Project Report

On

TIME SERIES ANALYSIS OF RAINFALL IN KERALA

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

By

KRISHNENDU P R

(Reg No. SM23AS007)

(2023-2025)

Under the Supervision of

VISMAYA VINCENT

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE(AUTONOMOUS)

ERNAKULAM, KOCHI–682011

APRIL 2025

ST. TERESA'S COLLEGE(AUTONOMOUS), ERNAKULAM

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **TIME SERIES ANALYSIS OF RAINFALL IN KERALA** is a bonafide record of the work done by **KRISHNENDU P R** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:

Place: Ernakulam

Vismaya Vincent
Assistant Professor, Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

Nisha Oommen
Assistant Professor & HOD
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

External Examiners

1. Sangeetha Chandran
30.04.2025

2. Anju. N. B
30.04.25

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **VISMAYA VINCENT**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date: 30/04/25

KRISHNENDU P R

SM23AS007

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage.

I am very grateful to my project guide Vismaya Vincent for the immense help during the period of work.

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

I am also very thankful to HOD for her valuable suggestions, critical examination of work during the progress.

ERNAKULAM

DATE: 30/04/25

KRISHNENDU P R
SM23AS007

ABSTRACT

In this study, monthly rainfall data of Kerala from July 2010 to December 2024 are analyzed and used for forecasting future rainfall value. The raw data were obtained from the official website of Department of Economics and Statistics, Thiruvananthapuram and Kerala Water Resource Irrigation System. Three time series model were deployed to forecast future monthly rainfall data of Kerala. SARIMA, Holt-Winter's Exponential Smoothening and prophet model were used to forecast the monthly rainfall of Kerala from January 2025 to December 2026. Comparison of these models were done using matrices like RMSE and MAE. Prophet model was selected as the best model with lowest RMSE value and MAE value.

ST.TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	KRISHNENDU P R
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. Vismaya Vincent
Department	PG Dept. of Mathematics & Statistics
Acceptable Maximum Limit	20
Submitted By	library@teresas.ac.in
Paper Title	TIME SERIES ANALYSIS OF RAINFALL IN KERALA
Similarity	6% AI - 7%
Paper ID	3408712
Total Pages	54
Submission Date	2025-03-17 14:54:00

Signature of Student

Signature of Guide

Checked By
College Librarian

ST. TERESA'S COLLEGE LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Rainfall is vital for all life on Earth and plays a significant role in ecosystem processes, crop production, and hydroelectric energy. Being the major source of water, rainfall has a critical impact on water resource management, agricultural planning, flood control, disaster preparedness, tourism, transportation, and environmental monitoring, among others. Knowledge of rainfall patterns and trends can improve decision-making in such areas. Rainfall distribution and intensity vary between regions based on latitude, elevation, and atmospheric conditions

Kerala's unique geography and topography, with the Western Ghats to the east and the Arabian Sea to the west, create significant rainfall variability across the state. The orographic effect of the Western Ghats results in heavy rainfall on the windward side during the Southwest Monsoon, while coastal areas receive more rainfall due to proximity to the sea. High-altitude regions like Wayanad, Idukki, and Munnar record more rainfall compared to lowlands, and the state's diverse geography leads to microclimates with localized rainfall patterns.

Kerala is the entry point of the Indian subcontinent summer monsoon. The major rainy seasons of the state are the southwest monsoon (June–September) and the northeast monsoon (October–November) (Raj & Azeez, 2012). The analysis of rainfall data from a century trend shows a statistically significant (99%) decreasing trend in many areas of Kerala, especially in January, July, and November (Nair et al., 2014). Climate change is now an undeniable phenomenon, leading to catastrophic weather occurrences globally. Numerous factors determine a region's climate, such as latitude, height, pressure and wind patterns, distance from the sea, ocean currents, and terrain. Differing rainfall patterns and unexpected heatwaves rank among the most severe impacts of climate change (Varghese & Vanitha, 2020). A historical analysis of rainfall in Kerala, carried out by the Indian Meteorological Department (IMD), shows changes in Kerala's rainfall pattern. The state records highest rainfall during July followed by June.

There has been a visible decrease in southwest monsoon rainfall, while post-monsoon rainfall has increased. The reduction in rainfall is most pronounced during June and July, whereas August and September have been relatively consistent during the monsoon season (Krishnakumar et al., 2009). Rainfall trends also differ from one region to another, with a rising trend noted in northern and eastern stations, whereas southern and western stations show a declining trend (Jagadeesh & Anupama, 2014). The occurrence of both Moderate Rain (MR)

and Heavy Rain (HR) events has decreased over Kerala, though some grids in the eastern region of the state have experienced a considerable rise in HR events during 1971-2019 (Surendran et al., 2020).

A number of research studies have investigated rainfall analysis and forecasting through different approaches. Construction of a precise forecasting system continues to pose problems for researchers, one of the problems being how to process past data and forecast future trends. Time series modeling presents a possible solution. For example, rainfall forecasting in Idukki district has been made with models like Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANN), and Exponential Smoothing State Space (ETS) (Kamath & Kamat, 2018). Further, artificial intelligence methods have also proven to predict Kerala's monsoon seasons with little prediction error (Dash et al., 2018). Comparative analyses using models such as SARIMA, Facebook Prophet, and Long Short-Term Memory (LSTM) networks have also been implemented (Sulasikin et al., 2021).

Time Series Analysis of Rainfall in Kerala, in its desire to determine the rainfall trends and patterns in Kerala, as well as predict rainfall in the future, sought to fill this need. This study helps to know about Kerala's shifting patterns of rainfall. Raw data for the study were obtained from the official website of the Department of Economics and Statistics, Thiruvananthapuram, and the Kerala Water Resource Irrigation System. Monthly rainfall data for the period from July 2010 to December 2024 were employed for predicting patterns of rainfall during January 2025 to December 2026.

1.1 OBJECTIVES

1. To perform EDA(Exploratory Data Analysis) to find pattern and trend of rainfall.
2. To model and forecast rainfall in Kerala using Seasonal ARIMA (Auto Regressive Integrated Moving Average).
3. To model and forecast rainfall in Kerala using Holt Winter’s Exponential smoothing technique
4. To model and forecast rainfall in Kerala using Prophet model
5. To compare the forecast of Seasonal ARIMA, Holt Winter’s Exponential smoothing technique and Prophet model.

CHAPTER 2

REVIEW OF LITERATURE

This chapter shows the results from the related research that analysed the various Rainfall datasets and made prediction using various statistical methods, data mining techniques, machine learning algorithms and so on.

Krishnakumar et al. (2009) studied Kerala's 20th-century trends in rainfall by analyzing 1871-2005 data based on Mann-Kendall statistics and linear trend analysis. A significant reduction in southwest monsoon rainfall, particularly in June and July, was observed, affecting hydropower generation and water supply during the summer season. In contrast, post-monsoon rainfall increased, and, if continued, would be beneficial for plantation crops.

Raj and Azeez (2012) investigated rainfall trends in the Bharathapuzha River basin, Kerala, India. From 34 years of rainfall data at 28 rain gauge stations, they analyzed patterns through Mann-Kendall statistics and wavelet analysis. The study revealed a notable decrease in annual, southwest monsoon, and pre-monsoon rainfall in recent years, which was likely due to global climate fluctuations and local environmental changes.

Jagadeesh and Anupama (2014) made a statistical and trend analysis of rainfall at the Bharathapuzha River basin. The 33-year study (1976–2008) indicated an increasing trend in southwest monsoon rainfall and annual rainfall at Eruthempathy and Malampuzha Dam, but northeast monsoon rainfall fell at all four stations. Sen's slope analysis placed Malampuzha Dam at having the highest year-to-year increase in rainfall (1.55 mm/year), and the maximum decrease was recorded by Thrithala (−5.80 mm/year). As a whole, the study recorded rising rainfall trends in the north and east and decreasing trends in the south and west, which is crucial information for managing future water resources.

Nair et al. (2014) conducted the Spatio-temporal analysis of rainfall trends over a coastal state (Kerala) of India for the past 100 years. This research examines rainfall variability and trends in Kerala for the last 100 years, which show strong (99%) declining trends, especially in January, July, and November, possibly due to global climate anomalies, urbanization, and deforestation. Regional variation indicates more rainfall variability in northern and southern Kerala, with changing seasonal means and rising asymmetry in rainfall distribution, as revealed by the seasonality index (SI).

Dash et al. (2018) also did research on the forecasting of rainfall over Kerala, India, applying artificial intelligence methods. Their work addressed shifting monsoon trends in Kerala, and results showed that the state experienced declining seasonal and post-monsoon rainfall, together with rising deficits in rainfall years, potentially enhancing water shortage in the state. The research utilized artificial intelligence models and established that the Extreme Learning Machine (ELM) model, with a reduced error rate of 3.8729%, performed better than the Single Layer Feed-Forward Neural Network (SLFN) in predicting summer monsoon rainfall.

Kamath and Kamat (2018) investigated time-series analysis and rainfall prediction for Idukki district, Kerala, based on a dataset obtained from Knoema, an open data website. Their work sought to compare the performance of different time-series models using January 2006 to December 2016 monthly rainfall data. The study applied ARIMA, Artificial Neural Network (ANN), and Exponential Smoothing State Space (ETS) models. Of these, ARIMA outperformed others, as was determined using Root Mean Squared Error (RMSE) and model fit criteria.

Surendran et al. (2020) examined the effect of climate change on heavy rainfall occurrences in Kerala during June to September from 1901 to 2019. They compared heavy rainfall (HR ≥ 100 mm) and moderate rainfall (MR between 5 mm and 100 mm) events, separating the period into two phases: 1901–1970 and 1971–2019. There was a notable declining trend in both MR (99% confidence level) and HR (95% confidence level) events, as well as a general decline in seasonal rainfall. Dekadal analysis showed a decrease in MR and HR events, especially in late July and mid-August, while HR events during early August had a notable rise (95%) in the second phase, which peaked in 2019 at 127 events. Although overall there has been a decline, some of the eastern regions saw an increase in HR events that triggered recent extreme rainfalls and landslides.

Varghese and Vanitha (2020) implemented time-series-based rainfall forecasting analysis for Idukki district, Kerala, using data from Knoema (2006–2016). These authors compared ARIMA, ANN, and ETS model performances to conclude ARIMA was best, according to RMSE and model fit.

Sulasikin and Nugraha (2021) studied the monthly rainfall forecasting with the Facebook Prophet model as a flood-mitigation effort in Central Jakarta. The researchers compared SARIMA, Facebook Prophet, and the LSTM models as rainfall predictors during a two-year period. Their findings revealed Facebook Prophet as the most effective model with the least MSE and RMSE.

Facebook Prophet successfully forecasted high levels of rainfall for January and February 2021, indicating likely flood threats. These results highlight the value of the model for evidence-based flood mitigation policy and offer a benchmark for further research.

CHAPTER 3

MATERIALS AND METHODOLOGY

3.1 DATA COLLECTION

The dataset utilized in this study comprises monthly rainfall measurements (in millimeters) for Kerala, spanning from July 2010 to December 2024. The data was sourced from the official website of the Department of Economics and Statistics, Thiruvananthapuram and Kerala Water Resource Irrigation System.

3.2 METHODOLOGY

The initial step in the analysis involved thoroughly examining the dataset. The primary objective of the study was to analyze rainfall patterns and trends in Kerala and forecast future rainfall using time series models. To begin, exploratory data analysis (EDA) was conducted to understand the dataset's characteristics. Subsequently, the SARIMA model, Holt-Winter's Exponential Smoothing technique, and the Prophet model were applied for forecasting. Finally, the performance of these three models was compared by evaluating their MAE and RMSE values.

3.3 TOOLS FOR ANALYSIS AND FORECASTING

1. EDA(Exploratory Data Analysis)
2. Seasonal ARIMA(Auto regressive Integrated Moving Average)
3. Holt-Winter's Exponential Smoothening Technique
4. Prophet model

3.4 TOOLS FOR COMPARISON

1. Mean Absolute Error (MAE)

The formula for MAE is:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

where, y_i is the i-th observed value, $\hat{y}_i$ is the corresponding predicted value,

n is the number of observations.

