Forecasting of Monthly Rainfall in Dir (L) KP Pakistan With ARMA Models

Aimal Khan[1], Habib Ullah Khan[2] and Muhammad Ismail[3]

https://doi.org/10.62345/jads.2025.14.1.57

Abstract

This study focuses on forecasting monthly mean rainfall in District Dir (Lower) using various ARMA models to identify the most suitable approach for accurate rainfall prediction. Rainfall forecasting is a critical and intriguing area of study, and the Box-Jenkins methodology, which employs ARMA models, is highly effective for analyzing and forecasting time series with diverse patterns of variation. The ARMA modeling process involves defining, estimating, diagnosing, and forecasting stages, making it well-suited for this application. Based on the analysis, the ARMA (4,4) model emerged as the best fit for the dataset, evaluated using SIGMASQ, AIC, and SC criteria, which yielded the smallest values and confirmed through the Ljung-Box Q-test. Using the ARMA (4,4) model, the study successfully forecasted monthly mean rainfall in District Dir (lower) from June 2022 to May 2028.

Keywords: Ljung-Box Test, Forecast Accuracy, Rainfall Forecasting, ARMA Models, Time Series Analysis Box-Jenkins Methodology, Time Series Analysis.

Introduction

Statistical forecasting involves using time series data—chronologically ordered observations collected over time—to predict future events accurately. It is critical in planning and decision-making across various disciplines, including business, economics, healthcare, engineering, environmental research, and finance (Box & Jenkins, 1970; Hyndman & Athanasopoulos, 2018). Accurate forecasting relies on scientific methodologies and the analysis of historical data rather than conjecture, ensuring more informed judgments and practical actions to address various challenges. Time series data, such as daily stock prices, hourly temperature readings, or annual growth rates, are extensively utilized in fields like medicine, environmental studies, and industrial processes. The primary goals of time series analysis include understanding the dynamic structure of a single series and identifying interactions among multiple series. This understanding helps improve the accuracy of predictions and supports the design of optimal control strategies (Hamilton, 1994). The success of forecasting hinges on selecting and fitting appropriate models to the data. Over the decades, researchers have developed advanced techniques and tools to enhance forecasting accuracy, aided by computational advancements (Makridakis et al., 1998). Forecasting has thus become a focal point in statistical research, intending to improve the predictive power of models used for dependent variables. Scientific forecasting fundamentally depends on thoroughly

[1]*Department of Statistics, Quaid-i-Azam University, Islamabad.*
Corresponding Author Email: aimalkhan@stat.qau.edu.pk
[2]*Department of Statistics, University of Malakand, Chakdara Dir Lower. Email: habibullah858@uom.edu.pk*
[3]*Department of Statistics, University of Peshawar. Email: Ismailfromdirdir@gmail.com*

evaluating historical patterns, identifying variables influencing the data, and leveraging mathematical and statistical modeling. This approach ensures a comprehensive understanding of past occurrences and enables predicting future outcomes with excellent reliability and precision.

Forecasting

Time series models are foundational for analyzing and forecasting processes or statistics over time. They are widely applied in various domains, such as inventory management, weather forecasting, and sales predictions, providing a reliable basis for decisions under future uncertainty. These models are particularly valuable in aiding organizations in anticipating and preparing for uncertain future conditions. Additionally, integrating time series models with data mining techniques enhances understanding of data behavior and facilitates the prediction of trends and patterns (Box & Jenkins, 1970; Hyndman & Athanasopoulos, 2018). Forecasting is essential to effective planning across management, business, economics, and administration. It supports decision-making where lead times vary significantly—from years to seconds. However, successful forecasting requires distinguishing between uncontrollable external events and controllable internal factors. While forecasting predicts external events, decision-making addresses internal processes, with planning as the critical link between the two (Makridakis et al., 1998). Organizations use diverse forecasting methods, from simple techniques like naive forecasts to advanced methods like neural networks and econometric models. Forecasting plays a crucial role in scheduling, resource acquisition, determining resource requirements, and addressing short-, medium---, and long-term organizational needs. Effective forecasting requires organizations to excel in four key areas: defining problems, applying various forecasting methods, selecting appropriate techniques for specific situations, and building robust support systems for implementing formalized forecasting processes (Holt, 2004; Armstrong, 2001).

Steps for Forecasting in Quantitative Data

Forecasting quantitative data involves five key steps: defining the problem, gathering information, conducting a preliminary exploratory analysis, selecting and fitting models, and using and evaluating the chosen forecasting model. An essential aspect of selecting an appropriate method is understanding the data patterns, which helps identify suitable forecasting techniques. Time series data typically exhibit one of four patterns: horizontal, seasonal, cyclical, or trend. Horizontal patterns occur when data values fluctuate around a constant mean. Periodic factors, such as months or days of the week, influence seasonal patterns. Cyclical patterns involve rises and falls without a fixed periodicity, while trends reflect the long-term movement of data, which can be increasing, decreasing, or stable. Recognizing these patterns is crucial for accurate and effective forecasting.

Introduction to Forecasting the Climate

Climate change for a region refers to long-term variations in meteorological factors such as temperature, precipitation, and wind speed. In Pakistan, an agro-based economy, changing rainfall patterns significantly impact crop success or failure. Global and regional studies indicate an increasing trend in mean surface air temperature, with human influence being dominant since the mid-20th century. This study uses the Box-Jenkins methodology to apply statistical methods in time series analysis to identify the best-fitted model for analyzing rainfall trends. The best-fitted model is selected based on criteria like the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC), which measure the relative quality of statistical models. AIC estimates model quality by balancing goodness-of-fit with complexity, making it a valuable tool for model selection. The forecasted rainfall trends are interpreted using the selected model, providing insights into future climatic conditions.

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

For a large set of models of data, the preferred model is the one with the minimum AIC value. AIC is a goodness of fit, but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting because increasing the number of parameters in the model almost always improves the goodness of fit. The Schwarz Criterion is a criterion for selecting among formal econometric models. The Schwarz Criterion is a number and it can be defined as

$SIC = \ln(n)k - 2\ln(L)$

where n is the number of data points in x, the number of observations, or equivalently, the sample size.

Objectives

- To evaluate the effectiveness and suitability of ARMA models or the BoxJenkins approach to determine the best model for predicting monthly mean rainfall in Dir (L)
- To determine the best-fitted model to forecast the mean rainfall in Dir (L)
- To evaluate and confirm the best ARMA model to predict average monthly rainfall
- Using the best-fitted model, projecting the average monthly rainfall for the future years in district Dir (L)

Literature Review

There are many previous studies for time series data that have used the method of Box Jenkins in the forecasting of rainfall. The Box-Jenkins methodology involves four steps: (i) identification (ii) estimation (iii) diagnostic checking and (iv) forecasting. First, the original series must be transformed to become stationary around its mean and its variance. Second, the appropriate order of p and q must be specified using autocorrelation and partial autocorrelation functions. Third, the value of the parameters must be estimated using some non-linear optimization procedure that minimizes the sum of squares of the errors or some other appropriate loss function. Diagnostic checking of the model adequacy is required in the fourth step. This procedure is continued until an adequate model is obtained (Holt, 2004).

Numerous meteorological elements, including climate change, air temperature, and atmospheric pressure, have an impact on rainfall, one of the most significant hydrological processes. Even though there are numerous stochastic and data-driven hydrological models is still difficult to predict rainfall accurately, especially at mallet time scales. The irregular nature of rainfall makes it difficult to model well (Lauritzen, 1974).

