

Project Report
On
TIME SERIES ANALYSIS OF RAINFALL IN ERNAKULAM
Submitted
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE
in
APPLIED STATISTICS AND DATA ANALYTICS
by
SANDHRAMOL S
(Reg No. SM22AS018)
(2022-2024)

Under the Supervision of
ANU MARY JOHN



DEPARTMENT OF MATHEMATICS AND STATISTICS
ST. TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM, KOCHI – 682011
APRIL 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

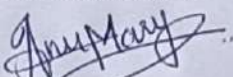


CERTIFICATE

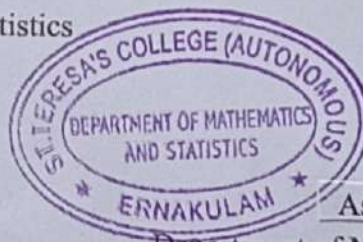
This is to certify that the dissertation entitled, **TIME SERIES ANALYSIS OF RAINFALL IN ERNAKULAM** is a bonafide record of the work done by **SANDHRAMOL S** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's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

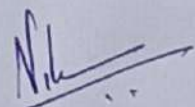
Date:

Place : Ernakulam


Anu Mary John

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

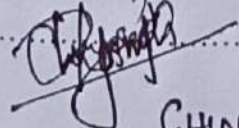


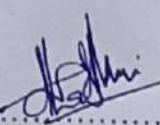


Nisha Oommen

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

External Examiners

1. 
CHINU JOSEPH
29/4/2024.

2. 
LAKSHMI SURESH
29/04/2024

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of ANU MARY JOHN, 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: 29/04/24

SANDHRAMOL S

SM22AS018

ACKNOWLEDGEMENTS

I take this opportunity to thank everyone who has encouraged and supported me to carry out this project.

I am very grateful to my project guide Ms. Anu Mary John for her immense help during the period of work.

In addition, I acknowledge with thanks to the Department for all the valuable support and guidance that has significantly contributed to the successful completion of this project.

I would also like to thank the HOD for her valuable suggestions and critical examinations of the project.

Ernakulam:

Date: 29/04/24

SANDHRAMOL S


SM22AS018

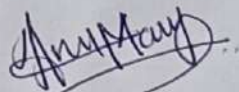
ABSTRACT


In this study, monthly rainfall data of Ernakulam district from January 2010 to December 2022 are analyzed and used for forecasting future rainfall value. The data was collected from Indian Meteorological Department (IMD). Two time series model were deployed to forecast future monthly rainfall data of Ernakulam. SARIMA and Holts-Winter's Exponential Smoothing were used. Comparison of these models were done using matrices like RMSE and MSE. SARIMA model was selected as the best model with lowest RMSE value.

**ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM****Certificate of Plagiarism Check for Dissertation**

Author Name	SANDHRAMOL S
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. ANU MARY JOHN
Department	Post Graduate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	TIME SERIES ANALYSIS OF RAINFALL IN ERNAKULAM
Similarity	0% AI 12%
Paper ID	1669608
Submission Date	2024-04-20 12:17:05


Signature of Student


Signature of Guide


Checked By
College Librarian



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

TABLE OF CONTENTS

S.NO	CONTENTS	PAGE NO
1	INDRODUCTION	7
2	LITERATURE REVIEW	9
3	MATERIALS AND METHADODOLOGY	14
4	EXPLORATORY DATA ANALYSIS	17
5	TIME SERIES	19
6	RESULTS AND DISCUSSIONS	27
7	CONCLUSION	52
8	REFERENCES	53

CHAPTER-1

INTRODUCTION

‘Time series Analysis of Rainfall in Ernakulam ’ was a study carried out to establish the trends and patterns of rainfall in Ernakulam as well as predict future rainfall. This project will assist in understanding rainfall patterns changing in the district. The raw data were obtained from the official website of India Meteorological Department, monthly rainfall data of Ernakulam district from 2010-2022 is used for the study.

Rainfall is important to all living organisms on the Earth. It plays crucial role on ecosystem functions, agriculture production etc. Rainfall is the primary source of water, it maintains the water cycle and give fresh water. Water management, agricultural planning etc., can benefit by studying patterns and trends in rainfall. Depending on their latitude, altitude and atmospheric conditions there are differences in rain distribution and intensity across regions.

In Kerala, a state in India that is famous for its extremely green landscapes and rich biodiversity, rainfall is not an ordinary natural phenomenon but rather a way of life. The Southwest Monsoon season usually starts in June and ends in September. It comes with heavy rain which maintains the state’s rich biodiversity and supports an active agricultural sector. However, due to Kerala’s unique geographical features such as its mountainous topography and proximity to the Arabian Sea, rains can vary from one area to another.

Ernakulam is a region within the state of Kerala that have different rainfall patterns. Ernakulam receives relatively high amounts of rainfall because it lies along the coastal belt of Kerala near to the sea which brings orographic uplift on Western Ghats’ slopes. The precipitation in Ernakulam enhances soil fertility thus making it favorable for various types of agricultural practices like horticulture and paddy cultivation.

For various stakeholders including farmers, urban planners and environmental

conservationists, understanding how rainfall dynamics work in Ernakulam is crucial. These regions may be delineated by precipitation patterns' relations with land use practices, socio-economic development if historical rainfall data is analyzed together with future trends are predicted. Such insights are also valuable for offering adaptive strategies against climate change-induced risks like drought or flooding which are major threats to resilience and sustainability of communities in Kerala.

In this project we conduct an extensive analysis of rainfall data for Ernakulam with a view of unraveling the intricacies underlying the livelihoods and local ecosystem processes related to precipitation regimes. Through rigorous statistical analysis and predictive modeling, we seek not only to explain the past trends, but We also want to have insights about how the precipitation behaves in a dynamically and what this means for the local societies and ecosystems.

1.1. OBJECTIVES

- 1)To perform EDA (Exploratory Data Analysis) to find pattern and trend of rainfall.
- 2) To model and forecast rainfall in Ernakulam using Seasonal ARIMA (Auto Regressive Integrated Moving Average).
- 3)To model and forecast rainfall in Ernakulam using HOLT-WINTER'S EXPONENTIAL SMOOTHENING TECHNIQUE.
- 4)To compare the forecast of Seasonal ARIMA and HOLT-WINTER'S EXPONENTIAL SMOOTHENING TECHNIQUE.

CHAPTER-2

LITERATURE REVIEW

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

1. A study by Kamath and Kamath (2018) on the basis a rainfall dataset collected from Knoema, an accessible web-based open data platform. The aim of this research was to test the accuracy of various time series model. Monthly rainfall data of Idukki district from January 2006 to December 2016 is used for the study. ARIMA, Artificial Neural Network (ANN), Exponential Smoothing State Space (ETS) are the models used in this study. At the end of the study, ARIMA modeling performed better than other models. Root Mean Squared Error (RMSE) and model fit were used as the evaluation metrics to determine how good the fitted models are.

2. Dash et al. (2018) using data of summer monsoon (June-September) and post-monsoon (October-December) rainfall in Kerala, India, from 2011 to 2016 conducted a study. Three artificial intelligence approaches were used: K-nearest neighbour (KNN), Artificial Neural Network (ANN) and extreme Learning Machine (ELM). In comparison, the ELM technique exhibited superior performance, achieving minimal mean absolute percentage error scores for both summer monsoon (3.075) and post-monsoon (3.149) compared to KNN and ANN. The study also highlighted that prediction accuracy was highly influenced by the number of hidden nodes in the hidden layer, with the ELM architecture (8-15-1) providing more accurate results. Consequently, the study revealed that the proposed artificial intelligence

approaches have significant potential for predicting both summer monsoon and post-monsoon rainfall in Kerala, India, with minimal prediction error scores.

3. Jayasree et al. conducted a study that addressed the importance of rainfall forecasting in economy and disaster management. A novel hybrid model that combines empirical model decomposition (EMD) and Random Forest (RF) to improve the accuracy of rainfall prediction was proposed. A dataset of annual rainfall of Kerala from 1871 to 2020 is used for this study. From comparing the hybrid RF-IMF model to traditional models like RF regression and ARMA model using Mean absolute Error (MAE), MSE, MAPE, RMSE and R-squared. The study got the clear evidence that the RF-IMS model has more efficiency over both the RF model and ARMA model for predicting rainfall.

4. Joshi and Tyagi (2021) emphasized the significance of rainfall as a crucial stochastic phenomenon affecting the Indian agricultural sector and contributing to the country's economic growth. However, predicting rainfall has become increasingly challenging due to climate changes linked to global warming. In their article, the researchers applied seasonal Naive, triple exponential smoothing, and seasonal ARIMA time series models to achieve accurate and timely rainfall predictions. They compared the accuracy of forecasts from these models using various scale-dependent error forecast methods and residual analysis. The empirical analysis utilized monthly rainfall data recorded from 2009 to 2018 in Bengaluru, Karnataka, India, to suggest the best-fitted time series model for monthly rainfall prediction in the region. The results indicated that the seasonal autoregressive moving average model (ARIMA (0,0,2) (1,1,1)₁₂) provided the most accurate rainfall forecasts for Bengaluru, surpassing the performance of other time series models.

5. Poornima and Pushpalatha (2019) address the critical concern of rainfall prediction in meteorology. Various techniques have been previously proposed for rainfall prediction, including statistical analysis, machine learning, and deep learning methods. Accurate time series data prediction in meteorology can significantly aid organizations responsible for

disaster prevention in their decision-making processes. The researchers present a novel approach called Intensified Long Short-Term Memory (Intensified LSTM) based Recurrent Neural Network (RNN) for rainfall prediction. The neural network is trained and tested using a standard data set of rainfall, producing predicted rainfall attributes. The model's performance and efficiency are evaluated based on parameters like Root Mean Square Error (RMSE), accuracy, number of epochs, loss, and learning rate. To showcase the improvement in rainfall prediction ability, the obtained results are compared with other models such as Holt-Winters, Extreme Learning Machine (ELM), Autoregressive Integrated Moving Average (ARIMA), Recurrent Neural Network, and Long Short-Term Memory models.

