounts. It also achieves the highest Nash-Sutcliffe Efficiency (NSE), a strong indicator of its predictive power. Moreover, the Mean Forecast Error (MFE) for the SD-based HMM-RF model is closest to zero, signifying that the model generates highly accurate predictions with minimal bias. The exceptional performance of the SD-based HMM-RF model is particularly noteworthy in its ability to accurately predict extreme rainfall events, which are critical for flood risk assessment and water resource management.

The robustness and reliability of the SD-based HMM-RF model make it an invaluable tool for hydrologists and climatologists who are engaged in environmental modeling and climate change research. Its capability to deliver precise predictions with minimal bias enhances the understanding of local climate dynamics and supports more informed decision-making in the context of climate change adaptation and mitigation

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

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

strategies. Consequently, the SD-based HMM-RF model emerges as a powerful asset in the ongoing efforts to improve climate resilience in the Selangor region and beyond.

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

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

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

CHAPTER 5

CONCLUSION

5.1 Overview

In this concluding chapter, we summarize the contributions of this thesis, examine its limitations, and suggest potential future research directions arising from this work.

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

5.2 Summary

Extreme weather events such as floods and tsunamis are likely to increase climate change's impact on the hydrological cycle. Continuous heavy rainfall over several days is the primary cause of flooding, which can disrupt daily life for weeks. This underscores the growing importance of projecting future climate at regional and local levels, given the unpredictability of extreme climate events in recent years. As a result, meteorologists and hydrologists are employing hybrid Statistical Downscaling-based Bayesian Approach-Machine Learning model (SD-based HMM-ML model) to evaluate predicted rainfall in Selangor, Malaysia, with a particular emphasis on severe rainfall events.

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

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

The first objective of this study is to evaluate the effectiveness of various imputation methods in dealing with missing data, specifically by determining the most suitable approach for handling the 2.1% of missing data in Selangor. The study rigorously applied several imputation techniques, including Mean Imputation, Median Imputation, k-Nearest Neighbor, EM Algorithm, MCMC, NIPALS, and Random Forest, to identify which method yields the most accurate and reliable results. The analysis revealed that Median Imputation emerged as the most effective method for addressing the missing daily rainfall data in Selangor. With a root mean square error (RMSE) of 0.146, a mean absolute error (MAE) of 0.021, and a Nash-Sutcliffe efficiency (NSE) of 1.000, Median Imputation provided the smallest estimation errors compared to the other methods tested. These findings highlight the robustness of the Median Imputation method in accurately estimating missing data, making it an essential tool for improving the quality and reliability of environmental data analysis in the region. This study not only emphasizes the importance of choosing the appropriate imputation method for accurate data recovery but also contributes valuable insights into the best practices for handling missing data in hydrological studies.

Next, the second objective of this study focuses on the reduction of data dimensionality and the selection of predictor variables through the use of Principal Component Analysis (PCA). Hydrologists frequently handle high-dimensional time-series data, which not only demands substantial computational resources but also presents challenges in accurately classifying patterns, recognizing trends, and visualizing the data when applying machine learning models. The complexity associated with high-dimensional datasets can hinder the effectiveness of these models, making it crucial to simplify the data without losing critical information. To tackle this

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

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

issue, PCA was employed as a dimensionality reduction technique aimed at identifying uncorrelated features, also known as principal components, thereby streamlining the dataset. By focusing on these principal components, the study was able to reduce the complexity of the data while retaining the most informative aspects. Specifically, PCA identified five key principal components, labeled PC1 through PC5, with a cumulative variance cut-off set at 93.254%. This substantial reduction in dimensionality resulted in the formation of a new dataset, structured as a matrix with 4,018 rows—each representing the number of days per station—and five columns corresponding to the extracted factors (4,018 × 5 matrix). This approach not only made the data more manageable and computationally efficient but also enhanced the accuracy of subsequent analyses by focusing on the most significant features of the dataset.

The third objective of this study centers on addressing the uncertainties that arise when dealing with zero-bounded rainfall data, a common feature in hydrology datasets. These datasets often contain a significant number of zero values, which can create challenges in data analysis, especially when the proportion of zeros is so high that standard statistical distributions fail to accurately represent the data. Such datasets are referred to as zero-inflated, and they require specialized methods to manage the excess zeros effectively. To tackle this issue, the study employed the Homogeneous Markov Model (HMM), a statistical tool designed to handle zero-inflated data by capturing the underlying patterns and distributions more accurately. The application of HMM in this context revealed a well-defined, symmetric distribution resembling a bell curve, which emerged when the data were analyzed at three hidden states. This symmetry indicates that the model provides reliable and consistent parameter estimates, crucial for accurately representing the data. Specifically, the density plots for K=3, with

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

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

an iteration count of 2000 and a Bayesian Information Criterion (BIC) value of 260,018.30, showed a clear peak at zero. This peak confirms that the model effectively captures the zero-inflation present in the data, accurately reflecting the underlying distribution. The fact that the density plots are both well-defined and symmetric, with a pronounced peak at zero, suggests that the HMM is not only capable of modeling the data's distribution but also of doing so with a high degree of precision. Consequently, the study was able to generate a new rainfall dataset that eliminates the excess zeros, thereby providing a more refined and accurate dataset for subsequent analysis. This improved dataset allows for more robust and reliable analyses, paving the way for better insights and predictions in hydrological studies.

The fourth objective is to investigate and compare hybrid Statistical Downscaling-based Bayesian and Machine Learning models for capturing non-linear relationships between predictor variables and rainfall patterns. This study implemented Statistical Downscaling-based Homogeneous Markov Model-Machine Learning model (SD-based HMM-ML model). The machine learning models used in this study are SD-based HMM-RF, SD-based HMM-SVM, SD-based HMM-DT, SD-based HMM-KNN and SD-based HMM-ANN for rainfall prediction. Next, comparing the five hybrid models by using four performance metrics namely, RMSE, NSE, MAE, and MFE. SD-based HMM-RF outperforms the other models across the majority of the metrics. Its superior performance in RMSE (2.298), NSE (0.752), and MAE (1.900), along with its near-zero MFE (-0.049), indicates that it offers the most accurate and reliable predictions with minimal bias. SD-based HMM-RF model holds potential benefits for hydrologists and climatologists, especially in downscaling studies aimed at projecting long-term rainfall trends. The research has successfully addressed challenges such as

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

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

missing data, high-dimensional datasets, zero-bounded data, and the processes involved in statistical downscaling approaches. The results of this analysis can enhance understanding of rainfall patterns in Selangor by providing an objective depiction of their primary downscaling predictions. Furthermore, these findings could aid in forecasting rainfall patterns, thereby mitigating the impact of disasters that threaten lives, cause property damage, and harm the environment.

