Cairo University

Faculty of Economics and Political Science

English Section

Statistics Department

3rd Year

Analysis and Forecasting of Rainfall Time Series

Submitted by:

Saif Essam Abdelmaboud (5200356)

Mohamed Aly Abdelrahman (5201010)

Under The Supervision of:

Dr. Eman Mahmoud

Dr. Yousra Hassan

Table of Contents

I. Introduction:

- Rainfall is an essential component of the Earth's climate system, and variations in rainfall patterns can have significant impacts on agriculture, water resources, and human livelihoods. Understanding the patterns and trends in rainfall over time is therefore of great importance for many applications, from drought monitoring to flood forecasting to water management.
- In this report, we will analyze a time series dataset of daily rainfall data from 16 locations in India from 2006 to 2020. We will explore the importance of analyzing this data, the interest in modeling it, and the importance of forecasting such a time series.

1- Importance of Analyzing the Data:

- Analyzing the rainfall time series data is essential for several reasons. First, it can provide insights into the patterns and trends in rainfall over time, including seasonal cycles, long-term trends, and extreme events. This information can be useful for understanding the impacts of climate change on rainfall patterns and for developing effective strategies for managing water resources in the face of changing conditions.

2- Interest in Modeling the Data:

- Modeling the rainfall time series data is interesting for several reasons. First, it can help us understand the underlying processes that govern rainfall variability, such as atmospheric dynamics, land surface processes, and feedbacks between the Earth's surface and atmosphere.
- Second, modeling the data can help identify the relative importance of different drivers and predictors of rainfall variability. For example, we can use statistical models such as autoregressive models to assess the contribution of different climate variables to rainfall variability, and to identify which variables are most important for predicting future rainfall patterns.

3- Importance of Forecasting the Time Series:

- Forecasting the rainfall time series data is essential for many applications, including agriculture, water management, and disaster risk reduction. By predicting future rainfall patterns, we can develop strategies for managing water resources, reducing the impacts of extreme events such as droughts and floods, and improving crop yields and food security.

- In conclusion, analyzing, modeling, and forecasting the rainfall time series data is of great importance for understanding the impacts of climate change on rainfall patterns, for managing water resources, and for reducing the risks of extreme events such as droughts and floods. The following sections of this report will describe the methods used to analyze, model, and forecast the data, and will present the results of our analysis.

II. Literature Review:

- There are several studies that have explored the patterns and trends in rainfall over time using time series analysis techniques. For example, *Sharma and Kumar (2020)* analyzed long-term rainfall data from India and found that there is a significant increasing trend in the annual rainfall, with a higher rate of increase in the monsoon season.
- In another study, *Wang et al. (2019)* analyzed daily rainfall data from China and found that the rainfall exhibited a strong seasonal cycle, with a peak in the summer months. They also identified significant trends in the rainfall over time, with increasing rainfall in some regions and decreasing rainfall in others.
- In addition to analyzing the patterns and trends in rainfall over time, several studies have also explored the drivers and predictors of rainfall variability. For example, *Xie et al. (2021)* used machine learning techniques to identify the key drivers of rainfall variability in China, including atmospheric circulation patterns, sea surface temperatures, and land surface conditions.
- Finally, there has been growing interest in developing tools for forecasting future rainfall patterns using time series analysis techniques. For example, *Rana et al. (2021)* developed a hybrid model based on autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN) to forecast monthly rainfall patterns in India. They found that the hybrid model performed better than either ARIMA or ANN alone.
- Overall, these studies highlight the importance of analyzing and modeling rainfall time series data for understanding the patterns and trends in rainfall over time, and for developing tools for forecasting future rainfall patterns. The following sections of this report will describe the methods used to analyze and model the rainfall time series data from India, and will present the results of our analysis.

III. Data Analysis:

1- Description of the Data Set:

- Source: The dataset used in this analysis is sourced from the Prediction of Worldwide Energy Resources (POWER) project, which is a research initiative by NASA. The project aims to provide meteorological data to support various applications such as renewable energy, building energy efficiency, and agriculture. The data is obtained from an assimilation model, which combines observational data with model predictions. The data is from Kaggle: https://www.kaggle.com/datasets/poojag718/rainfall-timeseries-data
- Variable Measured: The variable of interest in this dataset is precipitation. Precipitation refers to the amount of moisture that falls from the atmosphere to the Earth's surface, typically in the form of rain, snow, sleet, or hail. In this dataset, precipitation is measured as the monthly sum of rainfall. Monthly rainfall measurements in millimeters.
- Scale Unit: Millimeters.
- Sample Size: The dataset covers a period from 2000 to 2020, providing a total sample size of 21 years. The data is collected at a monthly frequency, resulting in 252 observations (12 months per year multiplied by 21 years). Each observation represents the monthly sum of rainfall recorded at a specific location in Mumbai, India.

2- Graphical Representation for the Time Series:

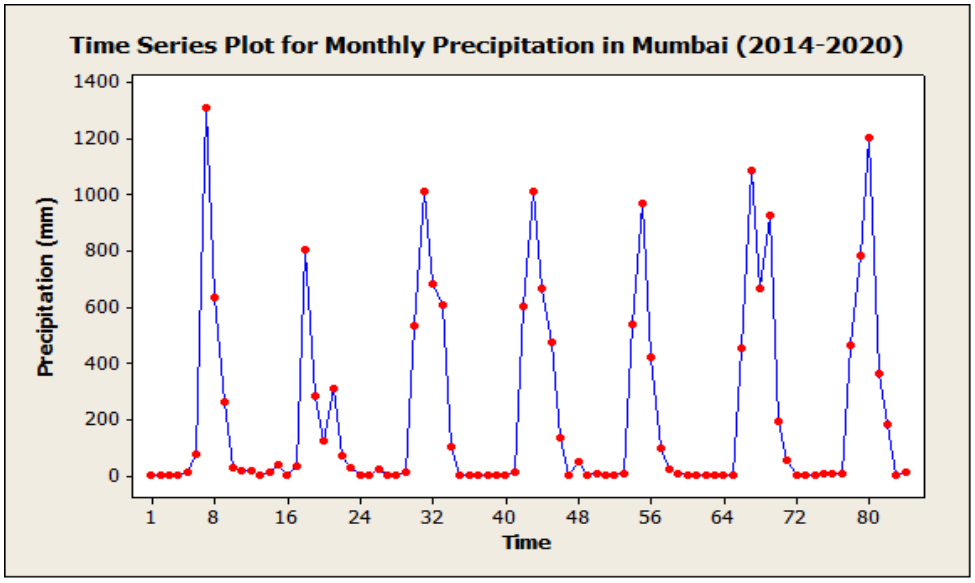


Figure (1): Time Series Plot for Precipitation

- Figure (1): Shows the monthly precipitation in Mumbai from 2014 to 2020. Overall, there is a noticeable variation in precipitation throughout the years.
- There appears that there is no clear upward or downward trend observed in the short term. However, upon closer examination, a subtle trend may be observed in the long term. While the precipitation levels show minor fluctuations throughout the years, there is a slight indication of a gradual increase over the period.
- Also, there appears to be a general seasonal pattern, with higher precipitation levels occurring during certain months, followed by drier periods.
- There is a significant increase in precipitation around the months [6, 7, and 8] and a notable decrease in [1, 2, and 3].
- Additionally, there are outliers, which may indicate extreme weather events or measurement anomalies.

3- Decomposition Method:

- We cannot start by analyzing the trend alone as the effect of seasonality will make it wrong so we will start by analyzing both of them together.
- Although there is a slight difference in the MAPE where the multiplicative is lower, we will **use the additive model**, as it has the lowest MAD and MSD.

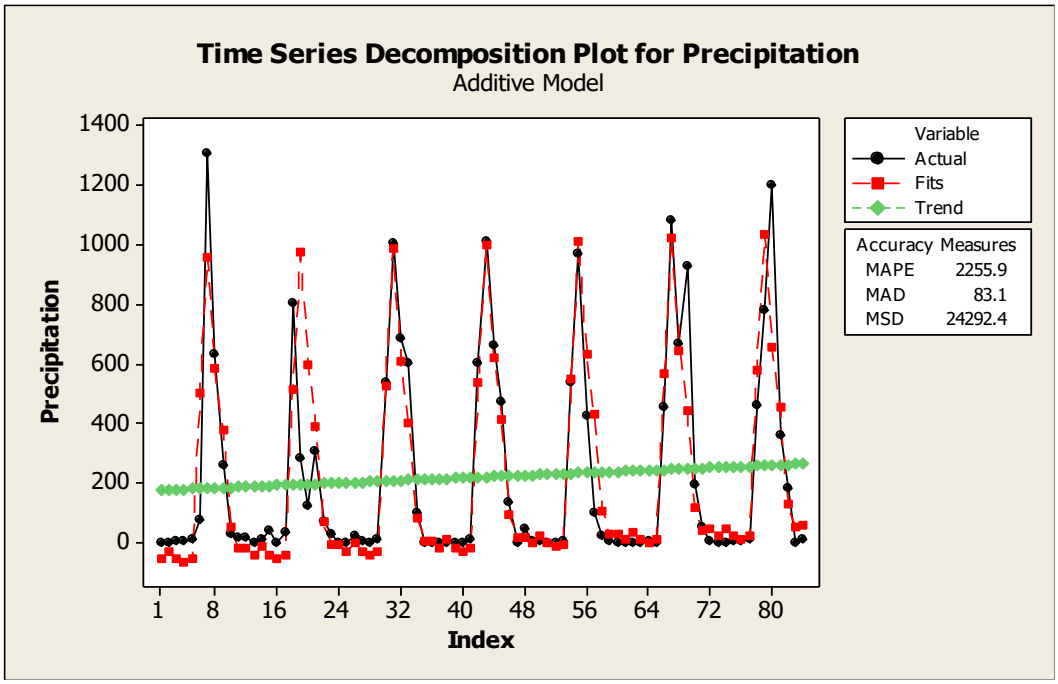


Figure (2): Time Series Decomposition Plot

- Figure (2): Shows that the decomposition of the time series reveals a small increase in the long-term trend of precipitation. The trend component shows a slight upward movement over the time period (from 2014 to 2020), indicating a gradual increase in precipitation levels over time.
- Although the trend is not pronounced, it suggests a potential shift towards higher precipitation amounts in Mumbai. This finding implies that there might be underlying factors contributing to the long-term increase in rainfall, which should be further investigated.