6. Ganapathy et al. (2021) in this study highlights the important role of rainfall in the agricultural sector, which affects the Indian economy. Rainfall is a blessing for farmers, but excessive or insufficient rainfall can adversely affect their hard work. The researchers forecast the rainfall data set of the Vellore region in Tamil Nadu, India, for the years 2021 and 2022 using various machine learning algorithms. Feature engineering is employed to eliminate auto-correlation in the data, enabling operations on time-series data that are typically performed on regular regression data. The forecasting techniques used include Autoregressive Integrated Moving Average (ARIMA) and exponential smoothing, with subsequent processing using Long Short-Term Memory (LSTM). Regression techniques are also applied to manipulate the data set. The study is benchmarked against various evaluation metrics, with Boost Regression technique delivering the best performance on the test data set. Notably, this work provides daily rainfall forecasts for the specified years in the Vellore region. In the future, this approach could be extended to predict rainfall over larger regions based on historical time-series data, offering valuable insights for farmers and the general public to plan and take precautionary measures.

7. Mitihya et al. (2020) in this study highlight the significant dependence of Indian agriculture on rainfall, as it influences both agricultural production and commodity prices.

In study monthly rainfall of India is forecasted using time series model. Linear and non linear models are used for this study . The study reveals that the non-linear model, specifically the Artificial Neural Network (ANN), outperforms linear models like simple seasonal exponential smoothing and Seasonal Auto-Regressive Integrated Moving Average in terms of diagnostic checking parameters (MAE, MSE, and RMSE). The selection of the ANN model aids in identifying the appropriate cropping pattern, enabling better planning and management for Indian agriculture.

8. Gowri et al. (2022) in this study emphasize the crucial importance of accurate rainfall prediction due to its potential to cause catastrophic events. Accurate forecasts enable individuals to take appropriate precautions and plan ahead effectively, especially in agriculture, which is essential for ensuring survival. However, predicting rainfall has been a significant challenge in recent years. To address this, the researcher employs Thanjavur Station rainfall data spanning 17 years from 2000 to 2016 to study the accuracy of rainfall forecasting. Three prediction models, ARIMA (Auto-Regression Integrated with Moving Average Model), ETS (Error Trend Seasonality Model), and Holt-Winters (HW), are compared using the R package to identify the most accurate forecasting model. The findings reveal that the HW and ETS models outperform the ARIMA model based on performance criteria such as Akaike Information Criteria (AIC) and Root Mean Square Error (RMSE). The study highlights the potential of machine learning techniques to improve rainfall prediction accuracy, contributing to disaster prevention and better crop preservation strategies.

9. Mortey (2011), in his study he addressed the growing concerns regarding climatic changes and the need to implement measures to mitigate drastic climate shifts. The aim of this study is to analyse and understand the trend of rainfall in four regions of Ghana. Here four models are used Linear Trend with Seasonal terms, Seasonal Exponential Smoothing, ARIMA, Single Exponential Smoothing, Linear (Holt) Exponential Smoothing. Among these models, the study identified Linear Trend with Seasonal Terms as the best

model for predicting rainfall in the regions under study. The selection criteria for this model included Means Square Error (MSE) and R-Square. Additionally, the study concluded that the rainfall levels in the four regions were projected to rise, at least for the year 2011. This research sheds light on the rainfall patterns in Ghana and emphasizes the importance of effective measures to address climate changes.

10. Jain and Kumar (2012) examined trends in rainfall, rainy days, and temperature across India. They employed Sen's non-parametric estimator along with the Mann-Kendall test to evaluate statistical significance. The findings revealed variations in trends across different spatial units, with only a few exhibiting statistical significance. In terms of annual rainfall, most basins showed decreasing trends, while a smaller number displayed increasing trends. Regarding temperature, mean maximum temperatures demonstrated a rising trend in numerous regions, while mean minimum temperatures showed both rising and falling trends. Interestingly, urban areas had a noticeable influence on the data, acting as heat islands. The study underscored the importance of establishing a robust network of baseline stations for more accurate and reliable climatic studies.

CHAPTER-3

MATERIALS AND METHODOLOGY

3.1 DATA COLLECTION

The dataset used for this study contains the monthly rainfall data in millimeter(mm) of Ernakulam from January 2010 to December 2022. Data used in this study is collected from the official website of Indian Meteorological Department (IMD).

3.2 METHODOLOGY

The first important step in the analysis was to study the data in detail. The main purpose of the study was to understand the pattern and trend of rainfall in Ernakulam district and to forecast future rainfall using time series models. As the first step EDA was executed to gain insight into the characteristics of the dataset. Then the model was forecasted using the SARIMA model. Then the model was forecasted using Holt-Winter's Exponential Smoothing Technique. Then the comparison of the two time series models is done by calculating the MSE and RMSE values of both models.

3.3 TOOLS FOR ANALYSIS AND FORECASTING

EDA (Exploratory Data Analysis)

Seasonal ARIMA (Auto regressive Integrated Moving Average)

Holt-Winter's Exponential Smoothing Technique

3.4 TOOLS FOR COMPARISON

Mean Squared Error (MSE)

The formula for MSE is:

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n}$$

Where,

y_i is the i th observed value.

\hat{y}_i is the corresponding predicted value.

n is the number of observations.

Root Mean Squared Error (RMSE)

The formula for RMSE is:

$$RSME = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N - P}}$$

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.

P is the number of parameter estimates, including the constant.

3.5 PYTHON

In this study, Python a popular programming language used to create websites, software's etc is used. Python is used to do EDA to find hidden patterns of the data set and to forecast using SARIMA and Holt-Winter's Exponential Smoothing Technique.

CHAPTER- 4

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) 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.

Descriptive Statistics:

Descriptive statistics describe the variability, distribution and central tendency of the data set by summarizing the mean, median, mode, standard deviation and quartiles. Measures like standard deviation and quartiles are used to identify outliers and extreme values of the dataset. Overall descriptive statistics is a complete summary of data's main characteristics, which will help to understand data more and do further analysis accurately.

Autocorrelation analysis:

Autocorrelation analysis is also known as Autocorrelation Function (ACF) Analysis, is used to find the correlation between observation at different time lags within a dataset. It helps to understand the dependencies of the dataset. It can also be used to identify seasonality in the dataset. It is also used as a diagnostic tool to check model accuracy.

Seasonal Decomposition:

Using seasonal decomposition technique time series data is decomposed in to its componets trend, seasonal and residual. Using this we can understand whether the dataset shows seasonality and also understand the trend of the data. Understanding the trend of the data and seasonality will helps to choose correct model and increase accuracy of the forecasting.

Trend: The long term decrease or increase data referred to as trend.

Seasonal: In time series data seasonality means a regular or predictable pattern that repeats over a short period. The interval period may be months, weeks or days. Rainfall is an example for this .

Residual: The difference between the observed value and the predicted value is called residuals. It helps to understand how well the data is fitted to the model.

Overall, EDA is the first and most important step in data analysis to understand the data and find the missing or null values. It also give an insight about which model are need for the dataset to forecast accurate results.

CHAPTER-5

TIME SERIES

5.1 TIME SERIES ANALYSIS

An approach of examining a succession of data points gathered over a period of time is called time series analysis. In time series analysis, data points are recorded at regular intervals throughout a predetermined time period, rather than randomly. But doing this kind of research involves more than just gathering data over time. The ability to analyse how factors change over time is what distinguishes time series data from other types of data. Put differently, time is an important component since it seems to affect how the information changes as it concentrates and the final outcome occurs. It provides an additional data source and a set of guidelines between the information.

In order to guarantee consistency and dependability, time series analysis usually needs a large number of data points. A large data collection ensures that your analysis can sift through erratic data and that your sample size is representative.

The underlying patterns and structure of a data is captured by the components of time series.

The main components of time series are :

- 1) Trend
- 2) Seasonality
- 3) Cyclic variations
- 4) Irregular variations

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 with out trend and seasonality.

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.

Autocorrelation 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.

Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function (PACF) is used to assess the direct link between two observations in a time series while taking additional data' effect into consideration. PACF is used to find the MA parameter of SARIMA and ARIMA model.

Augmented Dickey Fuller Test

Augmented Dickey Fuller Test (ADF) is test used to check the stationarity of a time series data. Only stationary data can be further processed for analysis.

5.2 SARIMA (Seasonal AutoRegressive Moving Average)

SARIMA (Seasonal Auto-Regressive Integrated Moving Average) is an extension of the ARIMA (Autoregressive Integrated Moving Average) model that incorporates seasonality in addition to the non-seasonal components. ARIMA models are widely used for time series analysis and forecasting, while SARIMA models are specifically designed to handle data with seasonal patterns. It is represented as,

$$\text{SARIMA } (p, d, q) \times (P, D, Q) m$$

where,

p: Trend autoregression order.

d: Trend difference order

q: Trend moving average order.

P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

m: The number of time steps for a single seasonal period

For the sections of the model that are seasonal, we use uppercase notation, and for the non-seasonal components, we use lowercase notation.

The following three elements are added to SARIMA models' seasonal component:

- o Seasonal Autoregressive (P): This component illustrates the correlation, particularly at seasonal delays, between the series' historical values and current value.
- o Seasonal Integrated (D): This component accounts for the differencing necessary to eliminate seasonality from the series, just like the non-seasonal differencing does.
- o Seasonal Moving Average (Q): This part simulates the relationship that exists between the present value and the seasonal lags of the residual errors of the prior forecasts.