5.3 Future Research

Several recommendations are put forward for future research based on the constraints identified, particularly concerning statistical downscaling methods.

i. The methodologies presented in this thesis were tested using a specific set of predictors and predictands in Selangor. Nevertheless, the techniques outlined in this study could potentially be applicable to the entire country of Malaysia and other tropical climate datasets, especially where rainfall variations are not significantly different from those in equatorial regions. Our hybrid model may offer improved generalization of rainfall projections produced through statistical downscaling for such datasets. Researchers are encouraged to explore these proposed methods in studies encompassing other tropical regions.

ii. The challenge of managing missing data in this study was addressed by employing the single imputation method, chosen for its superior performance. However, there exist numerous other imputation methods, including more advanced approaches, that could be explored for handling missing values.

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

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

Researchers are encouraged to investigate different imputation methods to identify the most effective approaches for addressing missing data.

iii. The machine learning-based methodology shows improved accuracy in predicting rainfall projections for this specific type of rainfall data. However, this study only presents visualizations of the rainfall projections. Therefore, the hybrid model proposed in this study is suggested for future rainfall projections covering periods of 20-30 years. Additionally, this approach could be applicable to other climate datasets beyond rainfall, provided they exhibit similar characteristics to the rainfall data analyzed in this study.

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

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

REFERENCES

Abdel Azeem, M., Dey, P., & Dev, S. (2023). A Multidimensionality Reduction Approach to Rainfall Prediction.

Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., & Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electronics*, *8*(3), 322.

Addi, M., Gyasi-Agyei, Y., Obuobie, E., & Amekudzi, L. K. (2022). Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Hydrological Sciences Journal*, *67*(4), 613-627.

Adnan, R. M., Mostafa, R. R., Dai, H. L., Heddam, S., Kuriqi, A., & Kisi, O. (2023). Pan evaporation estimation by relevance vector machine tuned with new metaheuristic algorithms using limited climatic data. Engineering Applications of Computational Fluid Mechanics, 17(1), 2192258.

Agarwal, S., Mukherjee, D., & Debbarma, N. (2023). Analysis of extreme annual rainfall in North-Eastern India using machine learning techniques. *AQUA—Water Infrastructure, Ecosystems and Society*, *72*(12), 2201-2215.

Aguilera, H., Guardiola-Albert, C., & Serrano-Hidalgo, C. (2020). Estimating extremely large amounts of missing precipitation data. *Journal of Hydroinformatics*, *22*(3), 578-592.

Agustian, H., Pujiastuti, A., & Sayoga, M. V. (2020). Comparison Of Simple Moving Average and Exponential Smoothing Methods to Predict Seaweed Prices. *CCIT J.*

Aieb, A., Madani, K., Scarpa, M., Bonaccorso, B., & Lefsih, K. (2019). A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. *Heliyon*, *5*(2).

Ali, S., & Shahbaz, M. (2020). Streamflow forecasting by modeling the rainfall–streamflow relationship using artificial neural networks. *Modeling Earth Systems and Environment*, *6*(3), 1645-1656.

Ali, S., Eum, H. I., Cho, J., Dan, L., Khan, F., Dairaku, K., ... & Fahad, S. (2019). Assessment of climate extremes in future projections downscaled by multiple statistical downscaling methods over Pakistan. *Atmospheric Research*, *222*, 114-133.

Alipour, M., Bejani, M., & Salehi, A. H. (2023). Statistical Downscaling of Global Climate Models for Temperature Trend Analysis in Calgary. *Engineering Proceedings*, *56*(1), 320.

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

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

Al-Mukhtar, M., & Qasim, M. (2019). Future predictions of precipitation and temperature in Iraq using the statistical downscaling model. *Arabian journal of geosciences*, *12*(2), 25.

Anaraki, M. V., Achite, M., Farzin, S., Elshaboury, N., Al-Ansari, N., & Elkhrachy, I. (2023). Modeling of monthly rainfall–runoff using various machine learning techniques in Wadi Ouahrane Basin, Algeria. *Water*, *15*(20), 3576.

Anaraki, M. V., Farzin, S., Mousavi, S. F., & Karami, H. (2021). Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods. *Water Resources Management*, *35*, 199-223.

Appiah-Badu, K., Anning, A. K., Eshun, B., & Mensah, G. (2022). Land use effects on tree species diversity and soil properties of the Awudua Forest, Ghana. *Global Ecology and Conservation*, *34*, e02051.

Asteris, P. G., & Mokos, V. G. (2020). Concrete compressive strength using artificial neural networks. *Neural Computing and Applications*, *32*(15), 11807-11826.

Avtar, S. S., Khuneswari, G. P., Abdullah, A. A., McColl, J. H., Wright, C., & GMS Team. (2019, November). Comparison between EM algorithm and multiple imputation on predicting children's weight at school entry. In *Journal of Physics: Conference Series* (Vol. 1366, No. 1, p. 012124). IOP Publishing.

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and quality of life outcomes*, *17*, 1-9.

Bacro, J. N., Gaetan, C., Opitz, T., & Toulemonde, G. (2020). Hierarchical space-time modeling of asymptotically independent exceedances with an application to precipitation data. *Journal of the American Statistical Association*, *115*(530), 555-569.

Baljon, M., & Sharma, S. K. (2023). Rainfall prediction rate in saudi arabia using improved machine learning techniques. *Water*, *15*(4), 826.

Banadkooki, F. B., Ehteram, M., Ahmed, A. N., Fai, C. M., Afan, H. A., Ridwam, W. M., ... & El-Shafie, A. (2019). Precipitation forecasting using multilayer neural network and support vector machine optimization based on flow regime algorithm taking into account uncertainties of soft computing models. *Sustainability*, *11*(23), 6681.

Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, *7*, 100204.

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

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

Benmakhlouf, M., El Kharim, Y., Galindo Zaldívar, J., & Sahrane, R. (2023). Landslide susceptibility assessment in western external rif chain using machine learning methods.

Bharadiya, J. P. (2023). A review of Bayesian machine learning principles, methods, and applications. *International Journal of Innovative Science and Research Technology*, *8*(5), 2033-2038.

