**Evaluating ARIMA Models for Predicting Monthly Rainfall in Dhaka, Bangladesh: A Study Using 2000-2022 Data**

**Abstract**

Bangladesh is predominantly an agricultural country. Majority of population depends on agricultural crop and the cultivation mainly depends on natural calamities like rainfall which is one of the most important and creeping phenomena than other climatic event and it is difficult to predict rainfall in a region where rainfall variability is higher. This paper attempted to identify the appropriate SARIMA model and checked the statistical properties of the fitted model and focused on rainfall forecast for a year lead time of Dhaka district. Seasonal Auto Regressive Integrated Moving Average (ARIMA) models was established and used to make short-term forecasts of average monthly rainfall in Dhaka, Bangladesh. Using the traditional Box-Jenkins method, SARIMA models for the rainfall recorded in Dhaka from 2000 to 2022 were found suitable. On the basis of an examination of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, the optimal orders of the ARIMA models were determined and assessed using the AIC and forecasting accuracy (RMSE and MAPE) criteria. The best fitted model SARIMA (1,0,0)(1,1,1)12 was selected with AIC and RMSE values 2951.324 and 111.2872 respectively compared with other possible models. Diagnostic check was then conducted for the best fitted model and the residuals were found as white noise. This SARIMA model can be used for rainfall forecasting in Dhaka city for the decision makers to adopt appropriate drainage systems and weather management strategies.

**Keywords:** Bangladesh, Dhaka, Rainfall, ARIMA, Forecast.

**Introduction**

Precipitation plays a crucial role in the water cycle, serving as the primary mechanism for delivering fresh water to the Earth's surface. The global average annual precipitation is approximately 100 cm (39 inches); however, its distribution is highly uneven across different regions. The areas experiencing the most significant rainfall are primarily located in the equatorial belt and the monsoon regions of Southeast Asia. Rainfall is a critical climatic factor that significantly influences forest biodiversity, water availability, food production, and livelihoods in many nations [1]. Thus, early predictions of heavy rainfall can help mitigate flood-related disasters by assessing the interaction between potential risks and the vulnerability of a specific region [2].

Rainfall can have both positive and detrimental effects on a nation's economy and agriculture, particularly when it is unfavorable. In Bangladesh, where crop agriculture remains the backbone of the economy, the sector contributes about 22% to the national GDP, with 60.1% of the land area devoted to agriculture [3]. Understanding the socio-economic context of rural communities, agricultural practices, projected changes in rainfall patterns, and the relationship between rainfall and crop growth is essential for sustainable development. Time series analysis of meteorological data offers a valuable tool for analyzing variability and forecasting short- and long-term rainfall trends.

Several studies have applied time series models to predict rainfall in different regions. For instance, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model was fitted to rainfall data from 1988 to 2013 in Arunachal Pradesh, India, predicting the subsequent 14 years of rainfall [4]. Similar models have been utilized for univariate time series analysis and climate forecasting in regions such as Dhaka, Bangladesh; the coastal areas of Teknaf, Bangladesh; and Allahabad, India [5-7]. In Amman, Jordan, the Box-Jenkins method was used to develop an ARIMA model, effectively forecasting peak rainfall values at the Amman airport station [8]. Moreover, in Batticaloa district, Sri Lanka, statistical models have been applied to study climatic trends and variability [9]. Machine learning approaches and other models, including Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANN), and Exponential Smoothing State Space (ETS) methods, have also been explored for rainfall forecasting, with ARIMA showing superior performance in certain cases [2, 10].