In the above equation, p , d , and q are represented as follows: Q is the order of the seasonal MA, s is the duration of the season (periodicity), D is the number of seasonal differencing, and P is the order of the seasonal AR model. Furthermore, the ω_t and B stand for the white noise value at time t and the reverse shift operator, in that order. Because of its relatively modest order, the SARIMA $(p, d, q) \times (P, D, Q)_s$ model may be used to a variety of time series while accounting for the connection between the variables. The period value of the time series s (seasonality) is calculated using the dataset.

5.3 Holt- Winters Exponential Smoothing

The Holt exponential smoothing technique is a development of Holt's approach that, at last, makes it possible to capture a seasonal component. Winter's technique is also referred to as triple exponential smoothing since it builds upon both single and double exponential smoothing.

Winter's method assumes that the time series has a level, trend and seasonal component. A estimate with Winter's exponential smoothing can be expressed as:

$$\mathbf{F_{t+k} = L_t + kT_t + S_t + k-M}$$

where,

L_t is the level estimate for time t ,

k is the number of estimates into the future,

T_t is the trend assess at time t ,

S_t is the seasonal estimate at time t

M is the number of seasons

The forecast condition is the result of combining the seasonal, S , component with the extenuation of the HES and SES procedures.

Just like with Holt's method, the forecasting equation has numerous varieties for each of

the types of time series - Additive and Multiplicative.

Additive Seasonality:

$$F_{t+k} = L_t + (k * T_t) + S_{t+k-M}$$

Multiplicative Seasonality:

$$F_{t+k} = [L_t + (k * T_t)] * S_{t+k-M}$$

Holt-Winters' additive method

The component for additive method is:

$$\begin{aligned}\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\ \ell_t &= \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m},\end{aligned}$$

Holt-Winters' multiplicative method

The component form for the multiplicative method is:

$$\begin{aligned}\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\ \ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \\ s_t &= \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1 - \gamma)s_{t-m}\end{aligned}$$

Akaike Information Criterion

This is a measure used to select and compare model. The model with lowest AIC is considered as the best model and chosen for the analysis.

Forecasting

Forecasting is the process of predicting or estimating values for future based on the past or historical data. Forecasting data using different models helps many organizations to understand the future of the business and take necessary actions .

CHAPTER-6

RESULTS AND DISCUSSION

This chapter discusses a comparative study of time series modelling and forecasting of monthly rainfall of Ernakulam using SARIMA and Holt-Winters forecasting Procedure. The rainfall data starting from January 2010 to December 2022.

6.1 EXPLORATORY DATA ANALYSIS

Descriptive Statistics

Table 6.1 shows descriptive statistics

count	156.000000
mean	270.488984
std	258.204683
min	0.000000
25%	50.350000
50%	202.766650
75%	437.908000
max	1244.183000

Table 6.1

The initial step of time series analysis is to draw a time series plot . Time series plot of monthly rainfall of Ernakulam from 2010 to 2022 is given in fig 6.1.

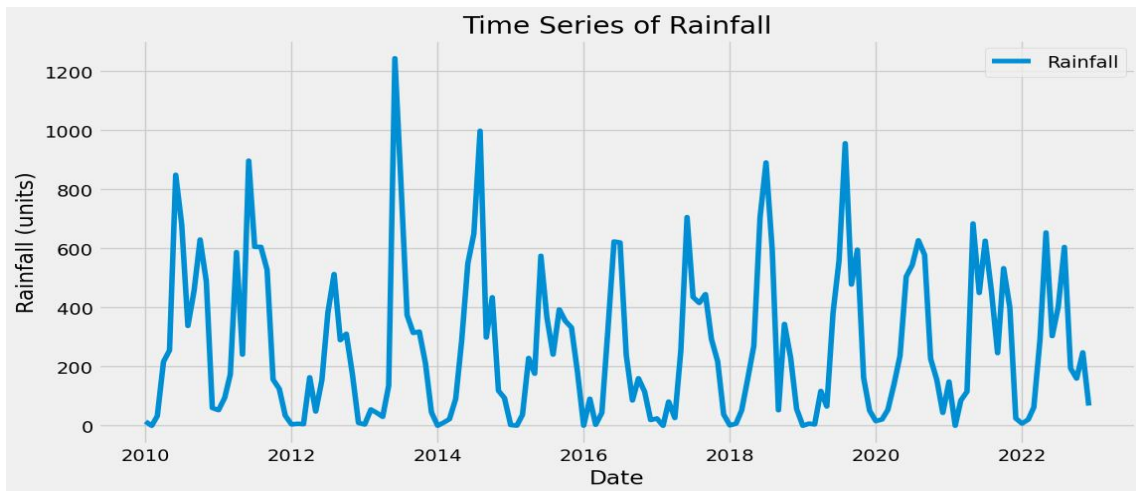


fig 6.1

SEASONAL DECOMPOSITION

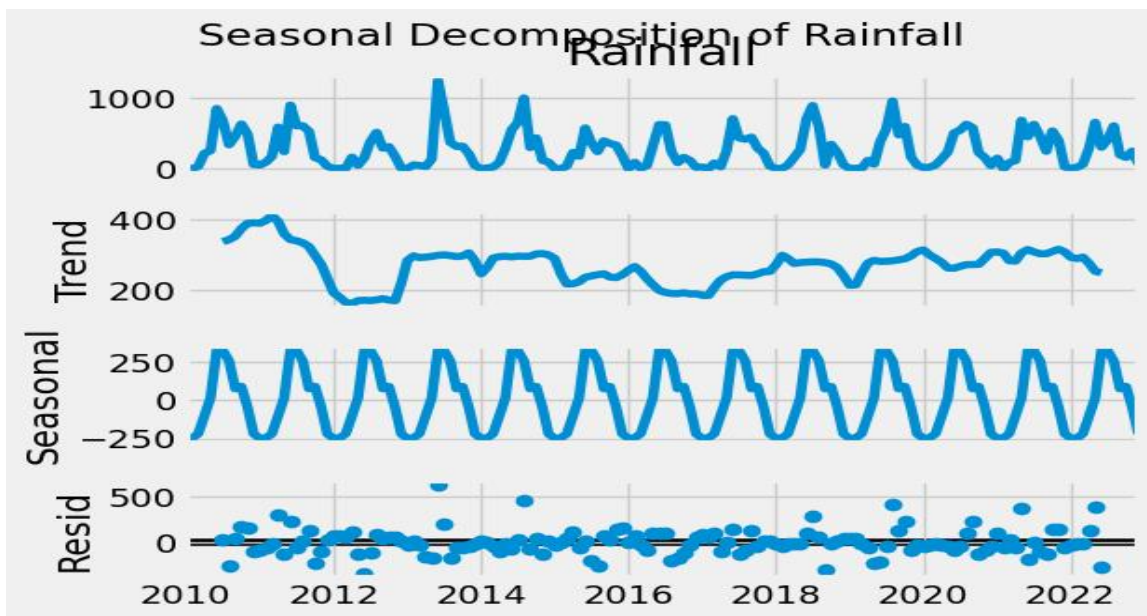


fig 6.2

Seasonal decomposition is performed for evaluation of trend, seasonality and random components.

Fig 6.2 shows the seasonal decomposition

6.2 Modelling of rainfall using SARIMA model

Fig 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 mean are statistically significant. So, to decide ACF and PACF is plotted. Fig 6.3 and 6.4 shows ACF and PACF.

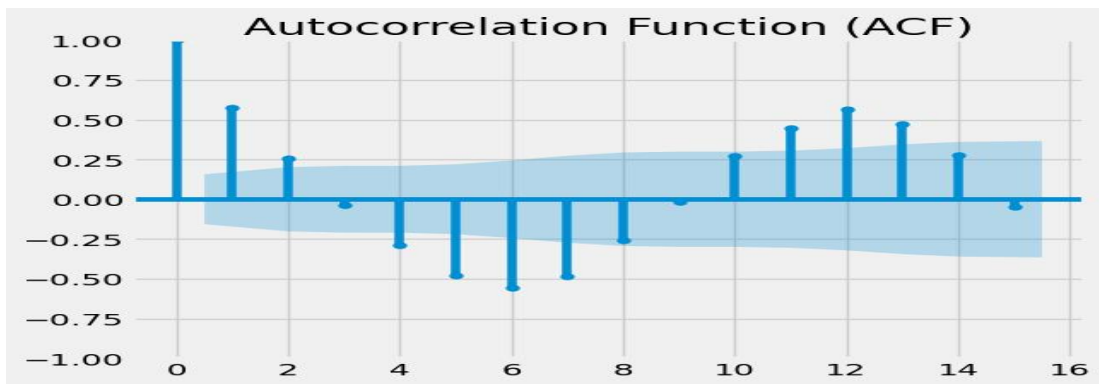


fig 6.3

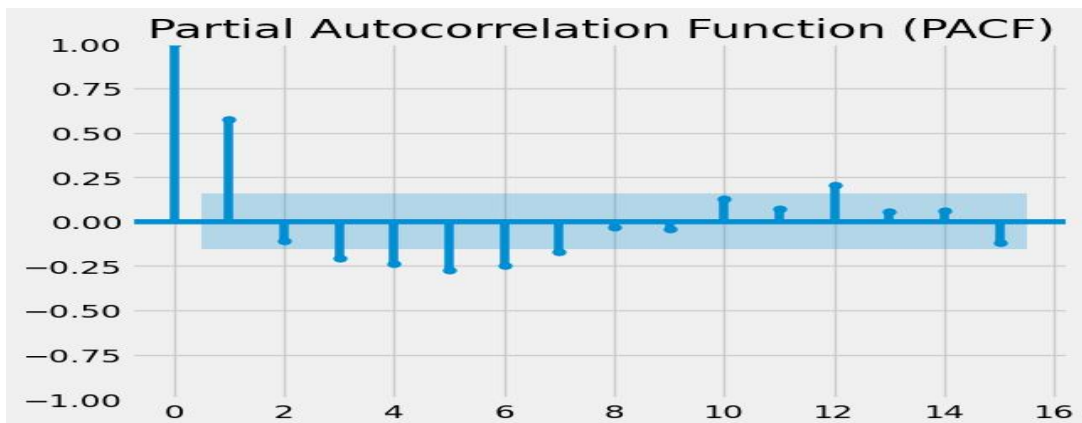


fig 6.4

From fig 6.3 and 6.4 it can be seen that ACF and PACF fails to die out rapidly towards zero, which is a typical pattern of non-stationary series. From fig 6.2 it is clear that the dataset exhibits seasonality the ACF and PACF plot supports that, so seasonal differencing is needed. On taking the difference between a value and a value with lag $S=12$ for transformed data. The time series plot of the seasonally differenced data is shown in fig 6.5.