2. Root Mean Squared Error (RMSE)

The formula for RMSE is:

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

where, y_i is the actual value for the i-th observation, $\hat{y}_i$ is the predicted value for the

i-th observation, n is the number of observations.

3.5 PYTHON

In this study, Python is a high-level, versatile, and easy-to-read programming language. Python was used to do EDA to find hidden patterns of the data set and to forecast using SARIMA and Holt-Winter's Exponential Smoothening Technique and prophet model.

CHAPTER 4

EXPLORATORY DATAANALYSIS (EDA)

4.1 EXPLORATORY DATA ANALYSIS (EDA)

It is the first step of data analysis process, it helps to understand the underlying structure and patterns of data. It also helps to understand the relationships within dataset. EDA can be also used to understand the quality of data, check null values, missing values etc. Using EDA after checking for null and missing values, next step is to summarize and visualize the data to understand it more. This may include calculating summary statistics such as mean, median, mode and standard deviation etc. Visualizing the data, decomposing the data to trend , seasonal and residual also include in this.

4.2 DESCRIPTIVE STATISTICS

Descriptive statistics summarize the key characteristics of a dataset, including its central tendency, variability, and distribution. Measures such as the mean, median, and mode represent the dataset's central values, while standard deviation and quartiles help assess the spread of data and detect outliers or extreme values. Overall, descriptive statistics provide a comprehensive summary of the dataset, facilitating a better understanding of the data and ensuring accurate further analysis.

4.3 TIME SERIES VISUALISATION

Time series visualization refers to the graphical representation of data recorded at consecutive time intervals. Various techniques, including line plots, seasonal subseries plots, autocorrelation plots, histograms, and interactive visualizations, help analysts detect trends, patterns, and anomalies within the data. These visual tools play a crucial role in interpreting time-dependent data and making informed decisions.

4.4 SEASONAL DECOMPOSITION

Seasonal decomposition is a technique used to break down time series data into three components: trend, seasonality, and residual. This process helps in identifying whether the data exhibits seasonal patterns and understanding the overall trend. Recognizing these components aids in selecting an appropriate forecasting model and improving prediction accuracy.

Trend: Represents the long-term increase or decrease in data over time.

Seasonal: Refers to a recurring and predictable pattern that repeats at regular intervals, such as monthly, weekly, or daily. Rainfall patterns are an example of seasonal behavior in time series data.

Residual: The difference between the actual observed values and the predicted values. Residuals help evaluate how well the model fits the data.

Overall, exploratory data analysis (EDA) serves as the initial and most critical step in data analysis. It helps identify missing or null values, understand data patterns, and determine the most suitable forecasting models to enhance prediction accuracy.

CHAPTER 5

TIME SERIES ANALYSIS

Time Series Analysis is a statistical process employed for analysis and interpretation of data at pre-specified intervals of time. It enables patterns, trends, seasonality, and intrinsic patterns in the underlying time-dependent data to be revealed, leading to predictions as well as analysis for future values. A time series is an interaction between a variable and time itself, measured over regular time periods like annual, monthly, weekly, daily, or hourly. Some examples of time series data are hourly temperatures, daily sales, and monthly production. Mathematically, a time series is defined by the function relationship $Y_t = f(t)$, where Y_t is the value of the variable under consideration at time t.

5.1 COMPONENTS OF TIME SERIES

Trend(T_t): The trend is the general long-run movement or tendency of the data over time. Icaptures whether or not the series follows a consistent rise, fall, or neither. Trends can be linear, with a steady rise or fall, or nonlinear, with more complicated fluctuations.

Seasonality (S_t): Seasonality refers to repeating cycles or variations occurring at regular times within a series of time data. These usually recur on annual, quarterly, monthly, or weekly cycles and are caused by seasonal changes, holidays, or business cycles.

Cyclic variations (C_t) : Cyclical variations are long-term fluctuations in a time series that donot have a fixed period such as seasonal patterns. These cycles generally last for a few years and are associated with economic or business cycles, capturing phases of growth and decline.

Irregularity or Noise (I_t) : Irregularity, alternatively referred to as noise or randomness, accounts for unexpected changes in the data that cannot be explained through trend, seasonality, or cyclical patterns. These fluctuations can be a result of random occurrences, errors in measurements, or unanticipated external forces, which increase the difficulty in observing underlying patterns within the time series.

5.2 MATHEMATICAL MODELS FOR TIME SERIES ANALYSIS

In time series analysis, there are supposed to be two models typically for the decompositions of a time series into its constituents.

a) Additive Model: As per the additive model, decomposition of time series is carried out with the assumption that the impact of different components is additive.

$$Y_t = T_t + S_t + C_t + I_t$$

Where $Y_t$is the time series value and T_t , S_t , $C_t$and I_t stands for trend, seasonal variations, cyclical variations and irregular variations respectively. In this model S_t, $C_t$and I_t are absolute quantities and can have positive or negative values. The model postulates that four elements of the time series are independent of one another.

b) Multiplicative Model: Under the multiplicative model, the breakdown of a time series is made on the premise that the impacts of the four parts of the time series are not mutually independent. Under the multiplicative model,

$$Y_t = T_t * S_t * C_t * I_t$$

In this model T_t , S_t , $C_t$and I_t are not absolute amounts as in the case of the additive model. There are relative variations and are expressed as rates or indices fluctuating above or below unity. The multiplicative model can be expressed in terms of the logarithm.

$$\log Y_t = logT_t + logS_t + logC_t + logI$$

5.3 MEASUREMENT OF SEASONAL VARIATIONS

Seasonal variation is measured in terms of an indicator, called a seasonal indicator. It's an normal that can be used to compare an factual observation relative to what it would be if there was no seasonal variation.

(a) Method of simple averages

(b) Ratio to Trend Method

(c) Ratio to Moving average method

5.4 TIME SERIES MODELING

5.4.1 BASIC DEFINITIONS

i. Stationary Time series

Stationary time series refers to series of observations where the mean, variance and the Autocorrelation remains constant over time. It is a time series data which exhibit a stable behaviour without trend and seasonality.

ii. Non-stationary Time series

Non-stationary time series refers to series of observation where the mean variance and the Autocorrelation varies over time. It is a time series data exhibit unstable behaviour with trend, seasonality and other patterns. Non-stationary data cannot be used for analysis.

iii. Auto Correlation Function (ACF)

Autocorrelation function in time series is a tool used to measure the correlation between a time series and its lagged value at different time intervals. ACF value of 1 or -1 indicate strong positive or negative autocorrelation. Patterns of ACF give idea about seasonality and other random behaviours. Stationarity can be assessed by ACF, ACF plots with lags dying to zero represents stationarity. The autocorrelation function of a stationary time series $\{Z_t\}$, $\rho(k)$ at lag k is defined as the correlation at lag k between Z_t and Z_{t+k}. Thus the autocorrelation function at lag k is given by,

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

$$\text{where } \gamma(k) = Cov\,(Z_t\,, Z_{t+k})$$

iv. Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function (PACF) comes to test the immediate connection between two observations of a time series with further data effect accounted for. PACF employed to determine the MA parameter for SARIMA and ARIMA model. Partial autocorrelation function, as with the autocorrelation function, carries important information about the dependence structure of a stationary process.

In the context of time series, a large portion of the correlation between Z_t and Z_{t+k}can be due to the correlation these variables have with $(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k+1})$. The partial autocorrelation of lag *k* can be thought of as the partial regression coefficients kk in the representation.

$$Z_t = \emptyset_{k1} Z_{t-1} + \emptyset_{k2} Z_{t-2} + \ldots + \emptyset_{kk} Z_{t-k} + \varepsilon_t$$

Thus the partial autocorrelation at lag k, $\emptyset_{kk}$, measures the correlation between $Z_t$and Z_{t-k} after adjusting for the effects of $(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k+1})$.

v. Augmented Dickey Fuller Test

The ADF test is part of a type of test referred to as 'Unit Root Test', which is the correct procedure for testing a time series for stationarity. Augmented Dickey-Fuller (ADF) test is a popular statistical test applied to test if a specific time series is stationary or not. It is one of the most popular statistical tests used in the context of analyzing the stationarity of a series. It is testing the following two null and alternative hypotheses:

H_0: The time series is non-stationary.

H_1: The time series is stationary.

Now, if the p-value from this test comes out to be less than a particular level (e.g. $\alpha = 0.05$) then in such cases the null hypothesis is rejected and concludes that the time series is stationary.

vi. Stationarity

The ARIMA technique is suitable only for a stationary series of data. Stationarity means that the AR coefficients should meet certain requirements for an ARIMA model to be stationary. There is a reason why we need stationarity: otherwise, we could not obtain meaningful estimates of the parameters of a process. When p=0, we have a pure MA model or a white noise series. All white noise and pure MA models are stationary, and so there are no stationarity conditions to test.
For an AR (1) or ARMA(1, q) process, the stationary requirement is that the absolute value of $\emptyset_1$ must be less than one: $|\emptyset_1| < 1$.
For an AR(2) or ARMA(2, q) process, the stationary requirement is a set of three conditions: $|\emptyset_2| < 1$, and $\emptyset_1 - \emptyset_2 < 1$.

vii. Invertibility

There is a second requirement that ARIMA models need to meet known as invertibility. This condition stipulates that the coefficients of the MA need to meet specific requirements. There is a rational basis for invertibility: a non-invertible ARIMA would mean that weights assigned to older Z observations do not decrease as we go back further in time, but rationality dictates that higher weight should be assigned to more recent observations. Invertibility guarantees that these outcomes persist. If q=0, then we have either a pure AR process or a white noise series. All white noise and pure AR processes are invertible and no additional checks are needed

For an MA (1) or ARMA(p, 1) process, invertibility requires that the absolute value of θ1 must be less than one, $|\theta_1| < 1$, $\theta_1+\theta_2 < 1$, and $\theta_1 - \theta_2 < 1$.

viii. Auto Regressive (AR) Process

A time series $\{Z_t\}$ is said to be an autoregressive process of order p, abbreviated as AR(p) if it is a weighted linear sum of the past p values plus a random shock so that

$$Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \ldots + \emptyset_p Z_{t-p} + \varepsilon_t$$

where $\{\varepsilon_t\}$ denotes a purely random process with 0 mean and constant variance σ^2. Using the backward shift operator B, such that $BZ_t = Z_{t-1}$, the AR (p) model may be written more succinctly in the form,

$$\emptyset(B) Z_t = \varepsilon_t$$

where $\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \ldots - \emptyset_p B^p$ is a polynomial in B of order p.

ix. Moving Average (MA) Process

A time series $\{Z_t\}$ is said to be a moving average process of order q, abbreviated as MA(q) if it is a weighted linear sum of the last q random shocks so that

$$Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q}$$

where $\{\varepsilon_t\}$ denotes a purely random process with 0 mean and constant variance σ^2 . Using the backward shift operator B the MA (q) model may be written in the form,

$$Z_t = \theta(B)\, \varepsilon_t$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$ is a polynomial in B of order q.

x. Auto Regressive Moving Average Process (ARMA)

Mixed autoregressive moving average model with p auto regressive terms and q moving average terms is abbreviated as ARMA(p,q) is given by

$$Z_t - \emptyset_1 Z_{t-1} - \emptyset_2 Z_{t-2} \text{ -...- } \emptyset_p Z_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q}$$

Using the backward shift operator B, the ARMA(p,q) can be written in the form,

$$Z_t - \emptyset_1 B Z_t - \emptyset_2 B^2 Z_t \text{ -...- } \emptyset_p B^p Z_t = \varepsilon_t - \theta_1 B \varepsilon_t - \theta_2 B^2 \varepsilon_t - \theta_q B^q \varepsilon_t$$

$$\emptyset(B)\, Z_t = \theta\,(B)\ \varepsilon_t$$

where $\emptyset(B)$ *and* $\theta\,(B)$ are polynomial in B of degree of order p and q respectively

xi. Akaike Information Criteria (AIC)

The Akaike Information Criterion (AIC) is a statistical metric used to evaluate and compare models in time series forecasting. It helps identify the best-fitting model by balancing goodness of fit and model complexity.