In a country like India which is heavily dependent on agriculture, accurate and efficient forecasting methods for varied climatic situations are essential. Due to the simultaneous occurrence of three patterns, namely temporal, spatial, and non-linear, forecasting rainfall is one of the most difficult problems in this environment. One of the promising and well-liked methods for modeling spatiotemporal time series data is the Space-Time Autoregressive Moving Average (STARMA) model (Box & Jenkins, 1976).

Using an auto-regressive integrated moving averages (ARIMA) approach, this work focuses on the spatial-statistical analysis of rainfall fluctuation, anomaly, and trend in the Hindu Kush region. In the studied area, variations in river discharge have a substantial impact on rainfall trends, which finally caused floods and a hydrological drought. Rainfall has been employed as a climatic parameter in this study (Bowerman & O'Connell, 1979).

Changes in precipitation and temperature, which are influenced by the dynamic structure of the climate, can be investigated via time series analysis. The Bhagirathi River basin, located in the Indian state of Uttarakhand, is the subject of this study's assessment of time series and seasonal analysis of the monthly mean minimum and maximum temperatures and precipitation. The data used ranges from 1901 to 2000. (100 years). Using the seasonal ARIMA (SARIMA) model, forecasting for the following 20 years was performed (2001–2020). The Box Jenkins technique is the foundation for the autoregressive (p) integrated (d) moving average (q) (ARIMA) model, which forecasts future trends by making the data stationary and eliminating seasonality (Box et al., 1994).

Rainfall is crucial for forecasting the weather, especially for the agriculture industry and the environment, both of which have a significant impact on the country's economy. Therefore, hydrologists must anticipate daily rainfall to support other individuals working in the agriculture sector with their harvesting schedules and ensure that their crops will produce satisfactory results. This project will use the ARIMA model and an Artificial Neural Network (ANN) model to anticipate the daily rainfall's future value (West & Harrison, 2006).

Due to the lack of data and simplicity of earlier methods, time series, and data-driven methods have been utilized as complementary tools for forecasting ISMR versus the intricate physically based dynamical models. The most recent advancement in the many methods for predicting rainfall is the use of hybrid decomposition data-driven models, however, these methods have quite varied frameworks. The Adaptive Ensemble Empirical Mode Decomposition-Artificial Neural Network (AEEMD-ANN) model for forecasting is a framework for adaptive hybrid modeling presented in this paper (Unnikrishnan & Jothiprakash, 2020).

The main goal of the current study is to fit a model to the automobile insurance data gathered from the insurance industry over 36 years to forecast the 10 amount of the personal damage claim. By taking into account the standard deviation (SD) statistic for the time series data from 1981 to 2016 and using the Box-Jenkins financial econometric approach, the Auto-Regressive Integrated Moving Average (ARIMA) model for the data has been constructed in this work (Saha et al., 2020).

India's agricultural methods and crop harvests are highly reliant on climate elements like rainfall. As a result, improved information for planning and developing agricultural strategies could be obtained from more accurate rainfall predictions. Numerous research teams have tried utilizing various ways to predict rainfall. In this work, single-order models of the Autoregressive Integrated Moving Average (ARIMA) class have been employed to forecast rainfall (Dawood et al., 2020).

As modeling tools, multivariate adaptive regression splines (MARS) and k-nearest neighbors (KNN) were used in two artificial intelligence (AI)-based models. Using minimum, maximum, and mean air temperatures, dew point temperature, station pressure, vapor pressure, relative humidity, wind speed, and antecedent precipitation data, nine single input scenarios under limited climatic data are implemented. The obtained findings show that while simulating the monthly precipitation, the performance of MARS and KNN alone is comparatively subpar. Additionally, by merging the MARS and KNN models with three other types of time series (TS) models, namely autoregressive (AR), moving average (MA), and autoregressive moving average, this study builds hybrid models to improve the precipitation modeling (ARMA) (Dimri et al., 2020).

Due to its significant influence on all aspects of human existence, including agriculture, air traffic management, and public health and safety, weather forecasting has drawn researchers from all over the world over the years. Although thorough research on weather forecasting dates back to the 19th century, research on weather forecasting tasks has considerably grown since weather-big data became readily accessible. This article suggests using the grid technique to forecast greater visibility for the varying values of the parameters p, d, and q using an auto-regressive integrated moving average (ARIMA) model (Masngut et al., 2020).

Data on rainfall have been forecast using a variety of forecasting techniques. One forecasting technique that could produce improved projections is the Kalman Filter. To our knowledge, rainfall data in Makassar, Indonesia, have not been forecast using the Kalman Filter approach. This study employs the Autoregressive Integrated Moving Average (ARIMA) and Kalman Filter methods to create more accurate rainfall forecasts for Makassar, Indonesia (Johny et al., 2020).

In the majority of the world, statistical and numerical analysis is still used in weather forecasting. Although statistical and numerical analysis yields better findings, it heavily depends on consistent historical relationships between the prediction and the predicted value as a predictor of future events. On the other side, machine learning investigates fresh, data-driven algorithmic approaches to prediction. A location's climate is affected by a variety of variable factors, including temperature, precipitation, atmospheric pressure, humidity, wind speed, and a mixture of other variables of this sort (Kumar et al., 2020).

Predicting precipitation accurately can help with preparing for various demands on water resources management and extend lead times for tactical and strategic planning of courses of action. This paper investigates the suitability of several wavelet packet decomposition (WPD)-based forecasting models for predicting yearly rainfall, and a unique hybrid WPD-ELM precipitation prediction framework is developed (Ghule et al., 2020).

Forecasting rainfall can help people live and produce more. However, the short-term rainfall forecasting accuracy of the present approaches is typically subpar. The geographic features of the rainfall area have no bearing on machine-learning approaches. The surface and high-altitude geographical peculiarities cause the prediction accuracy to constantly vary in different regions. A surface and high-altitude Combined Rainfall Forecasting model (ACRF) is suggested to increase the prediction accuracy of short-term rainfall forecasting (Mehdizadeh, 2020).

A technique created in a certain temporal order for prediction is called time series analysis. The ARIMA model developed by Box and Jenkins is one of the time series analysis models utilized for forecasting. The Kalman Filter technique was one of the algorithms used to create the ARIMA model over time. To forecast rainfall using ARIMA and ARIMA Kalman Filter, this work attempts to estimate the parameters of the ARIMA model used as the starting value of the Kalman Filter (Salman, & Kanigoro, 2021).

Forecasting rainfall is necessary to manage water resources and make prompt decisions to reduce the negative effects of unforeseen events. Given that the factors influencing rainfall might fluctuate over the year, one strategy for putting forecasting models into practice is to create a model for each period in which the mechanisms are essentially constant, such as every season. The final models perform better because the selected predictors can be more reliable. From a practical standpoint in the tropical Andean region, it has not been determined whether the approach indicated above delivers improved performance in forecasting models (Ananda & Wahyuni, 2021).