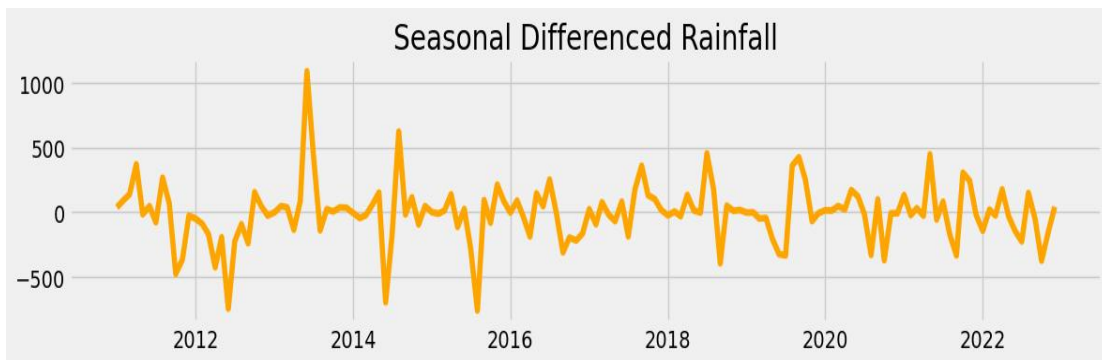


fig6.5

From the seasonal differenced fig 6.5 it is evident that the seasonal behavior is removed from the series. Now again plot the seasonal differenced ACF and PACF.

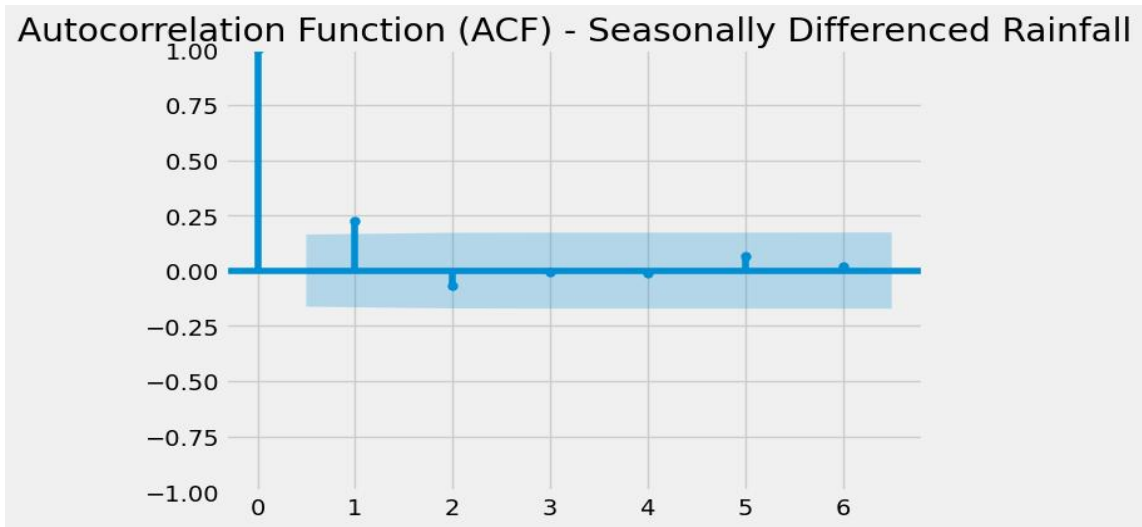


fig 6.6

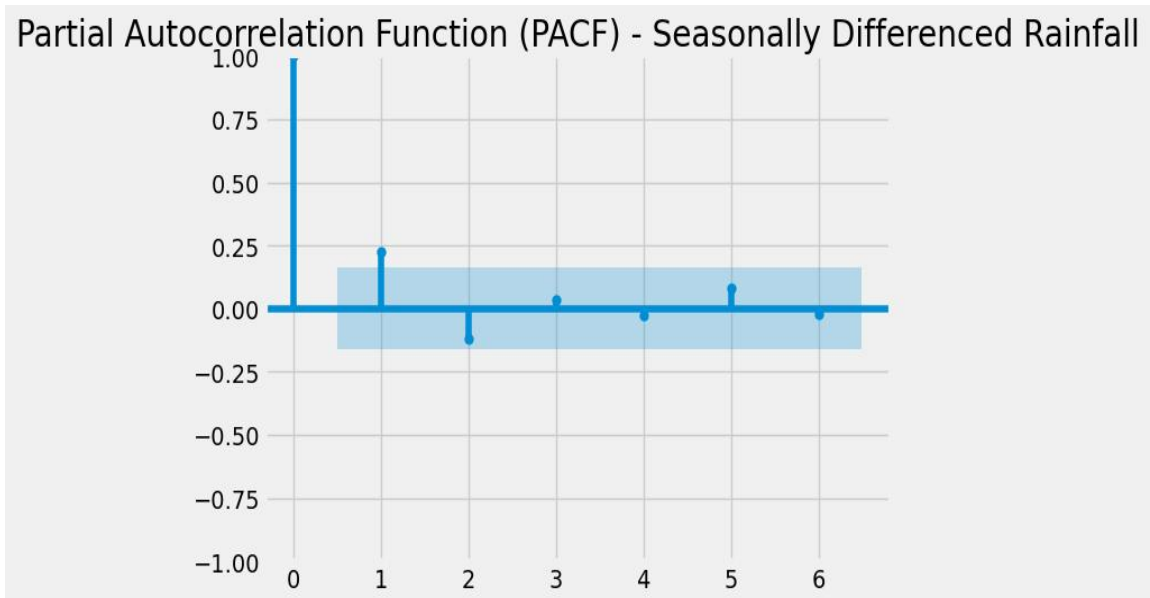


fig 6.7

Augmented Dickey-Fuller Test

To clarify whether the differenced series is stationary or not, Augmented Dickey – Fuller (ADF) test is performed. The result of the test is in table 6.2.

Dickey-Fuller	-4.061399300667899
P-value	0.0011210219225322236

Table 6.2

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 seasonal AR order (P) and the seasonal MA order (Q) have to plot the ACF and PACF of stationary data at seasonal lags.