The AIC is calculated using the formula:

$$AIC = 2k - 2\ln(L)$$

where k represents the number of parameters in the model, and L is the maximum likelihood of the model. A lower AIC value indicates a better model, as it suggests a good trade-off between accuracy and simplicity.

xii. Diagnostic Checking

In diagnostic checking, we need to test model adequacy by inspecting whether or not the assumptions of the model hold. The general assumption is that $\{\varepsilon_t\}$ is white noise. Therefore, model diagnostic checking is achieved via meticulous examination of residual series $\{\varepsilon_t\}$. To ensure that the errors are normally distributed, one would create a histogram of standardized residuals and compare the same with standard normal distribution. The assumption that random shocks have zero mean and constant variance can be tested using a residuals plot. To test if the residuals are roughly white noise, we calculate the sample ACF of the residuals to determine if they are all statistically insignificant. The three-stage UBJ procedure is iterative in nature. Estimation and diagnostic-checking stages offer warning indications when, and in what manner, a model must be reformulated. It is persisted in to re-identify, re-estimate, and recheck until a model is obtained satisfactory by a variety of criteria.

xiii. Q-Q Plot

The quantile-quantile(q-q plot) graph is a visual technique for deciding whether a data set follows some probability distribution or not, and whether two sets of data have been drawn from the same population or not. Q-Q plots are especially convenient for checking if a data set is normally distributed or if it follows some known distribution

xiv. Correlogram

A correlogram is a graphical plot of the autocorrelation function (ACF) for a time series against values of lags. It provides a measure of correlation between observations at various time lags. It is useful to detect patterns, autocorrelation, seasonality, and whether residuals are like white noise (no large autocorrelations).

xv. Histogram

A histogram displays the distribution of the residuals by frequency, KDE gives a smoothed estimate of the density, and the standard normal curve plots on top to compare the distribution of the residuals to a normal distribution.

xvi. Ljung-Box Test

The Ljung-Box test is a statistical test to see if a time series has significant autocorrelation (i.e., whether future values depend on past values). The test can be used to test if the residuals (errors) of a model are white noise or if there is still structure that the model has failed to explain. The test provides a p-value, and based on this, conclusions can be made:

If p-value > 0.05 , then the residuals are independent (no significant autocorrelation), the model adequately captures the time series structure and no significant patterns are left in the residuals.

5.5 SEASONAL AUTOREGRESSIVE MOVING AVERAGE (SARIMA)

Time series data are non-stationary, i.e., their statistical characteristics (mean, variance, autocorrelation) vary over time. AR, MA, and ARMA models presume stationarity, and hence cannot be used for non-stationary series directly. To deal with non-stationary data, one usual method used is differencing, which converts the series to a stationary form by calculating the difference between two consecutive observations. This is achieved by subtracting the last value

from the present value:

$$Z_t - Z_{t-1} = (1-B)\,Z_t$$

where B is the backward shift operator such that $BZ_t = Z_{t-1}$. If necessary, this differencing can be applied multiple times until stationarity is achieved. After differencing d times, an ARMA (p, q) model can be fitted to the transformed data. The resulting model is called ARIMA (p, d, q), where:

p is the order of the autoregressive (AR) process,

d is the number of differences applied to achieve stationarity,

q is the order of the moving average (MA) process.

Mathematically, an ARIMA (p, d, q) model is expressed as:

$$\emptyset(B)(1-B)^d Z_t = \theta(B)\varepsilon_t$$

where:

$\emptyset$ (B) = $1 - \emptyset_1 B - \emptyset_2 B^2 - \ldots - \emptyset_p B^p$ represents the AR component.

θ(B) = $1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$ represents the MA component.

$(1-B)^d$applies d^{th} order differencing.

ε_t is a white noise process with mean 0 and constant variance σ^2

Incorporating seasonality :

Most real-time series are seasonal, i.e., they have cycles that recur at fixed intervals s (e.g., monthly data with s =12 for an annual cycle). SARIMA (Seasonal ARIMA) is an extension of ARIMA that adds seasonal terms to capture cycles recurring every s time periods.

A SARIMA (p, d, q) × (P, D, Q)_s model accounts for both non-seasonal and seasonal components, where:

P is the order of the seasonal autoregressive, this term represents the relationship, especially at seasonal lags, between the series' past values and present

D is seasonal differencing order, this part covers the differencing that is required to remove seasonality from the series, similar to that of non-seasonal differencing.

Q is the seasonal moving average order, this component replicates the relationship that there is between the current value and the seasonal lags of the residual errors

s is the seasonal period

The mathematical expression of SARIMA model can be stated as :

$$\emptyset(B)\Phi(B^s)\ (1-B)^d(1-B^s)^D Z_t = \theta(B)\Theta(B^s)\ \varepsilon_t$$

where:

$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \ldots - \emptyset_p B^p$

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$

$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_p B^{Ps}$

$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$

$(1-B)^d$ applies non-seasonal differencing d times.

$(1-B^s)^D$ applies seasonal differencing D times.

ε_t is a white noise process.

This multiplicative model retains both short-term (non-seasonal) and long-term (seasonal) relationships in time series data, and hence SARIMA is a very effective method for forecasting data with trend and seasonal variations.

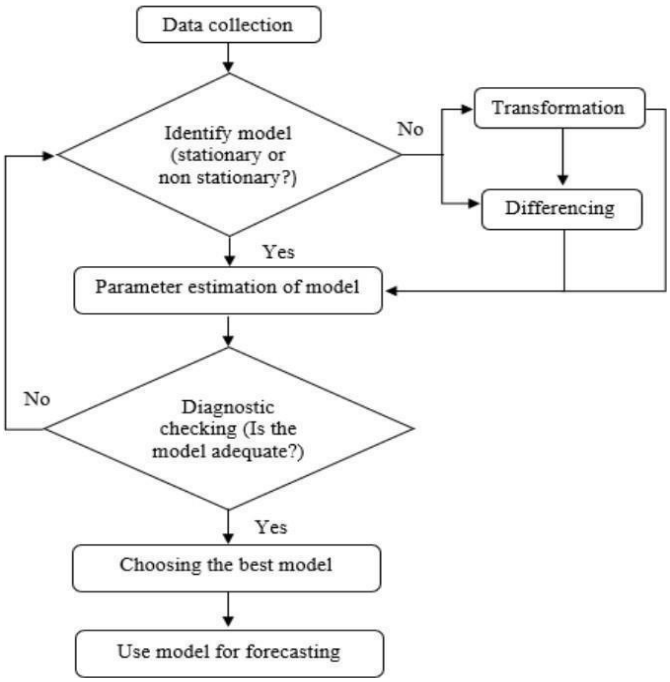


figure 5.1 Steps involved in forecasting using SARIMA model

5.6 HOLT WINTERS EXPONENTIAL SMOOTHING TECHNIQUE

Holt's method can be extended to deal with time series which contain both trend and seasonal variations.Holt-Winters can be in both additive and multiplicative forms depending on the exact nature of the time series data.. Let L_t, T_t, $I_t$denote the local level, trend and seasonal index, respectively at time *t*.

The interpretation of it depends on whether seasonality is thought to be additive or multiplicative. In the additive case I, $y_t - I_t$ is the deseasonalized value, while in the multiplicative class it is y_t/I_t. The values of the 3 quantities L_t, T_t, I_t all need to be estimated and so we need 3 updating equations with three smoothing parameters, say $\propto$, γ and δ. As before the smoothing parameters are usually chosen in the range (0, 1). The form of the updating equations is again intuitively plausible.

Suppose the seasonal variation is multiplicative. Then the (recurrence form) equations for updating L_t, T_t, I_t when a new observation y_t becomes available are

$$L_t = \alpha (y_t / I_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma (L_t - L_{t-1}) + (1-\gamma) T_{t-1}$$

$$I_t = \delta (y_t / L_t) + (1-\delta) I_{t-s}$$

and the forecasts from time (t) are then,

$\hat{y}_{t+h} = (L_t + hT_t) I_{t-s+h}$ *for* $h = 1,2,\ldots\ldots s$

where s denotes the seasonal period (For example: s = 4 for quarterly data and 12 for monthly data). If the seasonal variation is additive, the equations for updating L_t, T_t, I_t when a new observation y_t becomes available are

$$L_t = \alpha (y_t - I_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \gamma (L_t - L_{t-1}) + (1-\gamma) T_{t-1}$$
$$I_t = \delta (y_t - L_t) + (1-\delta) I_{t-s}$$

and the forecasts from time (t) are then

$\hat{y}_{t+h} = (L_t + hT_t) I_{t-s+h}$ *for* $h = 1,2,\ldots\ldots s$

For starting values, it seems sensible to set the level component L_0 , equal to the average observation in the first year that is,

$$L_0 = \sum_{t=1}^{s} y_t$$

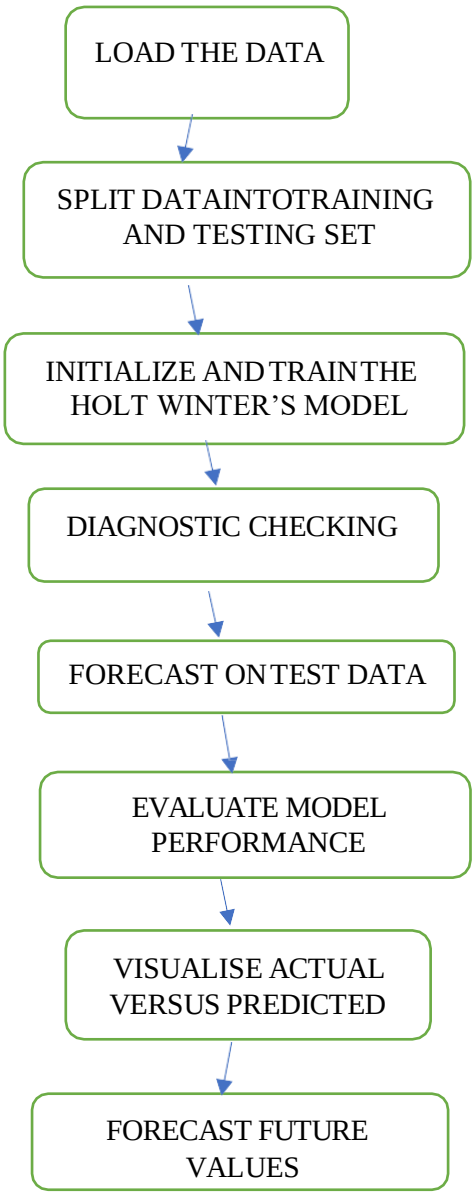
where s is the number of seasons. The starting values for the slope component can be taken from the average difference per period between the first and second-year averages. That is,

$$T_0 = \frac{\frac{\sum_{t=s+1}^{2s} y_t}{s} - \frac{\sum_{t=1}^{s} y_t}{s}}{s}$$

Finally, the starting value of S I can be calculated after allowing for a trend adjustment, as follows:

$$I_0 = (y_k - (k-1) T_0) / 2 \text{ (multiplicative)}$$
$$I_0 = (y_k - (L_0 + (k-1) T_0)/2) \text{ (additive)}$$

Where $k = 1, 2 \ldots\ldots s$. This will lead to (s) separate values for I_0, which is what is required to gain the initial seasonal pattern.



5.7 PROPHET MODEL

The Prophet model is a time series model, which was revealed by its creators, Taylor and Letham (2017) of the Facebook data science team. It is an open-source and Python and R- supported forecasting tool applied in forecasting. Although Facebook Prophet provides yearly, monthly, and daily forecasts in non-linear data, it also incorporates holidays as specified. It can pre-process data. The following equation is used to define the process of integrating the components:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Additivity y(t) in time series data trends are denoted by g(t), seasonality by s(t), holiday impact by h(t), and model error ε_t in the Regressive Model. The model is built with the python-based fbprophet API and takes only two inputs: the target variable to be forecasted, which is denoted as "y," and the timestamp, which is marked as "ds".

