Time Series Analysis Report on Monthly Rainfall Data of Delhi (1901-2021)

Table of Contents

-Sujal Yadav

1. Introduction

1.1 Background

Rainfall patterns are a critical component of the hydrological cycle, directly affecting agriculture, water resources, and urban planning. Understanding these patterns and predicting future rainfall is essential for mitigating the impacts of extreme weather events and ensuring sustainable development.

Delhi, the capital city of India, experiences significant variation in rainfall, impacting its large population and infrastructure. The city's climate is influenced by the monsoon season, which brings a substantial portion of the annual rainfall. Accurate forecasting of rainfall can help in various sectors, such as agriculture, water resource management, and urban infrastructure planning.

The availability of long-term historical rainfall data provides an opportunity to analyze trends, seasonal patterns, and anomalies. This study utilizes monthly rainfall data for Delhi from January 1901 to December 2021, covering over 120 years. This extensive dataset allows for a detailed analysis and robust modeling of rainfall patterns.

1.2 Objective

The primary objectives of this analysis are:

- To visualize and understand historical rainfall patterns in Delhi.
- To identify trends, seasonal effects, and other characteristics of the rainfall data.
- To develop and evaluate time series models for forecasting future rainfall.
- To provide insights that can aid in planning and decision-making processes.

1.3 Significance of Study

Rainfall forecasting plays a crucial role in various sectors:

- Agriculture: Helps farmers plan irrigation and crop selection. Timely and accurate rainfall forecasts can reduce crop losses and optimize the use of water resources.
- Water Resources Management: Aids in reservoir management and flood control. Predicting rainfall helps in the efficient allocation and conservation of water resources, ensuring adequate supply during dry periods and preventing overflow during heavy rains.
- Urban Planning: Assists in drainage design and managing urban water supplies. Rainfall predictions are essential for designing infrastructure that can handle heavy rainfall and prevent urban flooding.
- Disaster Management: Enables early warning systems for floods and droughts. Accurate forecasts can help in the preparation and implementation of measures to mitigate the impacts of extreme weather events.

Overall, the study aims to provide a comprehensive understanding of rainfall patterns and reliable forecasting models to support various sectors in Delhi.

2. Data Description

2.1 Data Source

The dataset used in this analysis was sourced from https://data.opencity.in/dataset/delhi-rainfall-data. It contains monthly rainfall measurements for Delhi from January 1901 to December 2021. This dataset provides a long-term perspective on rainfall patterns in Delhi.

2.2 Data Structure

The dataset comprises the following columns:

- **Month**: The date of the observation (formatted as YYYY-MM).
- **Rainfall**: The amount of rainfall recorded in millimeters.

Below is a sample of the dataset:

Month	Rainfall (in mm)
31-01-1901	34.54987216
28-02-1901	7.518799543
31-03-1901	4.8116045
30-04-1901	0
31-05-1901	0.671333253
30-06-1901	29.34695841

2.3 Data Summary

The dataset consists of 1452 observations, spanning over 120 years. Descriptive statistics for the rainfall data are as follows:

- **Mean: 52.59 mm**
  This indicates a moderate level of rainfall on average.

- **Median: 12.13 mm**
  The median rainfall is significantly lower than the mean, suggesting that the distribution of rainfall is skewed. Many months experience rainfall below the average, while a few months with extremely high rainfall increase the mean.

- **Standard Deviation: 89.59 mm**
  The high standard deviation indicates significant variability in monthly rainfall. This large spread implies that some months receive very little rain while others receive a lot, leading to a high degree of fluctuation.

- **Minimum: 0 mm**
  There are months with no rainfall at all, highlighting periods of complete dryness.

- **Maximum: 591.5991 mm** on 31st August, 1995
  This extreme value points to occasional but intense rainfall events, which can significantly affect the mean and standard deviation.

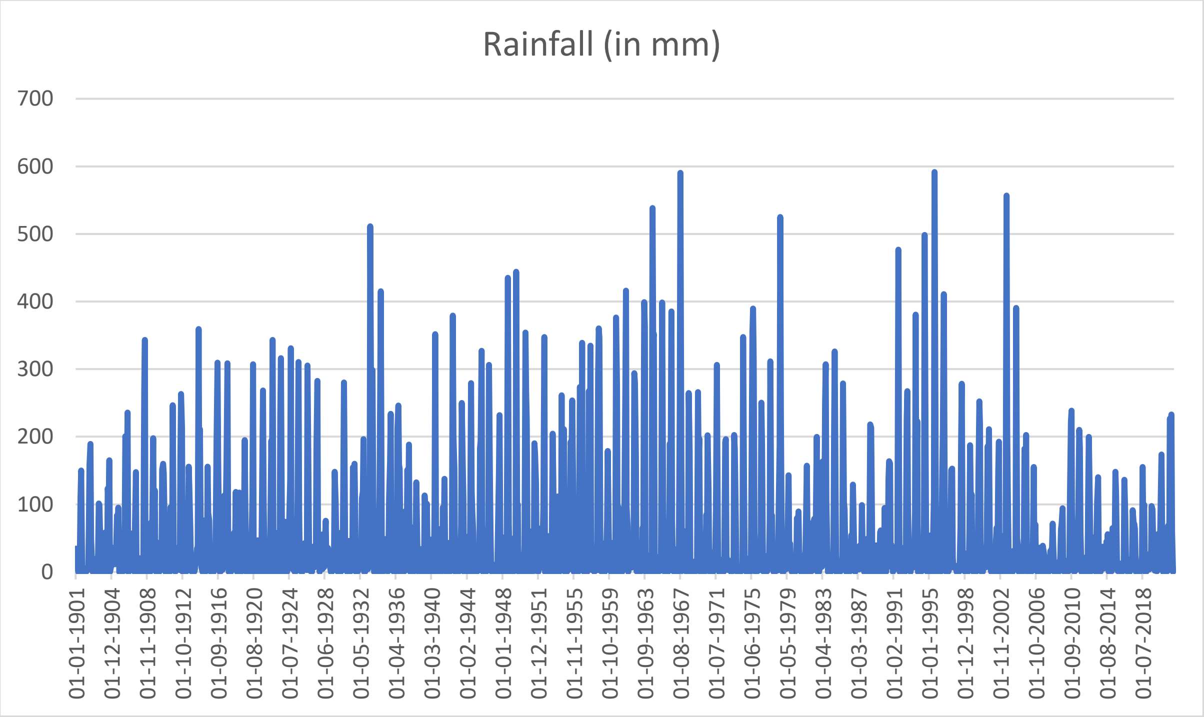
- **Coefficient of Variation: 1.703**
  This high CV underscores the high level of dispersion in the data relative to the mean, indicating that monthly rainfall in Delhi is highly inconsistent and unpredictable.

The data provides a comprehensive view of the rainfall trends and patterns over an extended period. Given the high variability and the occurrence of extreme values, it's likely that the data exhibits seasonal patterns, with certain months or seasons experiencing significantly more rainfall than others. Monsoon seasons, for example, could contribute to the higher values observed.

2.4 Data Visualization

A preliminary visualization of the data helps in understanding the distribution and trends. The plot below shows the monthly rainfall data from 1901 to 2021.

(Using EXCEL)



3. Methodology

3.1 Data Preprocessing

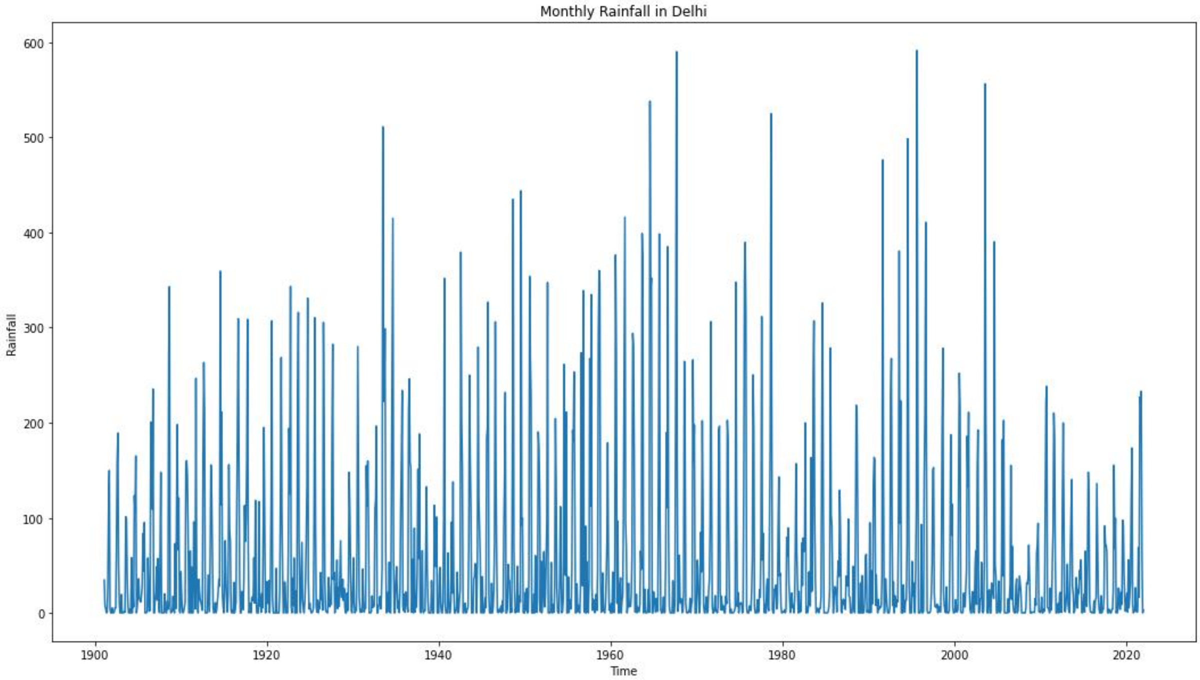
Date Indexing

The Month column was converted to a datetime format and set as the index for time series analysis. This step is crucial for leveraging time-based indexing and resampling capabilities.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Time Series Plot

A time series plot of the monthly rainfall data was generated to visualize trends and seasonal patterns. This plot provides an overview of the data distribution over time.