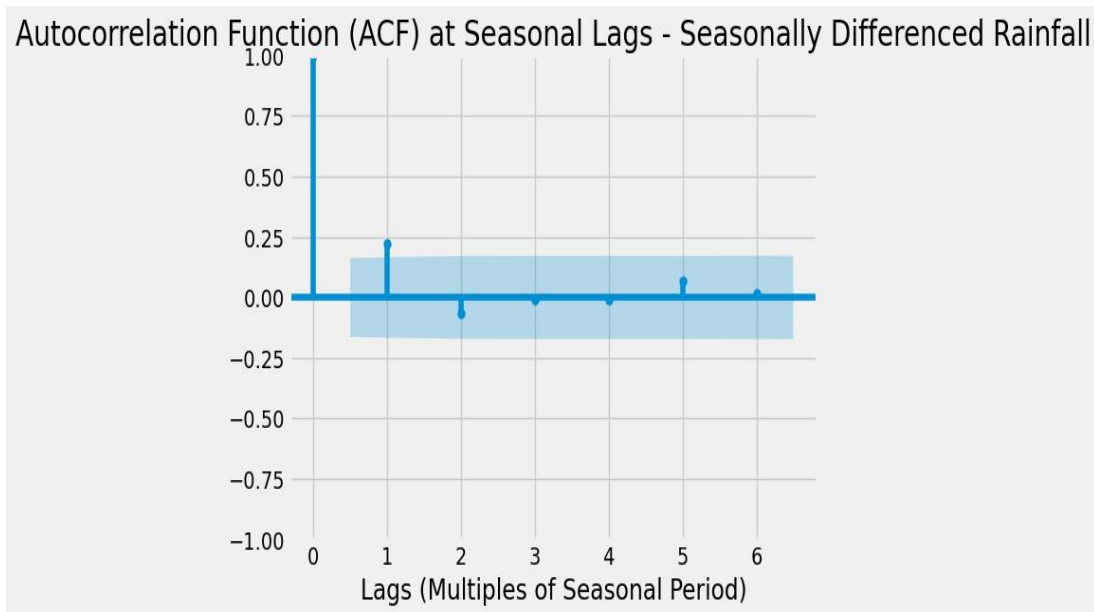


fig 6.8

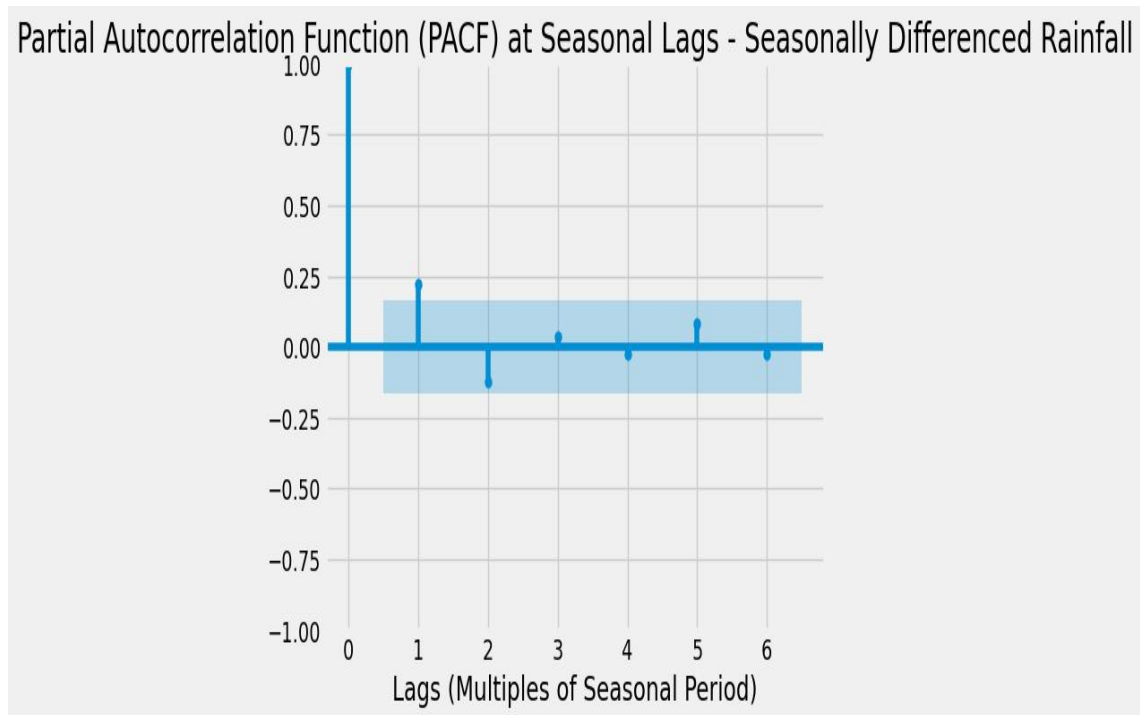


fig 6.9

Fig 6.8 and fig 6.6 show 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 is maximum $p=1$ and maximum $q=1$, and also know that $d=0$. From the ACF and PACF pf seasonally differenced data at seasonal lags maximum $P=1$ and maximum $Q=1$ and $D=1$.

Thus, the possible time series models and their corresponding AIC statistics for the monthly rainfall data of Ernakulam district are:

NO.	ARIMA(p,d,q)x(P, D, Q)	AIC
1	ARIMA (0, 0, 0) x (0, 1, 0, 12)	1961.745895210988
2	ARIMA (0, 0, 0) x (0, 1, 1, 12)	1721.621227427446
3	ARIMA (0, 0, 0) x (1, 1, 0, 12)	1763.012676482092
4	ARIMA (0, 0, 0) x (1, 1, 1, 12)	1719.982411695062
5	ARIMA (0, 0, 1) x (0, 1, 0, 12)	1941.822102104152
6	ARIMA (0, 0, 1) x (0, 1, 1, 12)	1706.101520382509
7	ARIMA (0, 0, 1) x (1, 1, 0, 12)	1752.822320918628
8	ARIMA (0, 0, 1) x (1, 1, 1, 12)	1704.089629416149
9	ARIMA (1, 0, 0) x (0, 1, 0, 12)	1956.400928648367
10	ARIMA (1, 0, 0) x (0, 1, 1, 12)	1718.874410472719

table 6.3

According to minimum AIC, ARIMA (0,0,1) x (1,1,1,12) model found to be more appropriate. The parameter estimates for the model are given in the table 6.4.

Parameter	Coefficient	Standard Error	z	P> z
MA lag1	0.2220	0.094	2.369	0.018
AR.S.lag12	-0.1828	0.083	-2.149	0.027
MA.S.lag12	-0.7453	0.100	-7.456	0.0
sigma2	2.617e+04	2313.996	11.309	0.0

Table 6.4

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.

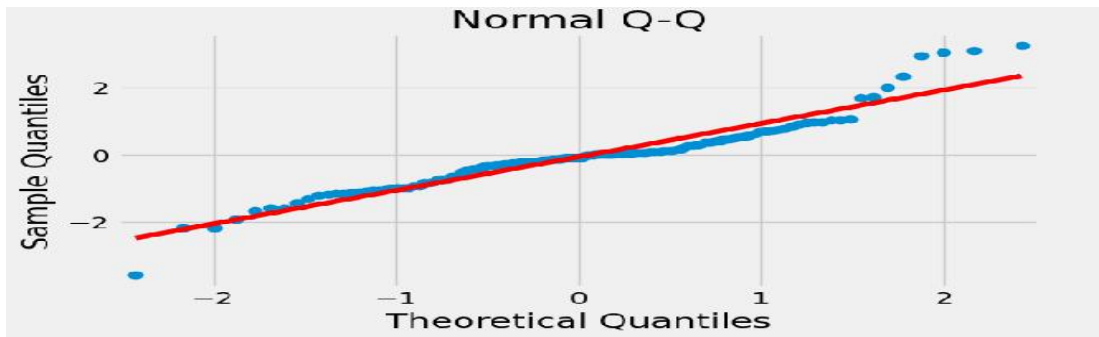


fig 6.10

The fig 6.10 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.

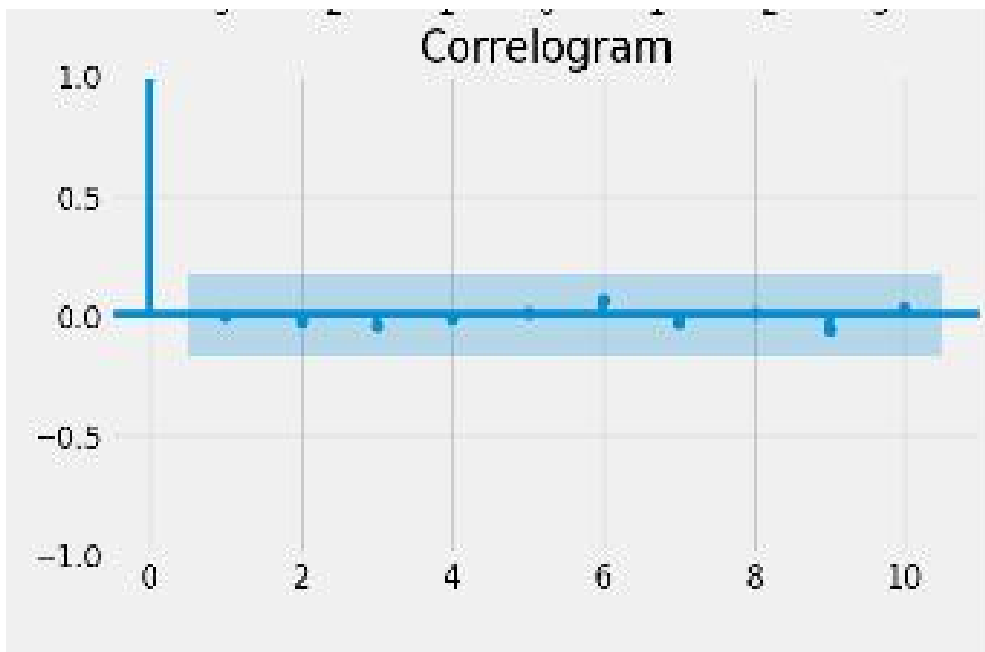


fig 6.11

The fig 6.11 shows correlogram, 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.

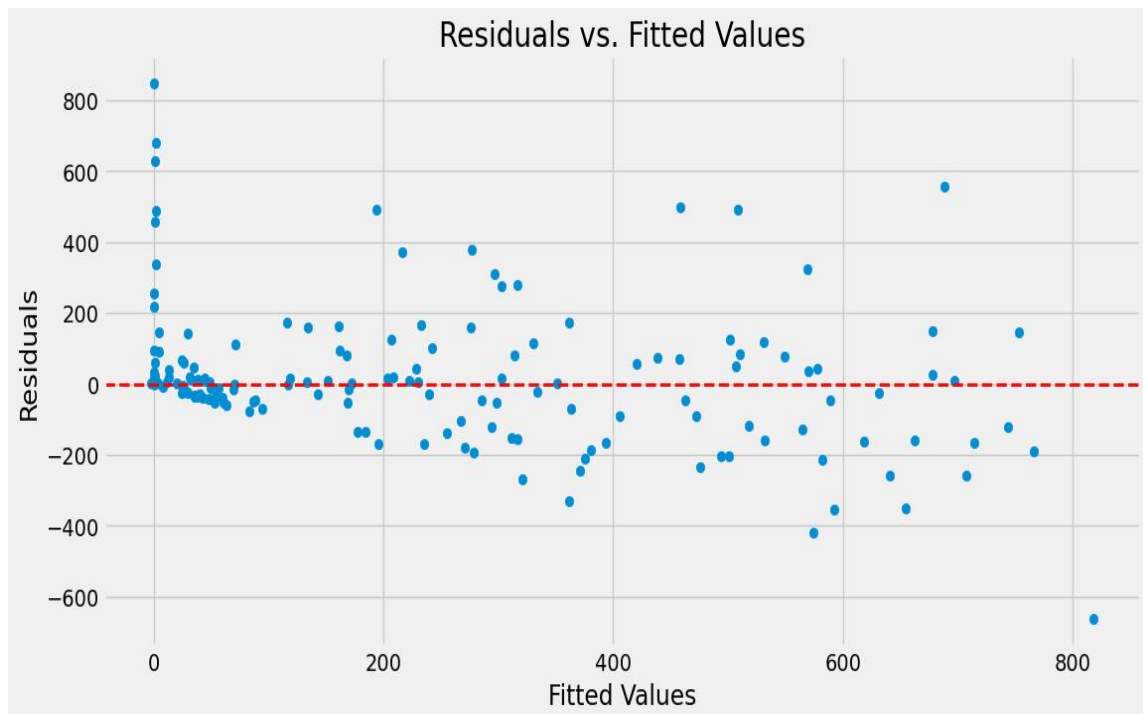


fig 6.12

Fig 6.12 shows the Residual vs Fitted value.