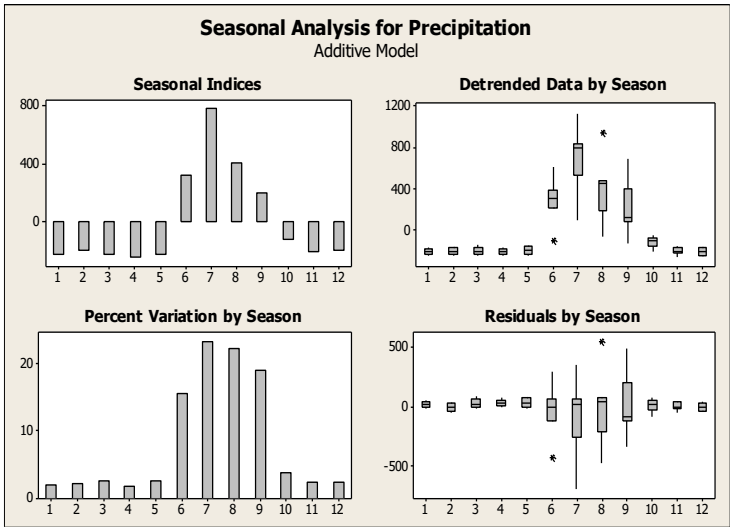


Figure (3): Seasonal Analysis for Precipitation

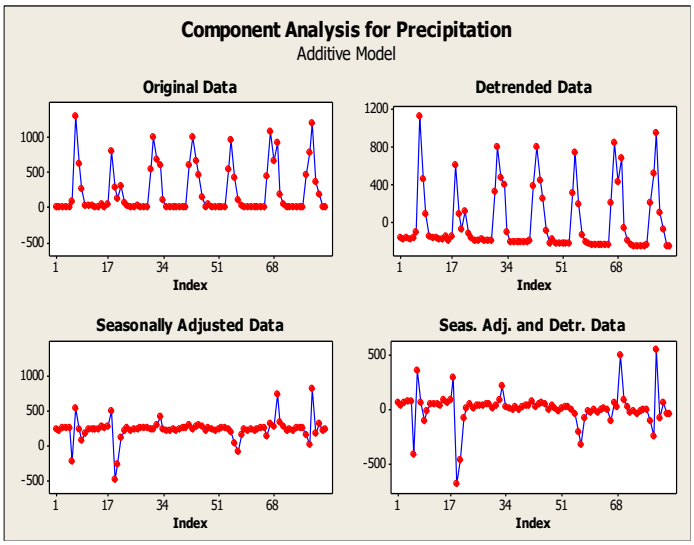


Figure (4): Component Analysis for Precipitation

- Figure (3): Shows that the seasonal component of the decomposition highlights the systematic variation in precipitation within each year. Specifically, there is a notable increase in the phenomenon during the months of June, July, and August (months 6, 7, and 8). Among these months, July exhibits the highest precipitation levels. This observation aligns with the monsoon season in Mumbai, where heavy rainfall is commonly experienced. Subsequently, the precipitation gradually decreases and returns to relatively normal levels during the remaining months. The distinct seasonal pattern suggests a significant influence of the monsoon season on the overall precipitation patterns in Mumbai.

- Figure (4): Shows that there is no visible difference between the original data and the detrended, suggesting the absence of a clear trend in the precipitation time series. This implies that the overall pattern of the data is relatively stable, without any notable long-term increasing or decreasing trend.
- Additionally, the plots reveal that the seasonally adjusted data closely resembles the seasonally adjusted and detrended data. This implies that the seasonal adjustment process effectively removes the seasonal patterns from the data, resulting in a time series that is relatively stable and free from seasonal variations.

✓ The Fitted Trend Equation for the Time Series:

- We estimate the trend by using the following linear model:

$$Y_i = \beta_0 + \beta_{1t} + u_i$$

- The estimated Model:

$$\hat{Y}_t = 174.8 + 1.06*t$$

- $\hat{Y}_t$: Represents the predicted value of the time series at time t.
- $\hat{\beta}_0$: The mean value of the trend will be 174.8, holding the time constant.
- $\hat{\beta}_1$: By increasing the time index by 1 unit, the predicted value of the time series trend increases, on average, by 1.06.

✓ The Seasonal Indices:

- The seasonal indices for the time series reveal the relative magnitude of the seasonal component for each period.
- We have found that:
  - S1= -230.399: Season 1 is reducing the precipitation by 230.399 of the trend.
  - S2= -204.527: Season 2 is reducing the precipitation by 204.527 of the trend.
  - S3= -232.349: Season 3 is reducing the precipitation by 232.349 of the trend.
  - S4= -245.623: Season 4 is reducing the precipitation by 245.623 of the trend.
  - S5= -235.741: Season 5 is reducing the precipitation by 235.741 of the trend.
  - S6= 320.468: Season 6 is increasing the precipitation by 320.468 of the trend.
  - S7= 779.437: Season 7 is increasing the precipitation by 779.437 of the trend.
  - S8= 399.858: Season 8 is increasing the precipitation by 399.858 of the trend.
  - S9= 194.133: Season 9 is increasing the precipitation by 194.133 of the trend.

- S10= -130.042: Season 10 is reducing the precipitation by 130.042 of the trend.
- S11= -208.062: Season 11 is reducing the precipitation by 208.062 of the trend.
- S12= -207.153: Season 12 is reducing the precipitation by 207.153 of the trend.

✓ Forecasting the Next Year (2021):

- The forecasted values provide an estimation of the precipitation for future periods in the time series as follows:

Period:	Forecast
85:	34.77
86:	61.70
87:	34.94
88:	22.73
89:	33.68
90:	590.95
91:	1050.98
92:	672.47
93:	467.81
94:	144.70
95:	67.74
96:	69.71

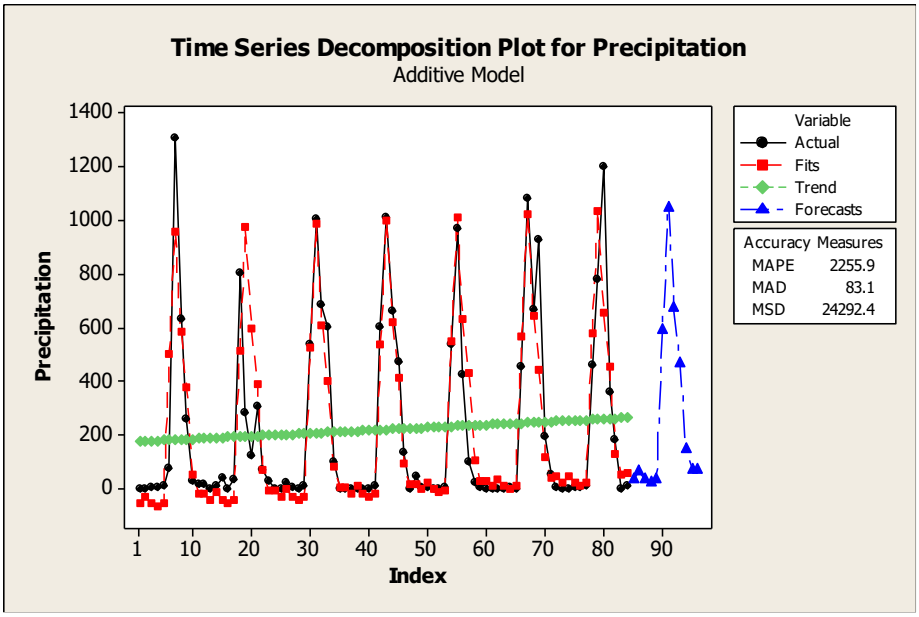


Figure (6): Time Series Decomposition Plot with Forecasts

- Figure (6): Represents the plot of the forecasts of the decomposition analysis.
- Examining the forecasted values, we observe varying levels of predicted precipitation for the upcoming periods. For instance, in Period 90, the forecast suggests a substantial increase in precipitation with a value of 590.95. This is followed by a significant surge in Period 91, where the forecasted precipitation reaches 1050.98. These high forecast values indicate a potential occurrence of heavy rainfall during these periods.