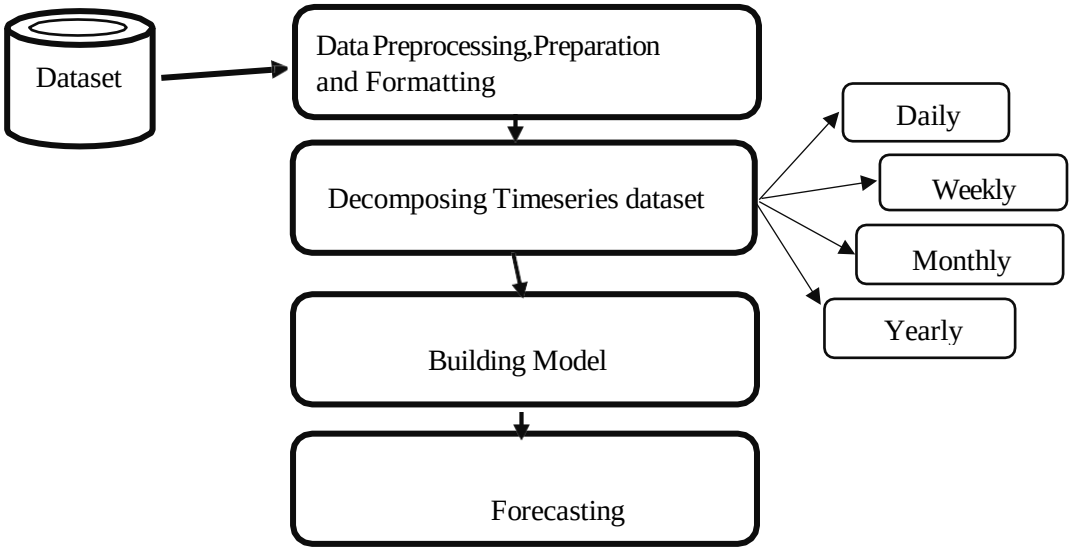


figure 5.2 Steps involved in forecasting using prophet model

The main process in prophet is:

1. Data Preprocessing: This was the part where data cleaning and parameter choice for analysis was taken into account.

2. Time Series Creation and Decomposition: This is where the use case associated with analysis was taken into account. These processes had required sub-factors like daily, weekly, monthly, and yearly choices.

3. Model Building: This step involves building a model for data prediction from chosen factors.

4. Prediction: The last step was predicting the performance of the model using testing data.

A time series decomposition is a significant task that helps in comprehending its very essence. It facilitates easier analysis and prediction of intricate time series with latent components like the trend, seasonal components, and periodic components. Two years of forecasting and just two variables are needed for the Python Prophet. The Facebook (FB) Prophet is the latest tool that has proved to have superior prediction accuracy. The Python fbprophet library is utilized to implement the Prophet approach on the dataset. Since Prophet is univariate, data was pre-cleaned to have date only and dependent factors.

CHAPTER 6

RESULTS AND DISCUSSION

This chapter discusses a comparative study of time series modeling and forecasting of monthly rainfall of Kerala using SARIMA, Holt-Winters Exponential Smoothing forecasting and Prophet model. The data comprises 174 observations from July 2010 to December 2024.

6.1 EXPOLATORY DATA ANALYSIS

6.1.1 Descriptive Statistics

Table 6.1 Descriptive Statistics

count	174.000000
mean	236.926437
std	230.067491
Min	0.300000
25%	47.350000
50%	168.050000
75%	377.425000
max	1041.100000

6.1.2 Time Series plot

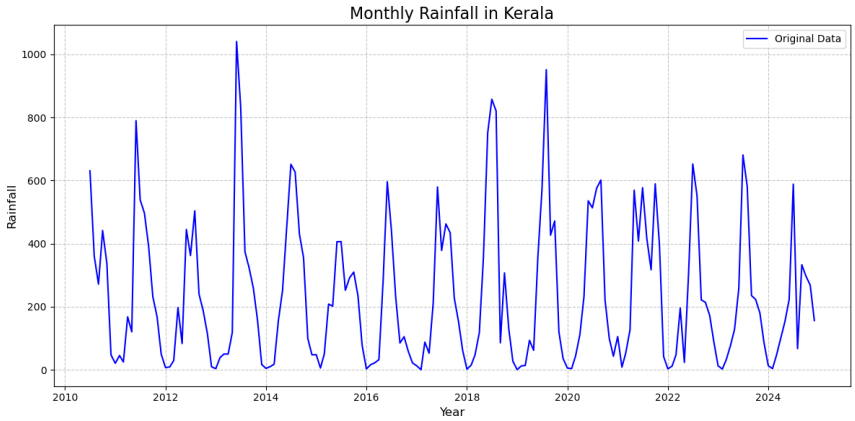


figure 6.1 Time series plot

6.1.3 Time Series Decomposition

Seasonal decomposition is performed for the evaluation of trend, seasonality and random components. From *figure 6.2* it is clearly visible that it shows seasonality.

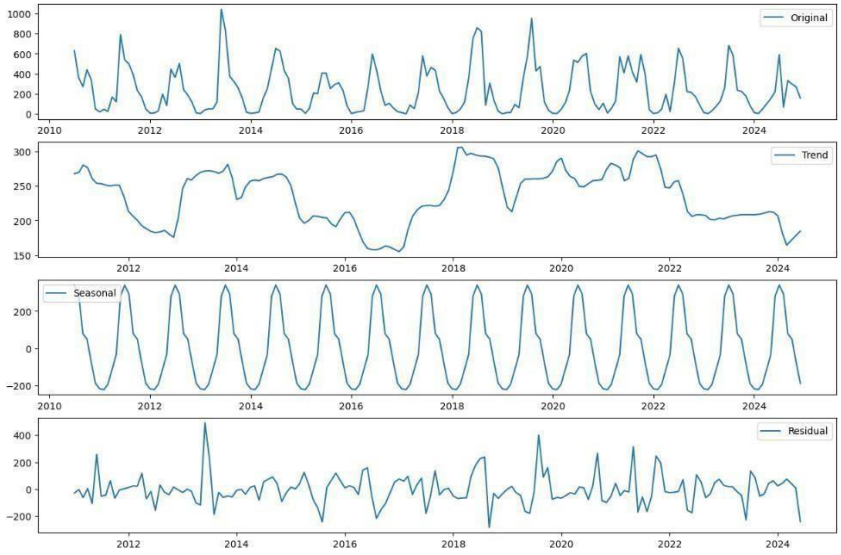


figure 6.2 Time Series Decomposition

Original Series (Top Panel): This plot represents the actual observed time series data. It shows significant seasonal fluctuations, with peaks and troughs occurring at regular intervals, which is typical of rainfall data.

Trend Component (Second Panel): This represents the long-term trend in the data. It smooths out short-term variations to show whether the overall rainfall levels are increasing, decreasing, or remaining stable. From the plot, the trend shows some fluctuations but generally appears to rise around 2018–2022 before slightly stabilizing.

Seasonal Component (Third Panel): This captures repeating seasonal patterns in the data. The regular peaks and troughs suggest strong seasonality, which is expected in rainfall data due to monsoon effects.

Residual Component (Bottom Panel): This represents the remaining variation in the data after removing the trend and seasonal components. It consists of irregular, random fluctuations that cannot be explained by the trend or seasonality alone.

6.2 MODELING AND FORECASTING OF RAINFALL USING SARIMA MODEL

figure 6.1 depicts the time series plot of monthly rainfall data. It is visible that there is seasonality in the data. The visual inspection alone is not enough to specify that the changes in the mean are statistically significant. So, to decide ACF and PACF is plotted. *figure 6.3* and *figure 6.4* shows ACF and PACF.

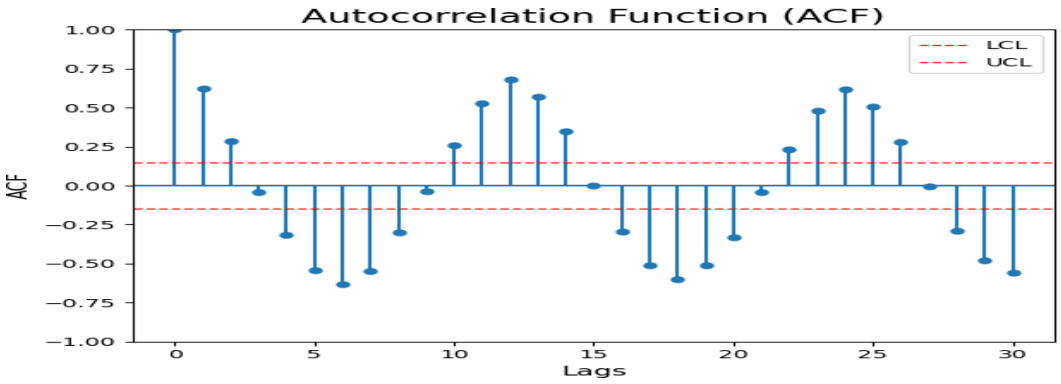


figure 6.3 ACF plot of time series data

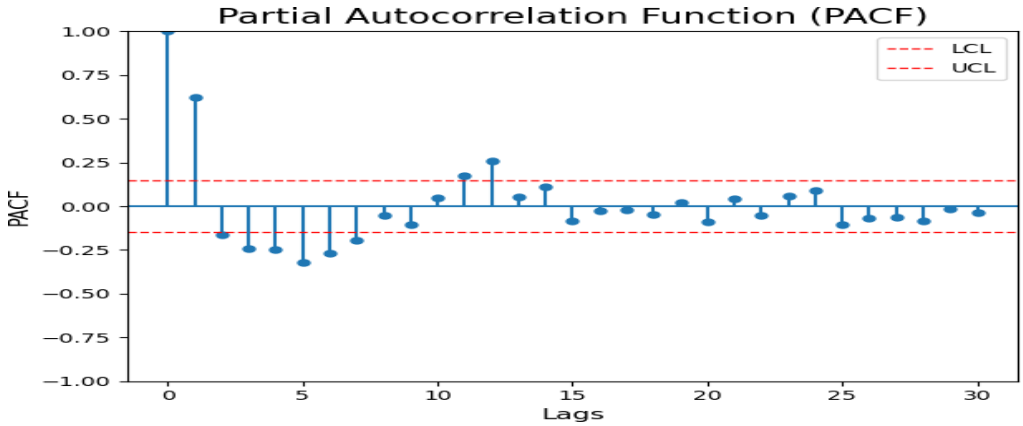


figure 6.4 PACF plot of time series data

The *figure 6.4* indicates that the ACF fails to die out rapidly towards zero. Rather than dying down along the first several lags, the graph displays a slow decrease in the size of ACF values, which is a typical pattern for a non-stationary series.

Now the seasonal index for various seasons can be obtained by ratio to moving average method. From *figure 6.5* it is obtained the seasonal indices for the data which shows the maximum rainfall is received during July and minimum rainfall is received during the month of February followed by January.