Thus, from the diagnostic checking it is evident that the fitted ARIMA (0,0,1) x (1,1,1,12) model is statistically adequate. So, the model can be used to forecast the monthly rainfall of Ernakulam district.

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 Jan 2022 to Dec 2022.

Months	Actual value	Predicted value
Jan 2022	7.4330	60.2244
Feb 2022	20.7833	68.3466
Mar 2022	62.2166	91.3155
Apr 2022	292.5660	219.8733
May 2022	654.3000	270.2586
Jun 2022	303.5660	705.3418
Jul 2022	402.2500	654.8745
Aug 2022	605.0833	551.5865
Sep 2022	194.3833	410.2039
Oct 2022	159.8330	408.1875
Nov 2022	248.2166	60.2244
Dec 2022	68.1833	68.3466

Table 6.5

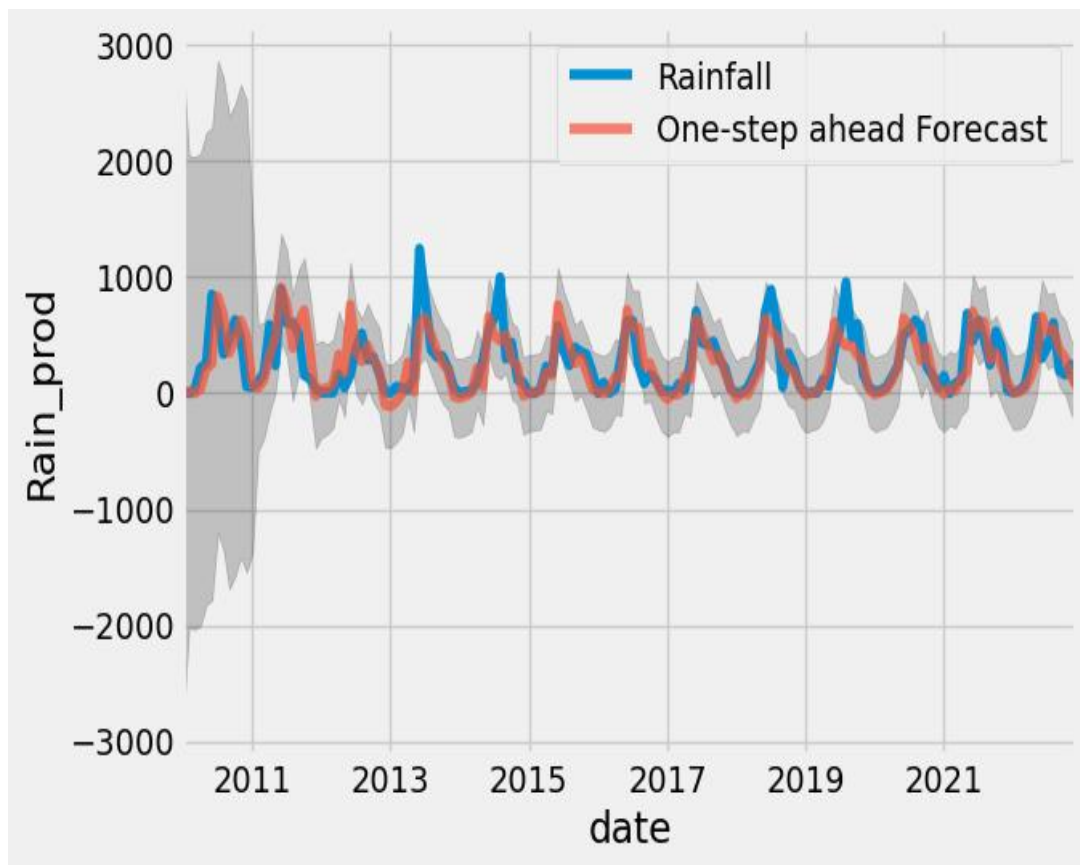


fig 6.13

From fig 6.13 we can see that the blue lines , that is the actual value and the red lines , the in-sample forecasting values.

Forecasting of rainfall using SARIMA model

The SARIMA model $ARIMA(0, 0, 1) \times (1, 1, 1, 12)$ can be used for forecasting. The rainfall from January 2023 to December 2026 is forecasted using the model fitted.

MONTHS	Forecasted values	LCL	UCL
Jan 2023	43.379009	0	362.3337
Feb 2023	7.366483	0	336.2051
Mar 2023	46.263778	0	375.8627
Apr 2023	130.42349	0	460.1142
May 2023	360.722862	31.01522	690.4303
Jun 2023	508.663134	178.9518	838.3744
Jul 2023	578.859875	249.1477	908.5720
Aug 2023	565.518469	235.8060	895.2308
Sep 2023	328.406918	0	658.1195
Oct 2023	371.786149	42.0730	701.4992
Nov 2023	239.004986	0	568.7200
Dec 2023	38.583384	0	368.3058
Jan 2024	29.092299	0	360.2299
Feb 2024	6.555056	0	337.8458
Mar 2024	46.768831	0	378.0860

Time series Analysis of Rainfall in Ernakulam

Apr 2024	161.217955	0	492.5409
May 2024	418.624007	87.2996	749.9483
Jun 2024	463.903524	132.5788	795.2282
Jul 2024	539.968116	208.6433	871.2928
Aug 2024	571.135407	239.8105	902.4602
Sep 2024	298.283997	0	629.6089
Oct 2024	325.618405	0	656.9437
Nov 2024	238.372717	0	569.6998
Dec 2024	42.148851	0	373.4823
Jan 2025	29.504940	0	373.2105
Feb 2025	4.193280	0	348.8944
Mar 2025	44.136006	0	388.9604
Apr 2025	152.348863	0	497.1975
May 2025	404.173939	59.3197	749.0281
Jun 2025	470.590210	125.7347	815.4456
Jul 2025	545.446675	200.5908	890.3024
Aug 2025	567.450100	222.5941	912.3060

Sep 2025	301.957145	0	646.8132
Oct 2025	332.595011	0	677.4516
Nov 2025	235.974054	0	580.8328
Dec 2025	38.885919	0	383.7529

Table 6.6

Table 6.6 is the forecasted values and its LCL and UCL of rainfall for Jan 2023 -Dec 2026.

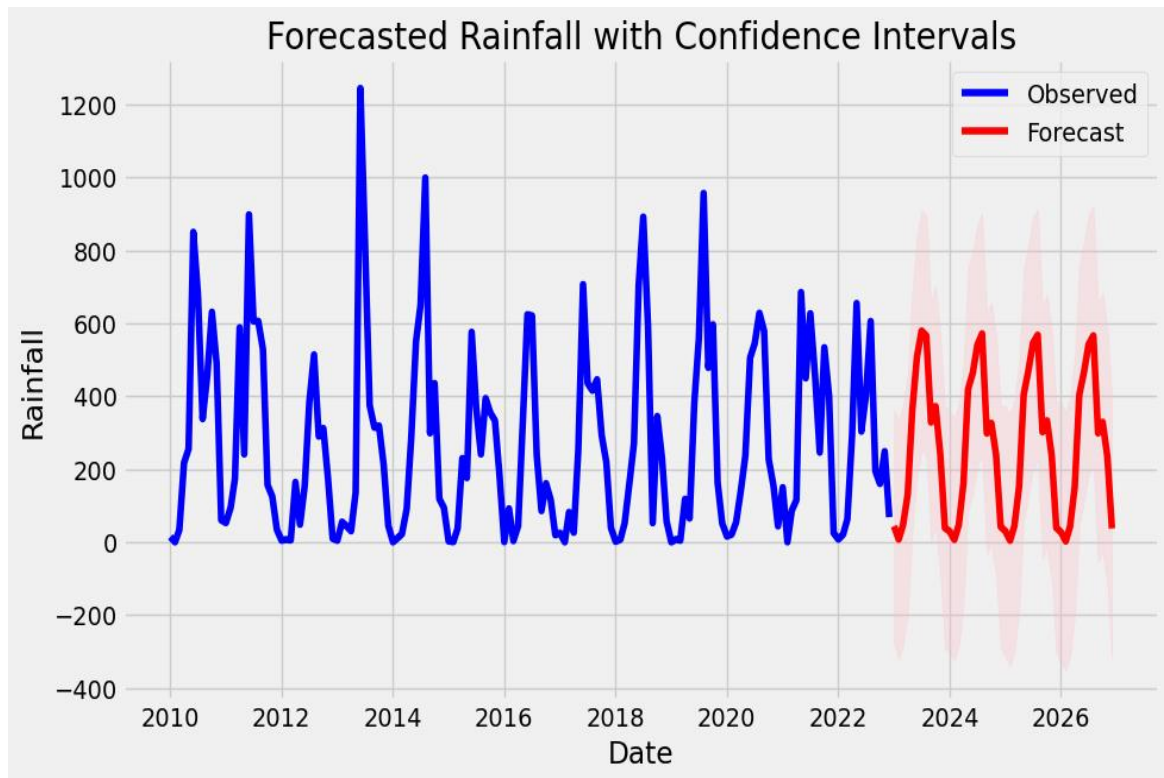


fig 6.14

fig 6.14 is the plot of forecasted values using SARIMA model.