Among the various time series analysis techniques, the univariate ARIMA model is one of the most effective and widely used [11]. The Box-Jenkins methodology is considered particularly suitable for predicting future trends and forecasting time series data [12]. In this study, we applied the Seasonal ARIMA model to forecast the monthly total rainfall for the Dhaka district in Bangladesh. While previous studies have predicted monthly rainfall in Dhaka, to the best of the researchers' knowledge, none have utilized such long-term data combined with more recent records. Against this backdrop, the present study aims to explore different ARIMA models using recent data (2000–2022) and identify those that perform best in predicting the monthly total rainfall in Dhaka. As Dhaka is the capital of Bangladesh, maintaining a healthy environment is crucial for both the well-being of its citizens and the smooth functioning of official operations. Understanding the nature and magnitude of potential climate changes in central Bangladesh is critical for policymakers and residents, enabling better preparation and adaptation strategies. Thus, developing a reliable forecasting model is essential for effective water resource management.

**Materials and methods**

**Climate data**

For this study, data on monthly average rainfall from 2000 to 2022 were obtained from the official reports of the Bangladesh Agricultural Research Council (BARC). These reports, publicly accessible via the BARC Climate Information Management System (<http://barcapps.gov.bd/climate/>), provide comprehensive and reliable meteorological data essential for agricultural research and climate analysis in Bangladesh.

**Statistical analyses**

The graphical analysis was conducted to get idea on the stationarity of monthly average rainfall data. It also gives insights on trend and seasonality components of the dataset. To further verify the results, the ADF test was employed to more rigorously assess stationarity and confirm whether the findings align with the graphical analysis. Once a time series was confirmed to be stationary, an ARIMA model has been fitted (or a SARIMA model if seasonality is present). The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were then examined to determine the orders (p, q) for the AR and MA terms, respectively. The ACF plot also provides insights into the seasonal orders (P, Q). To identify the best model from a set of plausible options, the Akaike Information Criterion (AIC) was evaluated. Models with the lowest AIC values were considered for further analysis. RMSE and MAPE values were compared, and the model with the lowest values for these metrics was chosen as the most suitable. After selecting the most suitable model, the model parameters are estimated using the least squares method. In the diagnostic checking step, the residuals from the fitted model should be evaluated for adequacy by conducting a correlation analysis using residual ACF plot (independence), Ljung-Box test (autocorrelation), and Jerque-Bera test (normality). At the last step, the model has been applied to predict the forthcoming monthly total rainfall in Dhaka.

**ARIMA model**

The well-known autoregressive integrated moving average (ARIMA) model for time series analysis is widely used to describe and predict meteorological variables for its accuracy and practicality. Developed from the ARIMA model, the ARIMA model that incorporates seasonal period performs better in the presence of an obvious seasonal pattern. An ARIMA (p, d, q)(P, D, Q)s model comprises 7 parameters: p and P are the orders of general and seasonal autoregression (AR) respectively; q and Q are the general and seasonal moving average (MA) orders respectively; d and D are the numbers of general and seasonal differencing respectively; s denotes periodicity [13]. A seasonal ARIMA (p, d, q) (P, D, Q)s model's general form can be expressed as follows:

Φ(B) MA(B)(1-B)d(1-Bs)DZt = θ(B) t

with Zₜ representing the value of time series at time t, and εₜ a white noise series. Here, B refers to a backward shift operator (e.g., BZₜ = Zₜ₋₁). (B) = 1 − ₁B − ₚ and Φ(B) = 1−Φ₁ − Φₚ are the general and seasonal autoregressive operators respectively. (B)= 1 – ₁B – q and (B) = 1−₁ – Q are the general and seasonal moving average operators respectively.

**Result**

From the decomposition of rainfall series (Fig 1), no obvious trend was found for the years 2000-2022. However, there is a declination from 2008 to 2015 and a strong indication of seasonality, and it occurs at a 12-months interval.

Initially, we can derive stationarity from the graph of monthly total rainfall against time series. From the Augmented Dickey Fuller (ADF) test, the total monthly rainfall was found to be stationary with test statistic value -11.279 and p-value 0.01. Thus, the null hypothesis that the data is non-stationary was rejected at 5% level of significance.