Trends:

1. **Intermittent Peaks**: The plot reveals several sharp peaks throughout the period, indicating occasional months with very high rainfall. These peaks do not show a consistent upward or downward trend over the long term, suggesting that extreme rainfall events have been sporadic.

2. **Fluctuations**: There is significant fluctuation in the rainfall amounts from month to month and year to year. This indicates high variability in Delhi's monthly rainfall over the years.

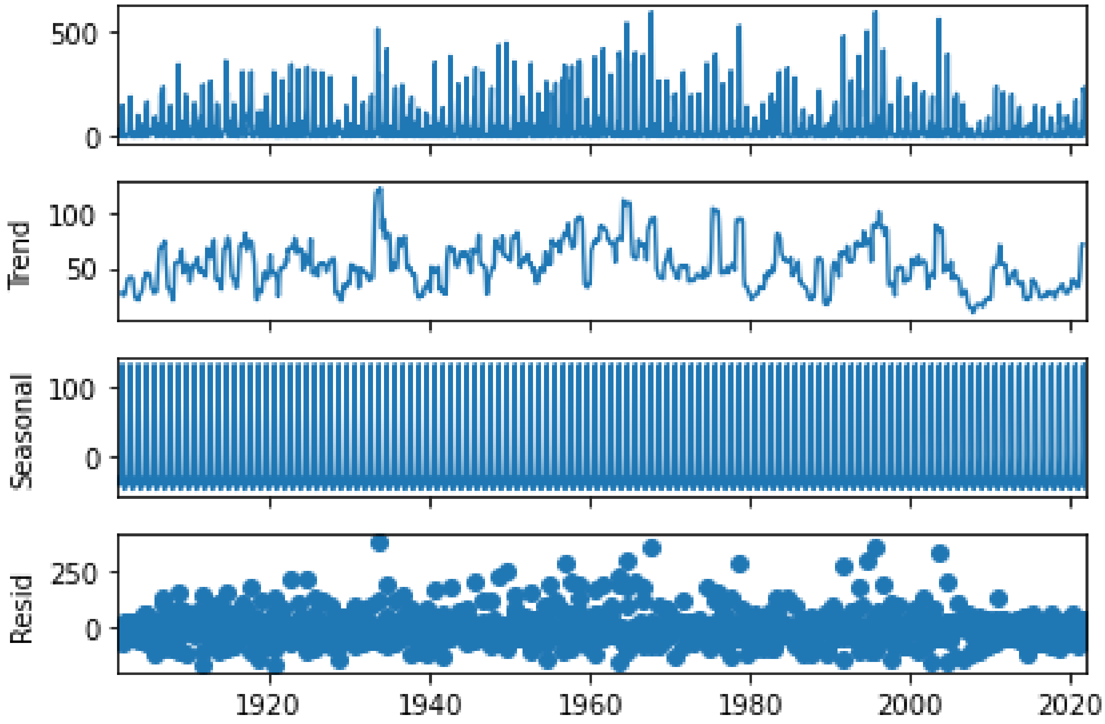
3. **No Clear Long-Term Trend**: The overall pattern does not exhibit a clear increasing or decreasing trend in rainfall over the 120 years. Instead, the data points are highly scattered, reflecting the irregular nature of rainfall.

Seasonal Patterns:

1. **Monsoon Seasonality**: The presence of clusters of high rainfall values at regular intervals suggests a seasonal pattern, likely corresponding to the monsoon season, typically occurring between June and September in Delhi. This can be inferred from the repeated occurrence of peaks within the same months each year.
2. **Dry Periods**: There are evident periods with very low or no rainfall, which can be attributed to the dry seasons in Delhi. These periods typically occur in the winter and early summer months.
3. **Annual Cycles**: The data shows annual cycles of rainfall with a notable increase during the monsoon months followed by relatively dry periods. This cyclical pattern aligns with the typical climatic behavior observed in Northern India.

3.2.2 Seasonal Decomposition

Seasonal decomposition using additive models was performed to separate the series into trend, seasonal, and residual components. This decomposition helps in understanding the underlying patterns and variations in the data.



- **Original Series** shows significant variability with intermittent high rainfall peaks, particularly noticeable during the monsoon season.

- **Trend Component** smooths out the short-term fluctuations to reveal the underlying long-term pattern. The trend line indicates varying rainfall over the years without a clear upward or downward long-term trend. Some periods show higher average rainfall, such as around the mid-20th century, while other periods exhibit lower averages. There are also noticeable fluctuations in the trend, reflecting periods of relatively higher or lower rainfall. This highlights the importance of understanding decadal or multi-decadal climate variability when planning for water resources and flood management.
- **Seasonal Component** captures the repeating patterns or cycles within each year. The seasonal pattern is highly regular and consistent, with peaks corresponding to the monsoon months and troughs during the dry seasons. This component confirms the strong seasonality in Delhi's rainfall, driven by the monsoon cycle. This information is crucial for agricultural planning, water storage management, and preparing for monsoon-related flooding.
- **Residual Component** which represents the irregularities or noise after removing the trend and seasonal effects. The residuals show random fluctuations around zero, indicating that the trend and seasonal components have effectively captured the main patterns in the data. There are some larger residuals, likely corresponding to anomalous or extreme rainfall events not explained by the trend or seasonal components. Analysing these residuals can help in identifying unusual patterns or extreme events that might require special attention.

3.3 Model Selection

3.3.1 Stationarity Check

The Augmented Dickey-Fuller (ADF) test was used to check the stationarity of the series. Stationarity is a key assumption for many time series models, indicating that the statistical properties of the series do not change over time.

H_0: The time series is not stationary.

H_1: The time series is stationary.

- **ADF Statistic**: -4.839032324456457
- **p-value**: 4.5767923526882895e-05

Given the p-value is much smaller than 0.01, we reject the null hypothesis at 1% level of significance. This implies that the monthly rainfall data in Delhi from January 1901 to December 2021 is stationary. Thus, it does not need any transformation and thus non-seasonal difference order d is zero.

The results of the Augmented Dickey-Fuller (ADF) test on the seasonally differenced data are as follows:

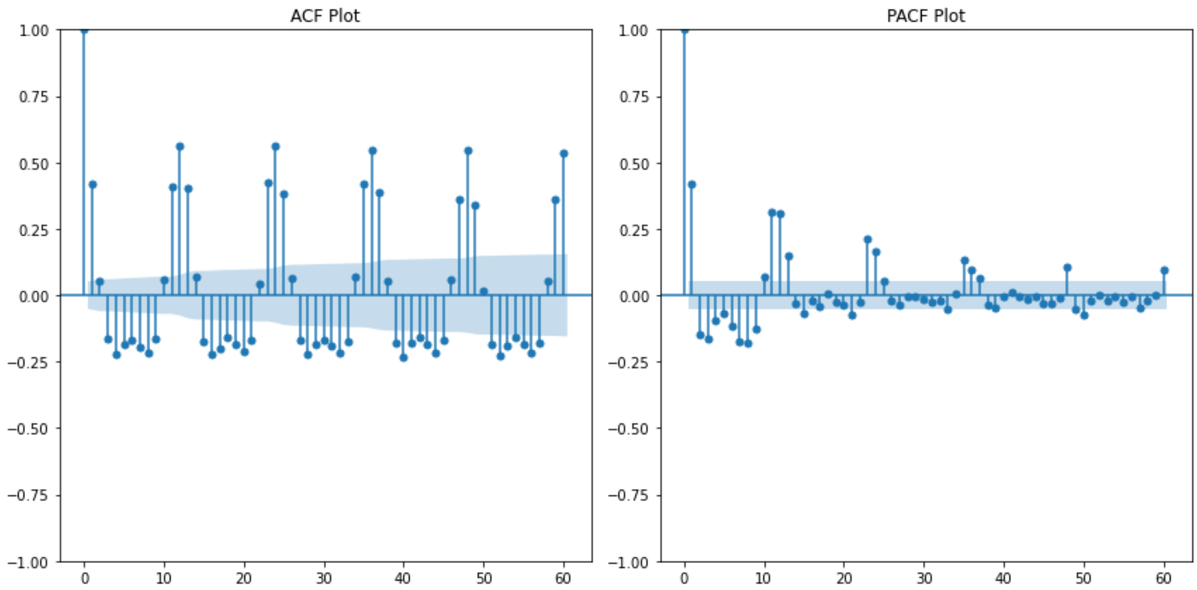
- **ADF Statistic:** -12.738705881282142
- **p-value:** 9.042418941416352e-24

Given the p-value is much smaller than 0.01, we reject the null hypothesis at 1% level of significance. This indicates that the seasonally differenced data is stationary. Thus, the seasonal differencing order can be **D = 0 or 1**.

3.3.2 Model Identification

SARIMA Models

Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots were used to identify the appropriate orders of AR and MA components. These plots help in determining the lag values for the models.



Since the PACF plot cuts off after a few lags, it suggests that non-seasonal AR component p could be anything between 0 to 5 and same goes for the non-seasonal MA Component q by looking at the ACF plot. Regarding the seasonal AR component P, the PACF plot shows that P could be chosen from 1 to 3 and the ACF plot shows that the seasonal MA component Q could be chosen from 1 to 4.

Holt-Winters Exponential Smoothing:

Holt-Winters exponential smoothing was also identified as a suitable model for the monthly rainfall data of Delhi. This method is particularly effective for data with strong seasonal patterns. Holt-Winters exponential smoothing is advantageous because it can accommodate both seasonal variations and trends, making it an appropriate choice for time series data like monthly rainfall, which exhibits clear seasonality and potential long-term trends.