The ability to predict days with heavy rain in advance is crucial for the effective management of weather-dependent activities because heavy rain harms ecosystems, causes flooding, makes up a major portion of the region's overall rainfall, and affects ecosystems. Numerical weather prediction models have historically been used to make weather predictions, but they are not without restrictions. Tools utilizing artificial intelligence and machine learning have grown in popularity

recently in this regard. The current study used a lengthy time series of rainfall data to determine the heavy and light rainfall days using the Gaussian Process Regression (GPR) approach, one of the machine learning approaches (Balamurugan & Manojkumar, 2021).
Farmers may benefit from accurate rainfall forecasts because the weather influences important decisions like crop selection and when to plant. The univariate time series ARIMA model may simulate stochastic processes using only historical data (Wang et al., 2021).
Since it is related to urban water management, rainfall forecasting in urban areas is a significant concern for city planners. In this study, the annual rainfall in Kolkata Municipal Corporation (KMC), West Bengal, was predicted using the ARIMA (auto-regressive integrated moving average) model as well as some regression techniques, including simple linear and second to sixth-degree polynomial regression equations (Zhang et al., 2021).
The current study focuses on one of the most significant climatic factors, precipitation, to analyze the rainfall pattern in five districts (Chhatarpur, Damoh, Panna, Sagar, and Tikamgarh) of Madhya Pradesh's Bundelkhand region, a semi-arid area. Despite the presence of multiple irrigation schemes, the quality of irrigation services remains subpar, and the majority of cultivated land still relies on rainfall. Since agriculture is the primary livelihood activity in this region, understanding the temporal fluctuations of rainfall is crucial for assessing climate-driven changes and proposing effective adaptation measures. Statistical analysis techniques, including the Mann–Kendall test, Sen's slope estimator, the MGCTI "Bertin matrix," and climate extreme indices (CDD, R95p, and RX1Day), were used to analyze seasonal, monthly, and annual rainfall trends and variability from 1951 to 2018. Additionally, the Autoregressive Integrated Moving Average Model (ARIMA) was applied to forecast annual rainfall in the study area for the period 2019 to 2050 (Mehta & Sukmawaty, 2021).
Rainfall and temperature are two key factors to examine while assessing climate change. Bangladesh has witnessed extremes in rainfall and temperature during the previous few decades, affecting both the environment and the agricultural economy. In this study, the ARIMA model is used to predict and forecast rainfall and temperature in Chattogram, Bangladesh from 1953 to 2070 considering seasonal variations (Amelia et al., 2021).

Methodology

Statistics is vital for forecasting and decision-making, providing numerical estimates that aid in planning. Time series analysis, a key statistical method, studies phenomena over time to predict future events with minimal error. It is essential in applied fields and sales forecasting, requiring stationary data for accurate predictions, where stationarity is determined by specific statistical properties.

Time Series

A time series consists of quantitative observations organized by time intervals, such as years, months, or days, and is used to predict future events. Mathematically, a time series is represented by a set of values. $z_1, z_{2,} z_3, z_{4,} z_{5,}$ ..., where each value Zt corresponds to a specific time t. These values are derived from random variables with their probability density functions. The joint probability density function describes the relationship between multiple time points in the series

Time Series Data

This research focuses on the time series data of average monthly rainfall in District Dir (L), Khyber Pakhtunkhwa. The data spans from January 2013 to May 2022 and was obtained from the Regional Meteorological Department of Dir (L).

Components of Time Series

Economic phenomena are influenced by a variety of factors, both directly and indirectly, leading to multiple variations over time. Time series analysis helps in understanding the behavior of these phenomena by studying their historical development. Statisticians identify four basic components within time series data. The first component is the Trend Component (T)**,** which refers to the long-term direction of the data, showing a gradual increase or decrease over time. This change is noticeable after a longer period compared to other components. The second is the Seasonal Component (S)**,** representing regular, predictable changes occurring at fixed time intervals (e.g., quarterly, monthly), often caused by external factors affecting the data. The third component is the Cyclical Component (C)**,** which involves changes occurring over longer periods than a year. These changes are irregular and do not follow a regular interval, distinguishing them from seasonal fluctuations.

Stationary and Non-Stationary Time Series

Time series data can be classified as stationary or non-stationary. Stationary time series exhibit fluctuations around a constant mean, with two types: strongly stationary, where the joint probability of values does not change over time, and weakly stationary, where the mean, variance, and autocovariance are constant. Non-stationary time series exhibit trends or cyclical changes that alter over time, making them difficult to analyze directly. Most real-world time series can be transformed into stationary series through methods like differencing, which removes trends and seasonality. To address non-stationarity, techniques such as first or second differencing are used, and transformations like taking logarithms or square roots can stabilize variance. Seasonal components can be removed using seasonal differencing. After these adjustments, the time series becomes stationary, enabling more accurate forecasting and modeling.

Test of Stationary Series

Economic variables are typically non-stationary time series due to their tendency to follow general trends, making them challenging to model. To address this, these time series need to be transformed into stationary time series. Various methods are used to test and ensure the stationarity of time series before further analysis or modeling.

Autocorrelation Function (ACF)

The autocorrelation function measures the correlation between neighboring observations in a time series. It serves two main purposes: detecting non-randomness in the data and identifying the appropriate time series model when the data are not random. The sample autocorrelation function at lag k is used to quantify this correlation.

$\gamma(k) = \sum n- (Zt-\bar{Z}\ k\ t=1\ )(\ Zt+k-\bar{Z})\ \sum n\ (Zt-\bar{Z}\ t=1\ )$, k=0,1,2,3,… ………………………….(1)

Where $\gamma(k)$ is an estimator for $\rho(k)$. This function allows the calculated autocorrelation coefficient between the observation of different periods, and the value of the autocorrelation coefficient would be $-1 \leq \rho(k) \leq 1$; with values near ±1 indicating stronger correlation. In the case of stationary the worth would be equal to zero $\rho(k) = 0$ as any correlation coefficient which means non-autocorrelation coefficients.

$\rho(k) = \gamma k\ \gamma 0$, k=0,±1, ±2, … ……………………………………………………………….(2)

Models of Time Series in Forecasting

Time series forecasting methods, such as Autoregressive (AR), Moving Average (MA), Mixed (ARMA), and Integrated Mixed (ARIMA) models, rely solely on the historical values of a variable, without considering other explanatory variables. These models apply to phenomena with suitable time series data. This research focuses on ARMA models, which will be used for forecasting in the study.

Autoregressive Models

The autoregressive model (AR) expresses the current value of a time series as a weighted sum of its previous values and a random error. The moving average model (MA) expresses the current value in terms of weighted past errors. The mixed model (ARMA) combines both AR and MA models, offering greater flexibility in representing time series data. ARIMA models (Autoregressive Integrated Moving Average) are an extension of ARMA models that address non-stationary time series. They include three components: AR (autoregressive), MA (moving average), and I (integration for making the series stationary). ARIMA is written as ARIMA(p, d, q), where p is the order of the AR model, q is the order of the MA model, and d represents the number of differences required to make the series stationary.