4- Modern Approach:

✓ Checking Stationarity:

- First, we have to check the stationarity, as it refers to the property of a time series where its statistical properties such as mean, variance, and autocorrelation remain constant over time. Thus, it is an important assumption that we have to check.
- From Figure (1): The time series plot appears to have many fluctuations which indicated that the series is not stationary in variance and there may also exist a small trend which indicates that the series is not stationary in mean.
- From Figure (7): We can see that the ACF function takes a sine wave pattern and dies slowly indicating that the series is non-stationary.

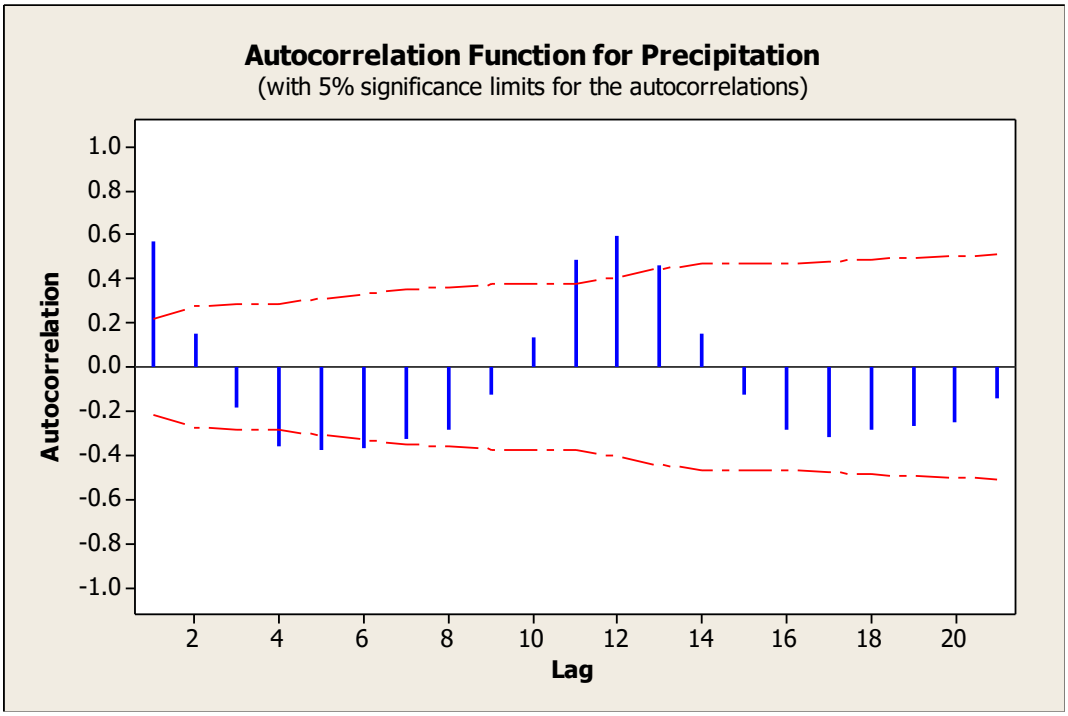


Figure (7): ACF for Precipitation

- **We have to convert the series into yearly observations to remove the effect of the seasonality.**

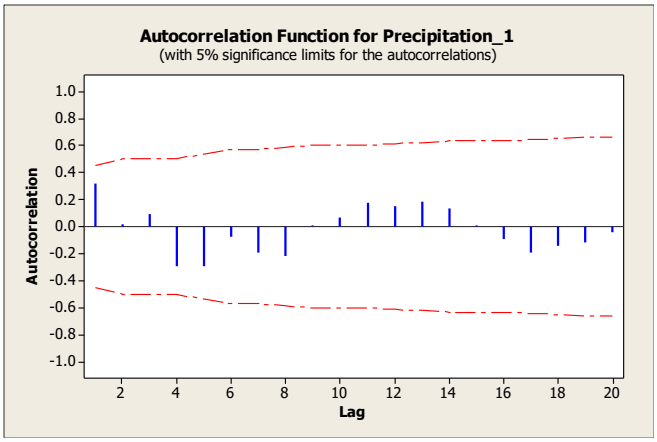


Figure (9): ACF of the New Yearly Precipitation Data

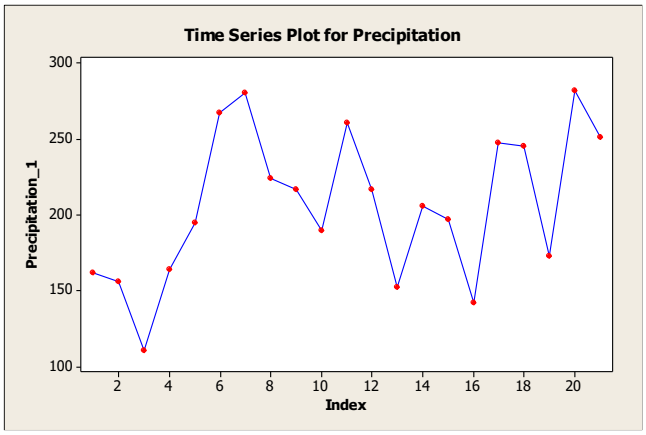


Figure (8): TS Plot for the Yearly Precipitation Data

- From Figure (8): The time series plot of the **new yearly data** has a trend then the series is not stationary in mean.
- From Figure (9): We can see that the ACF has a sine wave pattern and it dies slowly, therefore it is non-stationary.