6.3 Modelling of rainfall using Holt-Winters model

Here Holt's Winters Forecasting Procedure is used to forecast rainfall. Now table 6.7 shows the parameter estimates of the model.

Parameters	Parameter Estimates
Alpha (Level)	0.1519736
Gamma (Trend)	0.0284018
Delta (Season)	0.0385530

table 6.7

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.

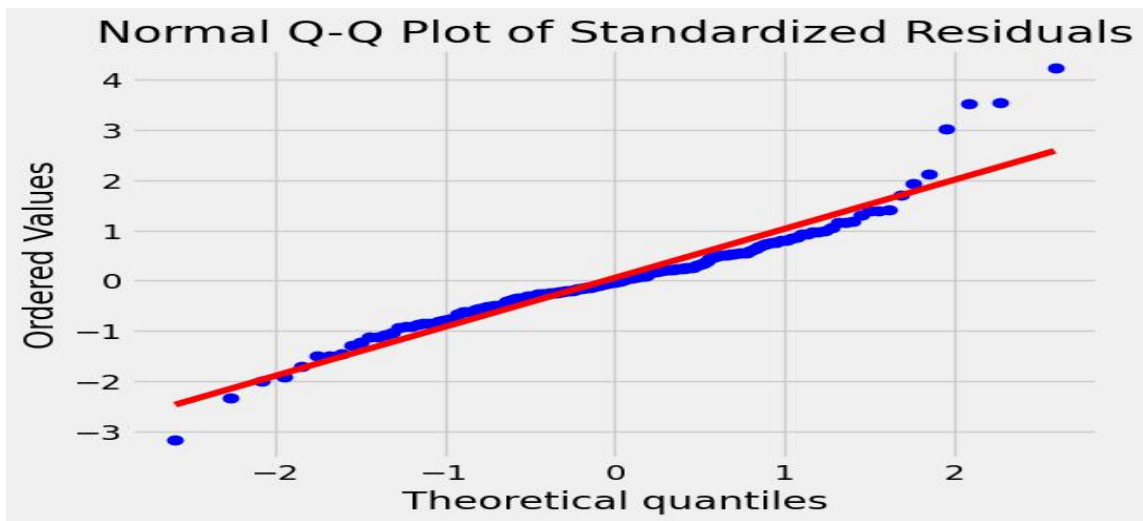


fig 6.15

The fig 6.15 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.

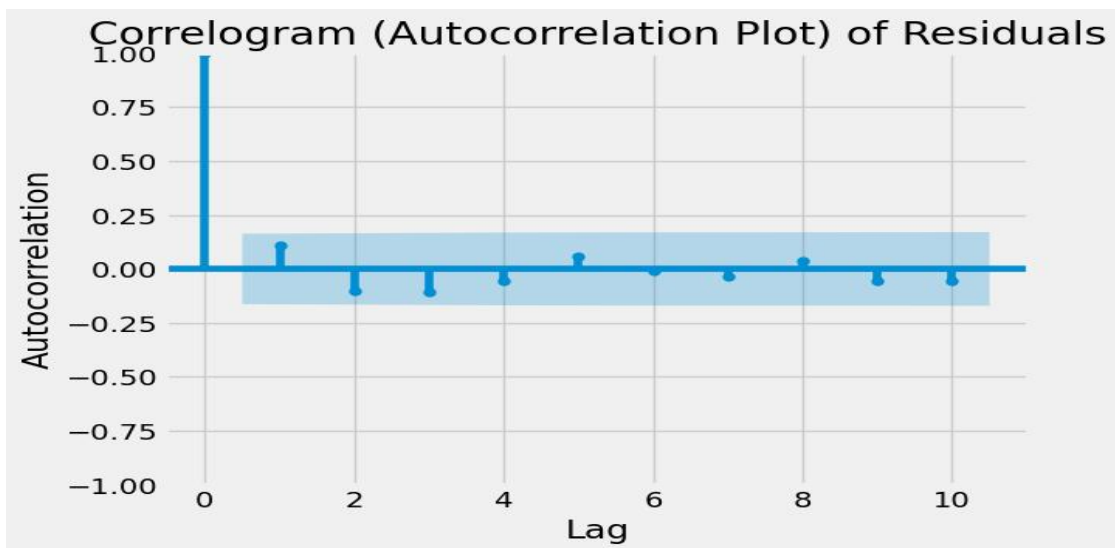


fig 6.16

The fig 6.16 shows correlogram, 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.

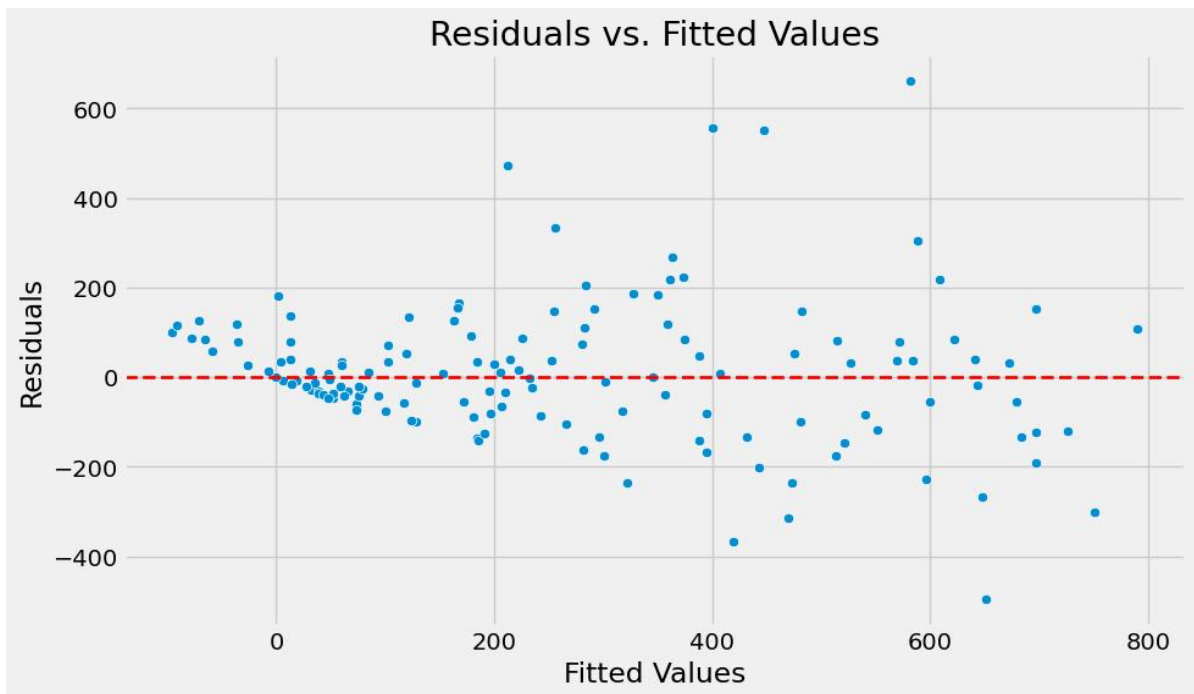


fig 6.17

Fig 6.17 shows the Residual vs Fitted value.

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

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 Jan 2022 to Dec 2022.

Months	Actual value	Predicted value
Jan 2022	7.4330	60.2244
Feb 2022	20.7833	68.3466
Mar 2022	62.2166	91.3155
Apr 2022	292.5660	219.8733
May 2022	654.3000	270.2586
Jun 2022	303.5660	705.3418
Jul 2022	402.2500	654.8745
Aug 2022	605.0833	551.5865
Sep 2022	194.3833	410.2039
Oct 2022	159.8330	408.1875
Nov 2022	248.2166	283.7468
Dec 2022	68.1833	101.4812