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *International Journal of Innovative Science and Research Technology*, *8*(5), 2028-2032.

Bhattacharya, S., & Mishra, S. (2018). Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. *Journal of Petroleum Science and Engineering*, *170*, 1005-1017.

Blackburn, G., Ashton, B. J., Thornton, A., Hunter, H., Woodiss-Field, S., & Ridley, A. R. (2024). Investigating the relationship between physical cognitive tasks and a social cognitive task in a wild bird. *Animal Cognition*, *27*(1), 52.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 045-076.

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

Bowden, C., Foster, T., & Parkes, B. (2023). Identifying links between monsoon variability and rice production in India through machine learning. *Scientific Reports*, *13*(1), 2446.

Cao, Q., Zhang, H., Lall, U., Holsclaw, T., & Shao, Q. (2024). The predictability of daily rainfall during rainy season over East Asia by a Bayesian nonhomogeneous hidden Markov model. *Journal of Flood Risk Management*, *17*(1), e12942.

Chan Chiu, P., Selamat, A., Krejcar, O., Kuok Kuok, K., Herrera-Viedma, E., & Fenza, G. (2021). Imputation of rainfall data using the sine cosine function fitting neural network.

Choi, J., Chapkin, R., & Ni, Y. (2020). Bayesian causal structural learning with zero-inflated Poisson Bayesian networks. *Advances in neural information processing systems*, *33*, 5887-5897.

Chuan, Z. L., Deni, S. M., Fam, S. F., & Ismail, N. (2020). The effectiveness of a probabilistic principal component analysis model and expectation maximisation algorithm in treating missing daily rainfall data. *Asia-Pacific Journal of Atmospheric Sciences*, *56*, 119-129.

Citakoglu, H., & Coşkun, Ö. (2022). Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of Sakarya Meteorological

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

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

Station in Turkey. *Environmental Science and Pollution Research*, *29*(50), 75487-75511.

Costa, J. A. D., Azevedo, C. F., Nascimento, M., Resende, M. D. V. D., & Nascimento, A. C. C. (2020). Genomic prediction with the additive-dominant model by dimensionality reduction methods. *Pesquisa Agropecuária Brasileira*, *55*, e01713.

Crisis24. 2017. https://crisis24.garda.com/alerts/2017/11/malaysia-floods-hit-selangor-state-november-7

Das, P., Sachindra, D. A., & Chanda, K. (2022). Machine learning-based rainfall forecasting with multiple non-linear feature selection algorithms. *Water Resources Management*, *36*(15), 6043-6071.

Di Nunno, F., Granata, F., Pham, Q. B., & de Marinis, G. (2022). Precipitation forecasting in Northern Bangladesh using a hybrid machine learning model. *Sustainability*, *14*(5), 2663.

Duarte, L. V., Formiga, K. T. M., & Costa, V. A. F. (2022). Comparison of methods for filling daily and monthly rainfall missing data: statistical models or imputation of satellite retrievals?. *Water*, *14*(19), 3144.

Egnyte. (2023). https://www.egnyte.com/guides/life-sciences/data-standardization

Ehteram, M., Ahmed, A. N., Sheikh Khozani, Z., & El-Shafie, A. (2023). Convolutional neural network-support vector machine model-gaussian process regression: a new machine model for predicting monthly and daily rainfall. *Water Resources Management*, *37*(9), 3631-3655.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, *8*, 1-37.

Fan, M., Xu, J., Chen, Y., & Li, W. (2022). Simulating the climate driven runoff in data-scarce mountains by machine learning and downscaling reanalysis data. *Stochastic Environmental Research and Risk Assessment*, *36*(11), 3819-3834.

Fan, M., Xu, J., Chen, Y., & Li, W. (2022). Simulating the climate driven runoff in data-scarce mountains by machine learning and downscaling reanalysis data. *Stochastic Environmental Research and Risk Assessment*, *36*(11), 3819-3834.

Fan, X., Jiang, L., & Gou, J. (2021). Statistical downscaling and projection of future temperatures across the Loess Plateau, China. *Weather and Climate Extremes*, *32*, 100328.

Farikha, E. F., Hadi, A. F., Anggraeni, D., & Riski, A. (2021, May). Projection pursuit regression in statistical downscaling model using artificial neural network for

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

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

rainfall prediction. In *Journal of Physics: Conference Series* (Vol. 1872, No. 1, p. 012021). IOP Publishing.

Farooq, R., Imteaz, M., & Ahmad, I. (2023, November). Prediction of seasonal rainfall using the hybrid machine learning algorithms. In *HWRS2023: Hydrology & Water Resources Symposium* (pp. 1-10). Sydney: Engineers Australia.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications*, *8*(1), 8.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications, 8*(1), 8.

Fitria, C., & Amalia, F. (2018). Society perception toward flood prevention in sentosa klang park selangor Malaysia. *Sumatra Journal of Disaster, Geography and Geography Education*, *2*(2), 39-46.

Geethalakshmi, V., Kokilavani, S., Ramanathan, S. P., Dheebakaran, G. A., Sathyamoorthy, N. K., & Maragatham, N. (2022). Monthly rainfall hind-cast using machine learning algorithms for Coimbatore, Tamil Nadu. *MAUSAM*, *73*(1), 19-26.

Ghazali, S. M., Shaadan, N., & Idrus, Z. (2020). Missing data exploration in air quality data set using R-package data visualisation tools. *Bulletin of Electrical Engineering and Informatics*, *9*(2), 755-763.

González-Abad, J., Baño-Medina, J., & Gutiérrez, J. M. (2023). Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches. *Journal of Advances in Modeling Earth Systems*, *15*(11), e2023MS003641.

Guo, L., Jiang, Z., Li, L., & Wang, H. (2022). Increase of future summer rainfall in the middle and lower reach of the Yangtze River basin projected with a nonhomogeneous hidden Markov model. *Geophysical Research Letters, 49*(7), e2021GL097325.

Hadi, F. A. A., Mohd Sidek, L., Ahmed Salih, G. H., Basri, H., Sammen, S. S., Mohd Dom, N., ... & Najah Ahmed, A. (2024). Machine learning techniques for flood forecasting. *Journal of Hydroinformatics*, *26*(4), 779-799.