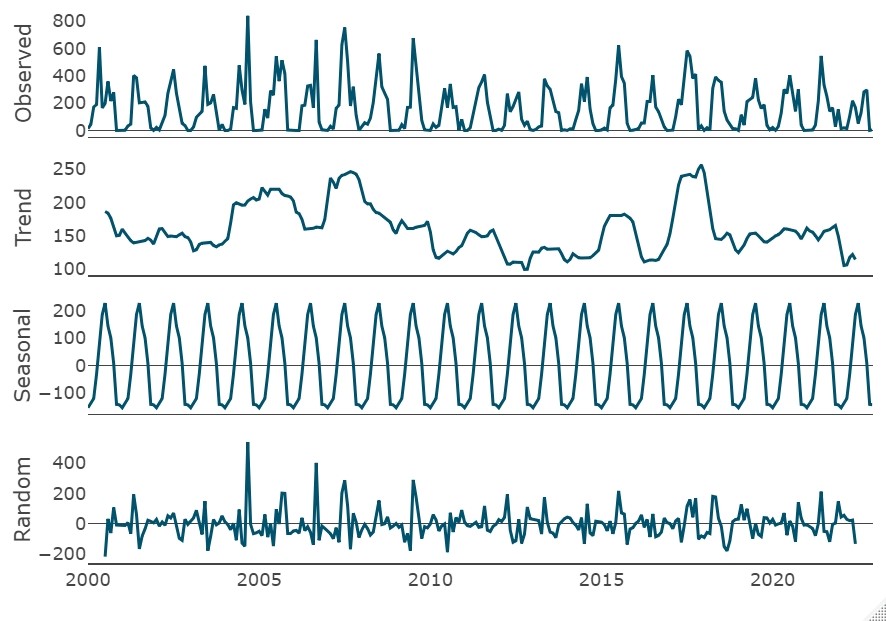


Figure 4.3: Decomposition of the total rainfall in Dhaka (2000-2022).

**Model Identification**

First, the rainfall dataset was partitioned into two samples: one for training and other for testing dataset. The rainfall data between January 2000 and December 2020 (252 months) were used to construct a SARIMA model as the training set, while those between January 2021 to 2022 (24 months) were then used to validate the model as the test set. Other validation methods (e.g., using 2- or 3-year data to validate the model) were also applied and revealed similar results.

The ACF and PACF plots were observed to determine the orders (p, q) of AR and MA terms respectively. ACF plot also gives notion for the seasonal orders (P, Q).

Red spike-for seasonal lag 12**,** Black spike-for non-seasonal lag

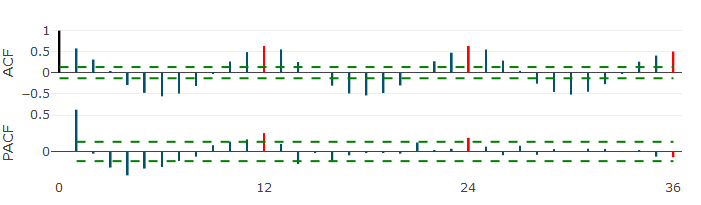


Figure 4.16: ACF and PACF plot of total monthly rainfall in Dhaka

From Fig 2, Almost all the spikes in ACF plot and more than four spikes in PACF-plot exceeded the confidence band, therefore, the maximum value of the non-seasonal parameter q was 3. As for the non-seasonal parameter p, the maximum value thereof was assumed to be 2. As the series is already stationary, non-seasonal difference (d) is zero. Now from the ACF plot, seasonality is clear to be there. So, seasonal difference (D) was assumed to be at most 1. Notice that the sample ACF values exceeded the critical value at lag 36, and that the sample PACF values exceeded the critical values at lag 24, the upper limits of the seasonal parameter Q and of the seasonal parameter P were 3 and 2 respectively. In order to search out the most suitable model, at least 20 ARIMA models under the following four conditions were carried out: p = 0, 1 or 2; q = 0, 1, 2 or 3; P = 0, 1 or 2; Q = 0, 1, 2 or 3. Table 1 lists the five ARIMA models with the lowest AIC value.