Table 6.2 Seasonal Indices

Month	Seasonal Indices
January	0.074224
February	0.056217
March	0.173290
April	0.514927
May	0.851742
June	2.184687
July	2.475431
August	2.072198
September	1.335291
October	1.266072
November	0.764395
December	0.231526

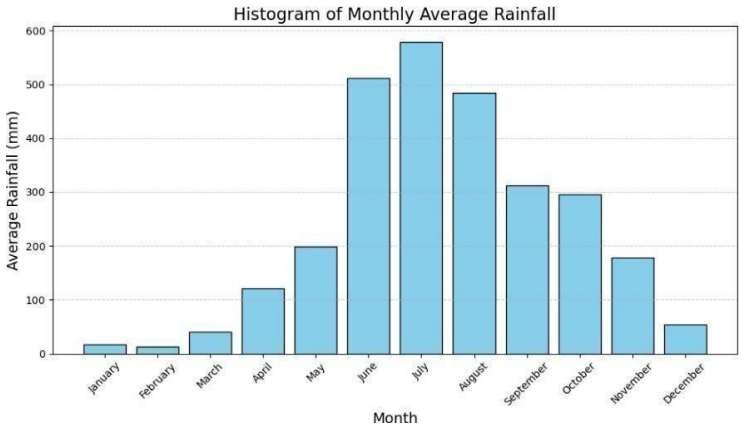


figure 6.5 Plot of Seasonal Indices

The seasonal behaviour of monthly data indicates the presence of a seasonal component.

$$H_0 = Data\ is\ stationary$$

$$H_1 = Data\ is\ non-stationary$$

On performing the ADF test we get,

Table 6.2 ADF test result of non stationary data

Dickey -Fuller	**-2.2482811418475026**
p-value	**0.18919088296251468**

Since p-value is less than 0.05, we reject the null hypothesis and the data is non stationary. So, perform seasonal differencing to make it stationary by taking the difference between a value and a value with lag S=12 for the transformed data. That is, the seasonal difference is $x_t = x_t - x_{t-12}$. The time series plot for the seasonally differenced data is shown in the *figure 6.6.*

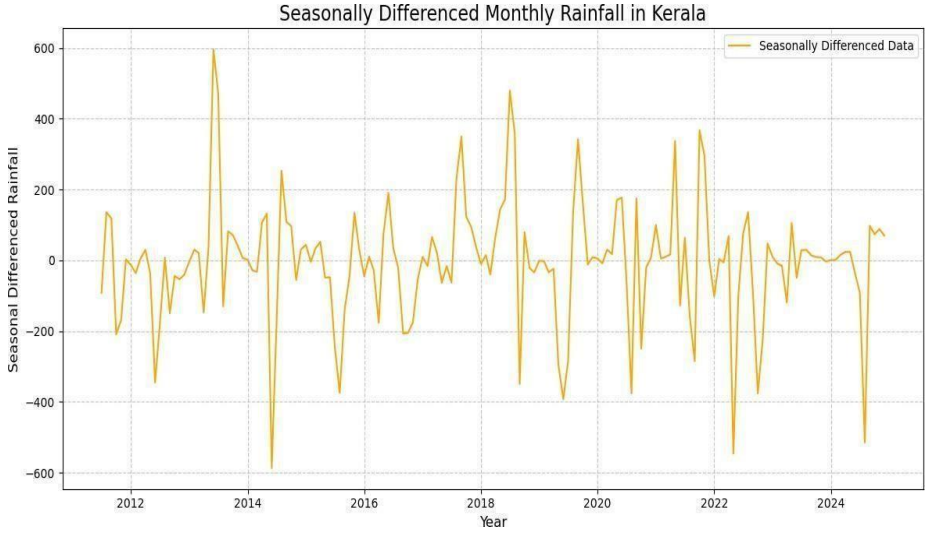


figure 6.6 Plot of seasonally differenced monthly rainfall in Kerala

From the seasonal differenced *figure 6.6* it is evident that the seasonal behaviour is removed from the series. Now again plot the seasonal differenced ACF and PACF.

ACF plot of Seasonally differenced data

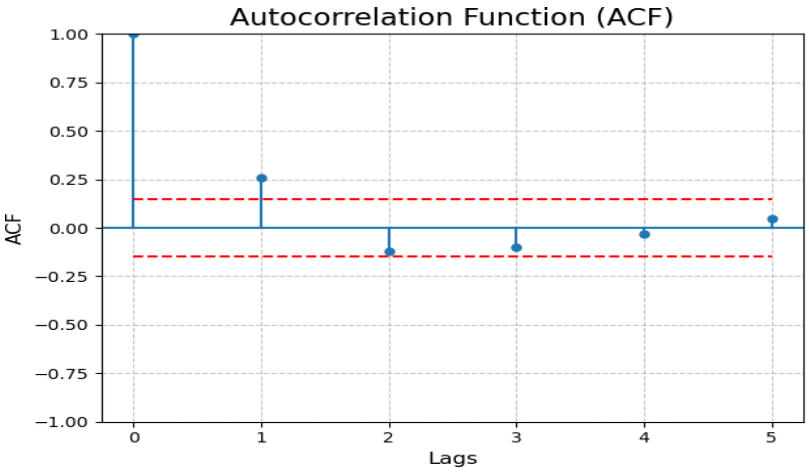


figure 6.7 ACF plot of seasonally differenced data

PACF plot of Seasonally differenced data

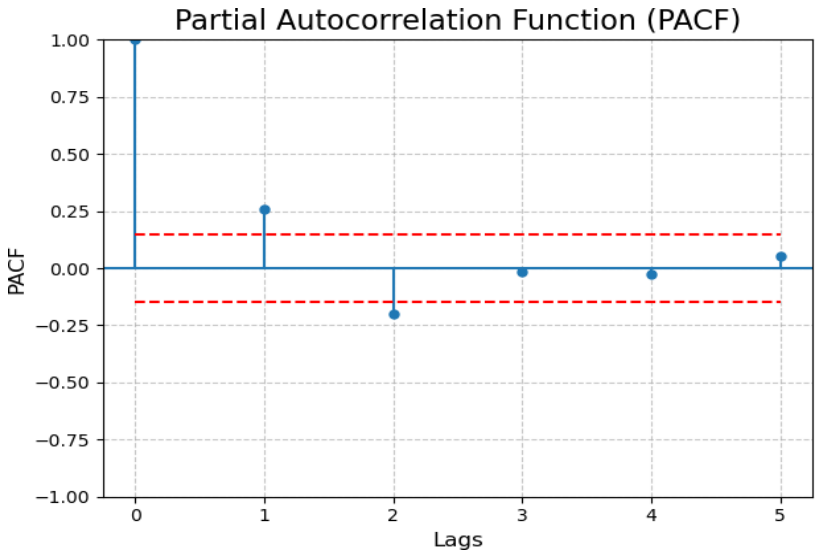


figure 6.8 PACF plot of seasonally differenced data

Augmented Dickey-Fuller (ADF) Test

To clarify whether the differenced series is stationary or not. ADF test is performed. The result of the test is in table 6.3.

Table 6 .3 ADF test result of stationary data

Dickey-Fuller	-4.272193193805211
p-value	0.0004961284892897813

Since p-value is less than 0.05, it is clear that the differenced data is stationary. No more seasonal differencing is needed. Now D=1 and d=0 , to find the seasonal AR order (P) and the seasonal MA order (Q) have to plot the ACF and PACF OF stationary data at seasonal lags.

ACF of seasonally differenced data at seasonal lags

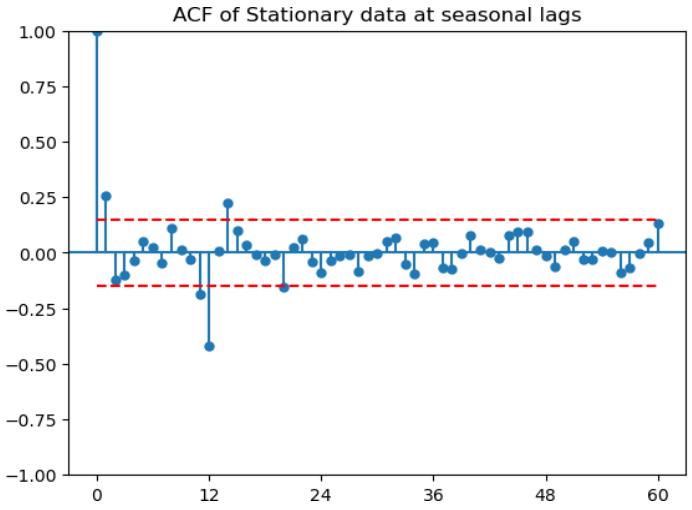


figure 6.9 ACF plot of seasonally differenced data at seasonal lags

PACF of seasonally differenced data at seasonal lags

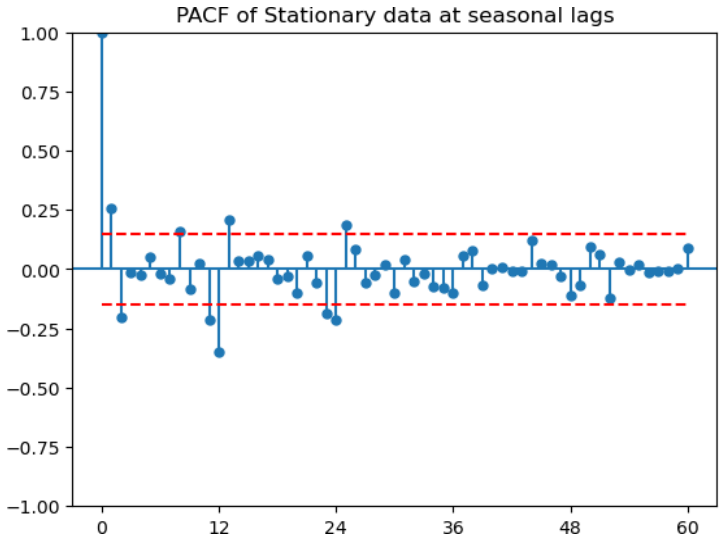


figure 6.10 PACF plot of seasonally differenced data at seasonal lags

figure 6.9 and *figure 6.10* shows the ACF and PACF of seasonally differenced data at seasonal lags. From the above ACF and PACF of seasonally differenced data non seasonal AR order and non- seasonal MA order is maximum p = 2 and maximum q = 1, and also know that d = 0. From the ACF and PACF of seasonally differenced data at seasonal lags maximum P = 8 and maximum Q = 4 and D = 1.

Thus, the possible time series models and their corresponding AIC statistics for the monthly

rainfall data is :

Table 6.4 ARIMA models and their corresponding AIC values

NO.	ARIMA(p,d,q) x (P,D,Q)	AIC
1	ARIMA(0,0,0)(0,1,0)[12]	2134.775
2	ARIMA(1,0,1)(1,1,0)[12]	2061.382
3	ARIMA(0,0,0)(0,1,0)[12]	2132.785
4	ARIMA(1,0,0)(0,1,0)[12]	2127.286
5	ARIMA(1,0,0)(2,1,0)[12]	2020.292
6	**ARIMA(0,0,2)(1,1,1)[12]**	**1869.696**
7	ARIMA(1,0,0)(3,1,0)[12]	2000.908
8	ARIMA(1,0,0)(4,1,0)[12]	1977.604
9	ARIMA(1,0,0)(5,1,0)[12]	1977.012
10	ARIMA(1,0,0)(6,1,0)[12]	1973.856
11	ARIMA(1,0,0)(7,1,0)[12]	1975.440
12	ARIMA(0,0,0)(6,1,0)[12]	1986.306
13	ARIMA(2,0,0)(6,1,0)[12]	1973.343
14	ARIMA(2,0,0)(5,1,0)[12]	1977.366
15	ARIMA(2,0,0)(7,1,0)[12]	1975.095
16	ARIMA(2,0,1)(6,1,0)[12]	1977.719
17	ARIMA(1,0,1)(6,1,0)[12]	1974.084

According to minimum Akaike Information Criteria, ARIMA(0,0,2) x (1,1,1)[12] model is found to be more appropriate. The parameter estimates for the model are given in the *table 6.5.*

Table 6.5 ARIMA(0,0,2)x(1,1,1)[12] model parameters

Parameter	Coefficient	Standard Error	z	P>|z|
ma.L1	0.2664	0.069	3.866	0.000
ma.L2	0.0496	0.071	0.695	0.487
ar.S.L12	-0.0598	0.077	-0.773	0.440
ma.S.L12	-0.7638	0.079	-9.659	0.000
sigma2	1.762e+04	1429.577	12.325	0.000

The fitted SARIMA model is :

$$(1 + 0.0598\,B^{12})(1 - B^{12})\,Z_t = (1 - 0.2664\,\mathrm{B} - 0.0496\,B^{12})(1 + 0.7638\,B^{12})\,\varepsilon_t$$

Diagnostic Checking

Diagnostic checking is a crucial step to ensure the reliability, effectiveness and validity of statistical models. It helps to understand how precise the model is and to improve prediction accuracy.