3.4 Model Fitting

3.4.1 SARIMA Models

SARIMA models were fitted to account for both the autoregressive and moving average components along with differencing to achieve stationarity. Several models were tested. (The tested model evaluation metrics are in the enclosed Excel file with same filename.) Each model's parameters were chosen based on the ACF and PACF plots.

3.4.2 Holt-Winters Exponential Smoothing

Holt-Winters Exponential Smoothing models were also considered for comparison. These models are particularly effective for capturing trends and seasonality in the data.

3.5 Model Evaluation

Based on the AIC, BIC, Ljung Box test on residuals, MAE, MSE and RMSE of the tested models, we evaluated and selected the following models to be used for forecasting:

1. **Model 1:** SARIMA (1, 0, 1) (3, 0, 1, 12)
2. **Model 2:** SARIMA (1, 0, 2) (1, 1, 1, 12)
3. **Model 3:** SARIMA (3, 0, 2) (2, 1, 1, 12)
4. **Model 4:** SARIMA (3, 0, 3) (1, 0, 1, 12)
5. **Model 5:** SARIMA (2, 0, 2) (2, 0, 0, 12) (Auto-selected by auto_arima in Python)
6. Holt-Winters Exponential Smoothing

(Showing the following analysis only for the selected models)

3.5.1 Residual Analysis

Residuals of the fitted models were analysed to check for any patterns or autocorrelations. Residual analysis helps in validating the model assumptions.

3.5.2 Ljung-Box Test

The Ljung-Box test was performed on the residuals to check for the absence of autocorrelation. A significant p-value indicates that the residuals are white noise, validating the model.

3.5.3 Forecast Accuracy

Model accuracy was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide a quantitative measure of the model's predictive performance.

4. Results

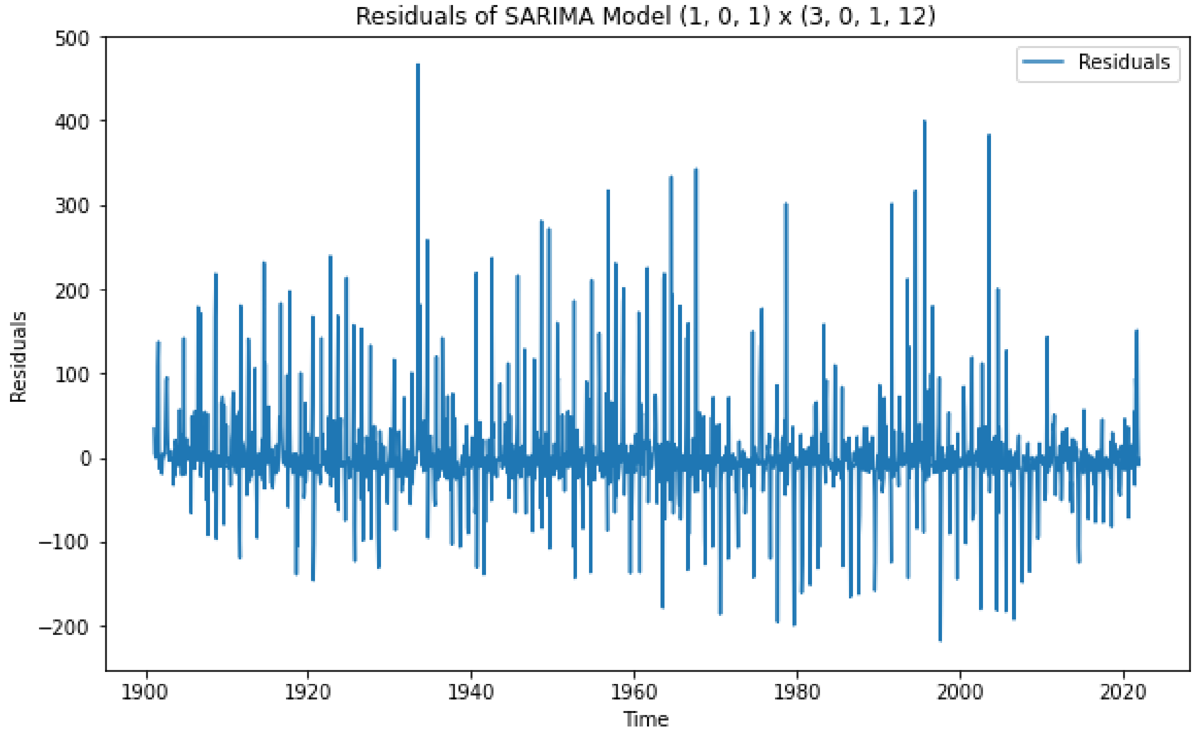
4.1 Model Comparison

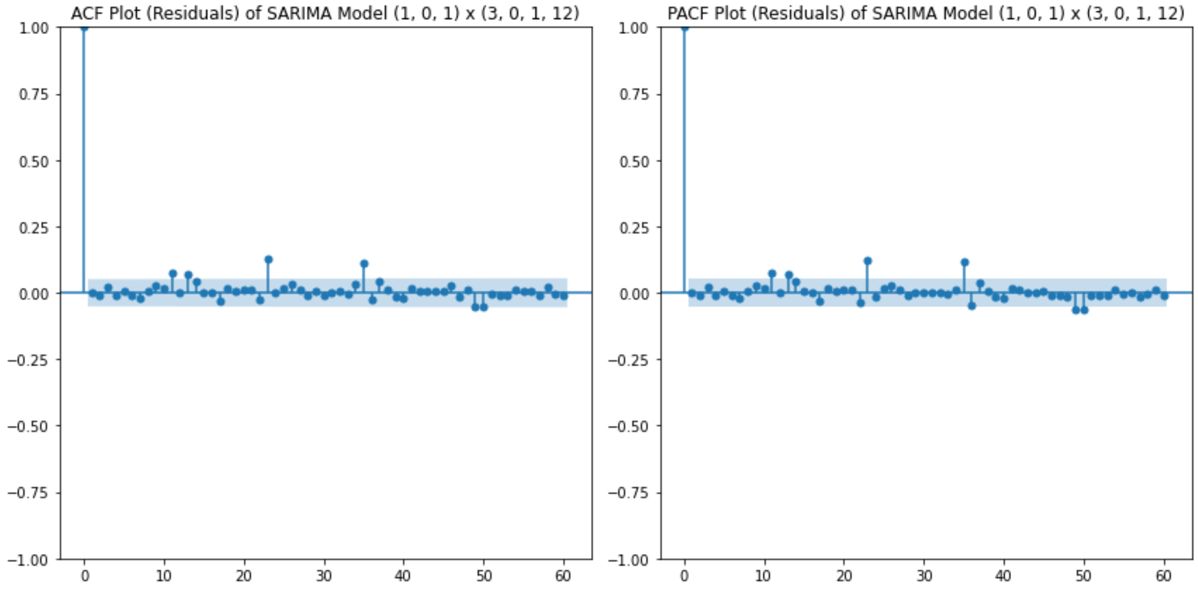
A summary of the AIC, BIC, HQIC, log-likelihood, and forecast accuracy metrics for the different models considered.

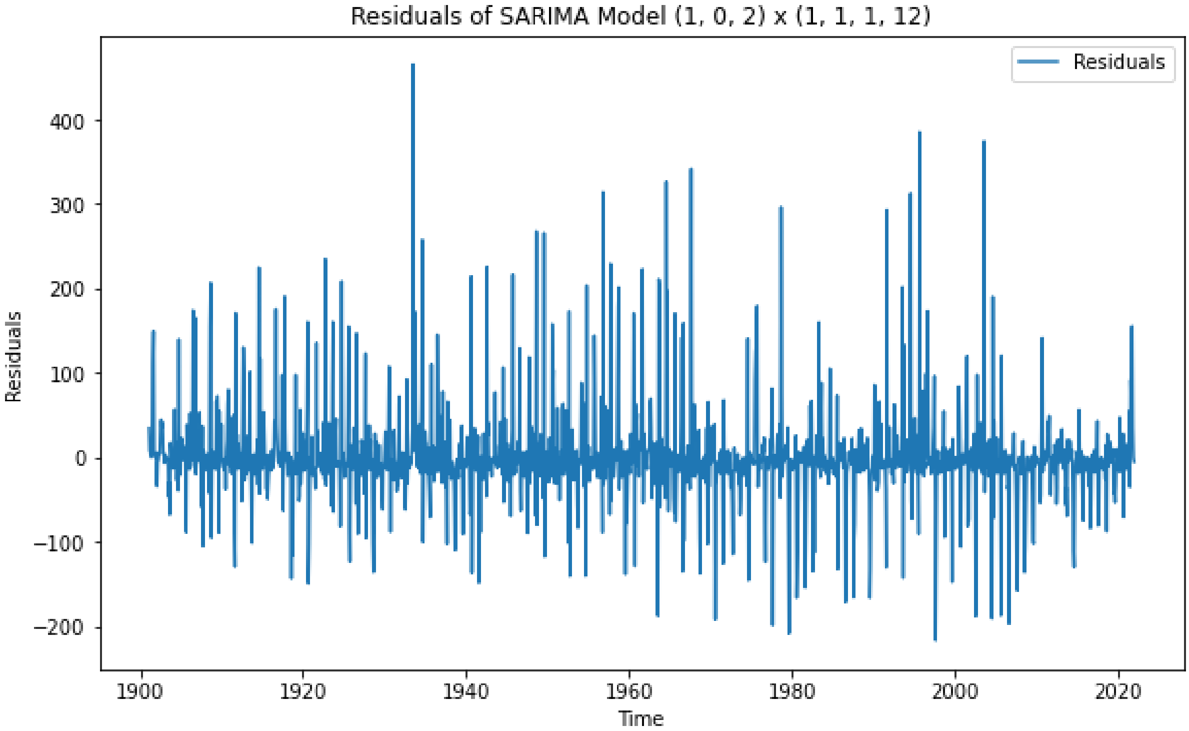
Model Number	Order	Seasonal Order	AIC	BIC	HQIC	Log-Likelihood	Ljung-Box p-value	MAE	MSE	RMSE
1	(1, 0, 1)	(3, 0, 1, 12)	16107.59	16144.56	16121.38	-8046.80	0.0097	33.7	3779.96	61.48
2	(1, 0, 2)	(1, 1, 1, 12)	15967.78	15999.41	15979.59	-7977.89	0.04	34.18	3758.15	61.30
3	(3, 0, 2)	(2, 1, 1, 12)	15976.94	16024.39	15994.65	-7979.47	0.05	34.44	3764.99	61.36
4	(3, 0, 3)	(1, 0, 1, 12)	16100.18	16147.7	16117.92	-8041.09	0.04	34.34	3752.41	61.26
5	(2, 0, 2)	(2, 0, 0, 12)	16387.13	16429.38	N/A	-8185.57	0.00	40.83	4599.84	67.82
6	Exponential Smoothing	Additive	12003.38	12087.87	N/A	N/A	0.0016	36.53	3807.59	61.70