Model Building Stage of Time-Series Data

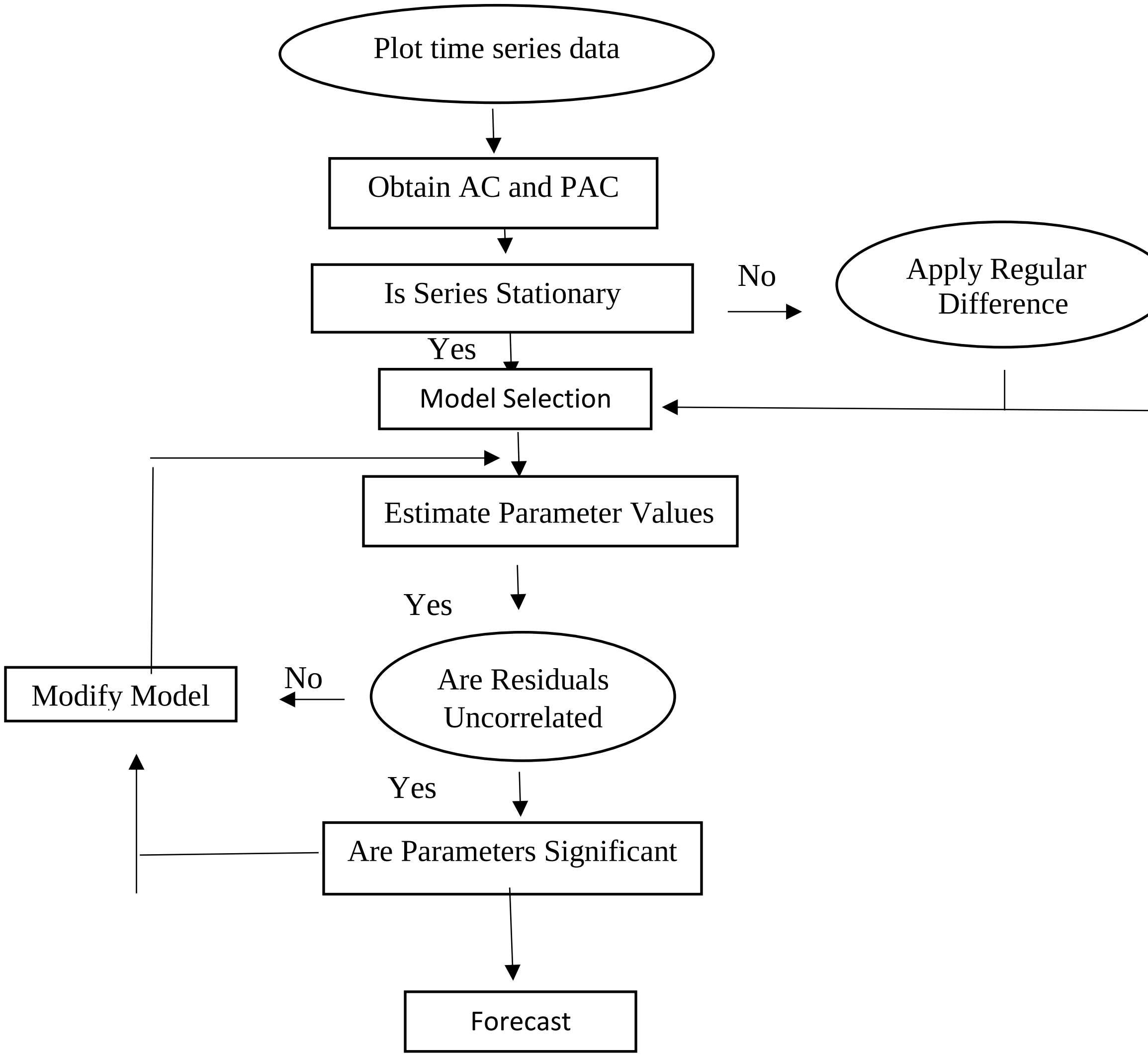
The Box-Jenkins method for time series forecasting consists of four primary stages. The first stage, model identification, involves analyzing the time series to determine the appropriate model, such as AR, MA, ARMA, or ARIMA, by examining patterns like autocorrelation and stationarity. In the second stage, estimation of model parameters, the parameters of the identified model are estimated using techniques like maximum likelihood estimation. The third stage, diagnostic check of the residuals, involves checking the residuals (errors) to ensure they resemble white noise, which indicates that the model has successfully captured the underlying patterns in the data. The final stage, model adequacy and forecasting**,** evaluates the model's accuracy and uses it to forecast future values. If the model is found to be inadequate, adjustments may be needed, requiring the process to be revisited or re-estimated.

Diagnostic and Graphical Analysis

In the diagnostic phase of developing a Box-Jenkins model, the first step is to determine whether the time series is stationary or non-stationary. Graphical analysis is the initial stage, where the time series data is plotted over time. This visualization helps identify patterns such as trends, seasonality, or abnormal values. By observing the plot, it becomes clear if the series is non-stationary, as it may show fluctuations in mean or variance. If non-stationary, further transformations are needed to stabilize the series before proceeding with analysis.

Stages of building model of the time series: After identifying the statistical properties of stationary time series and transforming the method of non-stationary series to the stationary series, we will come to know the forecasting method by using the random method that is known as the Box-Jenkins method. There are four primary stages in building a Box-Jenkins time series model. These are model identification, estimation of the model parameters, diagnostic check of the residuals, model adequacy, and forecasting. These stages can be summarized as follows:

Figure 1: Box-Jenkins Modeling Approach



The diagnostic phase in developing a Box–Jenkins model involves determining whether the time series is stationary or non-stationary. This begins with a graphical analysis**,** where the series is plotted over time. The plot provides insights into key characteristics such as trends, seasonal components, or outliers, helping to identify non-stationarity. If the series shows instability in its mean or variance, appropriate transformations are applied to stabilize it before proceeding with further analysis.

Autocorrelation Function and Partial Autocorrelation Function Plots

The autocorrelation function (ACF) and partial correlation function (PACF) are used to detect the stationary or non-stationary time series and a test of significant autocorrelation coefficients using the Ljung-Box Chi-Square test (Q–test).

The Ljung-Box test can be defined as follows.

H0: The series is stationary.

H1: The series are non - stationary.

The test statistic is:

$Q_{L.B} = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}^2{}_k}{n-k} \sim x^2(h)$.. (3)

Where, $\hat{\rho}^2{}_k$ =Autocorrelation at lag k., n = Sample size., h = Number of time lags includes in the test.

Forecasting

Forecasting is the final and most crucial stage of time-series analysis, aimed at predicting future values based on historical data. This process is only possible if the initial model passes all diagnostic tests, such as autocorrelation and partial autocorrelation functions. If the model fails, it must be refined, and the process is repeated until a suitable model is obtained. Once an adequate ARIMA model is established, forecasting future values Zn+1, Zn+2,…Z relies on using historical data Z1, Z2,…, Zn. Statistical inference for future values requires the conditional probability density function based on past observations, termed the predictive distribution**.** The best forecast point is the conditional expectation, which minimizes the mean square error and ensures the smallest possible variance. An optimal forecast uses the correct model to achieve minimal errors, enabling accurate predictions of future values for the time series. If we symbolize the present value of the time series as Z_t And we want to forecast the value of the time series in period (t + L), and we suppose that. $\hat{Z}_t$ This value will be represented at the time (t), we can obtain the forecasting by taking the conditional expected at the time of the original (t) of the model and after writing it when the period (t+ L), i.e. $E_t(Z_{t+L}/\ Z_t, Z_{t-1}, Z_{t-2}, \ldots)$ By using the conditional expected we will get the forecasting $\hat{Z}_t(L)$ With Mean Square Error Forecasting (MSEF), the least can be possible. The calculation of the forecasting is possible after a few steps (L) according to the formula:

$\hat{Z}_{t+L} = E_t(Z_{t+L}/\ Z_t, Z_{t-1}, Z_{t-2}, \ldots)$ for L≥ 1............ .... (4)