✓ So, we should try the **first difference** for the TS:

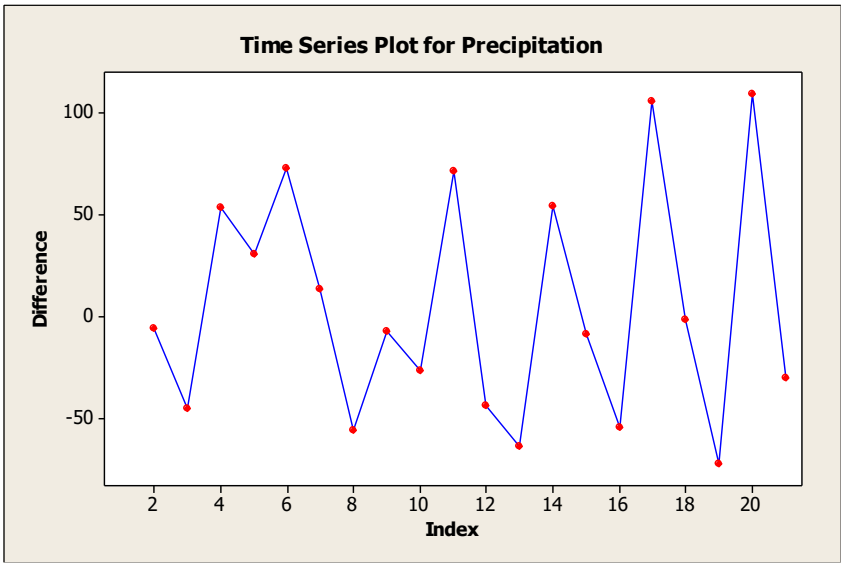


Figure (10): Time Series Plot After the First Difference

- Figure (10): Shows that the observed data shows no trend after taking the first difference and the series is stationary in mean.
- Now after we got summation for the variable then we had 1st order difference; **now our model is stationary as it became constant in mean and variance.**

✓ Checking ACF and PACF:

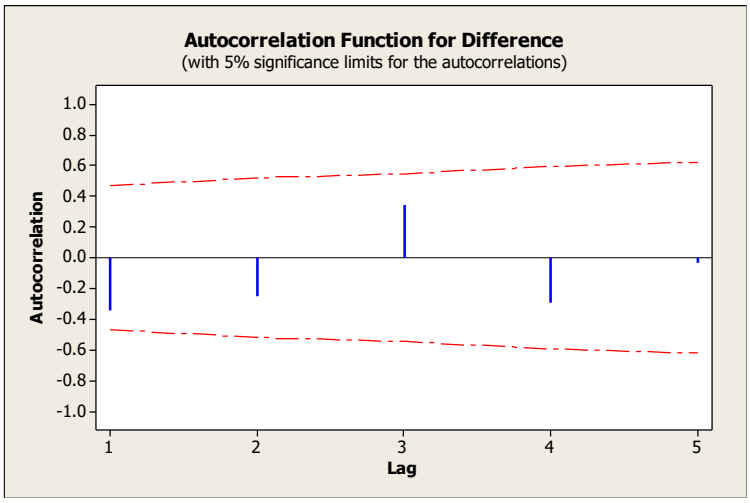


Figure (11): ACF of the Stationary Series

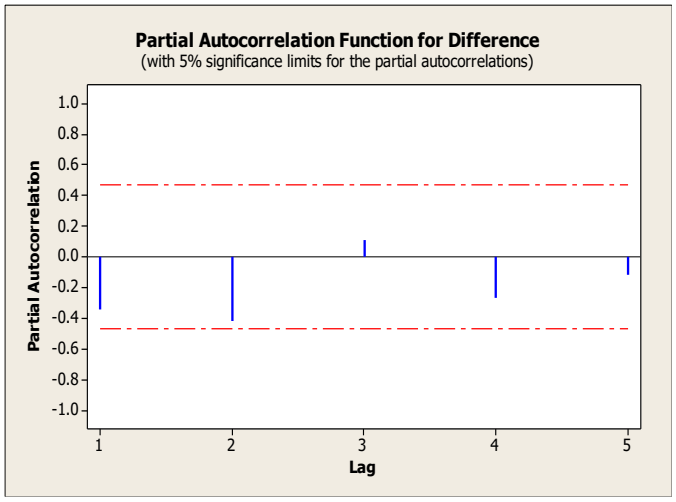


Figure (12): PACF of the Stationary Series

- Figure (11) and Figure (12): Shows that the ACF and PACF don't have any known features and are approximately zeroes which gives **no indication about the initial model to start with.**

5- Box and Jenkins Analysis:

The Box-Jenkins approach helps in selecting an appropriate ARIMA model that captures the autocorrelation and time-dependent structure in the stationary time series data.