4.2 Residual Analysis

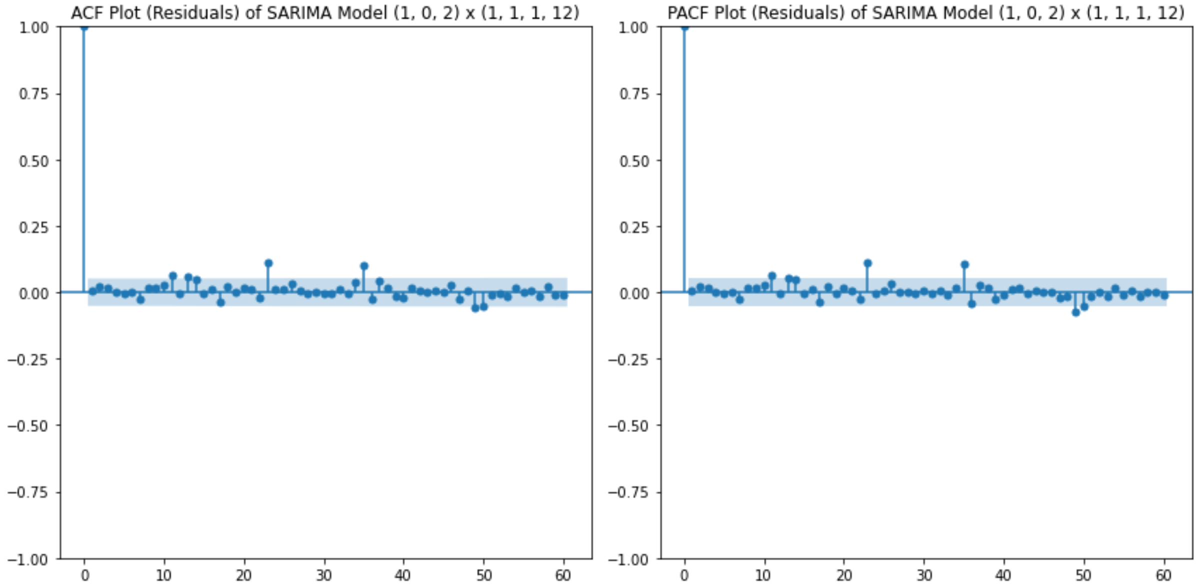
The residual analysis for the models showed no significant autocorrelation, indicating that the models adequately captured the underlying patterns in the data.





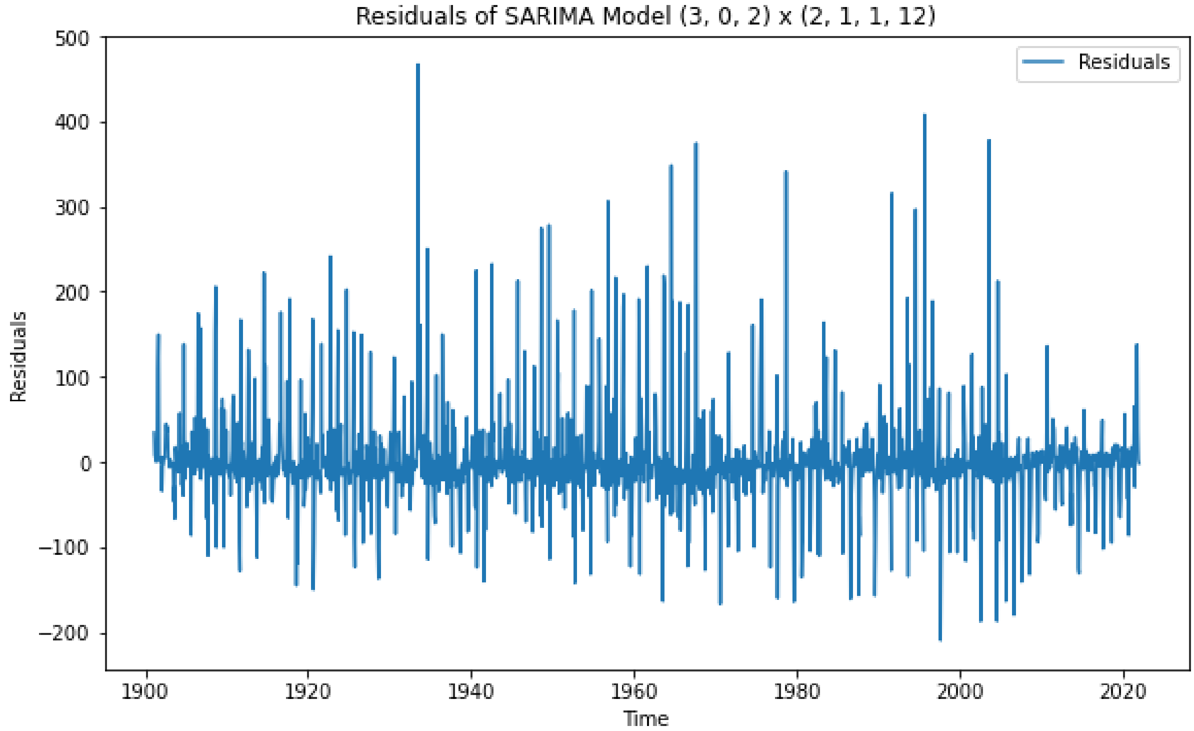


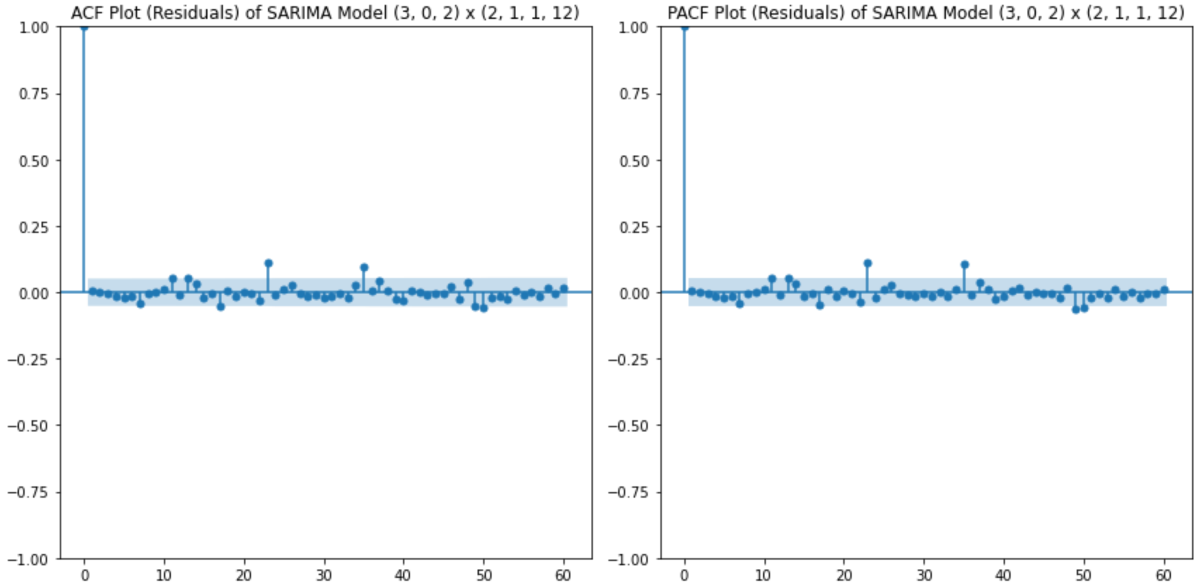
Residuals of SARIMA Model (1, 0, 2) x (1, 1, 1, 12)