Auto regression model AR (p)

It is possible to calculate the forecasting after steps (L) according to the formula:

$E(Z_{t+\tau}) = \phi_1 E(Z_{t+L-1}) + \phi_2 E(Z_{t+L-2}) + \cdots + \phi_{p+d} E\big(Z_{t+L-p-d}\big)$ L≥ 1

And the best formula for forecasting model AR(p) after a few steps (L) is:

$\hat{Z}_{t+L} = \phi_1{}^L Z_{t+L-1} + \phi_2{}^L Z_{t+L-2} + \cdots + \phi_p{}^L E(Z_{t+L-p})$ L≥ 1 ....(5)

Moving Average Model MA (q)

It is possible to calculate the forecasting after steps (L) according to the formula:

$E(Z_{t+\tau}) = \varepsilon_{t+L} - \theta_1 E({}_1\varepsilon_{t+\tau-1}) - \theta_2 E(\varepsilon_{t+\tau-2}) - \cdots - \theta_q E(\varepsilon_{t+\tau-q})$ L≥ 1

And the best formula for forecasting model MA(q) after a few steps of (L) is:

$\hat{Z}_{t+L} = \varepsilon_{t+L} - \theta_1{}^L \varepsilon_{t+L-1} - \theta_2{}^L \varepsilon_{t+L-2} - \cdots - \theta_q{}^L \varepsilon_{t+L-q}$ L≥ 1...............................(6)

Mixed Models ARIMA (p,d,q)

The best formula for forecasting model ARIMA (p,d,q) after a few steps of (L) is:

$E(Z_{t+L}) = \quad \hat{Z}_L = \phi_1 E(Z_{t+L-1}) + \phi_2 E(Z_{t+L-2}) + \cdots + \phi_{p+d} E\big(Z_{t+L-p-d}\big) - \theta_1 E({}_1\varepsilon_{t+L-1}) - \theta_2 E(\varepsilon_{t+L-2}) - \cdots - \theta_q E(\varepsilon_{t+L-q}) + E(\varepsilon_{t+L})$ L≥ 1.. (7)

Result and Discussion

The study analyzes the time series of monthly mean rainfall in District Dir (L) using various ARMA models. The dataset consists of 113 monthly observations from January 1, 2013, to May 31, 2022. The goal is to identify the appropriate ARMA (p, q) model for the data and estimate its parameters. After selecting the model, its efficiency is evaluated using different tests. The chosen ARMA model is then used for forecasting. The analysis is conducted using the statistical software EViews 12 to generate the results.

Study of the Series Stationary

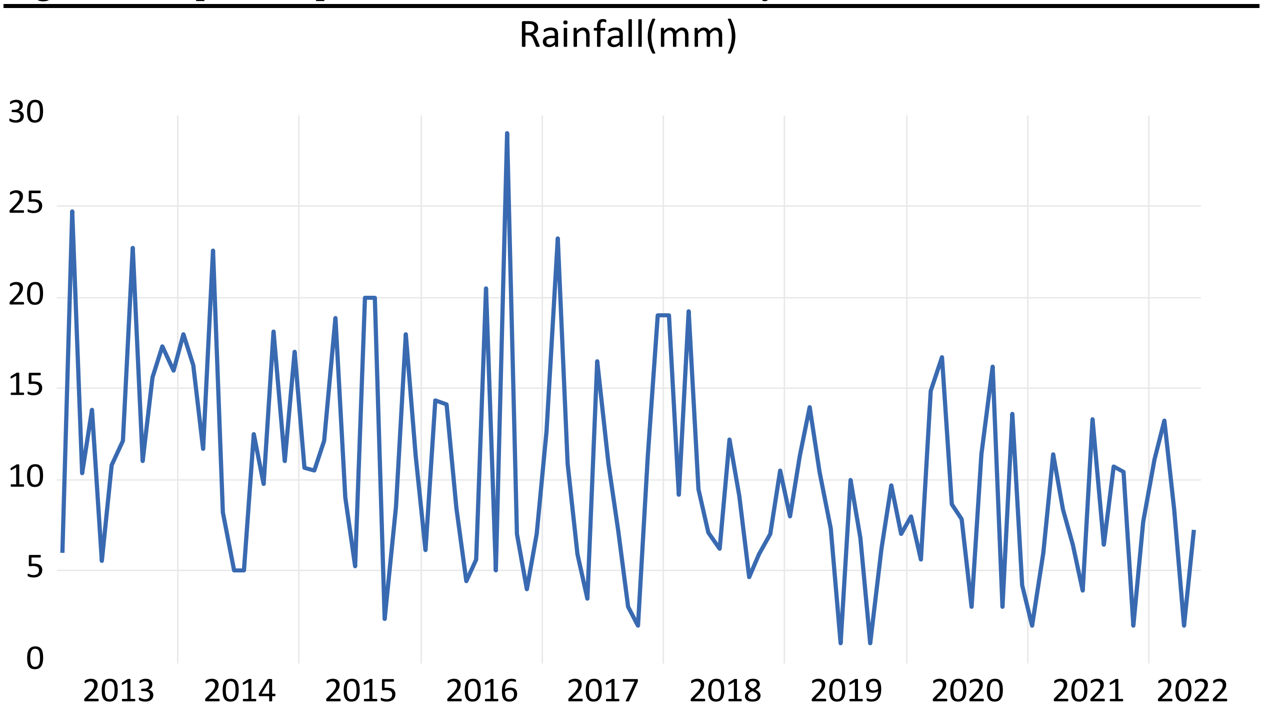
At this stage the time series for original data is drawn to know initially about some characteristics of this series, the following graph represents:

$$\left(\frac{-1.96}{\sqrt{n}} \leq \rho_k \leq \frac{1.96}{\sqrt{n}}\right)$$
$$\left(\frac{-1.96}{\sqrt{113}} \leq \rho_k \leq \frac{1.96}{\sqrt{113}}\right)$$

Study of the series of stationary

At this stage, the time series for original data is drawn to know initially about some characteristics of this series, the following graph represents that,

Figure 2: Graphical representation of the series monthly mean rainfall

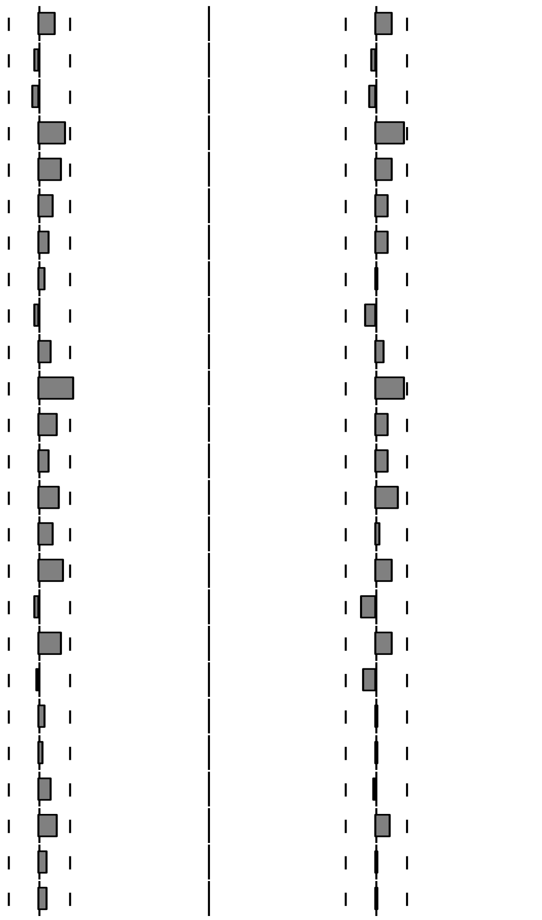


Through figure 2 of the monthly mean rainfall series, we see that the series spread randomly, and therefore, this graphical presentation gives us no answers as to whether the series is stationary or not. So, we draw the autocorrelation function (ACF), and the partial autocorrelation function (PACF) of data and draw the confidence interval of (ACF) and (PACF) to detect the stationary or non-stationary time series, as well as the use of the Ljung Box test (Q test) to ensure stationary of the series, as in the following table:

Figure 3: Stationary of the series

Date: 07/22/22 Time: 23:31
Sam~~ple:~~ 2013M01 2022M05
Included observations: 113

Autocorrelation Partial Correlation



	AC	PAC	Q-Stat	Prob
1	0.094	0.094	1.0177	0.313
2	-0.022	-0.031	1.0753	0.584
3	-0.038	-0.033	1.2426	0.743
4	0.171	0.179	4.7256	0.317
5	0.133	0.101	6.8433	0.233
6	0.083	0.072	7.6885	0.262
7	0.060	0.071	8.1336	0.321
8	0.037	0.013	8.3036	0.404
9	-0.025	-0.061	8.3805	0.496
10	0.076	0.055	9.1145	0.521
11	0.211	0.174	14.776	0.193
12	0.118	0.069	16.567	0.167
13	0.063	0.076	17.086	0.195
14	0.124	0.140	19.110	0.161
15	0.089	0.022	20.156	0.166
16	0.153	0.105	23.306	0.106
17	-0.020	-0.082	23.357	0.138
18	0.140	0.090	26.048	0.099
19	-0.009	-0.078	26.058	0.129
20	0.041	-0.003	26.291	0.156
21	0.024	-0.004	26.371	0.193
22	0.074	-0.011	27.163	0.205
23	0.111	0.083	28.943	0.182
24	0.055	0.004	29.382	0.206
25	0.045	0.005	29.688	0.236

Through table (2) of the original series of Correlation Coefficients and figures of ACF and PACF, we note that there is stationary in the data of the series and most of the values within the confidence interval, and that the Significant value of autocorrelation coefficients by using the Ljung – Box test was:

$Q = 29.688 < x^2_{25,0.05} = 37.652$

Table 1: ADF test results for the original series (Monthly mean rainfall)

ADF – Test	T -Statistic	Critical Values 5%	Prob.
With constant	−9.557766	−2.887425	0.0000
With constant trend	−10.98563	−3.450436	0.0000

Through the data of the table (4.3) we conclude. The Statistical values calculated for the Dickey-Fuller test in the case (with constant), and (with constant and trend) are less than the corresponding table value. i.e., we reject the hypothesis of a unit root. The p-value is also highly significant. The results of this test indicate the stationary of the series.

Table 2: Compared to a set of values of AIC, SIC, SIGMASQ

Models	**AR**	MA	SIGMASQ	AIC	SC
ARMA (4,4)	**4**	4	30.469	6.330	6.426
ARMA (2,2)	2	2	30.474	6.330	6.427
ARMA (1,4)	1	4	30.922	6.341	6.437
ARMA (3,4)	3	4	31.003	6.344	6.440
ARMA (3,1)	3	1	31.745	6.366	6.463
ARMA (3,3)	3	3	31.737	6.367	6.464
ARMA (1,2)	1	2	31.833	6.369	6.465
ARMA (2,3)	2	3	32.089	6.377	6.473

From the table 2, the minimum values for SIGMASQ, AIC and SC are given under the model ARMA (4,4). Thus, the ARMA (4,4) is the most suitable model for the monthly mean rainfall.

Figure 4: All the roots for MA and AR process are inside the unit circle so the ARMA (4,4) process is stationary and invertible

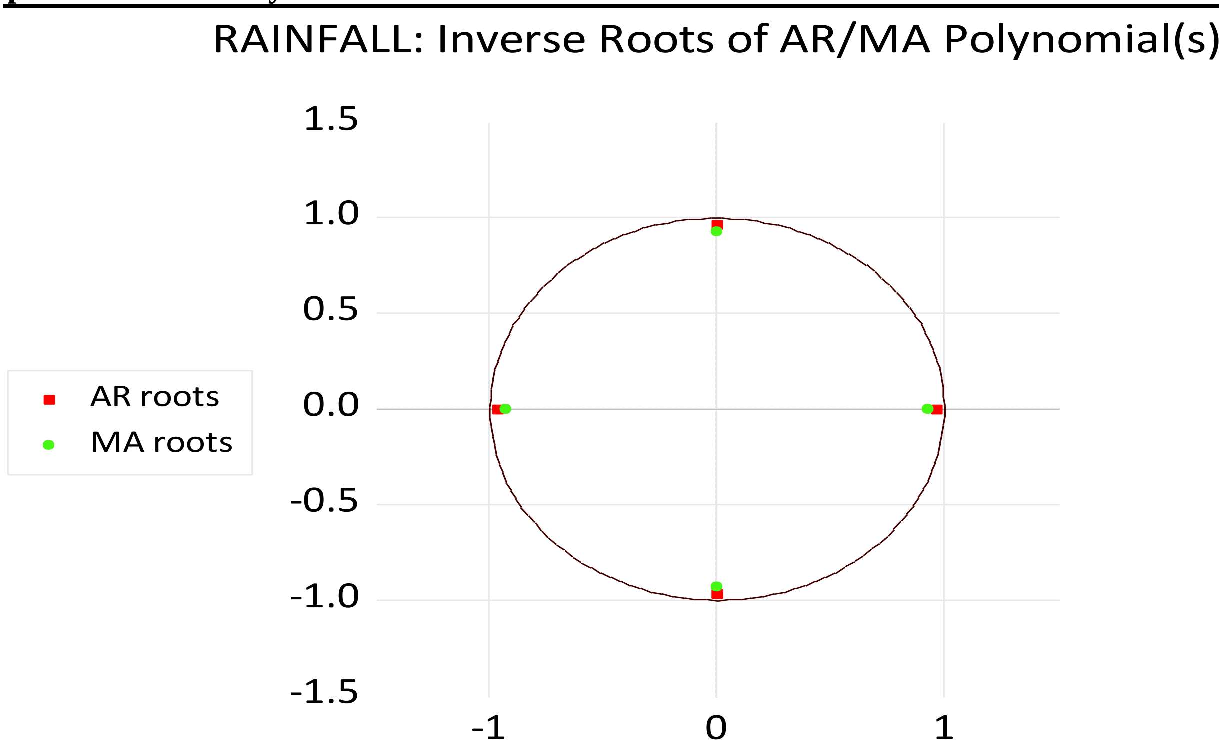
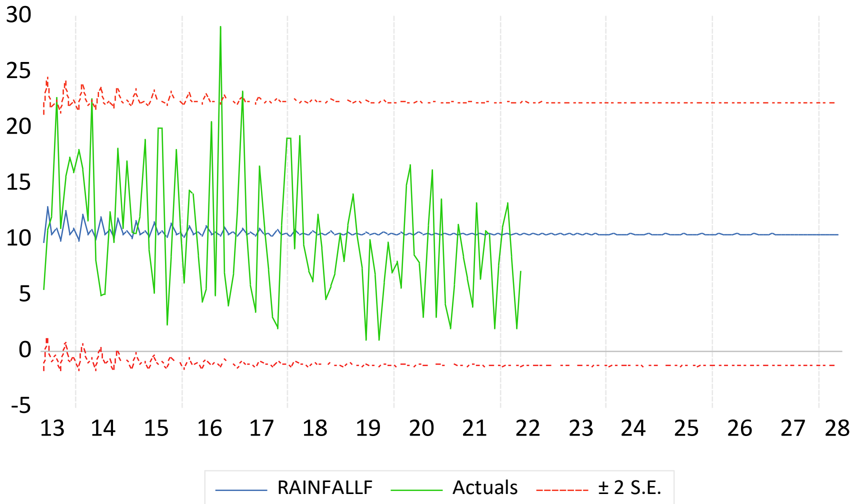