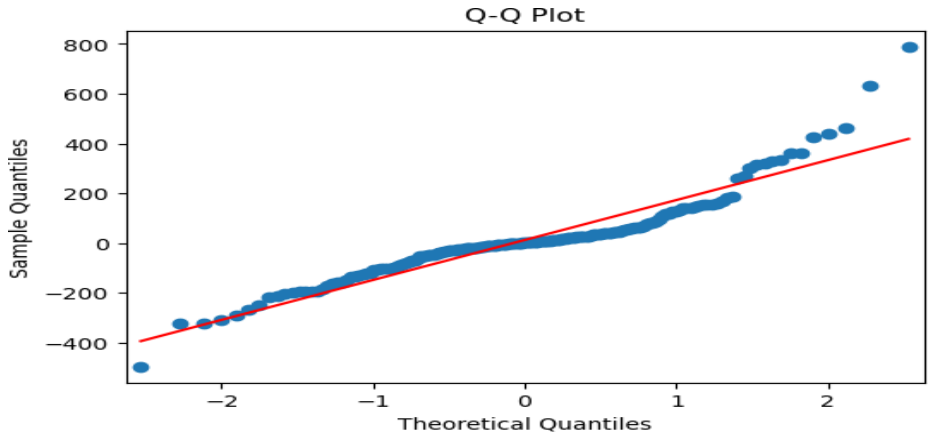


figure 6.11 Q-Q plot of residuals

The *figure 6.11* depicts the quantile-quantile plot comparing the distribution of residuals with the normal distribution. It is clear that the most of the residual values lie on the straight line , which indicates the residuals are approximately normally distributed.

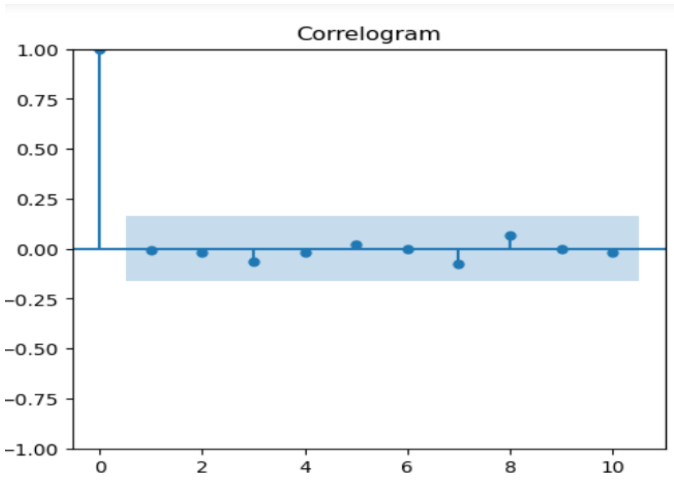


figure 6.12 Correlogram of residuals

The *figure 6.12* shows the correlogram, examination of correlogram can be evidently seen that all the lags die to zero means that there is no significant autocorrelation present in the data.

Since all autocorrelations fall within the confidence intervals (shaded area), the residuals are not significantly autocorrelated, indicating a good model fit.

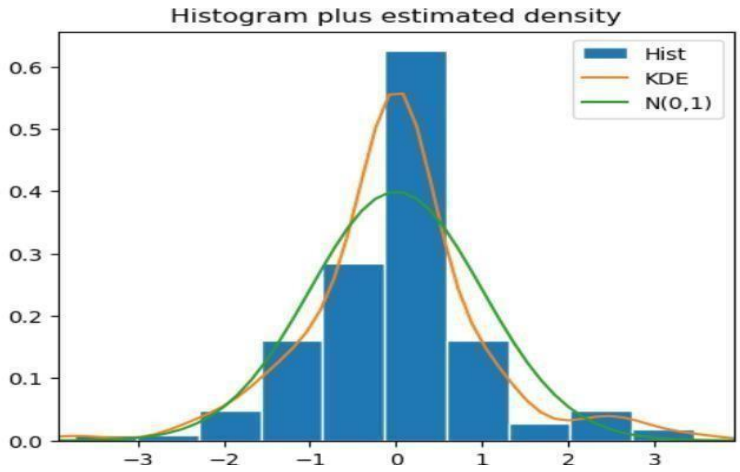


figure 6.13 Histogram of residuals

The *figure 6.13* shows the histogram of standardized residuals overlaid with the kernel density estimate (KDE) and the standard normal distribution N(0,1). Since the histogram matches the normal distribution curve (green line), the residuals are approximately normal.

Table 6.6 Ljung box test result

Test statistic	p_value
15.748041	0.107076

H_0 : There is significant autocorrelation in the residuals

H_1 : There is no significant autocorrelation in the residuals

Table 6.6 indicates the Ljung box test result. Since the p-value (0.107) is greater than 0.05, we fail to reject the null hypothesis. This means there is no significant autocorrelation in the residuals.

Thus the diagnostic checking reveals that the fitted ARIMA (0,0,2) x (1,1,1)[12] model is statistically adequate. Also, the model statisfies stationary and invertibility requirements. So the model can be used to forecast the monthly Rainfall data.

In-sample forecast

Now, the fitted time series model is used to do In-sample forecasting. In-sample forecsting is done for the last year in the dataset that is from January 2024 – December 2024.

Table 6.7 In-sample forecast using the SARIMA model

Months	**Actual value(mm)**	**Predicted value(mm)**
Jan	13.0	32.9674
Feb	3.9	5.8867
Mar	46.8	40.4900
Apr	100.2	123.5940
May	152.7	204.2318
Jun	223.0	419.1762
Jul	588.7	548.3822
Aug	67.3	562.6908
Sep	333.1	179.8283
Oct	297.0	323.8057
Nov	268.6	186.3770
Dec	156.2	141.7653

Fitted Versus Actual values

The actual and fitted values are plotted in the *figure 6.14.* The red line shows the predicted (fitted) values and blue line shows the actual values.

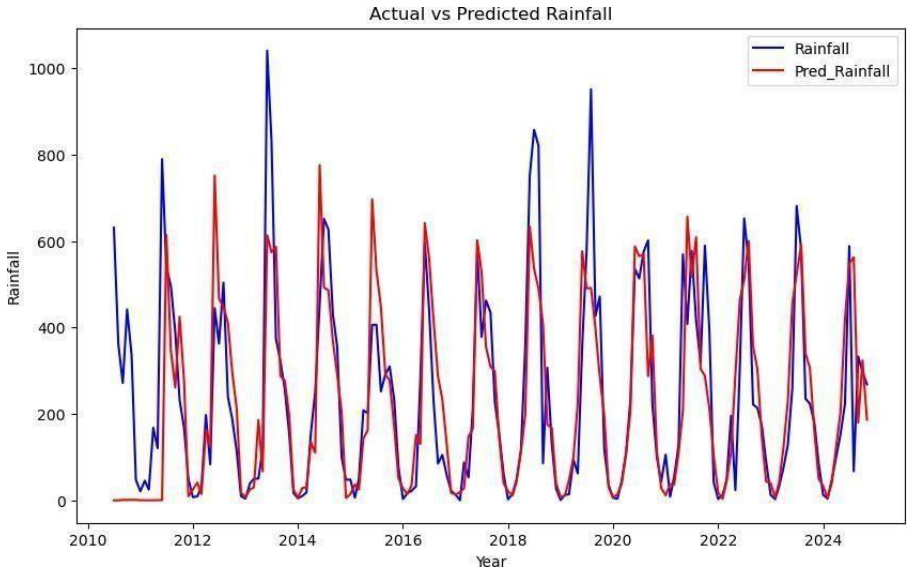


figure 6.14 Fitted versus actual values using the SARIMA model.

Forecasting of rainfall using the SARIMA model

The model can be used for forecasting. The rainfall from January 2025 to December 2026 is forecasted using the SARIMA model fitted. The forecasted values are computed and plotted (*Table 6.8 and figure 6.15*). The green dotted line indicates the forecasted rainfall.

Table 6.8 SARIMA model forecast

Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)	Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)
Jan 2025	44.9081	0	290.73	Jan 2026	24.1579	0	290.41
Feb 2025	11.8728	0	268.05	Feb 2026	8.4641	0	274.49
Mar 2025	42.2341	0	302.76	Mar 2026	42.5078	0	309.19
Apr 2025	115.4348	0	374.71	Apr 2026	114.5235	0	381.17
May 2025	195.1394	0	452.46	May 2026	192.6009	0	458.96
Jun 2025	386.3188	101.35	636.84	Jun 2026	376.5499	91.4913	643.53
Jul 2025	605.3150	329.08	864.61	Jul 2026	604.3212	318.9825	871.07
Aug 2025	475.5263	169.79	705.31	Aug 2026	451.1083	160.2971	712.39
Sep 2025	309.4874	35.24	570.76	Sep 2026	310.8998	25.0732	577.16
Oct 2025	300.6518	25.31	560.83	Oct 2026	300.4334	25.2879	567.28
Nov 2025	200.4659	0	465.49	Nov 2026	204.5414	0	471.82
Dec 2025	77.3011	0	342.90	Dec 2026	82.0205	0	349.21

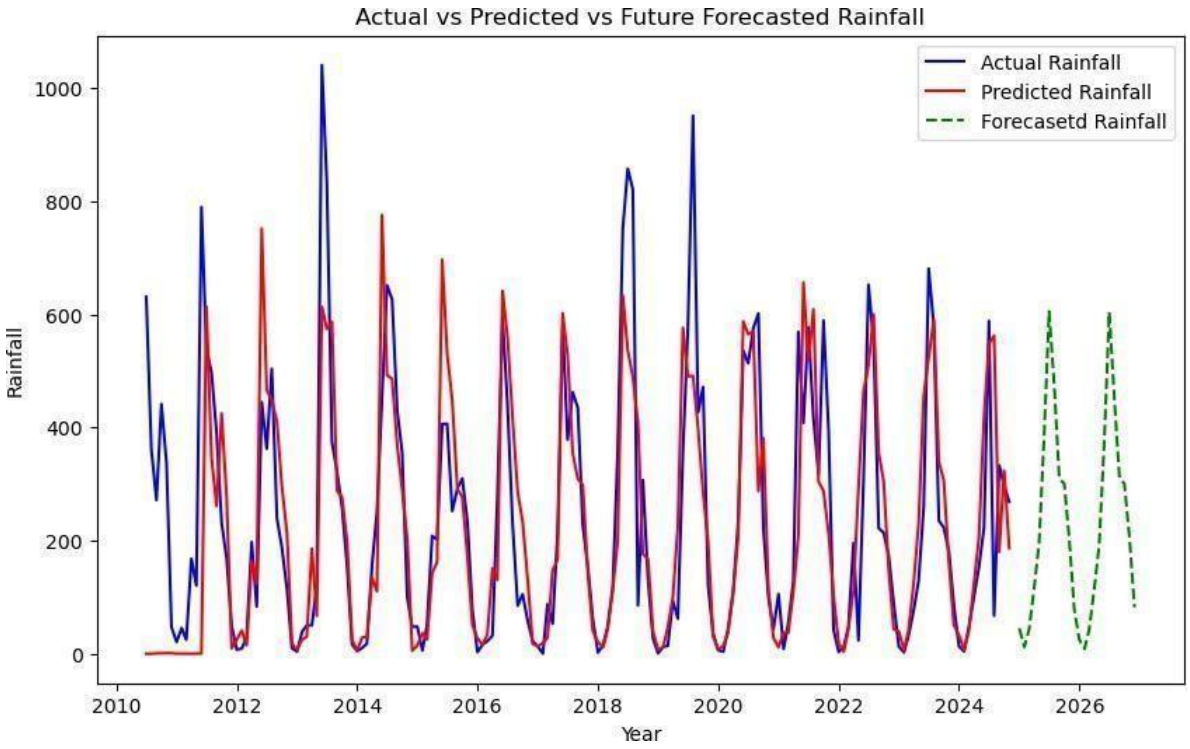


figure 6.15 A plot of forecasted values using SARIMA model

6.3 MODELING AND FORECASTING OF RAINFALL USING HOLT-WINTERS METHOD

Here Holt Winters Forecasting procedure is used to forecast rainfall. Now *table 6.9* shows the parameter estimates of the model.