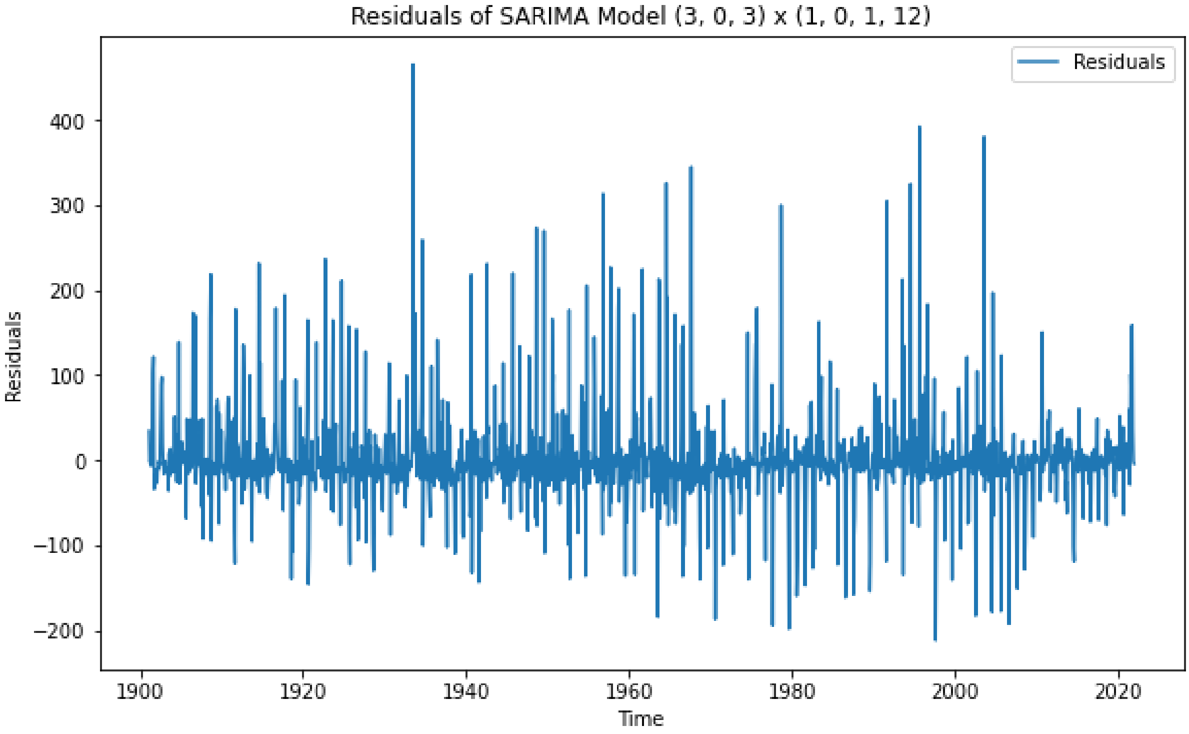


ACF Plot (Residuals) of SARIMA Model (1, 0, 2) x (1, 1, 1, 12)

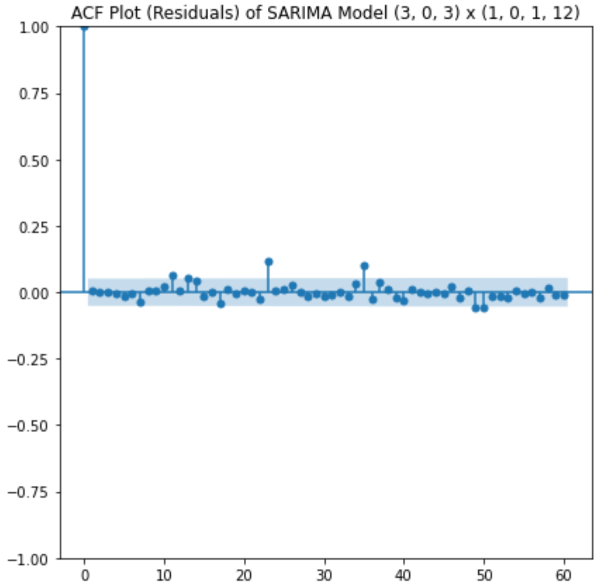
PACF Plot (Residuals) of SARIMA Model (1, 0, 2) x (1, 1, 1, 12)


Residuals of SARIMA Model (3, 0, 2) x (2, 1, 1, 12)
Residuals
Residuals
Time
500
400
300
200
100
0
−100
−200
1900
1920
1940
1960
1980
2000
2020

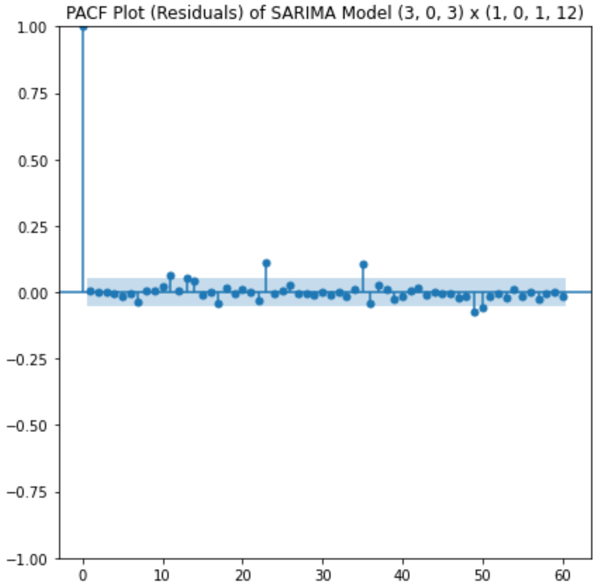

ACF Plot (Residuals) of SARIMA Model (3, 0, 2) x (2, 1, 1, 12)
PACF Plot (Residuals) of SARIMA Model (3, 0, 2) x (2, 1, 1, 12)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60



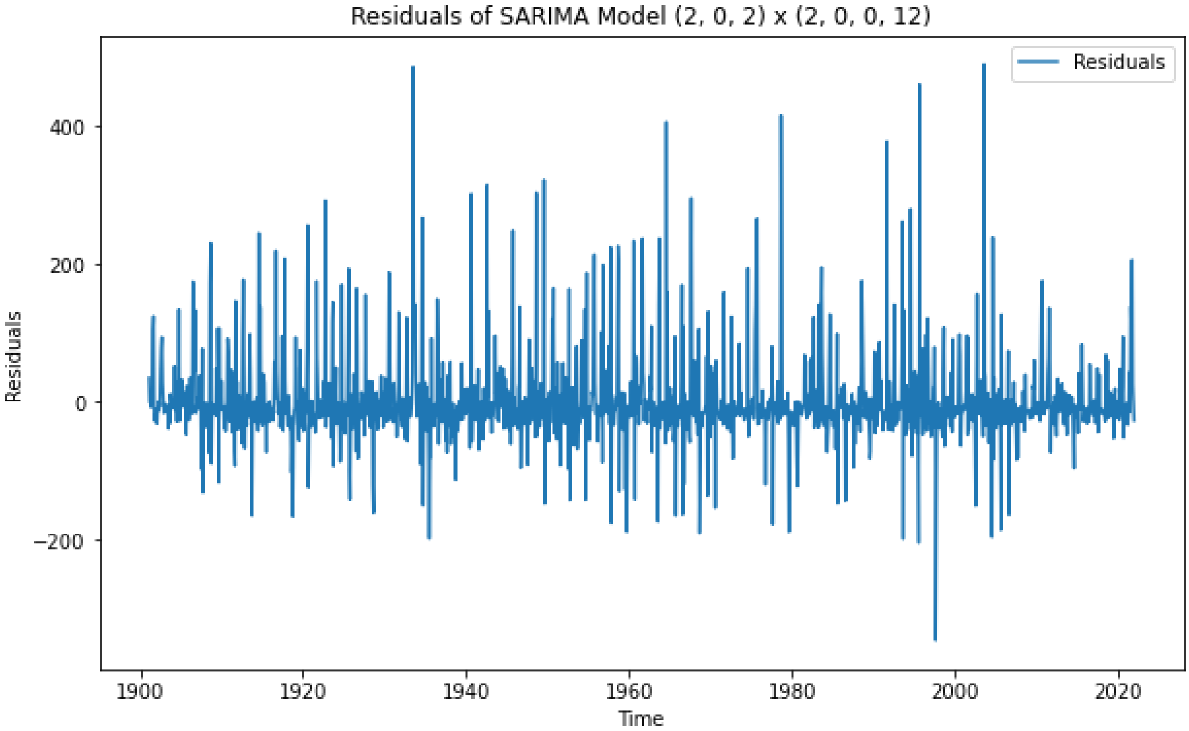
Residuals of SARIMA Model (3, 0, 3) x (1, 0, 1, 12)

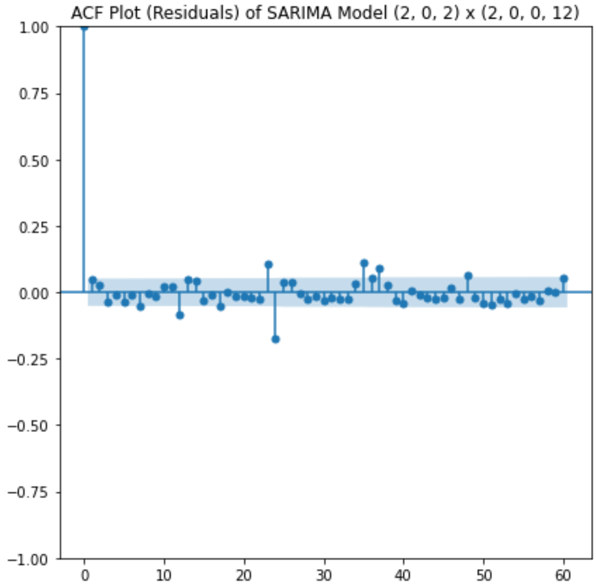


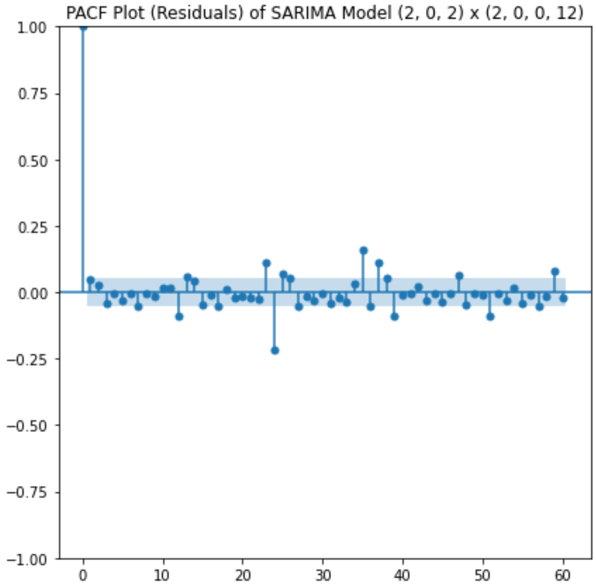
ACF Plot (Residuals) of SARIMA Model (3, 0, 3) x (1, 0, 1, 12)