- **Starting with ARMA (1,1) as an initial model:**

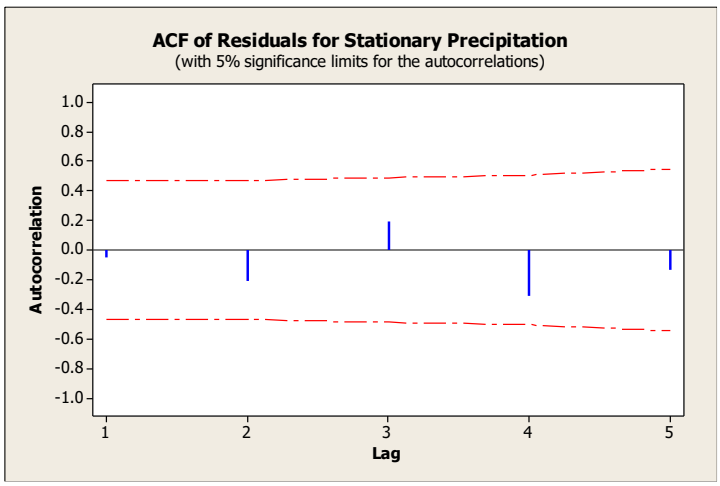


Figure (13): ACF of the residuals of ARMA (1,1) Model

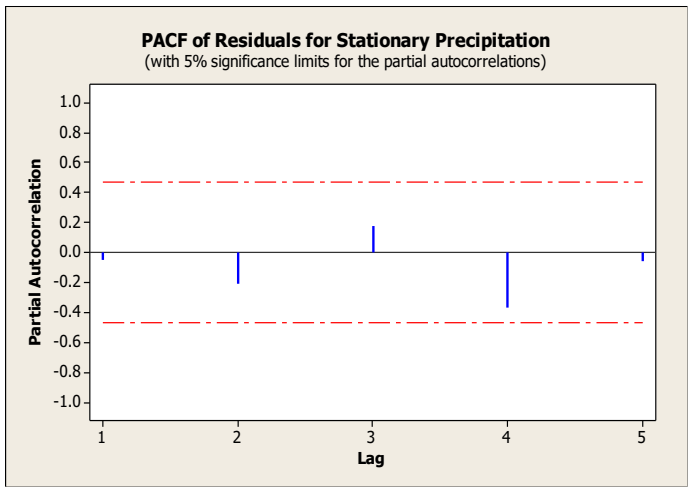


Figure (14): PACF of the residuals of ARMA (1,1) Model

- We can see from Figure (13) and Figure (14) that the ACF and PACF of the residuals are all insignificant.

```
Final Estimates of Parameters
Type      Coef  SE Coef     T      P
AR   1  0.3241   0.2985  1.09  0.292
MA   1  0.9094   0.2288  3.97  0.001
```

- While, we can find that the p-value for the coefficient of AR is 0.292 which is greater than 0.05, this implies that the coefficient is insignificant.

- **We should try different model orders:**
  - After trying many different order combinations, I have found that: ARMA (2,1), MA(1), and AR (2) are all satisfying the diagnostic tests.
  - So, we have to compare between them all using the measures of accuracy of each model.

ARMA (2,1)	AR (2)	MA (1)
MAD = 32.2614	MAD = 38.0224	MAD = 41.3101
MSE = 1793.46	MSE = 2207.14	MSE = 2464.42
MAPE = 108.354%	MAPE = 166.156%	MAPE = 217.938%

  - By comparing the values for the MAD, MSE, and MAPE, we have found that the model of ARMA (2,1) is the appropriate one as it has the least errors, and therefore the best measures of accuracy.
  - Thus, we will consider ARMA (2,1) as it is the suitable model for forecasting our time series data.

- **The Model Diagnoses:**

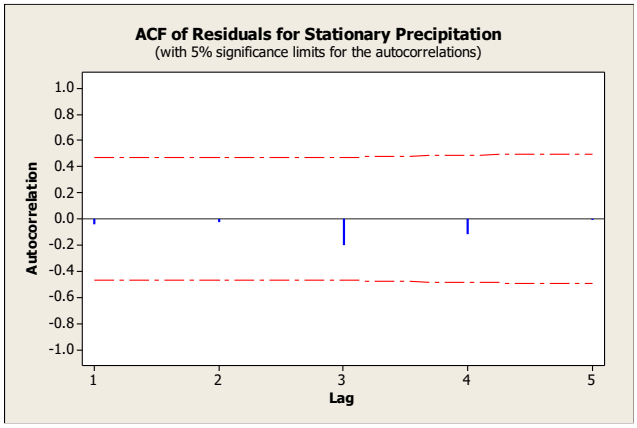


Figure (15): ACF of the residuals of ARMA (2,1) Model

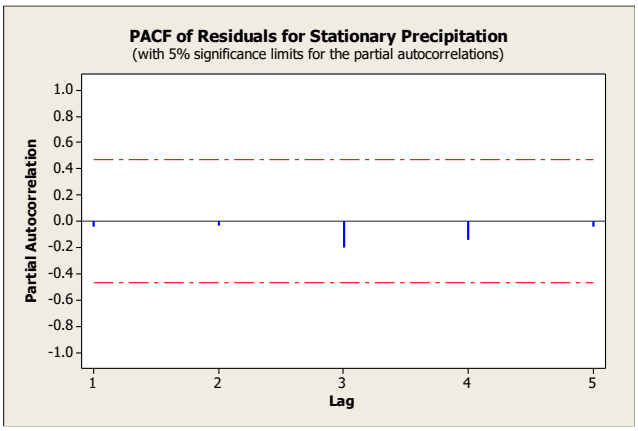


Figure (16): PACF of the residuals of ARMA (2,1) Model

- Figure (15) and Figure (16): Shows that the ACF and PACF of the residuals are all insignificant, which is a good indication that the model is fitting the data well and that the residuals are behaving as **white noise**.

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag             12   24   36   48
Chi-Square     8.1    *    *    *
DF               9    *    *    *
P-Value      0.529    *    *    *
```

- The **Ljung-Box test** is a diagnostic test used to assess the presence of autocorrelation in the residuals of a time series model. The test statistic follows a Chi-Square distribution, and the p-value indicates the probability of observing the test statistic under the assumption of no autocorrelation.
- In this case, for the lag of 12, the obtained p-value of 0.529 is greater than the significance level of 0.05. This suggests that there is no significant evidence of autocorrelation in the residuals at that lag.