Table 6.9 Holt Winter's model parameters

Parameter	Parameter Estimates
Alpha (Level)	0.0401474
Gamma (Trend)	0.0393733
Delta (Season)	0.0677946

Fitted Holt Winters Additive model is :

$$L_t = 0.0401474(y_t - I_{t-s}) + (1 - 0.0401474)(L_{t-1} + T_{t-1})$$

$$T_t = 0.0393733\,(L_t - L_{t-1}) + (1 - 0.0393733)\,T_{t-1}$$

$$I_t = 0.0677946\,(y_t - L_t) + (1 - 0.0677946)\,I_{t-s}$$

Diagnostic Checking

Diagnostic checking is a crucial step to ensure the reliability, effectiveness, and validity of statistical models. It helps to understand how precise the model is and to improve the prediction accuracy.

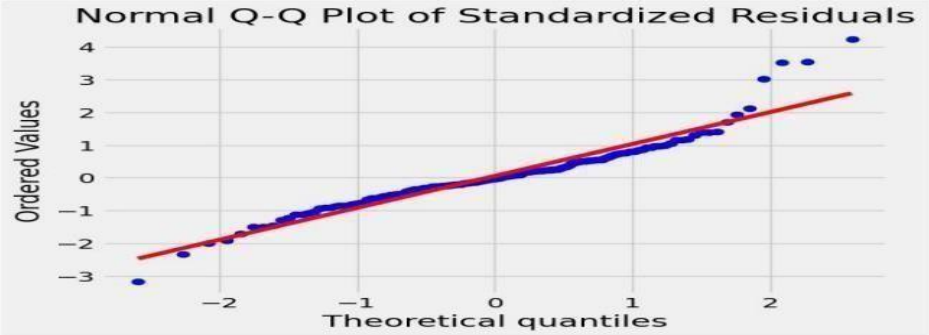


figure 6.16 Q-Q plot of residuals

The *figure 6.16* depicts the Q-Q plot , it is clear that the most of the residual values lie on the straight line, which indicates that residuals are approximately normally distributed.

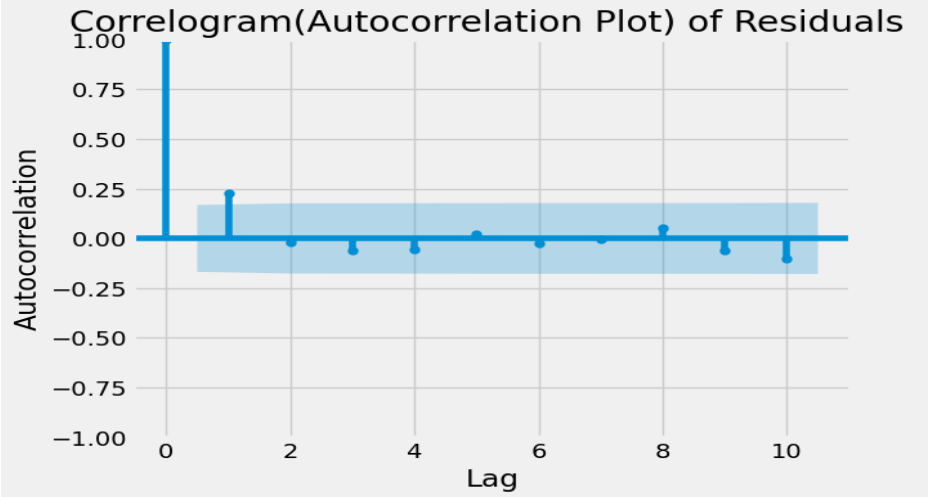


figure 6.17 Correlogram of residuals

The *figure 6.17* shows the correlogram , from the examination of correlogram it can be evidently seen that all the lags die to zero means that there is no significant autocorrelation present in the data.

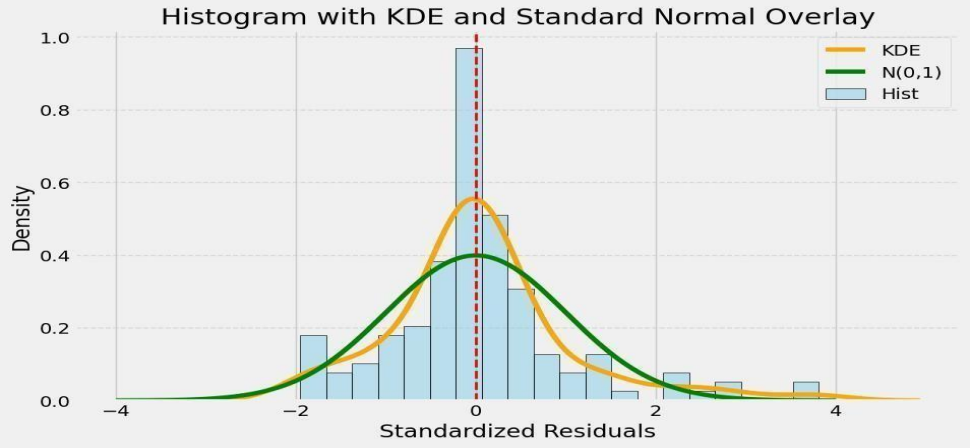


figure 6.18 Histogram of residuals

The *figure 6.18* shows the histogram of standardized residuals. Since the histogram matches the normal distribution curve, the residuals are approximately normal.

Table 6.10 Ljung box test result

Test statistic	p_value
10.759955	0.376523

H_0 : There is significant autocorrelation in the residuals

H_1 : There is no significant autocorrelation in the residuals

Table 6.10 indicates the Ljung box test result. Since the p-value (0.107) is greater than 0.05, we fail to reject the null hypothesis. This means there is no significant autocorrelation in the residuals.

Thus from the diagnostic checking it is evident that the fitted Holt-Winters model is statistically adequate. So, the model can be used to forecast the monthly rainfall of Kerala.

In-sample Forecasting

Now the fitted time series model is used to do In-sample forecasting. In-sample forecasting is done for the last year in the dataset that is from January 2024 to December 2024.

Table 6.11 In-sample forecasting using Holt winter's model

Months	Actual value(mm)	Predicted value(mm)
Jan	13.0	90.7842
Feb	3.9	86.5774
Mar	46.8	112.3429
Apr	100.2	193.0212
May	152.7	303.0725
Jun	223.0	652.8139
Jul	588.7	642.3624
Aug	67.3	590.6612
Sep	333.1	406.5909
Oct	297.0	371.9924
Nov	268.6	229.7777
Dec	156.2	123.1724

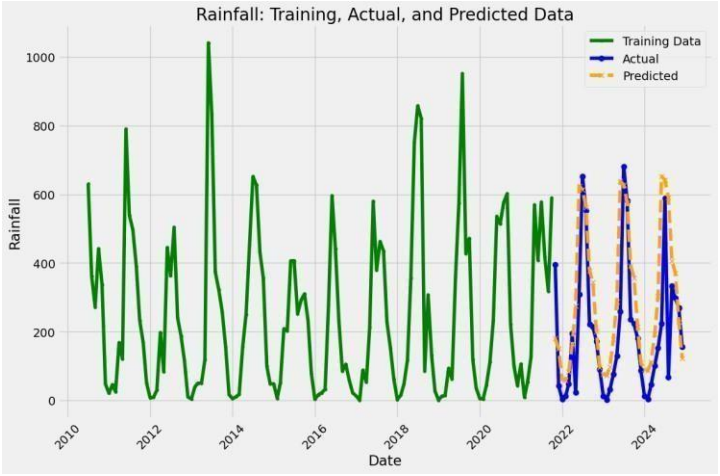


figure 6.19 Fitted versus actual values using the Holt winter's exponential smoothing technique

From *figure 6.19* it given that green lines are the training data, blue line is the actual testing data and orange line indicated testing predicted data.

Forecasting of rainfall using Holt-Winters model

Rainfall from January 2025 to December 2026 is forecasted using the model fitted.

Table 6.12 Holt winter's model forecast

Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)	Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)
Jan 2025	104.8621	0	241.71	Jan 2026	114.7331	0	279.02
Feb 2025	100.6553	0	228.45	Feb 2026	140.4981	0	267.54
Mar 2025	126.4208	0	202.89	Mar 2026	221.1770	0	388.93
Apr 2025	207.0991	0	344.17	Apr 2026	331.2282	0	393.04
May 2025	317.1503	0	473.64	May 2026	680.9696	0	788.32
Jun 2025	666.8918	101.35	687.84	Jun 2026	670.5182	91.4913	844.34
Jul 2025	604.7391	329.08	864.61	Jul 2026	618.8170	318.9825	817.84
Aug 2025	420.6687	169.79	815.19	Aug 2026	434.7466	160.2971	512.39
Sep 2025	386.0702	35.24	560.93	Sep 2026	400.1481	25.0732	542.22
Oct 2025	243.8556	25.31	593.38	Oct 2026	257.9335	25.2879	367.82
Nov 2025	137.2503	0	465.49	Nov 2026	151.3281	0	271.18
Dec 2025	118.9400	0	423.80	Dec 2026	133.0179	0	249.01

Table 6.12 is the forecasted values and its LCL and UCL of rainfall for January 2025 to December 2026

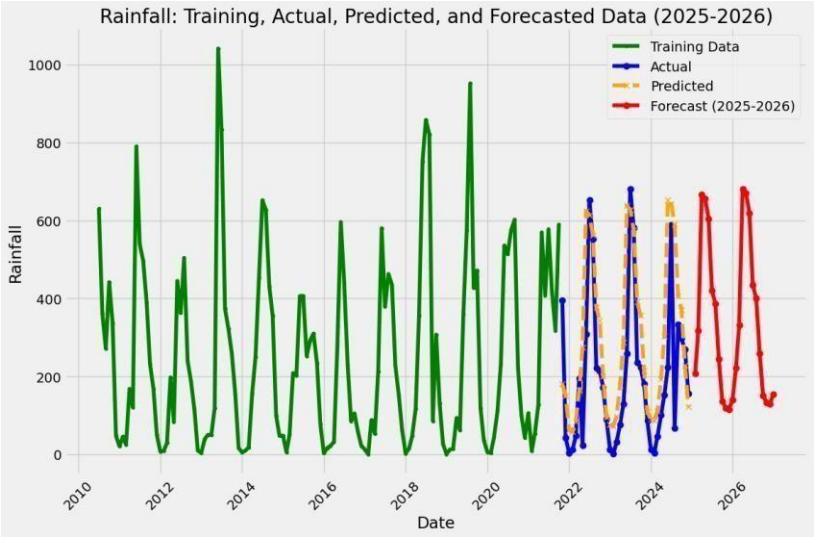


figure 6.20 A plot of forecasted values using Holt winter's exponential smoothing technique

The *figure 6.20* is the plot of forecasted values using Holt Winter's model.

6.4 MODELING AND FORECASTING OF RAINFALL USING PROPHET MODEL

Here Prophet model is used to forecast the rainfall from January 2025 to December 2026.

Diagnostic checking

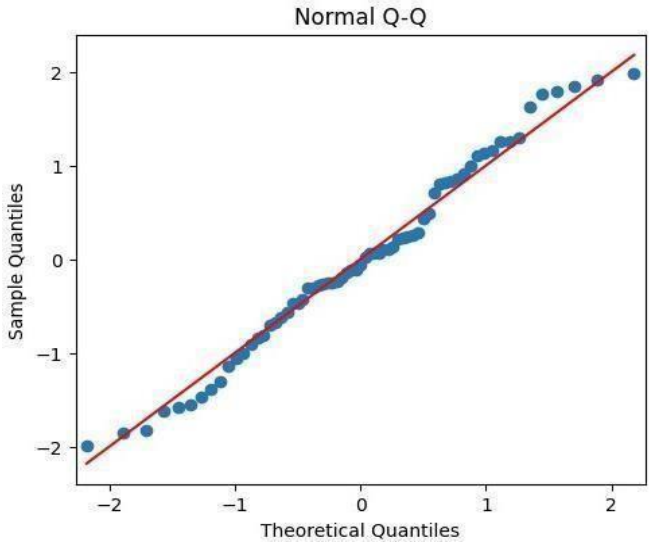


figure 6.21 Q-Q plot of residuals

It is clear from the *figure 6.21* most ofthe residual values lie onthe straight line, which indicates the residuals are approximately normally distributed.