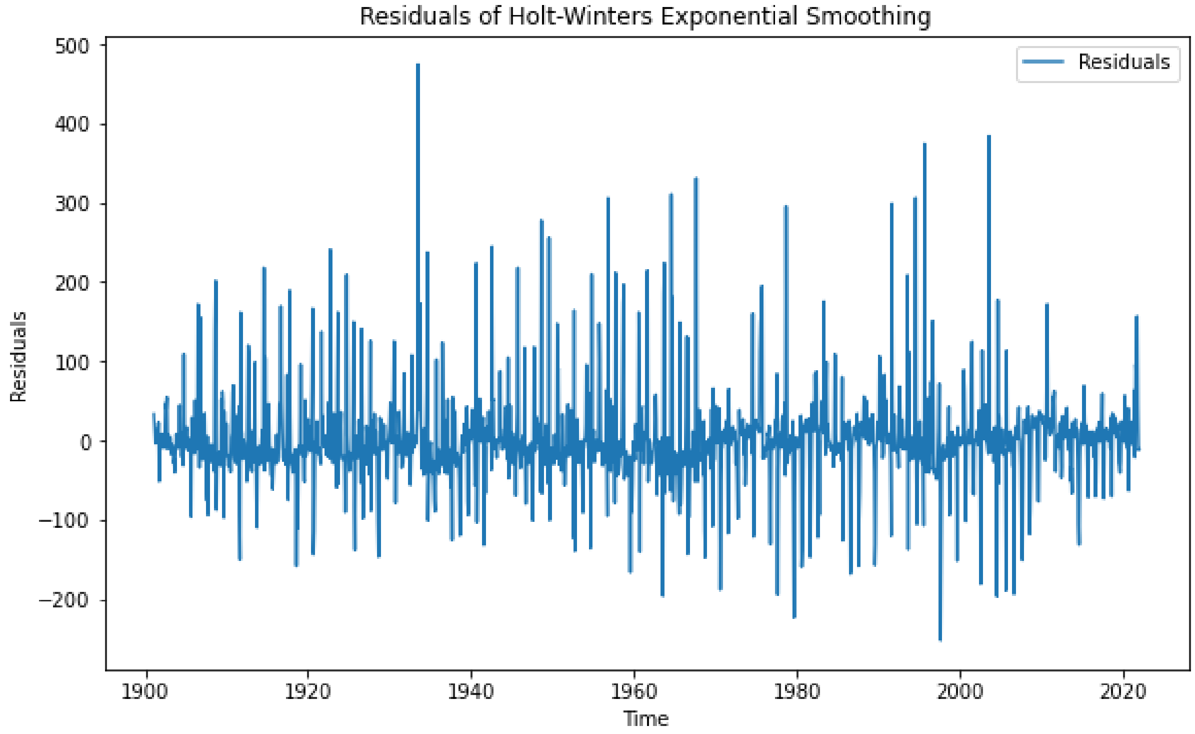


PACF Plot (Residuals) of SARIMA Model (3, 0, 3) x (1, 0, 1, 12)

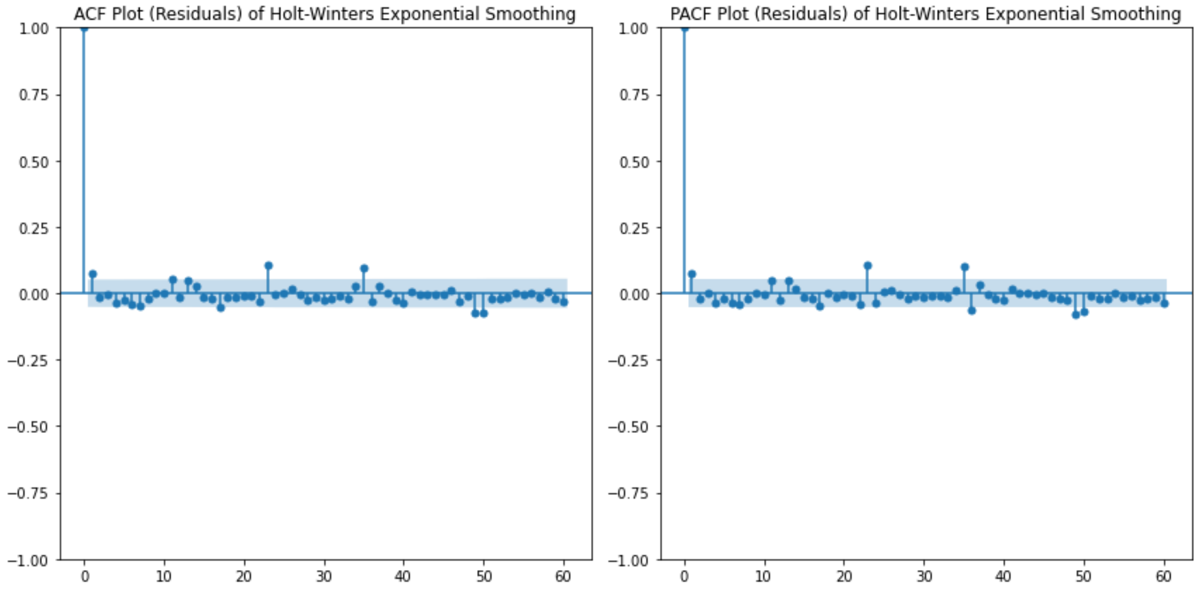

Residuals of SARIMA Model (2, 0, 2) x (2, 0, 0, 12)
Residuals
Residuals
400
200
0
−200
1900
1920
1940
1960
1980
2000
2020
Time


ACF Plot (Residuals) of SARIMA Model (2, 0, 2) x (2, 0, 0, 12)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60


PACF Plot (Residuals) of SARIMA Model (2, 0, 2) x (2, 0, 0, 12)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60



Residuals of Holt-Winters Exponential Smoothing



ACF Plot (Residuals) of Holt-Winters Exponential Smoothing

PACF Plot (Residuals) of Holt-Winters Exponential Smoothing

5. Forecasting

5.1 Forecasting with SARIMA

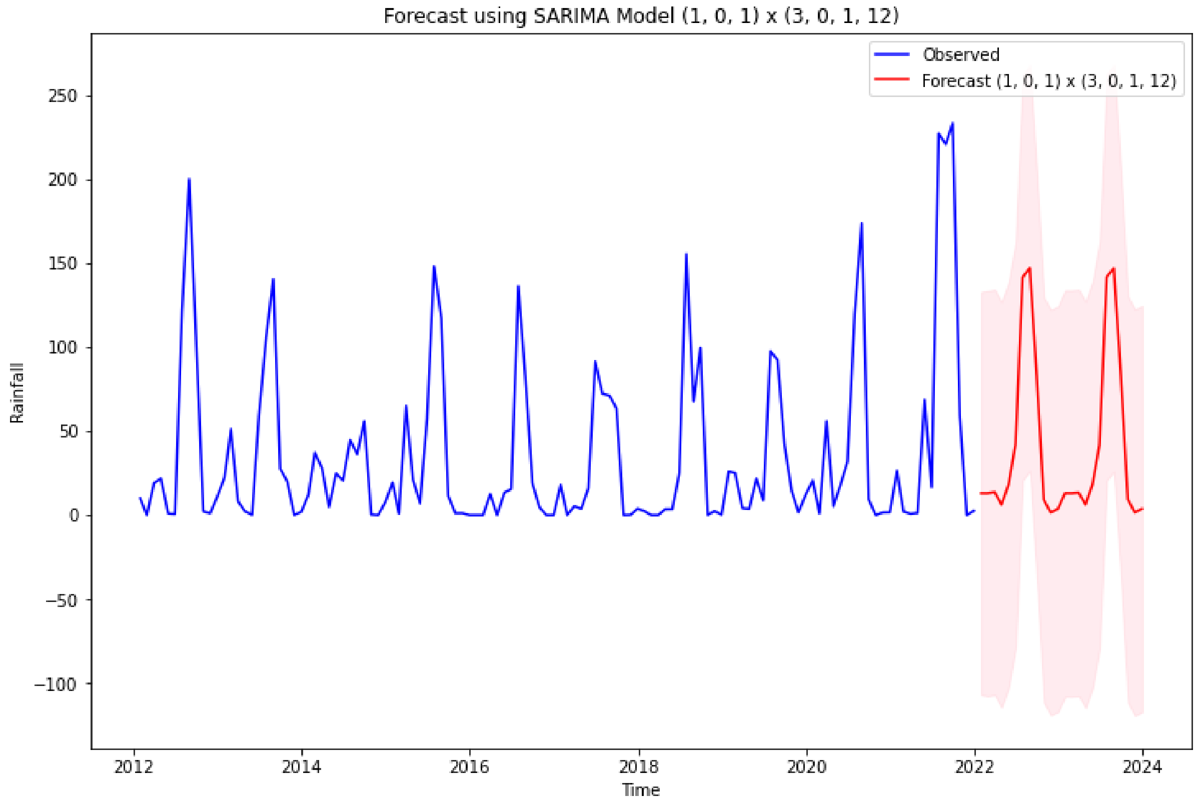
The SARIMAX model was used to generate forecasts for the year 2022. The model parameters were chosen based on the lowest AIC, BIC, and HQIC values. The forecast accuracy was evaluated using MAE, MSE, and RMSE.