Hadi, S. J., Abba, S. I., Sammen, S. S., Salih, S. Q., Al-Ansari, N., & Yaseen, Z. M. (2019). Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. *IEEE Access*, *7*, 141533-141548.

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

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

Hamzah, F. B., Hamzah, F. M., Razali, S. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civil Engineering Journal*, *7*(9), 1608-1619.

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20-30.

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20-30.

He, S., Sang, X., Yin, J., Zheng, Y., & Chen, H. (2023). Short-term runoff prediction optimization method based on BGRU-BP and BLSTM-BP neural networks. *Water Resources Management*, *37*(2), 747-768.

Hırca, T., & Eryılmaz Türkkan, G. (2024). Assessment of Different Methods for Estimation of Missing Rainfall Data. *Water Resources Management*, 1-28.

Holsclaw, T., Greene, A. M., Robertson, A. W., & Smyth, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling.

Hsu, S. C., & Sharma, A. K. (2023). Forecast Rainfall Density by Utilizing Machine Learning Models.

Huang, G. (2021, February). Missing data filling method based on linear interpolation and lightgbm. In *Journal of Physics: Conference Series* (Vol. 1754, No. 1, p. 012187). IOP Publishing.

Husin, N. A., Ngah, F., Yulia, A., Azim, Z. A. Z., & Ghazali, A. S. M. (2020). Critical Dimensions of Well-being in Selangor, Malaysia. *Indian Journal of Finance and Banking*, *4*(3), 39-52.

Ian, V. K., Tse, R., Tang, S. K., & Pau, G. (2023). Bridging the Gap: Enhancing storm surge prediction and decision support with bidirectional attention-based LSTM. *Atmosphere*, *14*(7), 1082.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913-933.

Jahan, F., Sinha, N. C., Rahman, M. M., Rahman, M. M., Mondal, M. S. H., & Islam, M. A. (2019). Comparison of missing value estimation techniques in rainfall data of Bangladesh. *Theoretical and Applied Climatology*, *136*, 1115-1131.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

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

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

Jiang, Q., Cioffi, F., Conticello, F. R., Giannini, M., Telesca, V., & Wang, J. (2023). A stacked ensemble learning and non-homogeneous hidden Markov model for daily precipitation downscaling and projection. *Hydrological Processes*, *37*(9), e14992.

Jun, S. (2023). Zero-Inflated Text Data Analysis using Generative Adversarial Networks and Statistical Modeling. *Computers*, *12*(12), 258.

Kanani, S., Patel, S., Gupta, R. K., Jain, A., & Lin, J. C. W. (2023). An AI-enabled ensemble method for rainfall forecasting using long-short term memory.

Kang, Y., Wang, S., Wang, D., & Zhu, F. (2023). Analysis of zero-and-one inflated bounded count time series with applications to climate and crime data. *Test*, *32*(1), 34-73.

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. *Journal of big Data*, *7*(1), 37.

Khooriphan, W., Niwitpong, S. A., & Niwitpong, S. (2022). Bayesian estimation of rainfall dispersion in Thailand using gamma distribution with excess zeros. *PeerJ*, *10*, e14023.

Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science*, *11*(12), 190.

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

Kumar, B., Chattopadhyay, R., Singh, M., Chaudhari, N., Kodari, K., & Barve, A. (2021). Deep learning–based downscaling of summer monsoon rainfall data over Indian region. *Theoretical and Applied Climatology*, *143*, 1145-1156.

Kuznietsov, P. M., & Biedunkova, O. O. (2024). Multivariate regression studies for the investigation of the COD, BOD, and TOC concentrations in the water of the Styr River within the zone of influence of the power plant discharge. *Water Practice & Technology*, *19*(4), 1412-1426.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, *70*(4), 407-411.

Lee, M. H., & Chen, Y. J. (2021). Precipitation modeling for extreme weather based on sparse hybrid machine learning and Markov chain random field in a multi-scale subspace. *Water*, *13*(9), 1241.

Li, W., Liu, C., Hu, C., Niu, C., Li, R., Li, M., ... & Tian, L. (2024). Application of a hybrid algorithm of LSTM and Transformer based on random search optimization for improving rainfall-runoff simulation. *Scientific Reports*, *14*(1), 11184.

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*, 1487-1509.

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

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

Loecher, A., Bruyns-Haylett, M., Ballester, P. J., Borros, S., & Oliva, N. (2023). A machine learning approach to predict cellular uptake of pBAE polyplexes. *Biomaterials Science*, *11*(17), 5797-5808.

Long, Y., Tang, R., Wang, H., & Jiang, C. (2019). Monthly precipitation modeling using Bayesian non-homogeneous hidden Markov chain. *Hydrology Research*, *50*(2), 562-576.

Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z., & Yan, S. (2019). Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE transactions on pattern analysis and machine intelligence*, *42*(4), 925-938.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Mao, G., Wang, M., Liu, J., Wang, Z., Wang, K., Meng, Y., ... & Li, Y. (2021). Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. *Physics and Chemistry of the Earth, Parts a/b/c*, *123*, 103026.

Maqsood, J., Afzaal, H., Farooque, A. A., Abbas, F., Wang, X., & Esau, T. (2023). Statistical downscaling and projection of climatic extremes using machine learning algorithms. *Theoretical and Applied Climatology*, *153*(3), 1033-1047.

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

Masngut, M. A., Ismail, S., Mustapha, A., & Yasin, S. M. (2020). Comparison of daily rainfall forecasting using multilayer perceptron neural network model. *IAES International Journal of Artificial Intelligence*, *9*(3), 456.

Mazher, A. (2020). Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques. *Water*, *12*(2), 590.

Meshram, S. G., Singh, V. P., Kisi, O., Karimi, V., & Meshram, C. (2020). Application of artificial neural networks, support vector machine and multiple model-ANN to sediment yield prediction. *Water Resources Management*, *34*(15), 4561-4575.

Mihalik, A., Chapman, J., Adams, R. A., Winter, N. R., Ferreira, F. S., Shawe-Taylor, J., ... & Alzheimer's Disease Neuroimaging Initiative. (2022). Canonical correlation analysis and partial least squares for identifying brain–behavior associations: A tutorial and a comparative study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(11), 1055-1067.

Mijwel, M. M. (2021). Artificial neural networks advantages and disadvantages. *Mesopotamian Journal of Big Data*, *2021*, 29-31.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308-2316.

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

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

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308-2316.