Table 4.10: Comparison of AIC values for selected candidate models

|  |  |  |
| --- | --- | --- |
| **Model** | **AIC** | **RMSE** |
| ARIMA(1,0,0)×(0,1,1)12 | 2951.229 | 112.9266 |
| ARIMA(1,0,0)×(1,1,1)12 | 2951.324 | 111.2872 |
| ARIMA(0,0,1)×(0,1,1)12 | 2951.799 | 112.8857 |
| ARIMA(0,0,1)×(1,1,1)12 | 2951.88 | 111.2328 |
| ARIMA(0,0,1)×(0,1,2)12 | 2952.68 | 111.9599 |

The AIC values for these five candidate models are the lowest. So, in order to come to a decision, the model that generates the most accurate predictions was selected based on its performance on the test set. ARIMA(1,0,0)(1,1,1)12 was regarded as the optimum model because having the lowest AIC and RMSE values. Even while ARIMA(0,0,1)(1,1,1)12 has a similarly low AIC and RMSE values, the Ljung-Box test reveals residual autocorrelation.

**Parameter Estimation**

The preliminary estimation of the parameter is done by using the training dataset, estimated on identification stage. Parameter estimation is done by maximum likelihood method. From Table 2, all of the seasonal and non-seasonal parameters were found statistically significant at 5% level of significance.

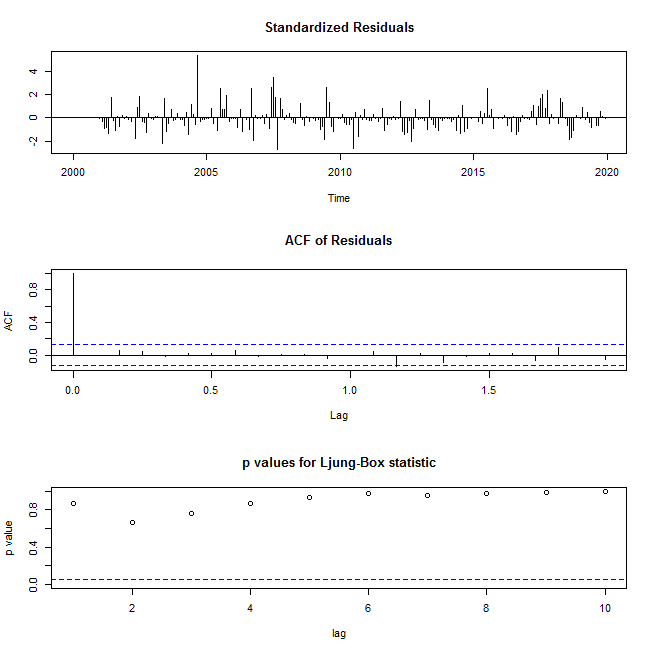
Table 2: Estimation of parameters of SARIMA model.

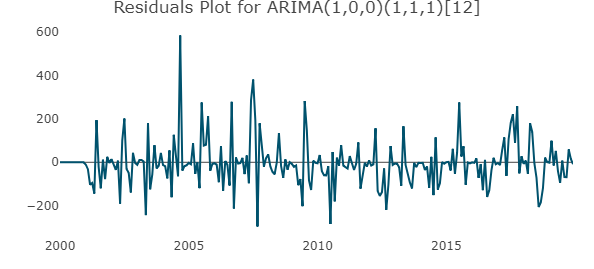
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficient** | **Estimate** | **Standard Error** | **z** | **p-value** |
|  | 0.164191 | 0.060603 | 2.7093 | 0.006743 \*\* |
|  | 0.135628 | 0.066836 | 2.0293 | 0.042431 \* |
|  | -0.999732 | 0.385683 | -2.5921 | 0.009539 \*\* |

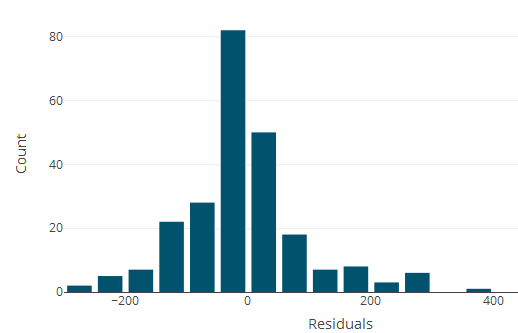
:1-order auto-regressive coefficient, :1-order seasonal auto-regressive coefficient,1-order seasonal moving average coefficient.