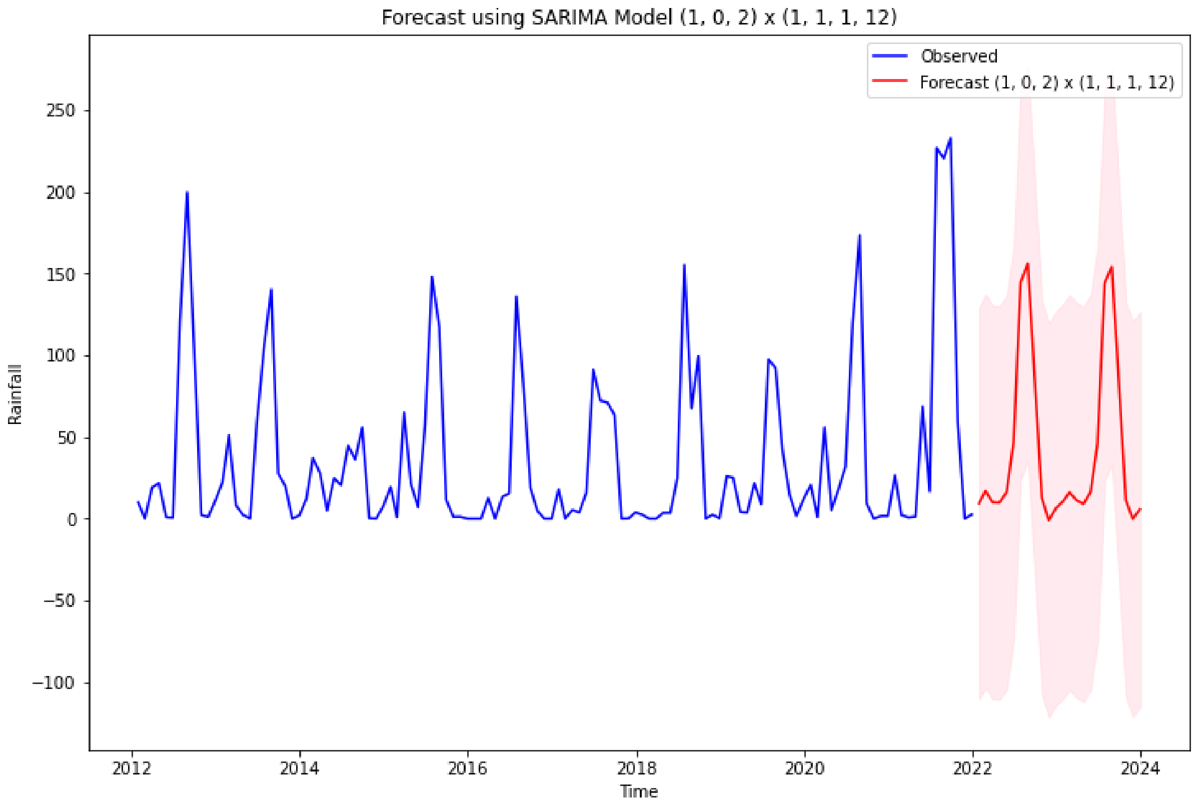
5.2 Forecasting with Exponential Smoothing

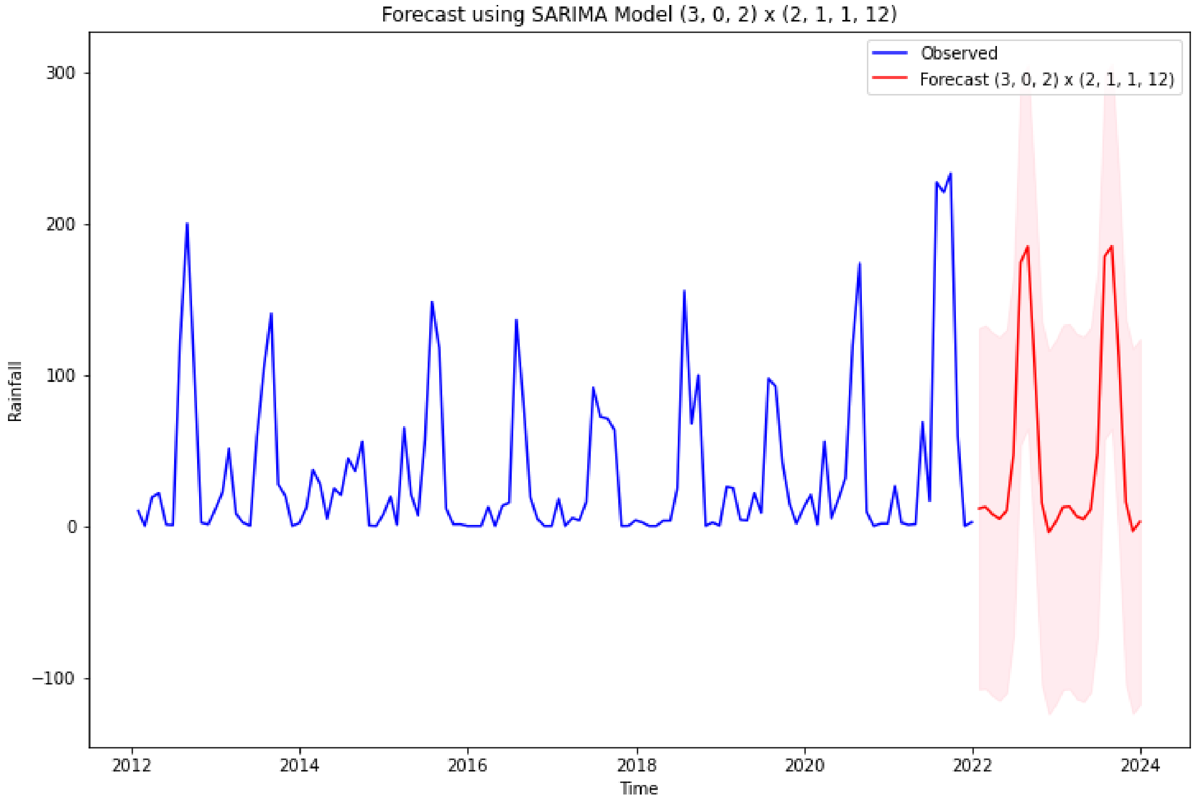
Similarly, forecasts for 2022 were generated using the Exponential Smoothing model. The forecast accuracy was evaluated using MAE, MSE, and RMSE.

