Proceedings of the 6th International Conference on Civil Engineering for Sustainable Development (ICCESD 2022), 10~12 February 2022, KUET, Khulna, Bangladesh (ISBN-978-984-35-1972-6)

FORECASTING MAXIMUM AND MINIMUM TEMPERATURE: A STUDY OF THE NORTHEASTERN BANGLADESH

Nirupam Deb

Student, MEDM Program, Department of Geography and Environment, Jagannath University, Bangladesh, e-mail: 10205006@uap-bd.edu

ABSTRACT

Climate change is now a widely discussed phenomenon in today's life, with one of its multidimensional aspects, changes in temperature, a significant issue. Along with the temperature change, the issue of seasonal variation is also directly involved. Temperature forecasting is crucial for the survival and development of plants and animals at all stages of the ecosystem and the nature, environment, and climate that depends on it. This study attempts to analyze and forecast both maximum and minimum monthly temperature data from BMD's Sreemangal station from 1962 to 2010 in two models - linear regression (LR) and seasonal autoregressive integrated moving average (SARIMA). It is hard to predict the raw behavior of the temperature with a linear regression model because of too many fluctuations and the seasonality in the data set, with a poor coefficient of determination (R^2) value and the more significant value of the standard error of the estimate (i.e., root means square error) (RMSE). No normal distribution with less homoscedasticity for maximum and minimum temperature in the linear regression models observed with the normal P-P plot and the scatterplot of regression standardized predicted and residual value. The linear regression model results as 0.642 and 0.404 outside the range of 1.5 to 2.5 for the maximum and minimum temperature dataset in the Durbin-Watson test statistic, which means a higher-level statistical analysis is required. Since seasonality is involved, SARIMA$(2,0,1)(1,1,2)_{12}$ and SARIMA$(1,0,1)(1,1,1)_{12}$ with an R^2 value of 0.869 and 0.964 for the maximum and minimum temperature time series results RMSE as 0.979 and 1.092. The residual ACF and PACF interpret- there has no significant correlation being all spikes within the confidence limits and Q-Q plot of the normal probability of residuals of both time series identifies the normal distribution of the predicted dataset. This SARIMA model predicts monthly maximum and minimum temperature data for 2011, which is entirely consistent with the recorded temperature of the concerned station in that year. Forecasting monthly temperature has the opportunity to draw the issue of climate change more elaborately and can help the authorities expedite the process of taking necessary steps to tackle climate change impacts.

Keywords: *Temperature, Linear Regression, Seasonal ARIMA, Time series, Forecasting*

1. INTRODUCTION

The earth planet confronts climate change every moment, the most severe threat. The warming planet impacts the weather system, hydrology, ecology, and various aspects of the environment (Rahman & Lateh, 2015). Though developing nations are more sensitive to the consequences of climate change than developed ones (Rahman & Lateh, 2015), one of the newest developing countries of South Asia, Bangladesh, has a diverse climatic condition with subtropical in the center-north and tropical in the south. The hot spring between March and May (i.e., pre-monsoon), a long rainy season from June to October due to the summer monsoon, and the warm and sunny winter from November to February (i.e., post-monsoon) circulations have a significant influence on the warm, humid climatic condition of the country (World Climate Guide, n.d.). The country's historical climate has experienced temperature ranges between 15ºC and 34ºC throughout the year, with an average of around 26ºC (The World Bank, n.d.). Though human activities are the primary cause of climate change, statistical tests can identify a climate state change using more comprehensive data and sophisticated analyses (Intergovernmental Panel on Climate Change, 2007). Using the linear regression model, (Paul & Roy, 2020) try to analyze Bangladesh's 100 years' temperature rise data to determine the temperature rise trend for this region.