**Diagnostic Check**

After selecting the most suitable model based on AIC and RMSE values, the next step is to evaluate its residuals. A good forecasting model should have residuals that meet certain criteria: they should be uncorrelated, follow a normal distribution and exhibit stationarity. These properties indicate that the model has captured the underlying structure of the data and the residuals represent random noise [14].

1. Standardized Residuals: Here from the first plot of the Figure 2 (a), the residuals fluctuate randomly so it satisfies the model well.
2. ****ACF of Residuals: From the second plot of the same figure, it can be observed that the residual spikes are not touching the confidence band at all. That means there is no auto-

****(b)



(a) (c)

Figure 2: Diagnostic check for the residuals of the ARIMA(1,0,0)×(1,1,1)12 model.

correlation. Hence, residuals are independent. It can be confirmed by checking the Ljung-Box Statistic in which the null hypothesis is that no serial correlation exists [15].

iii) Ljung-Box statistic: All p-values for different lags are greater than the significance threshold, α = 0.05, as shown in the final plot. The null hypothesis is accepted. The residuals are, therefore, not auto correlated for the ARIMA(1,0,0)×(1,1,1)12 model.

From Figures 2 (b) & 2 (c), residuals seem to have zero mean and constant variances and they seem to be normally distributed which is a good sign. As the residuals of the selected model have the four properties, ARIMA(1,0,0)×(1,1,1)12 model is a well fitted model and adequate for forecasting.

**Comparison of actual vs fitted values:**

The graph in Figure 3 depicts a comparison between the actual values and those fitted by the SARIMA model between 2000 and 2020. Using the ARIMA (1,0,0) (1,1,1)12 model, we obtain a graphical plot of the observed pitch versus the predicted series for the period of 2021–2022 and it is evident from a quick examination of the plot that the selected model is adequate as the predicted series closely resembles the observed series. Therefore, the model could be applied to predict the forthcoming total monthly rainfall in Dhaka.

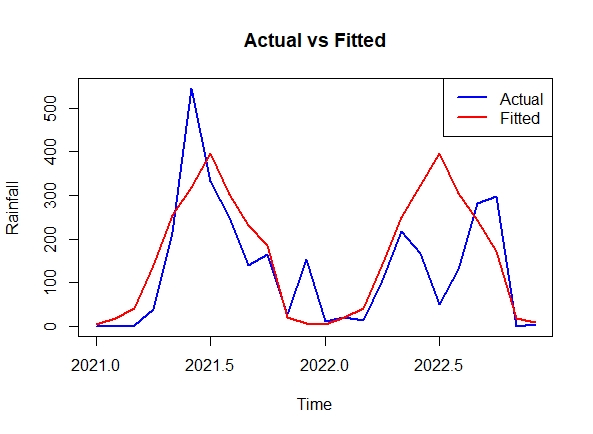


Figure 3: Graph of actual vs fitted and predicted values of the rainfall series

**Rainfall Forecasting:**

Based on the proposed model, Table 3 illustrates the predicted monthly total rainfall along with 95% confidence intervals for 2023.