Figure 5: Forecasting Stage



Forecast: RAINFALLF	
Actual: RAINFALL	
Forecast sample: 2013M01 2028M05	
Adjusted sample: 2013M05 2028M05	
Included observations: 181	
Root Mean Squared Error	5.545160
Mean Absolute Error	4.443590
Mean Abs. Percent Error	82.96094
Theil Inequality Coef.	0.248731
Bias Proportion	0.003608
Variance Proportion	0.846157
Covariance Proportion	0.150236
Theil U2 Coefficient	0.610735
Symmetric MAPE	46.01248

Figure 6: The mean monthly rainfall forecasting results for district Dir(L), it has been according to the model ARMA (4,4) over the period (June 2022 – May2028)

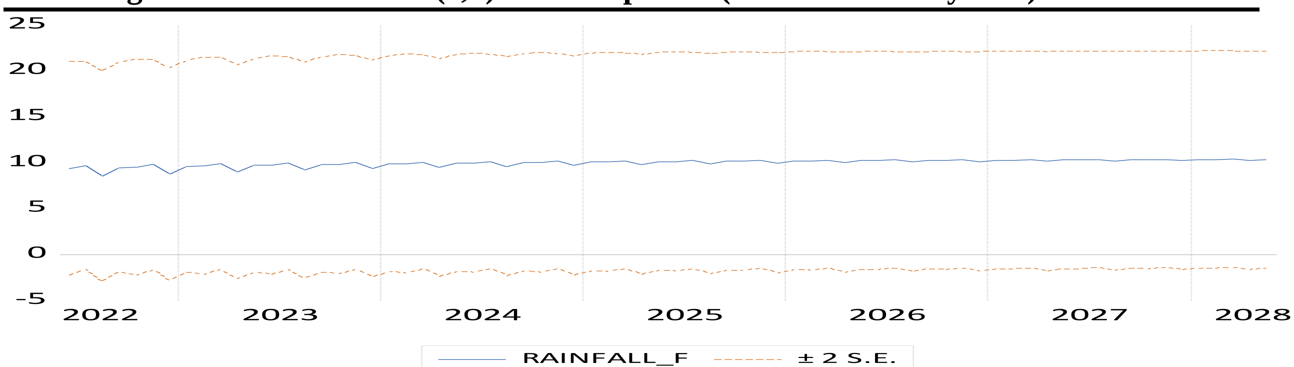


Figure 7: Actual and forecasted values

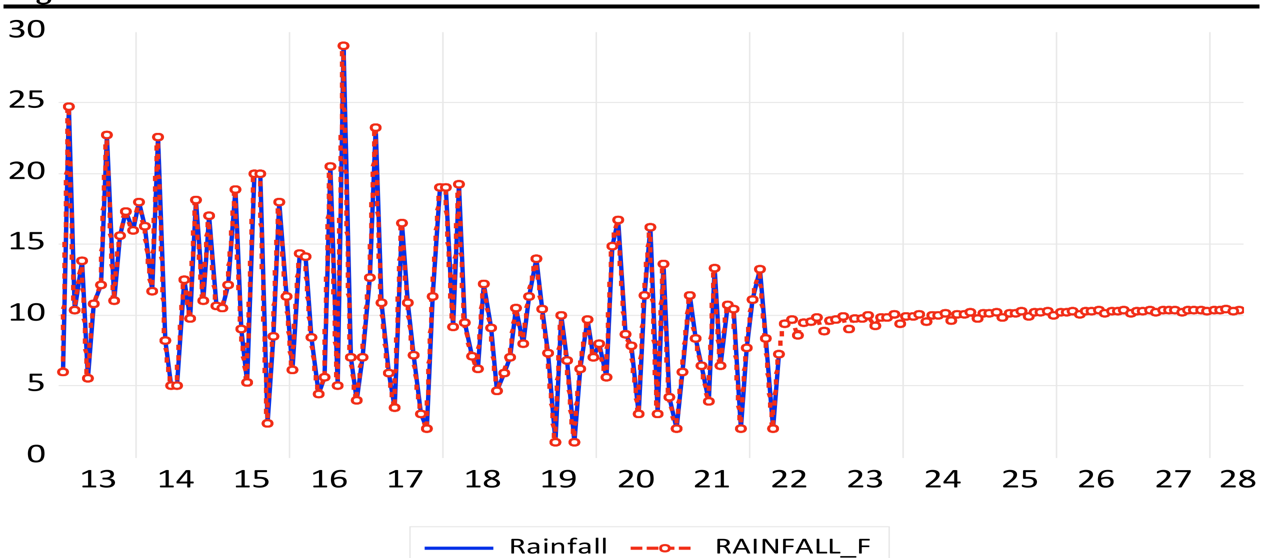


Table 3: Forecast the mean monthly rainfall (mm)

Point	Forecast	Point	Forecast	Point	Forecast	Point	Forecast
Jun-22	9.396	Dec-23	9.366	Jun-25	10.140	Dec-26	10.131
Jul-22	9.699	Jan-24	9.885	Jul-25	10.228	Jan-27	10.282
Aug-22	8.557	Feb-24	9.916	Aug-25	9.901	Feb-27	10.291
Sep-22	9.476	Mar-24	10.069	Sep-25	10.163	Mar-27	10.336
Oct-22	9.553	Apr-24	9.504	Oct-25	10.179	Apr-27	10.171
Nov-22	9.794	May-24	9.556	Nov-25	10.256	May-27	10.303
Dec-22	8.818	Jun-24	9.984	Dec-25	9.971	Jun-27	10.331
Jan-23	9.600	Jul-24	10.117	Jan-26	10.199	Jul-27	10.35
Feb-23	9.648	Aug-24	9.624	Feb-26	10.213	Aug-27	10.206
Mar-23	9.878	Sep-24	10.019	Mar-26	10.280	Sep-27	10.321
Apr-23	9.025	Oct-24	10.043	Apr-26	10.031	Oct-27	10.328
May-23	9.708	Nov-24	10.159	May-26	10.230	Nov-27	10.362
Jun-23	9.750	Dec-24	9.729	Jun-26	10.243	Dec-27	10.237
Jul-23	9.950	Jan-25	10.074	Jul-26	10.301	Jan-28	10.337
Aug-23	9.207	Feb-25	10.095	Aug-26	10.084	Feb-28	10.343
Sep-23	9.802	Mar-25	10.196	Sep-26	10.258	Mar-28	10.373
Oct-23	9.839	Apr-25	9.821	Oct-26	10.269	Apr-28	10.263
Nov-23	10.014	May-25	10.121	Nov-26	10.320	May-28	10.351

Results and Discussion

It is concluded that the most suitable model for the analysis of monthly mean rainfall in district Dir(L) is ARMA (4,4). We have found that the most suitable time series model is ARMA (4,4) because this model has lower SIGMASQ, AIC, and SC as compared to other fitted time series models. The tests of diagnosis for residuals series for the model ARIMA (4,4) are independent and random. All the roots for the MA and AR process are inside the unit circle so the ARMA (4,4) process is stationary and invertible. The forecasts were constant for the last five months of the mean monthly rainfall series (using the Chow test), which means that the forecasts after May 2022 will be accurate. The forecasted mean monthly rainfall for district Dir(L) using the above-mentioned model for the next seventy-two months is shown in figure (4.14). The forecasting is based on sound statistical methods, so it is adequate forecasting.