Mishra, S., & Datta-Gupta, A. (2017). *Applied statistical modeling and data analytics: A practical guide for the petroleum geosciences*. Elsevier.

Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. *International Journal of Scientific and Technology Research*, *9*(1), 3236-3240.

Mohd Zahari, H., Zainol, N. A. M., & Ismail, A. (2022). Media information, flood images, and perceptions in times of flood. *Sustainability*, *14*(17), 10623.

Muhamad, N., & Shaidin, H. A. (2022). Assessment of Flood Events in Selangor, Malaysia. *Journal of Advanced Geospatial Science & Technology*, *2*(2), 42-55.

Nakhaei, M., Nakhaei, P., Gheibi, M., Chahkandi, B., Wacławek, S., Behzadian, K., ... & Campos, L. C. (2023). Enhancing community resilience in arid regions: A smart framework for flash flood risk assessment. *Ecological Indicators*, *153*, 110457.

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M., & Meyer, N. (2019). Determining the number of components in PLS regression on incomplete data set. *Statistical Applications in Genetics and Molecular Biology*, *18*(6), 20180059.

New Straits Times. 2018. https://www.nst.com.my/news/nation/2018/11/431992/monsoon-flooding-hits-selangor-222-evacuated#google_vignette

Nikfalazar, S., Yeh, C. H., Bedingfield, S., & Khorshidi, H. A. (2020). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, *62*, 2419-2437.

Nor, S. M. C. M., Shaharudin, S. M., Ismail, S., Zainuddin, N. H., & Tan, M. L. (2020). A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. *Bulletin of Electrical Engineering and Informatics*, *9*(2), 635-643.

Nourani, V., Baghanam, A. H., & Gokcekus, H. (2018). Data-driven ensemble model to statistically downscale rainfall using nonlinear predictor screening approach. *Journal of Hydrology*, *565*, 538-551.

Nourani, V., Baghanam, A. H., & Gokcekus, H. (2018). Data-driven ensemble model to statistically downscale rainfall using nonlinear predictor screening approach. *Journal of Hydrology*, *565*, 538-551.

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

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

Nurwatik, N., Ummah, M. H., Cahyono, A. B., Darminto, M. R., & Hong, J. H. (2022). A comparison study of landslide susceptibility spatial modeling using machine learning. *ISPRS International Journal of Geo-Information*, *11*(12), 602.

Nyongesa, A. M., Zeng, G., & Ongoma, V. (2020). Non-homogeneous hidden Markov model for downscaling of short rains occurrence in Kenya. *Theoretical and Applied Climatology*, *139*(3), 1333-1347.

Ocampo-Marulanda, C., Fernández-Álvarez, C., Cerón, W. L., Canchala, T., Carvajal-Escobar, Y., & Alfonso-Morales, W. (2022). A spatiotemporal assessment of the high-resolution CHIRPS rainfall dataset in southwestern Colombia using combined principal component analysis. *Ain Shams Engineering Journal*, *13*(5), 101739.

Ochoa-Muñoz, A. F., & Contreras-Reyes, J. E. (2023). Multiple factor analysis based on Nipals algorithm to solve missing data problems. *Algorithms*, *16*(10), 457.

Ochoa-Muñoz, A. F., González-Rojas, V. M., & Pardo, C. E. (2019). Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm. *Dyna*, *86*(211), 249-257.

Oh, M., Lee, J., Kim, J. Y., & Kim, H. G. (2022). Machine learning-based statistical downscaling of wind resource maps using multi-resolution topographical data. *Wind Energy*, *25*(6), 1121-1141.

05-4506832 Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Parviz, L., Rasouli, K., & Torabi Haghighi, A. (2023). Improving hybrid models for precipitation forecasting by combining nonlinear machine learning methods. *Water Resources Management*, *37*(10), 3833-3855.

Patel, A., Keriwala, N., Soni, N., Goel, U., Bhoj, R., Adhyaru, Y., & Yadav, S. M. (2023). Rainfall Prediction using Machine Learning Techniques for Sabarmati River Basin, Gujarat, India. *Journal of Engineering Science & Technology Review*, *16*(1).

Peng, Y., & Nagata, M. H. (2020). An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos, Solitons & Fractals*, *139*, 110055.

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: what's one to do?. *Journal of Statistical Computation and Simulation*, *83*(9), 1671-1683.

Polson, N. G., Scott, J. G., & Windle, J. (2014). The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(4), 713-733.

Poornima, S., Pushpalatha, M., Jana, R. B., & Patti, L. A. (2023). Rainfall forecast and drought analysis for recent and forthcoming years in India. *Water*, *15*(3), 592.

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

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

Portnova-Fahreeva, A. A., Rizzoglio, F., Nisky, I., Casadio, M., Mussa-Ivaldi, F. A., & Rombokas, E. (2020). Linear and non-linear dimensionality-reduction techniques on full hand kinematics. *Frontiers in bioengineering and biotechnology*, *8*, 429.

Pour, S. H., Shahid, S., Chung, E. S., & Wang, X. J. (2018). Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh. *Atmospheric research*, *213*, 149-162.

Praveena, R., Babu, T. G., Birunda, M., Sudha, G., Sukumar, P., & Gnanasoundharam, J. (2023, March). Prediction of Rainfall Analysis Using Logistic Regression and Support Vector Machine. In *Journal of Physics: Conference Series* (Vol. 2466, No. 1, p. 012032). IOP Publishing.

Prottasha, N. J., Tahabilder, A., Kowsher, M., Mia, M. S., & Kobra, K. T. (2023) Short-Term Rainfall Prediction Using Supervised Machine Learning. *Advances in Technology Innovation, 111-120*, 8(2).

Psychogyios, K., Ilias, L., Ntanos, C., & Askounis, D. (2023). Missing value imputation methods for electronic health records. *IEEE Access*, *11*, 21562-21574.

Putri, C. D., Farikha, E. F., Hadi, A. F., Dewi, Y. S., Tirta, I. M., Ubaidillah, F., & Anggraeni, D. (2022, February). Projection Pursuit Regression on Statistical Downscaling Using Artificial Neural Network and Support Vector Regression for Rainfall Forecasting in Jember. In *International Conference on Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021)* (pp. 204-210). Atlantis Press.

Quadros Gramosa, A. H., Ferraz do Nascimento, F., & Castro Morales, F. E. (2020). A Bayesian approach to zero-inflated data in extremes. *Communications in Statistics-Theory and Methods*, *49*(17), 4150-4161.