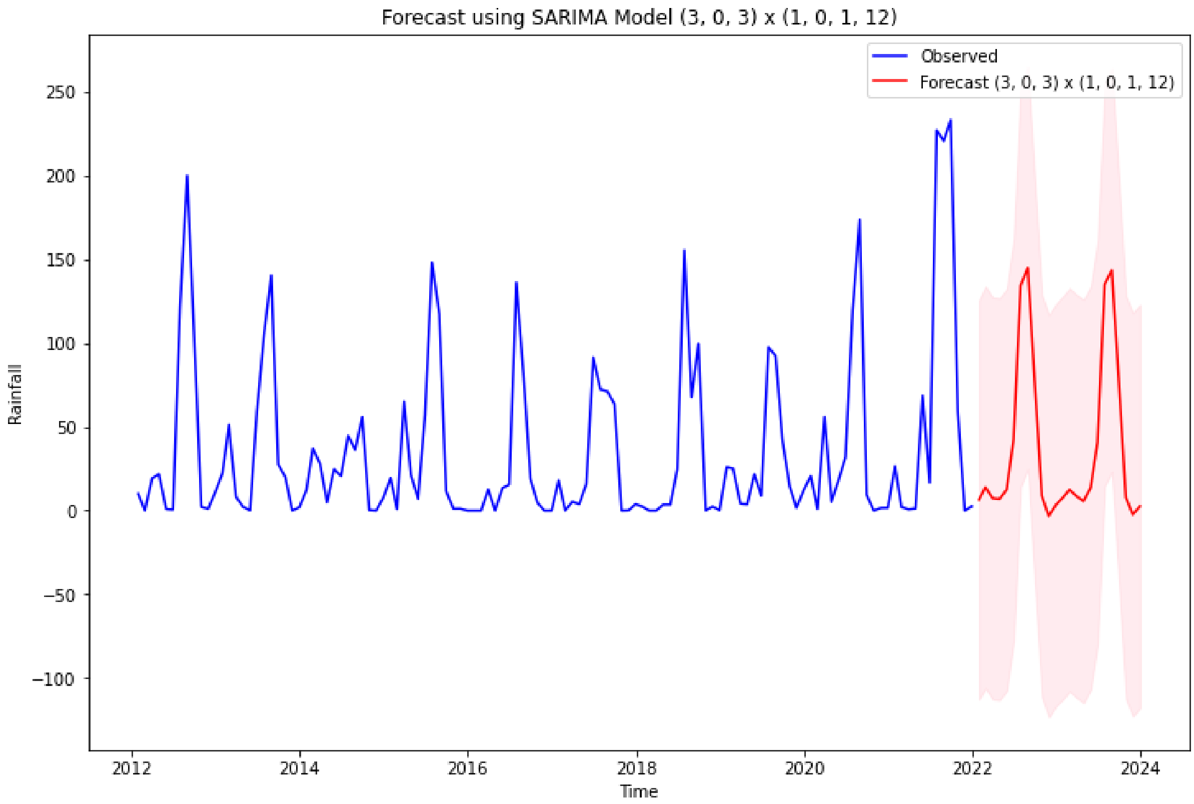
5.3 Visualizing Forecasts

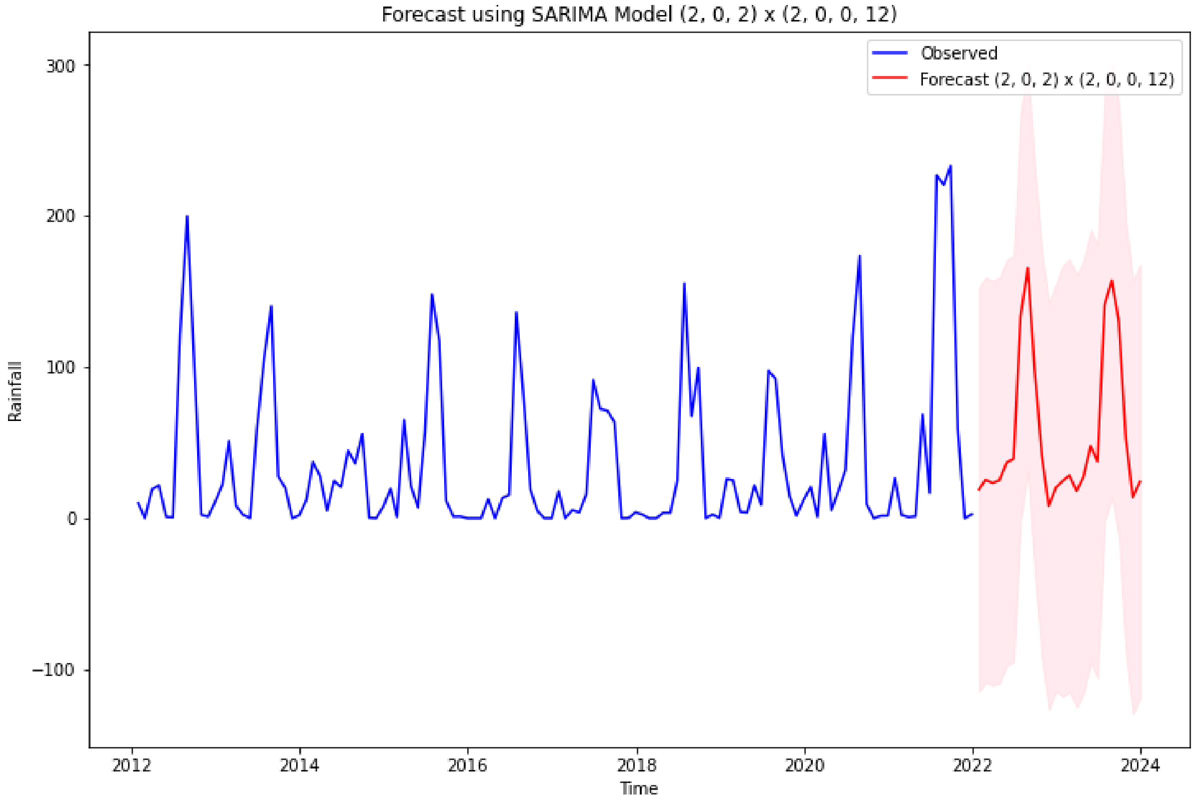
The forecasts values were plotted to visualize the performance of the models.

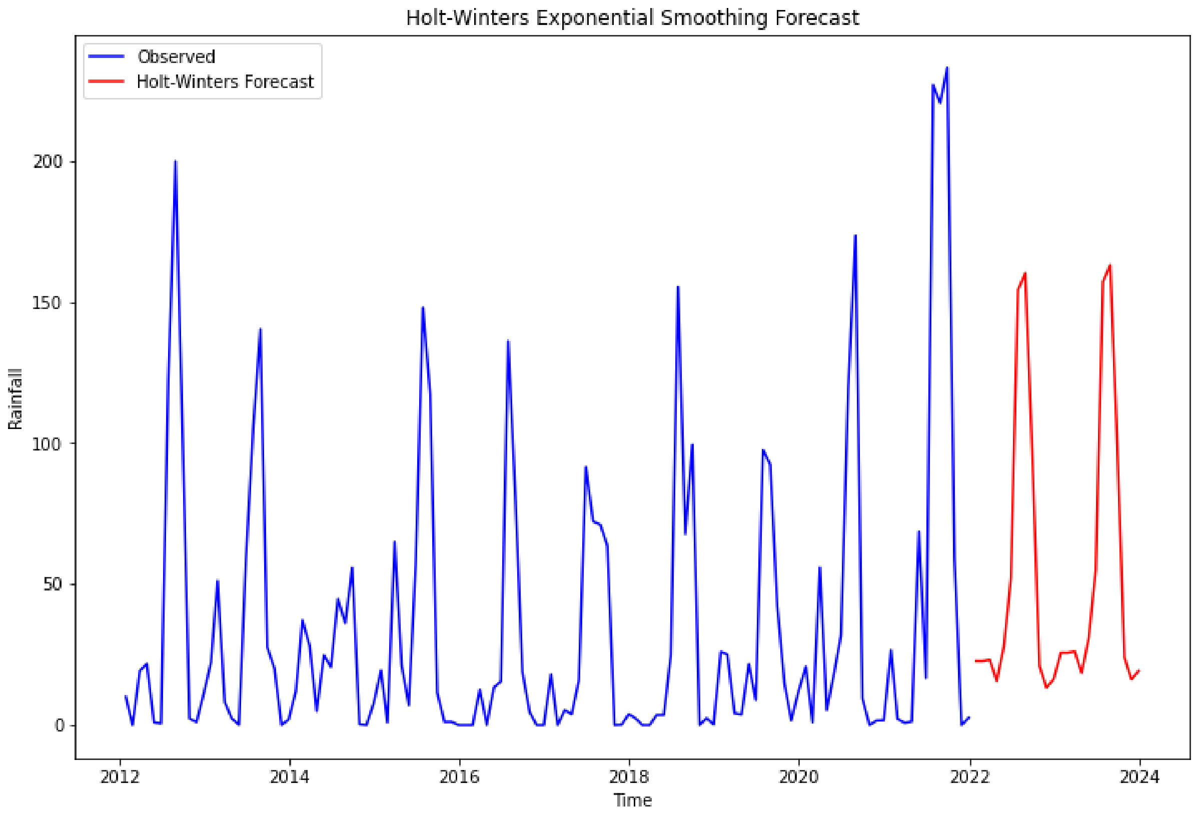



Forecast using SARIMA Model (1, 0, 2) x (1, 1, 1, 12)
Observed
Forecast (1, 0, 2) x (1, 1, 1, 12)
Rainfall
250
200
150
100
50
0
−50
−100
2012
2014
2016
2018
2020
2022
2024
Time


Forecast using SARIMA Model (3, 0, 2) x (2, 1, 1, 12)
Observed
Forecast (3, 0, 2) x (2, 1, 1, 12)
Rainfall
300
200
100
0
−100
2012
2014
2016
2018
2020
2022
2024
Time


Forecast using SARIMA Model (3, 0, 3) x (1, 0, 1, 12)
Observed
Forecast (3, 0, 3) x (1, 0, 1, 12)
250
200
150
100
50
0
−50
−100
Rainfall
2012
2014
2016
2018
2020
2022
2024
Time


Forecast using SARIMA Model (2, 0, 2) x (2, 0, 0, 12)
Observed
Forecast (2, 0, 2) x (2, 0, 0, 12)
300
200
100
0
−100
Rainfall
2012
2014
2016
2018
2020
2022
2024
Time


Holt-Winters Exponential Smoothing Forecast
Observed
Holt-Winters Forecast
Rainfall
200
150
100
50
0
2012
2014
2016
2018
2020
2022
2024
Time

6. Discussion

6.1 Interpretation of Results

The analysis revealed that both SARIMA and Exponential Smoothing models are effective in capturing the seasonal and trend components of the rainfall data. The SARIMA model slightly outperformed Exponential Smoothing in terms of accuracy metrics, making it a preferable choice for forecasting.

6.2 Implications for Policy and Planning

The findings of this study have significant implications for various sectors in Delhi:

- Agriculture: Accurate rainfall forecasts can help farmers optimize irrigation schedules and crop planning.
- Water Management: Forecasting can aid in reservoir management, ensuring adequate water supply during dry periods.
- Urban Planning: Rainfall predictions can inform the design and maintenance of urban drainage systems, reducing the risk of flooding.

6.3 Limitations and Future Work

While the models provided satisfactory forecasts, there are limitations to consider:

- Data Quality: The accuracy of forecasts is dependent on the quality and granularity of the historical data. Missing values and measurement errors can affect model performance.
- Model Complexity: More complex models, such as machine learning approaches, could potentially improve forecast accuracy. However, these models require more computational resources and expertise.
- Climate Change: The impact of climate change on rainfall patterns should be considered in future analyses. Climate models and additional environmental variables can provide a more comprehensive understanding of future rainfall trends.

Future work could involve incorporating additional climatic variables, exploring advanced forecasting techniques, and extending the analysis to other regions. Collaboration with meteorological experts and access to high-quality, real-time data can further enhance the accuracy and reliability of rainfall forecasts.

7. Conclusion

The time series analysis of monthly rainfall data in Delhi from 1901 to 2021 revealed significant seasonal patterns and trends. SARIMA and Exponential Smoothing models were used to forecast future rainfall. The models' performances were evaluated using various criteria, with SARIMA providing slightly better accuracy. These forecasts can aid in planning and decision-making processes in sectors affected by rainfall variability.

This analysis underscores the importance of robust statistical methods in understanding and predicting climatic variables. Further research can extend this work by incorporating additional climatic factors and exploring more advanced machine learning models for improved accuracy.

8. References

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

2. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

3. Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples. Springer.

4. [Online Documentation for `statsmodels`]
(https://www.statsmodels.org/stable/index.html)

5. [Python Data Science Handbook]
(https://jakevdp.github.io/PythonDataScienceHandbook/)