Conclusion

This study focused on forecasting mean monthly rainfall in District Dir (L) using time series analysis. The analysis revealed that the raw data series of mean monthly rainfall was stationary, as confirmed by ACF and PACF plots and the Ljung-Box test (Q = 29.688, which is less than the critical value of 37.652). The Dickey-Fuller test further validated the absence of a unit root in the series, ensuring its stationarity. The ARMA (4,4) model was identified as the most appropriate for forecasting the series, with statistically significant coefficients ($p < 0.05$) as determined by the T-student and F-statistics. The model's parameters were estimated using the least squares method, and model adequacy was confirmed through various criteria such as SIGMASQ, AIC, and SC values. Residual analysis indicated that the model's residuals were white noise, further validating the model's reliability. After model diagnostics and efficiency tests, the ARMA (4,4) model was selected for forecasting. The forecast for the period June 2022 to May 2028 was generated using this model, with expected high accuracy for future predictions.

References

- Alam, M. J., & Majumder, A. (2022). *A Comparative Analysis of ARIMA and Other Statistical Techniques in Rainfall Forecasting: A Case Study in Kolkata (KMC), West Bengal.*
- Amelia, R., Dalimunthe, D. Y., Kustiawan, E., & Sulistiana, I. (2021, November). ARIMAX model for rainfall forecasting in Pangkalpinang, Indonesia. *In IOP 51 Conference Series: Earth and Environmental Science, 926*(1), p. 012034). IOP Publishing.
- Ananda, E. Y. P., & Wahyuni, M. S. (2021, November). Rainfall Forecasting Model Using ARIMA and Kalman Filter in Makassar, Indonesia. *In Journal of Physics: Conference Series 2123*(1), p. 012044). IOP Publishing.
- Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer.
- Balamurugan, M. S., & Manojkumar, R. (2021). Study of short-term rain forecasting using machine learning-based approach. *Wireless networks, 27*(8), 5429-5434.
- Bowerman, B. L., & O'Connell, R. T. (1979). *Time series and forecasting: an applied approach. Brooks*/Cole.
- Box, G. E. P., Jenkins, G. M and Reinsel, G. C. (1994). *Time Series Analysis Forecasting and Control.* 3rd ed., Englewood Cliffs, N.J. Prentice Hall.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- Box, G.E.P and G.M Jenkins, (1976). *Time series Analysis forecasting and control.*, 2nd ed, San Francisco., Holden day, Inc.
- Dawood, M., Rahman, A. U., Ullah, S., Mahmood, S., Rahman, G., & Azam, K. (2020). Spatio-statistical analysis of rainfall fluctuation, anomaly, and trend in the Hindu Kush region using the ARIMA approach. *Natural Hazards, 101*(2), 449-464.
- Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, *129*(1), 1-16.
- Dwivedi, D. K., & Shrivastava, P. K. (2022). Rainfall Probability Distribution and Forecasting Monthly Rainfall of Navsari using ARIMA Model. *Indian Journal of Agricultural Research, 56*(1), 47-56.
- Ghule, P. S., Popale, P. G., Gorantiwar, S. D., & Pandit, D. S. (2020). Forecasting of Weekly Rainfall using ARIMA Models in Ghod Catchment of Pune District (MS). *Journal of Agriculture Research and Technology*, *45*(2), 115.
- Holt, C. C. (2004). *Forecasting Trends and Seasonalities in Time Series*. International Journal of Forecasting.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.
- Johny, K., Pai, M. L., & Adarsh, S. (2020). Adaptive EEMD-ANN hybrid model for Indian summer monsoon rainfall forecasting. *Theoretical and Applied Climatology, 141*(1), 1-17.
- Kumar, V. S., Satpathi, D. K., Kumar, P. P., & Haragopal, V. V. (2020, July). Forecasting motor insurance claim amounts using the ARIMA model. *In AIP Conference Proceedings 2246*(1), p. 020005). AIP Publishing LLC.

- Lauritzen, S. L. (1974). Note on a Case-Study in Box-Jenkins Seasonal Forecast of Time Series. Stanford univ ca dept of statistics.
- Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting Methods and Applications*. Wiley.
- Masngut, M. A., Ismail, S., Mustapha, A., & Yasin, S. M. (2020). Comparison of daily rainfall forecasting using multilayer perceptron neural network model. *IAES International Journal of Artificial Intelligence*, *9*(3), 456.
- Mehdizadeh, S. (2020). Using AR, MA, and ARMA time series models to improve the performance of MARS and KNN approaches in monthly precipitation modeling under limited climatic data. *Water Resources Management*, *34*(1), 263-282.
- Mehta, W. A., & Sukmawaty, Y. (2021, November). Rainfall prediction climatological station of Banjarbaru using arima Kalman filter. *In Journal of Physics: Conference Series*, *2106*(1). 1, p. 012003). IOP Publishing.
- Saha, A., Singh, K. N., Ray, M., & Rathod, S. (2020). A hybrid spatio-temporal modeling: An application to space-time rainfall forecasting. *Theoretical and Applied Climatology*, *142*(3), 1271-1282.
- Salman, A. G., & Kanigoro, B. (2021). Visibility forecasting using autoregressive integrated moving average (ARIMA) models. *Procedia Computer Science*, *179*, 252- 259.
- Som, K. S., & Dey, M. (2022). Analysis and forecasting of rainfall trends in semiarid Bundelkhand region of Madhya Pradesh, India: using statistical methods. *Environmental Earth Sciences*, *81*(3), 1-16.
- Subrahmanyam, K. V., Ramsenthil, C., Girach Imran, A., Chakravorty, A., Sreedhar, R., Ezhilrajan, E., & Jha, C. S. (2021). Prediction of heavy rainfall days over a peninsular Indian station using machine learning algorithms. *Journal of Earth System Science*, *130*(4), 1-9.
- Unnikrishnan, P., & Jothiprakash, V. (2020). Hybrid SSA-ARIMA-ANN model for forecasting daily rainfall. *Water Resources Management*, *34*(11), 3609-3623.
- Wang, H., Wang, W., Du, Y., & Xu, D. (2021). Examining the Applicability of Wavelet Packet Decomposition on Different Forecasting Models in Annual Rainfall Prediction. *Water*, *13*(15), 1997.
- West, M., & Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.
- Zhang, P., Cao, W., & Li, W. (2021). Surface and high-altitude combined rainfall forecasting using convolutional neural network. *Peer-to-Peer Networking and Applications*, *14*(3), 1765-1777.