Quadros Gramosa, A. H., Ferraz do Nascimento, F., & Castro Morales, F. E. (2020). A Bayesian approach to zero-inflated data in extremes. *Communications in Statistics-Theory and Methods*, *49*(17), 4150-4161.

Rafik, A., Brahim, Y. A., Amazirh, A., Ouarani, M., Bargam, B., Ouatiki, H., ... & Chehbouni, A. (2023). Groundwater level forecasting in a data-scarce region through remote sensing data downscaling, hydrological modeling, and machine learning: A case study from Morocco. *Journal of Hydrology: Regional Studies*, *50*, 101569.

Rafique, D., & Velasco, L. (2018). Machine learning for network automation: overview, architecture, and applications [Invited Tutorial]. *Journal of Optical Communications and Networking*, *10*(10), D126-D143.

Rahman, A. U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., ... & Mosavi, A. (2022). Rainfall prediction system using machine learning fusion for smart cities. *Sensors*, *22*(9), 3504.

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

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

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, *8*, 54776-54788.

Robertson, C., Sawford, K., Gunawardana, W. S., Nelson, T. A., Nathoo, F., & Stephen, C. (2011). A hidden Markov model for analysis of frontline veterinary data for emerging zoonotic disease surveillance. *PLoS One*, *6*(9), e24833.

Rosmadi, H. S., Ahmed, M. F., Mokhtar, M. B., & Lim, C. K. (2023). Reviewing challenges of flood risk management in Malaysia. *Water*, *15*(13), 2390.

Rumaling, M. I., Chee, F. P., Dayou, J., Chang, J. H. W., Kong, S. S. K., & Sentian, J. (2020). Missing value imputation for PM10 concentration in Sabah using nearest neighbour method (NNM) and expectation-maximization (EM) algorithm. *Asian Journal of Atmospheric Environment*, *14*(1), 62-72.

Saeipourdizaj, P., Sarbakhsh, P., & Gholampour, A. (2021). Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbor methods. *Environmental Health Engineering and Management Journal*, *8*(3), 215-226.

Saha, S. (2023). An empirical comparison of linear and non-linear classification using support vector machines. *Int J of Comput Sci Eng*, *11*, 120-126.

Sahoo, A., & Ghose, D. K. (2022). Imputation of missing precipitation data using KNN, SOM, RF, and FNN. *Soft Computing*, *26*(12), 5919-5936.

Sahoo, A., Samantaray, S., & Ghose, D. K. (2021). Prediction of flood in Barak River using hybrid machine learning approaches: a case study. *Journal of the Geological Society of India*, *97*(2), 186-198.

Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer networks*, *148*, 164-175.

Sammen, S. S., Ehteram, M., Sheikh Khozani, Z., & Sidek, L. M. (2023). Binary coati optimization algorithm-multi-kernel least square support vector machine-extreme learning machine model (BCOA-MKLSSVM-ELM): a new hybrid machine learning model for predicting reservoir water level. *Water*, *15*(8), 1593.

Sani, N. S., Abd Rahman, A. H., Adam, A., Shlash, I., & Aliff, M. (2020). Ensemble learning for rainfall prediction. *International Journal of Advanced Computer Science and Applications*, *11*(11).

Satrio, C. B. A., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, *179*, 524-532.

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

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

Shaharudin, S. M., Andayani, S., Kismiantini, N. B., Kurniawan, A., Basri, M. A. A., & Zainuddin, N. H. (2020). Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia. *International Journal*, *9*(1.4).

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308-6325.

Shirmard, H., Farahbakhsh, E., Beiranvand Pour, A., Muslim, A. M., Müller, R. D., & Chandra, R. (2020). Integration of selective dimensionality reduction techniques for mineral exploration using ASTER satellite data. *Remote Sensing*, *12*(8), 1261.

Someetheram, V., Marsani, M. F., Mohd Kasihmuddin, M. S., Mohd Jamaludin, S. Z., & Mansor, M. A. (2024). Double decomposition with enhanced least-squares support vector machine to predict water level. *Journal of Water and Climate Change*, *15*(6), 2582-2594.

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, *10*(1), 20410.

Stoner, O., & Economou, T. (2020). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, *76*(3), 789-798.

Sulaiman, N. A. F., Shaharudin, S. M., Ismail, S., Zainuddin, N. H., Tan, M. L., & Abd Jalil, Y. (2022). Predictive modelling of statistical downscaling based on hybrid machine learning model for daily rainfall in east-coast peninsular malaysia. *Symmetry*, *14*(5), 927.

Sumayli, A. (2023). Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. *Arabian Journal of Chemistry*, *16*(7), 104833.

Sun, L., & Lan, Y. (2021). Statistical downscaling of daily temperature and precipitation over China using deep learning neural models: Localization and comparison with other methods. *International Journal of Climatology*, *41*(2), 1128-1147.

Susianto, Y., Notodiputro, K. A., Kurnia, A., & Wijayanto, H. (2017, March). A comparative study of imputation methods for estimation of missing values of per capita expenditure in central java. In *IOP Conference Series: Earth and Environmental Science* (Vol. 58, No. 1, p. 012017). IOP Publishing.

Tahvili, Sahar, and Leo Hatvani. "Artificial intelligence methods for optimization of the software testing process." (2022).

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

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

Tan, W. L., Yusof, F., & Yusop, Z. (2017). Subseasonal to multidecadal variability of northeast monsoon daily rainfall over Peninsular Malaysia using a hidden Markov model. *Theoretical and Applied Climatology, 129,* 577-586.

Traore, S., Wang, Y. M., & Kerh, T. (2010). Artificial neural network for modeling reference evapotranspiration complex process in Sudano-Sahelian zone. *Agricultural water management*, *97*(5), 707-714.

Truong, V. H., & Pham, H. A. (2021). Support vector machine for regression of ultimate strength of trusses: A comparative study. *Engineering Journal*, *25*(7), 157-166.

Tuysuzoglu, G., Birant, K. U., & Birant, D. (2023). Rainfall Prediction Using an Ensemble Machine Learning Model Based on K-Stars. *Sustainability, 15*(7), 5889.

Ulkir, O., Bayraklılar, M. S., & Kuncan, M. (2024). Raster Angle Prediction of Additive Manufacturing Process Using Machine Learning Algorithm. *Applied Sciences*, *14*(5), 2046.