Table 3: Forecasted monthly total rainfall in Dhaka, Bangladesh during Jan 2023-Dec 2023

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | **Forecasted Total Rainfall**  **(in millimeters)** | **95% CI of Forecasted Total Rainfall** | |
| **LCL** | **UCL** |
| Jan-2023 | 0.784185 | -208.35136 | 209.9197 |
| Feb-2023 | 17.861715 | -194.06898 | 229.7924 |
| Mar-2023 | 39.104007 | -172.90153 | 251.1095 |
| Apr-2023 | 125.776793 | -86.23076 | 337.7843 |
| May-2023 | 243.110428 | 31.10282 | 455.1180 |
| Jun-2023 | 311.056205 | 99.04860 | 523.0638 |
| Jul-2023 | 327.060599 | 115.05299 | 539.0682 |
| Aug-2023 | 273.887407 | 61.87980 | 485.8950 |
| Sep-2023 | 264.820793 | 52.81318 | 476.8284 |
| Oct-2023 | 191.448477 | -20.55913 | 403.4561 |
| Nov-2023 | 14.581066 | -197.42641 | 226.5885 |
| Dec-2023 | 13.383085 | -198.61970 | 225.3859 |

**Discussion**

In our study, a rainfall data from 2000-2022 of Bangladesh was used in order to fit an ARIMA model to forecast the rainfall amount (mm) in 2023. A SARIMA model was found to be best fit based on AIC and RMSE values among 20 possible ARIMA models. According to a study in Kerala, among several time series models: Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN) and Exponential Smoothing State Space (ETS), ARIMA(0, 0, 2)(1,1,1)12 outperformed while fitting on the monthly rainfall at Idukki district for the period from January 2006 to December 2016 and the model was evaluated with reference to Root Mean Squared Error (RMSE) and model fit [3]. Another study in Arunachal Pradesh found that the monthly rainfall time series data (1988–2007) was non-stationary, therefore, Box-Cox transformation was applied and minimum values of AIC and Schwarz Bayseian Information (SBC) is the reason to select the SARIMA (0,1,1) (0,1,1)12 model [4]. A monthly rainfall data was taken for Amman airport station in Jordan for the period of 1922-1999 with a total of 936 readings in order to forecast rainfall for the upcoming 10 years, for this purpose, different ARIMA models were compared based on sum of squared residuals values, further, ARIMA (1, 0, 0) (0, 1, 1)12 model was developed and it’s goodness-of-fit was tested by means of Chi-square statistics [5]. A study on monthly rainfall data from 1981-2013 in coastal area (Teknaf), to forecast monthly rainfall for 12 months lead time the best fit ARIMA (0, 0, 0) (1, 1, 1) model was selected based on Normalized BIC (Bayesian Information Criteria) and R-squared, then diagnostic check was conducted and the predicted rainfall amount from the best fitted model was compared with the observed data [6]. A study in China reported that best model was fitted to bacillary dysentery (BD) cases of Beijing where analyzing ACF and PACF plots of the series, it was transformed using logarithms to make the series stationary and ARIMA(1,1,1)(2,1,1)12 model was fitted based on AIC value, significance of the coefficients of the model and autocorrelation test named Ljung-Box Q test [7].

Predictive accuracy of different SARIMA models was evaluated by RMSE. However, it was the SARIMA that predicted the rainfall most accurately than the model ARIMA(1,0,0)×(0,1,1)12 with lowest AIC value or ARIMA(0,0,1)×(1,1,1)12 with lowest RMSE value. The selected model was the one with the optimum AIC and RMSE values. The diagnostic check revealed that the standardized residuals of the model were white noise and normally distributed. From the Fig 4, it is clear that the predicted series closely resembles the observed series which assure the adequacy of the model.

Yet this study is not free from limitations. Our analysis was conducted on ground of monthly data while weekly data or daily data could probably improve the accuracy of lagged time estimate. Also due to the dynamic nature of the atmosphere, it is a challenging task to predict rainfall very precisely. Understanding the nature of fluctuations in this variable is crucial. As time progresses, the uncertainty around the predictions increases; therefore, further research on the dynamic nature of atmospheric characteristics is essential. We predicted the rainfall data for the forthcoming year 2023.

In conclusion, this study successfully demonstrated the application of SARIMA model to time series data and to forecast the upcoming. The effectiveness of such a model was evident from an accurate goodness of fit and high predictive accuracy especially in the short term. This study may help the decision-makers to adopt the strategies towards the optimal planning of agriculture, drainage systems and water resources management in Dhaka city. Also, the proposed forecasting models in this study may be used as a guidance for further research on the causal association among the atmospheric parameters and more accurate prediction for successfully handling the hazards caused by their individual or joint effects in the region.