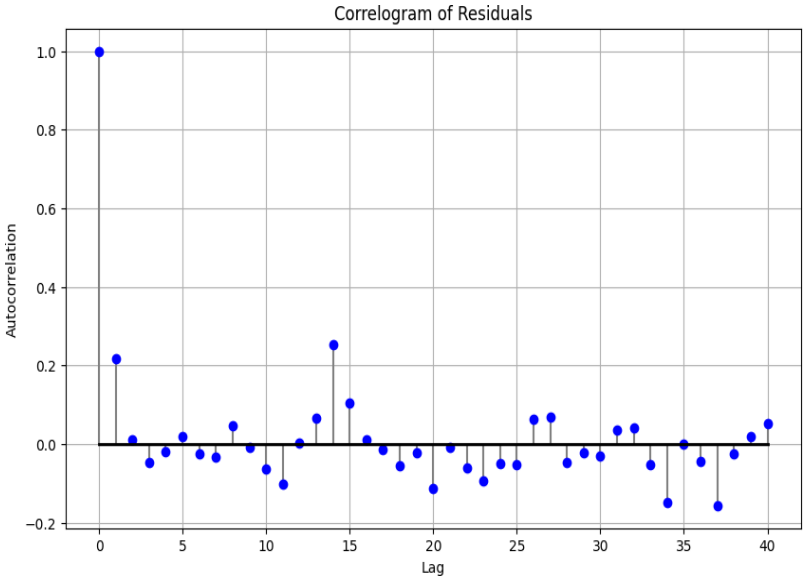


figure 6.22 Correlogram of residuals

From the examination of correlogram in *figure 6.22* can be evidently seen that all the lags die to zero means that there is no significant autocorrelation present in the data.

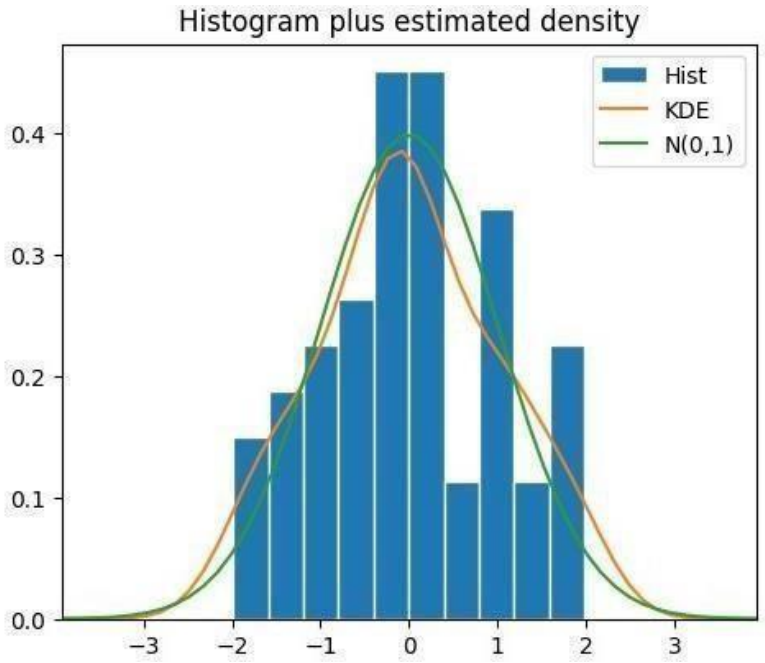


figure 6.23 Histogram of residuals

In the *figure 6.23*, the histogram matches the normal distribution curve (green line), the residuals are approximately normal.

Table 6.13 Ljung box test result

Test statistic	p_value
10.448435	0.402069

H_0 : There is significant autocorrelation in the residuals

H_1 : There is no significant autocorrelation in the residuals

Table 6.13 indicates the Ljung box test result. Since the p-value (0.107) is greater than 0.05, we fail to reject the null hypothesis. This means there is no significant autocorrelation in the residuals.

Thus fromthe diagnostic checking it is evident that the fitted Prophet model is statistically adequate. So, the model can be used to forecast the monthly rainfall of Kerala.

In-sample Forecasting

Now the fitted time series model is used to do In-sample forecasting. In-sample forecasting is done for the last year in the dataset that is from January 2024 to December 2024.

Table 6.14 In-sample forecast using Prophet model

Months	Actual value(mm)	Predicted value(mm)
Jan	13.0	10.5111
Feb	3.9	10.7698
Mar	46.8	35.8474
Apr	100.2	108.9124
May	152.7	204.9871
Jun	223.0	514.3679
Jul	588.7	568.2219
Aug	67.3	458.1893
Sep	333.1	317.6038
Oct	297.0	285.2710
Nov	268.6	173.2909
Dec	156.2	48.2215

Forecasting of rainfall using Prophet model

Rainfall from January 2025 to December 2026 is forecasted using the model fitted.

Table 6.15 Prophet model forecast

Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)	Months	Forecasted Values(mm)	LCL (mm)	UCL (mm)
Jan 2025	12.9929	0	182.186	Jan 2026	15.2341	0	179.800
Feb 2025	29.0089	0	188.206	Feb 2026	26.7806	0	197.684
Mar 2025	125.7613	0	290.795	Mar 2026	127.9675	0	277.950
Apr 2025	164.7044	0	322.764	Apr 2026	155.7066	0	316.843
May 2025	481.5497	11.8124	637.446	May 2026	474.5642	311.296	644.044

Jun 2025	580.3397	322.800	738.761	Jun 2026	582.5591	423.187	743.535
Jul 2025	521.7161	427.299	671.773	Jul 2026	533.1738	373.395	705.078
Aug 2025	277.7921	259.505	437.007	Aug 2026	268.6594	113.836	437.665
Sep 2025	297.4294	116.657	459.532	Sep 2026	299.0927	135.202	450.107
Oct 2025	167.6165	137.067	334.067	Oct 2026	165.6124	4.857	324.094
Nov 2025	43.7777	9.649	196.812	Nov 2026	42.2809	0	199.794
Dec 2025	8.7047	0	172.672	Dec 2026	22.45	0	61.987

Table 6.15 is the forecasted values and its LCL and UCL of rainfall for January 2025 to December 2026

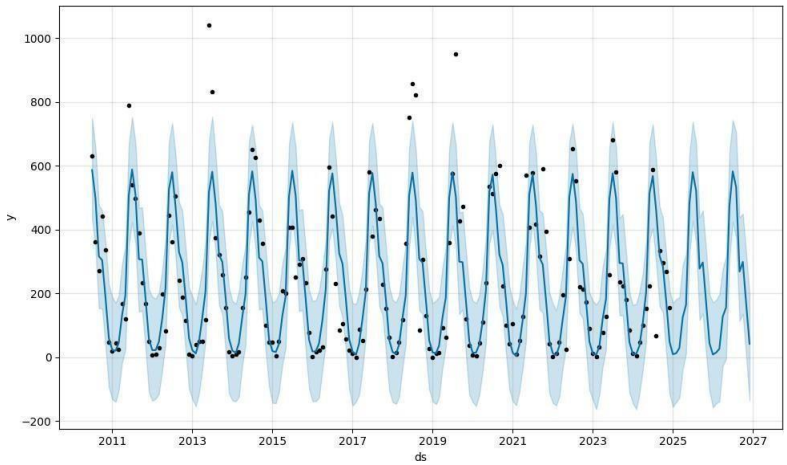


figure 6.24 A plot of forecasted values using Prophet model

The *figure 6.24* is the plot of forecasted values using Prophet model.

6.5 COMPARISOMN BETWEEN SARIMA, HOLT-WINTERS AND PROPHET MODEL

To determine the best model, the performance metrices should be compared and considered. Here the Mean Absolute Error (MAE) and Root Mean Square Error were provided for the three models.

Table 6.16

	SARIMA Model	**Holt Winter's Model**	**Prophet Model**
MAE	**103.16**	**113.23207798875812**	**85.54657338176848**
RMSE	**161.4308216265108**	**164.12057903418**	**147.8211734743146**

From *table 6.16* it is evident that Prophet model has the lower MAE and RMSE value compared to SARIMA and Holt Winter's model. That is Prophet model appears to be perform better than the other model.

CONCLUSION

Three time series models were used for the forecasting of monthly rainfall data of Kerala for the year January 2025 to December 2026.The historical rainfall data of Kerala from January 2010 to December 2024 were analysed and forecasted using SARIMA, Holt winter's exponential smoothing technique and Prophet model. Prophet model was concluded as the best model with a low RMSE of 147.821 and MAE of 85.546. SARIMA model had a RMSE of 161.430 and MAE of 103.16 and Holt winter's exponential smoothing had a RMSE of 164.120 and MAE of 113.232. All the three models offer viable forecasting solutions, the Prophet model is recommended for its superior accuracy and flexibility, particularly for datasets with seasonality or missing values. However, if computational efficiency is critical, Holt-Winter's Exponential Smoothing could be considered for simpler datasets. For data with clear statistical properties and stationarity, SARIMA remains a robust option but may require significant effort in tuning parameters.

Even though prophet is the best fitted model, we chose SARIMA as our best model, due to the prophet's black box approach. SARIMA have strong theoretical foundation and it also provides explicit control over seasonal and non-seasonal differencing, AR, and MA components, allowing for greater transparency in model behavior and the parameters can be explicitly analysed. Unlike Prophet is a black-box approach due to its automated detection of seasonality, trends, and holiday effects, While Prophet simplifies forecasting by automating many aspects, it may not always align well with domain-specific requirements, making SARIMA a preferred choice when interpretability and control are essential.

From the analysis of the 174 months of past rainfall data and the new forecasted data it is evident that there is a change in the rainy months. June was the most raining month. However, in the current dataset, the trend has shifted, July emerges as the most rainy month followed by August. This shift is visible from 2018 and it also indicate a changing climate pattern or other external factors influencing the rainfall distribution over the years.

In conclusion, this study uncovers past rainfall patterns in Kerala and hints at exciting possibilities for more research. By using advanced models, gained insights into historical data, underlining the importance of ongoing climate and environmental studies. This research not only tells about the past but also opens doors for future exploration in the realm of climate science.

REFERENCES

Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Computers & Electrical Engineering, 70*, 66–73.

Jagadeesh, P., & Anupama, C. (2014). Statistical and trend analyses of rainfall: A case study of Bharathapuzha River basin, Kerala, India. *ISH Journal of Hydraulic Engineering*. Taylor & Francis.

Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. *Disaster Advances, 11*(11), 27–33.

Krishnakumar, K. N., Rao, G. S. L. H. V. P., & Gopakumar, C. S. (2009). Rainfall trends in twentieth century over Kerala, India. *Atmospheric Environment, 43*(11), 1940–1944.

Nair, A., Joseph, K. A., & Nair, K. S. (2014). Spatio-temporal analysis of rainfall trends over a maritime state (Kerala) of India during the last 100 years. *Atmospheric Environment, 88*, 123–132.

Raj, P. P. N., & Azeez, P. A. (2012). Trend analysis of rainfall in Bharathapuzha River basin, Kerala, India. *International Journal of Climatology, 32*(4), 533–539.

Sulasikin, A., Nugraha, Y., Kanggrawan, J. I., & Suherman, A. L. (2021, August). Monthly rainfall prediction using the Facebook Prophet model for flood mitigation in Central Jakarta. In *2021 International Conference on ICT for Smart Society (ICISS)* (pp. 1–5). IEEE

Surendran, D. E., Sridhar, L., Kumari, A., Sreejith, O. P., & Pai, D. S. (2020). Impact of climate change on the heavy rainfall events during June to September over Kerala (1901–2019). *Current Science, 119*(5), 829–836

Varghese, L. R., & Vanitha, K. (2020). Atime-series based prediction analysis of rainfall detection. *2020 International Conference on Inventive Computation Technologies (ICICT)*, 26-28 February 2020, Coimbatore, India. IEEE.