Usman, C. D., Widodo, A. P., Adi, K., & Gernowo, R. (2023). Rainfall prediction model in Semarang City using machine learning. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(2), 1224-1231.

Van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of personality assessment*, *102*(3), 297-308.

Van, N. H., Van Thanh, P., Tran, D. N., & Tran, D. T. (2023). A new model of air quality prediction using lightweight machine learning. *International Journal of Environmental Science and Technology*, *20*(3), 2983-2994.

Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology, 137,* 557-570.

Velasco, L. C., Aca-ac, J. M., Cajes, J. J., Lactuan, N. J., & Chit, S. C. (2022). Rainfall forecasting using support vector regression machines. *International Journal of Advanced Computer Science and Applications*, *13*(3).

Wang, H., Asefa, T., & Sarkar, A. (2021). A novel non-homogeneous hidden Markov model for simulating and predicting monthly rainfall. *Theoretical and Applied Climatology, 143*(1), 627-638.

Workie, M. S., & Azene, A. G. (2021). Bayesian zero-inflated regression model with application to under-five child mortality. *Journal of big data*, *8*(1), 4.

Wu, S., Wang, Z., Du, Z., Huang, B., Zhang, F., & Liu, R. (2021). Geographically and temporally neural network weighted regression for modeling spatiotemporal non-stationary relationships. *International Journal of Geographical Information Science*, *35*(3), 582-608.

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

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

Xu, S., Zhao, Q., Yin, K., He, G., Zhang, Z., Wang, G., ... & Zhang, N. (2021). Spatial downscaling of land surface temperature based on a multi-factor geographically weighted machine learning model. *Remote Sensing*, *13*(6), 1186.

Yamamoto, H., Fujimori, T., Sato, H., Ishikawa, G., Kami, K., & Ohashi, Y. (2014). Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC bioinformatics*, *15*, 1-9.

Yang, R., Yu, L., Zhao, Y., Yu, H., Xu, G., Wu, Y., & Liu, Z. (2020). Big data analytics for financial Market volatility forecast based on support vector machine. *International Journal of Information Management*, *50*, 452-462.

Yang, S., Du, P., Feng, X., He, D., Chen, Y., Zhong, L. L., ... & Luo, J. (2023). Propensity score analysis with missing data using a multi-task neural network. *BMC Medical Research Methodology*, *23*(1), 41.

Yao, Q., Cheng, L., Chen, W., & Mao, T. (2024). Some Generalized Entropy Ergodic Theorems for Nonhomogeneous Hidden Markov Models. *Mathematics*, *12*(4), 605.

Zaidi, S. M., Gisen, J. I. A., Eltahan, M., Yu, Q., Misbari, S., & Ngien, S. K. (2022). Assessment of weather research and forecasting (WRF) physical schemes parameterization to predict moderate to extreme rainfall in poorly gauged basin. *Sustainability*, *14*(19), 12624.

Zamani, H., & Bazrafshan, O. (2020). Modeling monthly rainfall data using zero-adjusted models in the semi-arid, arid and extra-arid regions. *Meteorology and Atmospheric Physics*, *132*(2), 239-253.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(1), 56-70.

Zhang, X., Yin, Q., Liu, F., Li, H., & Qi, Y. (2023). Comparative study of rainfall prediction based on different decomposition methods of VMD. *Scientific Reports*, *13*(1), 20127.

Zhou, W., Yan, Z., & Zhang, L. (2024). A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. *Scientific Reports*, *14*(1), 5905.

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

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

APPENDICES

A R codes – HMM

```
library("HMM")

#Function to calculate transition matrix
trans.mat=function(K,qq)
{
 t.qq=matrix(0,K,K)
 for(i in 1:K){
  for(j in 1:K)
  {
   t.qq[i,j]=round(mean(qq[i,j,]),2)
  }
 }
 return(t.qq)
}


#Import data
rain.data=read.table("wet_5.txt",header=TRUE)
rain.data=data.matrix(rain.data)

#Model
system.time(my.hmm1<-HMM(y=rain.data, K=3, iters=2000, burnin=200,
        emdist="gamma", nmix=3, delta=TRUE))
OBIC(my.hmm1)
plot(ts(my.hmm1$loglik))
plot(density(my.hmm1$loglik))   #density plot
zzTest=Oz(my.hmm1)  #compare with the truth zgamma
table(zzTest)
qq=OQQ(my.hmm1)
qqmat=trans.mat(K=3,qq) #K=number of states
pp=OWcoef(my.hmm1,FALSE)
tt=Oemparams(my.hmm1,FALSE)
write.csv(zzTest, file = "zz.xlsx")
write.csv(qqmat, file = "qq.xlsx")
write.csv(tt, file = "tt.xlsx")
write.csv(pp, file = "pp.xlsx")

#Prediction
filelocation="C:\\Users\\User\\"
system.time(my.hmm1<-HMM(y=rain.data, W=NULL, K=3, iters=2000, burnin=200,
        emdist="gamma", nmix=1, delta=TRUE,
        outdir=filelocation, pT=1664, yrep=1,Wp=NULL, ypred=1))
```

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

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

B R code – HMM-RF

```
#Import necessary libraries
library(randomForest)
library(caret)
library("readxl")

# Read data from Excel file
data <- read_excel("MEDIAN.xlsx")

# Number of rows in the dataset
n_rows <- nrow(data)  #calling all rows

# Number of rows for training
n_train <- floor(0.7 * n_rows)  #70% from the rows as training sequentially

# Split the data into training and test sets
train_data <- data[1:n_train, ]  # 70% data
test_data <- data[(n_train + 1):n_rows, ]  # 30% data

#random forest model
model <- randomForest(LOCAL ~ ., data = train_data)
print(model)

# Predict the target variable for the train data
predtrain = predict(model, train_data)

# Predict the target variable for the test data
predtest = predict(model, test_data)

print(predtrain)
plot(predtrain)

write.csv(predtrain, file = "rfpredtrain.xlsx")
write.csv(predtest, file = "rfpredtest.xlsx")
```