**Conclusion**

Prediction of rainfall is crucial for climate monitoring, drought detection, severe weather forecasting, agricultural production, energy and industrial planning, communication and pollution dispersion, among other applications. In this study, the ARIMA (1,0,0)(1,1,1)12 models are found to be suitable for predicting the monthly total rainfall in Dhaka district for the following year. This model chosen based on model selection criteria such as AIC and predictive accuracy (RMSE and MAPE). Autocorrelation is evaluated using the Ljung-Box test resulted as there is no autocorrelation between the residuals of the fitted model at 5% level of significance and the residuals are found to be approximately normally distributed from the histogram of residuals. The graphical comparison reveals that the projected series differed from the original series by a negligible amount, indicating that the stochastic model developed to forecast rainfall found effective. Hence the model developed for Dhaka can be applied to determine possible future strategy and to develop sustainable water resource planning and improve drainage system.

**References**

1. Chung, E. S., Park, K., & Lee, K. S. (2011). The relative impacts of climate change and urbanization on the hydrological response of a Korean urban watershed. \*Hydrological Processes, 25\*(4), 544-560.
2. Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. Disaster Advances, 11(11), 27-33.
3. Hossain, M. M. (2016). On identifying the SARIMA model to forecast the humidity of some selected stations in Bangladesh. International Journal of Environmental Sciences, 6(6), 1-11. ISSN 0976-4402.
4. Dabral, P. P., & Murry, M. Z. (2017). Modelling and Forecasting of Rainfall Time Series Using SARIMA. *Environmental Processes*, *4*(2), 399–419. https://doi.org/10.1007/s40710-017-0226-y.
5. Mahmud, I., Bari, S., & Hussain, M. (2015). Year-Long Monthly Rainfall Forecasting for a Coastal Environment of Bangladesh. *Journal of Natural Sciences Research*, *5*(21), 52–57. https://iiste.org/Journals/index.php/JNSR/article/download/27121/27804
6. Mahsin, M., Akhter, Y., & Begum, M. (2012). Modeling rainfall in Dhaka Division of Bangladesh using time series analysis. Journal of Mathematical Modelling and Application, 1(5), 67-73. ISSN: 2178-2423.
7. Graham, A., & Mishra, E. P. (2017). Time series analysis model to forecast rainfall for Allahabad region. \*Journal of Pharmacognosy and Phytochemistry, 6\*(5), 1418-1421.
8. Naill, P. E., & Momani, M. (2009). Time series analysis model for rainfall data in Jordan: Case study using time series analysis. American Journal of Environmental Sciences, 5(5), 599-604.
9. Partheepan, K., Jeyakumar, P., & Manobavan, M. (2005). Development of a time-series model to forecast climatic data in the Batticaloa District, Sri Lanka. Water Professionals' Day Symposium.
10. Parmar, A., Mistree, K., & Sompura, M. (2017). Machine learning techniques for rainfall prediction: A review. In Proceedings of the International Conference on Innovations in Information Embedded and Communication Systems.
11. Johnson, L. A., & Montgomery, D. C. (1976). Forecasting and time series analysis (7th ed.). McGraw-Hill Book Company.
12. Gerretsadakan, A., & Sharma, M. K. (2011). Modeling and forecasting of rainfall data of Mekele for Tigray region (Ethiopia). Material and Methods, 9(1976), 31–53.
13. Cryer JD, Chan K-S. Time Series Analysis With Applications in R. 2nd ed. New York: Springer; 2008.
14. Box GE, Jenkins GM. \*Time series analysis: forecasting and control\*. Revised ed. San Francisco: Holden-Day; 1976.
15. Jung GM, Box GEP. On a measure of lack of fit in time series models. Biometrika. 1978;65(2):297–303. doi:10.2307/2335207.