```
Final Estimates of Parameters

Type        Coef  SE Coef      T      P
AR    1  -1.0451   0.2100  -4.98  0.000
AR    2  -0.6717   0.2187  -3.07  0.007
MA    1  -0.8728   0.1882  -4.64  0.000
```

- Moreover, the estimates of the parameters of the model are **all significant** according to the p-value which is less than 0.05 for each coefficient. Which indicates that each coefficient has a statistically significant effect on the model's predictions.

○ The Estimated Model:

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\widehat{Y}_t = -1.0451 * Y_{t-1} - 0.6717 * Y_{t-2} - 0.8728 * \varepsilon_{t-1} + \varepsilon_t$$

○ Checking Stationarity and Invertibility:

- Since, θ is between -1 and 1, **therefore the model is invertible.**
- Since:

✓ $\emptyset 2 - \emptyset 1 = -0.6717 - (-1.0451) = 0.3734$, which is less than 1.

✓ $\emptyset 2 + \emptyset 1 = -0.6717 + (-1.0451) = -1.7168$ which is less than 1.

✓ $\emptyset 2 = -0.6717$ which is between -1 and 1.

- **Therefore, the model is Stationary.**

The stationarity and invertibility of the ARMA (2,1) model are important properties that ensure the model is appropriate for capturing the patterns and dynamics of the time series data.

○ Forecasting Using Box and Jenkins Analysis:

- First Year Forecast (for the Year 2021) = 214.073833
- Second Year Forecast (for the Year 2022) = 273.903833
- Third Year Forecast (for the Year 2023) = 236.588833

○ The Yearly Forecasts of the Decomposition Method:

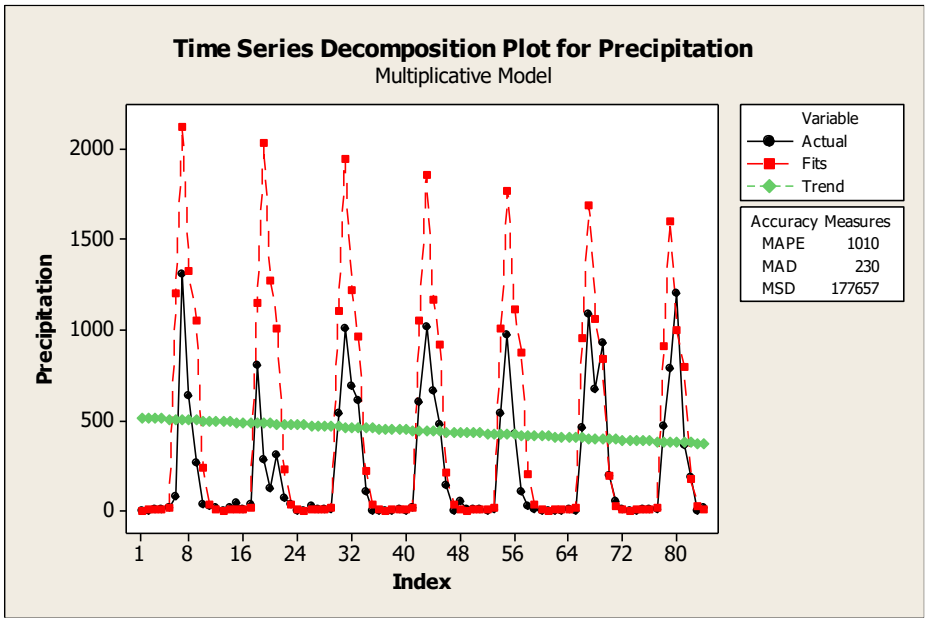
- First Year Forecast (for the Year 2021) = 271.015
- Second Year Forecast (for the Year 2022) = 283.78
- Third Year Forecast (for the year 2023) = 296.5425

IV. Conclusion:

- The analysis of the rainfall time series data has provided valuable insights into the underlying patterns and forecasted values of the phenomenon.
- First, in terms of the trend component, it was observed that there is a small increase in the long-term trend of the rainfall data, indicating a potential upward movement over time. However, it is important to note that the trend is relatively stable and not significant.
- The Box and Jenkins analysis was conducted to select an appropriate model for forecasting. After considering various combinations, the ARMA (2,1) model was identified as the most suitable model based on diagnostic tests and accuracy measures. The estimated coefficients of the model were found to be significant, indicating their contribution to explaining the variations in the data. The model was found to be both stationary and invertible, satisfying the necessary conditions.
- The forecasts obtained from the ARMA (2,1) model were compared to the forecasts derived from the decomposition method. It was observed that there were slight differences between the two sets of forecasts.
- In real-life applications, the findings of this analysis can have significant implications. Accurate and reliable forecasting of rainfall patterns is crucial for various sectors, including agriculture, water resource management, and urban planning. The ability to predict future rainfall values can help in making informed decisions, such as crop planning, water allocation, and infrastructure development. The ARMA (2,1) model, along with the decomposition method, provides a framework for understanding and forecasting rainfall patterns, which can contribute to effective decision-making and planning in these sectors.

V. Appendix:

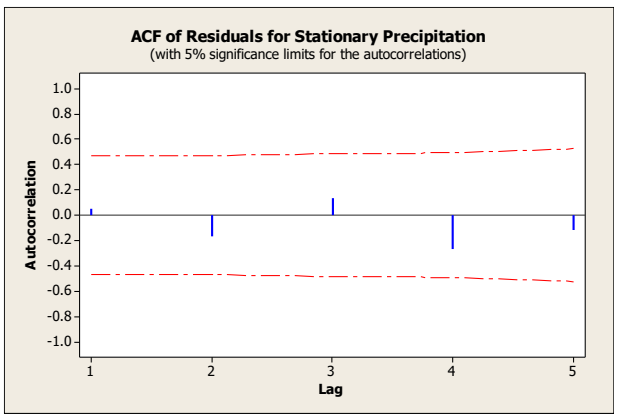
× Decomposition Using Multiplicative Model:

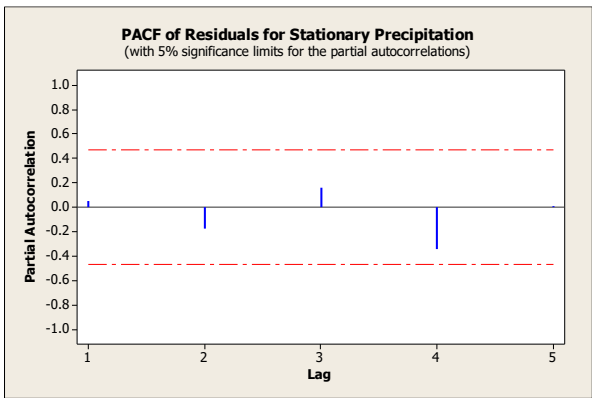


× Other Tried Models:

- ***MA (1):***

ACF and PACF:





```
Final Estimates of Parameters
Type      Coef  SE Coef     T      P
MA   1  0.6006   0.1909  3.15  0.005


Number of observations:  20
Residuals:    SS =  49288.3 (backforecasts excluded)
              MS =  2594.1  DF = 19


Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag            12  24  36  48
Chi-Square    8.2   *   *   *
DF             11   *   *   *
P-Value     0.691   *   *   *
```

- ***ARMA (2,2):***

```
Final Estimates of Parameters
Type       Coef  SE Coef      T      P
AR   1  -0.9843   0.3502  -2.81  0.013
AR   2  -0.6211   0.3619  -1.72  0.105
MA   1  -0.7881   0.4139  -1.90  0.075
MA   2   0.0918   0.4475   0.21  0.840


Number of observations:  20
Residuals:    SS =  35890.4 (backforecasts excluded)
              MS =  2243.2  DF = 16


Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag            12  24  36  48
Chi-Square    8.2   *   *   *
DF               8   *   *   *
P-Value     0.411   *   *   *
```

- ***AR (1):***

```
Final Estimates of Parameters
Type       Coef  SE Coef      T      P
AR   1  -0.3329   0.2182  -1.53  0.144


Number of observations:  20
Residuals:    SS =  55044.2 (backforecasts excluded)
              MS =  2897.1  DF = 19
```

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag             12   24   36   48
Chi-Square    10.2    *    *    *
DF              11    *    *    *
P-Value      0.510    *    *    *
```

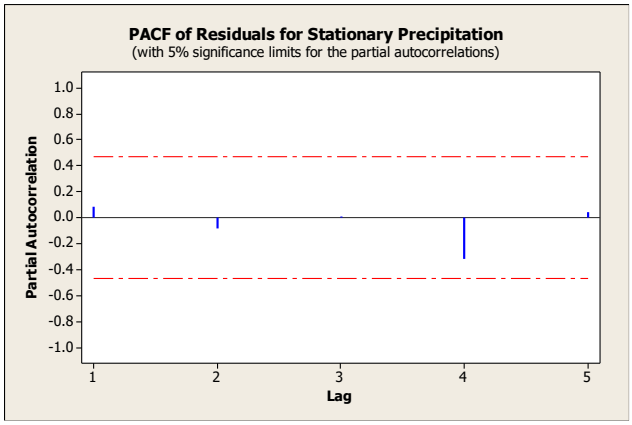
- **MA (2):**

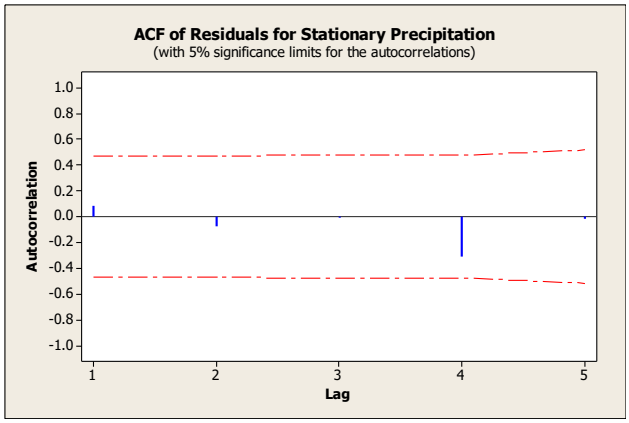
```
Final Estimates of Parameters
Type      Coef  SE Coef     T      P
MA   1  0.0304   0.1689  0.18  0.859
MA   2  0.8887   0.1741  5.11  0.000


Number of observations:  20
Residuals:    SS =  40350.2 (backforecasts excluded)
              MS =  2241.7  DF = 18


Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag             12   24   36   48
Chi-Square    12.7    *    *    *
DF              10    *    *    *
P-Value      0.239    *    *    *
```

- **AR (2):**






```
Final Estimates of Parameters
Type       Coef  SE Coef      T      P
AR   1  -0.4739   0.2102  -2.25  0.037
AR   2  -0.4993   0.2339  -2.13  0.047


Number of observations:  20
Residuals:    SS =  44142.9 (backforecasts excluded)
              MS =  2452.4  DF = 18
```

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag             12   24   36   48
Chi-Square     7.8    *    *    *
DF              10    *    *    *
P-Value     0.644    *    *    *
```

VI. List of References:

- Our data is from: https://www.kaggle.com/datasets/poojag718/rainfall-timeseries-data
- Sharma, P., and Kumar, A. "Trend Analysis of Long-term Rainfall Data in India." **Journal of Hydrology**, vol. 589, 2020, p. 125043.
- Wang, Z., et al. "A Time Series Model for Monthly Rainfall Prediction Using LSTM Networks." **IEEE Access**, vol. 7, 2019, pp. 44143-44153.
- Rana, S., Prasad, N. K., and Singh, D. K. "Analysis of Rainfall Trends, Variability and Prediction of Extreme Rainfall in the Betwa River Basin, India Using Statistical and Machine Learning Techniques." **Water Resources Management**, vol. 35, no. 2, 2021, pp. 529-550.
- Xie, M., Feng, Z., Guo, Y., and Zhang, Y. "Long-term Spatiotemporal Rainfall Analysis Using Hybrid Ensemble Empirical Mode Decomposition, Wavelet Transform, and Extreme Value Theory." **Atmospheric Research**, 2021.