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

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

C R codes – tuning parameter

```
library(kernlab)
library(readxl)

# Read data from Excel file
data <- read_excel("S1.xlsx")

# Number of rows in the dataset
n_rows <- nrow(data)

# Number of rows for training
n_train <- floor(0.7 * n_rows)

# Split the data into training and test sets
train_data <- data[1:n_train, ]
test_data <- data[(n_train + 1):n_rows, ]

library(kernlab)

# Define parameter grid
param_grid <- expand.grid(sigma = seq(0.1, 1, by = 0.1),
                          C = c(0.1, 1, 10))

# Perform cross-validation
cv_results <- lapply(1:nrow(param_grid), function(i) {
  rvm_model <- ksvm(LOCAL ~ .,
                    data = train_data,
                    kernel = "rbfdot",
                    kpar = list(sigma = param_grid$sigma[i]),
                    C = param_grid$C[i])
  cv_result <- crossval(rvm_model, folds = 10)
  mean(cv_result@error)
})
```

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

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

D R codes – HMM-SVM

```
library(e1071)
library(caTools)
library("readxl")

# Read data from Excel file
data <- read_excel("MEDIAN.xlsx")

# Number of rows in the dataset
n_rows <- nrow(data)  #calling all rows

# Number of rows for training
n_train <- floor(0.7 * n_rows)  #70% from the rows as training data sequentially

# Split the data into training and test sets
train_data <- data[1:n_train, ]  # 70% data
test_data <- data[(n_train + 1):n_rows, ]  # 30% data

#Regression with SVM
#LOCAL represents the response variable (Y data)
#"." denotes all other variables in the dataset to be used as predictor variables (X data)
modelsvm = svm(formula = LOCAL~.,data= train_data, type= 'eps-regression',
               kernel= "radial", cost= 5, gamma=0.5, epsilon= 0.1)
```

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

```
modelsvm


# Get prediction values for the training data
train_predictions <- predict(modelsvm, train_data)
plot(train_predictions)

#Predict using SVM regression
test_prediction = predict(modelsvm, test_data)


write.csv(train_predictions, file = "trainrad.xlsx")
write.csv(test_prediction, file = "testrad.xlsx")
```

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

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

E R codes – HMM-DT

```
# Load the required package
library(stats)
library("readxl")

# Read data from Excel file
data <- read_excel("MEDIAN.xlsx")

# Number of rows in the dataset
n_rows <- nrow(data)

# Number of rows for training
n_train <- floor(0.7 * n_rows)

# Split the data into training and test sets
train_data <- data[1:n_train, ]
test_data <- data[(n_train + 1):n_rows, ]

# Train the decision tree model
tree_model <- lm(LOCAL ~ ., data = train_data)
print(tree_model)

# Get prediction values for the training data
train_predictions <- predict(tree_model, newdata = train_data)
plot(train_predictions)

# Visualize the decision tree (optional)
plot(tree_model)
text(tree_model)

# Make predictions on the test set
predictions <- predict(tree_model, newdata = test_data)

write.csv(train_predictions, file = "traindt.xlsx")
write.csv(predictions, file = "testdt.xlsx")
```

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

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

F R codes – HMM-KNN

```
library(caret)
library(readxl)

# Read data from Excel file
data <- read_excel("MEDIAN.xlsx")

# Number of rows in the dataset
n_rows <- nrow(data)

# Number of rows for training
n_train <- floor(0.7 * n_rows)

# Split the data into training and test sets
train_data <- data[1:n_train, ]
test_data <- data[(n_train + 1):n_rows, ]

# Train the k-NN model
knn_model <- train(LOCAL ~ ., data = train_data, method = "knn")
plot(knn_model)
print(knn_model)
# Get prediction values for the training data
train_predictions <- predict(knn_model, newdata = train_data)
plot(train_predictions)

# Make predictions on the test set
test_predictions <- predict(knn_model, newdata = test_data)
print(predictions)

write.csv(train_predictions, file = "knntrainmed.xlsx")
write.csv(test_predictions, file = "knntestmed.xlsx")
```

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

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

G R codes – HMM-ANN

```
library(forecast)
library(caTools)
library(neuralnet)
library(caret)
library("readxl")

# Read data from Excel file
data <- read_excel("MEDIAN.xlsx")

# number of rows in the dataset
n_rows <- nrow(data) #calling all rows

# number of rows for training
n_train <- floor(0.7 * n_rows) #70% from the rows as training data sequentially

# split the data into training and test sets
train <- data[1:n_train, ] # 70% data
test <- data[(n_train + 1):n_rows, ] # 30% data

#neural network:
n <- neuralnet(LOCAL ~ ., data=train, hidden=c(5,10), linear.output=FALSE); n
plot(n)
print(n)

# Get prediction values for the training data
train_predictions <- predict(n, newdata = train)
plot(train_predictions)
# Make predictions on the test set
test_predictions <- predict(n, newdata = test)

write.csv(train_predictions, file = "train_ann.xlsx")
write.csv(test_predictions, file = "test_ann.xlsx")
```

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

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

H R codes – PCA

```
#read data from excel
library("readxl")

# Example dataset (replace this with your dataset)
data <- read_excel("ATMOSPHERE.xlsx")

#install package for run PCA
install.packages(c("FactoMineR", "factoextra"))

library(FactoMineR)
PCA(my_data, scale.unit = FALSE, ncp = 4, graph = TRUE)
my_data.active <- my_data[1:4018, 1:6]
res.pca <- PCA(my_data.active, graph = FALSE)
print(res.pca)

#for the PCA result
library("factoextra")
# Eigenvalues
eig.val <- get_eigenvalue(res.pca)
eig.val
#screenplot
fviz_screeplot(res.pca, addlabels = TRUE, ylim = c(0, 100))
#cumulative
res.pca <- prcomp(my_data.active, scale = TRUE)
summary(res.pca)
# loadings
head(res.pca$rotation)
# PCs score
res.pca$x[1:4018,1:5]

write.csv(res.pca$x, file = "PC_SCORES.xlsx")
```

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

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

I Percentage of Missing Local Data

Station	Percentage of Missing Data (%)
22-334	0.722
23-333	4.579
23-334	0.000
24-332	13.216
24-333	0.000
24-334	0.548
24-335	0.274
25-332	2.688
25-333	3.111
25-335	4.803
25-336	0.025
26-332	0.050
26-333	0.025
27-331	0.000
27-332	0.000
27-333	3.609
27-334	3.136
28-330	0.996
28-332	0.025
28-333	3.683
28-334	3.534
29-331	0.025
29-332	1.269
29-335	3.683
30-329	0.324
30-330	0.050
30-331	3.509
30-332	4.430
30-333	3.708
30-334	3.758
30-335	3.683
31-328	0.672
31-329	0.025

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