table 6.8

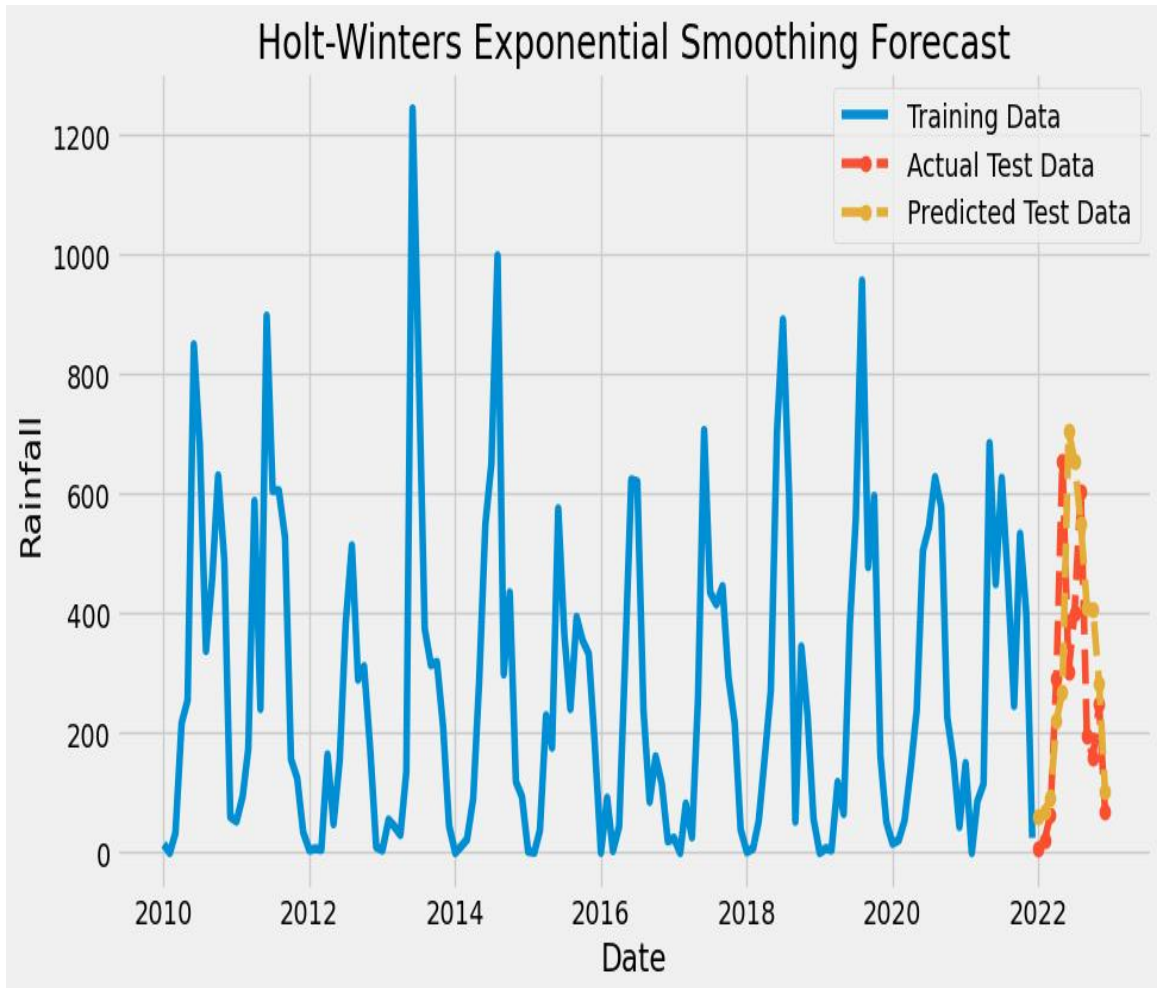


fig 6.17

From fig 6.17 we can see that the blue lines , that is the actual value and the red lines , the in-sample forecasting values.

Forecasting of rainfall using Holt's-Winters model

Rainfall from January 2023 to December 2025 is forecasted using the model fitted.

MONTHS	Forecasted values	LCL	UCL
Jan 2023	72.5074	0	380.270972
Feb 2023	80.6269	0	388.393123
Mar 2023	103.5985	0	411.36207
Apr 2023	232.1563	0	539.919872
May 2023	282.5416	0	590.305177
Jun 2023	717.6247	409.8612	1025.388337
Jul 2023	667.1574	359.3939	974.921028
Aug 2023	563.8694	256.1058	871.633000
Sep 2023	422.4868	114.7233	730.250443
Oct 2023	420.4704	112.7069	728.234013
Nov 2023	296.0297	0	603.793346
Dec 2023	113.7641	0	421.527706
Jan 2024	84.7903	0	392.553918

Time series Analysis of Rainfall in Ernakulam

Feb 2024	92.9125	0	400.676069
Mar 2024	115.8814	0	423.645016
Apr 2024	244.4392	0	552.202818
May 2024	294.8245	0	602.588123
Jun 2024	729.9077	422.1441	1037.671283
Jul 2024	679.4404	371.6768	987.203974
Aug 2024	576.1523	268.3888	883.915946
Sep 2024	434.7698	127.0068	742.533389
Oct 2024	432.7534	124.9898	740.516959
Nov 2024	308.3127	0.54918	616.076292
Dec 2024	126.0470	0	433.810652
Jan 2025	97.0733	0	404.836864
Feb 2025	105.1954	0	412.959015
Mar 2025	128.1644	0	435.927962
Apr 2025	256.7222	0	564.485764
May 2025	307.1075	0	614.871069
Jun 2025	742.1906	434.4271	1049.954229

Jul 2025	691.7233	383.9598	999.486920
Aug 2025	588.4353	280.6717	896.198892
Sep 2025	447.05278	139.2892	754.816335
Oct 2025	445.036	137.2727	752.799905
Nov 2025	320.5956	12.8321	628.359238
Dec 2025	138.3300	0	446.093598

table 6.8

Table 6.8 is the forecasted values and its LCL and UCL of rainfall for Jan 2023 -Dec 2025.

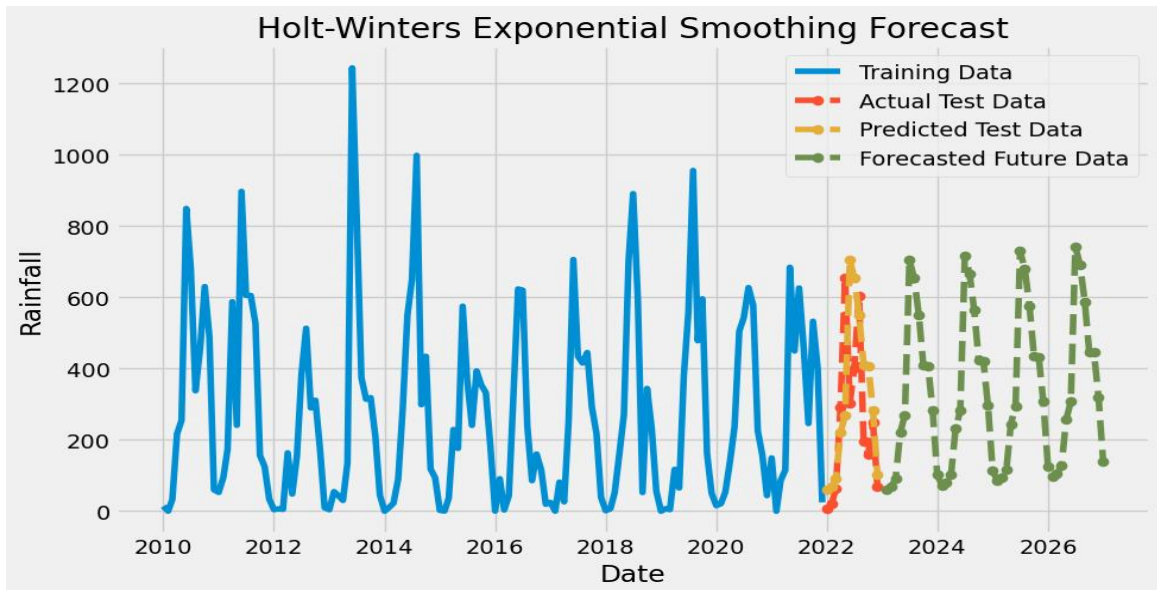


fig 6.18

fig 6.18 is the plot of forecasted values using SARIMA model.

6.4 Comparison between SARIMA and Holt's-Winter's model.

To determine the best model, the performance , metrices should be compared and considered . Here the Mean Square Error (MSE) and Root Mean Squared Error (RMSE) were provided for both models.

	SARIMA Model	Holt's-Winters Model
MSE	30947.611509403967	41450.205077156
RMSE	175.9193323924462	203.5932343599776

table 6.9

From table 6.9 it is evident that SARIMA model has the lower MSE and RMSE compared to Holts -Winters model. That is SARIMA model appears to perform better than the decision tree model.

CHAPTER-7

CONCLUSION

The analyses of two time series model resulted forecasting of future 3 years monthly rainfall values of Ernakulam district. The historical rainfall data of Ernakulam district from Jan 2010 to Dec 2022 were analyzed and forecasted using SARIMA and Holt-Winter's Exponential Smoothing. SARIMA model was got as the best model with a low RMSE of 175.92 and MSE of 30947.61. Holt's winter model had a RMSE of 203.59 and MSE of 41440.99. Both models offer viable forecasting solutions, with ARIMA emphasizing precision and Holt-Winters providing a balanced trade-off between accuracy and simplicity.

From the analysis of the 12 years past rainfall data of Ernakulam district it is evident that there is change in the most rainy months. June- July was the most rainy months in past according to data. However, in the current dataset, the trend has shifted, and August emerges as the most rainy month. This shift could 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 Ernakulam 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

1. 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.
2. G Ganapathy, G. P., Srinivasan, K., Datta, D., Chang, C. Y., Purohit, O., Zaalishvili, V., & Burdzieva, O. (2022). Rainfall Forecasting Using Machine Learning Algorithms for Localized Events. *Comput. Mater. Contin*, 71, 6333-6350.
3. Gowri, L., Manjula, K. R., Sasireka, K., & Deepa, D. (2022). Assessment of Statistical Models for Rainfall Forecasting Using Machine Learning Technique. *Journal of Soft Computing in Civil Engineering*, 6(2), 51-67.
4. Jain, S. K., & Kumar, V. (2012). Trend analysis of rainfall and temperature data for India. *Current Science*, 37-49.
5. Jayasree, A., Sasidharan, S. K., Sivadas, R., & Ramakrishnan, J. A. (2023). Hybrid EMD-RF Model for Predicting Annual Rainfall in Kerala, India. *Applied Sciences*, 13(7), 4572.
6. Joshi, H., & Tyagi, D. (2021). Forecasting and Modeling Monthly Rainfall in Bengaluru, India: An Application of Time Series Models. *Int. J. Sci. Res. in Mathematical and Statistical Sciences* Vol, 8(1).
7. Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. *Disaster Adv*, 11(11), 27-33.

8. Mithiya, D., Mandal, K., & Bandyopadhyay, S. (2020). Time series analysis and forecasting of rainfall for agricultural crops in India: An application of artificial neural network. *Research in Applied Economics*, 12(4), 1-21.
9. Morte, K. K. (2011). A Times Series Analysis into the Rainfall Patterns in Four Selected Regions of Ghana (Doctoral dissertation).
10. Poornima, S., & Pushpalatha, M. (2019). Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. *Atmosphere*, 10(11), 668.