As the dumpy nature and too much fluctuation of the temperature dataset, the linear regression model predicts within a significant error level and can draw only the upwards trend of temperature (Paul & Roy, 2020). (Doulah, 2018), (Nury et al., 2017) accomplished temperature forecasting with the highest accuracy using seasonal ARIMA models. In Bangladesh, fluctuating monthly minimum temperatures and comparatively steady monthly maximum temperatures exhibited high seasonality (Doulah, 2018). A tropical climatical condition cited in Sreemangal, the northeast of Bangladesh, consists of an average annual temperature of 24.5 °C. August is the warmest month with an average of 27.5 °C, and with an average of 18.2 °C, January is the wintry month in the whole year (CLIMATE-DATA.ORG). However, global warming demonstrates with the monthly maximum temperature increasing by 2.97°C and 0.59°C per hundred years and the minimum temperature by 2.17°C and 2.73°C per hundred years at the Sylhet and Sreemangal stations, having an apparent growing trend(Nury et al., 2017). However, the increasing temperature trend in this region might affect the evaporation rate, rainfall, streamflow, crop yields, and climate forecasting can assist policymakers in determining the kind and magnitude of potential temperature changes in northeastern Bangladesh (Nury et al., 2017); this study has taken to accomplishment for forecasting the monthly maximum and minimum temperature at Sreemangal station.

2. METHODOLOGY

Sreemangal, one of the 7th Upazilas of the Moulvibazar district, accumulated an area of 450.74 sq km, located in between 24°08' and 24°28' north latitudes and in between 91°36' and 91°48' east longitudes (banglapedia, n.d.). The country's tea capital has diverse land coverage having Rema-Kalenga wildlife sanctuary in its southern boundary, Lawachara national park on the eastern gate, and Hail haor on the western side (google maps, n.d.). Analyzing the temperature variations in the studied area, treated as a regression task, and experimented with the following algorithms to identify the best model fitted with the data of the concerned area. The IBM SPSS Statistics 26 software used for completing the statistical interpretations consists of identification, estimation, diagnostic check, and application or forecast the maximum and minimum monthly temperature of the studied area within the following methods. Both the coefficient of determination (R^2) and standard error of the estimate (i.e., root means square error) (RMSE) accounted for the performance analysis of the best fit model. The study's primary limitation is the maximum and minimum temperature data recordation in only one coordinate of BMD's Sreemangal station (latitude 24.30ºN and longitude 91.73ºE) (Mannan et al., 2015). With more gauge points, the best fit model will be able to predict temperature with more precision.

2.1 Data Description

From Bangladesh Agricultural Research Council (BARC), maximum and minimum temperature data of the Sreemangal station were collected (BARC, n.d.). BARC receives weather data from Bangladesh Meteorological Department (BMD) (BMD, n.d.). The dataset contains data of monthly maximum and minimum temperature from 1962 to 2010. After the screening of missing values, the dataset prepares to analyze in several statical approaches.

2.2 Linear Regression (LR)

As the simplest possible model, the linear regression used to find a good fit model takes the relationship as linear between the independent and dependent variables of the study area's maximum and minimum temperature dataset. The formula of linear regression stands as

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \varepsilon_t \tag{1.1}$$

The betas are constants, and the epsilons are prediction errors, typically assumed to be independently and identically normally distributed (Nau, n.d.). The predicted value from the model can be above or below the actual value, and it is the reason behind the scene of using the least square error (i.e., squared of the error between the predicted value and the actual value) instead of the plain error (Paul & Roy, 2020). Now, the least-squares estimate the coefficients and intercept (Nau, n.d.) (i.e., setting them equal to the unique values that minimize the sum of squared errors) within the sample of data to identify the best fit model in equation (1.2).

$$\hat{y}_t = b_0 + b_1 X_{1t} + \cdots + b_k X_{kt} \quad (1.2)$$

Where "$\hat{y}_t$" is the predicted variable at time t, "b_0" means intercept constant, b_1 is the slope coefficients, "X_{1t}" refers to the independent variable to observed at time t, "K" is the time lag number.

2.3 Seasonal Autoregressive Integrated Moving Average (SARIMA)

The acronym ARIMA(p,d,q) stands for an auto-regressive integrated moving average model, where according to (Bibi et al., 2014), component "p" has used for prediction, whereas component "d" eliminates non-stationarity, and component "q" has the responsibility to alter future predictions. The ARIMA equation stands as,

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \quad (1.3)$$

Following the convention introduced by Box and Jenkins, with negative (-) signs, the moving average parameter (θ) has defined in equation (1.3). According to Box and Jenkins, both non-seasonal (p, d, q) and seasonal (P, D, Q) factors incorporate a multiplicative model titled SARIMA (i.e., Seasonal ARIMA model). The shorthand notation for the model (PennState Eberly College of Science, n.d.-b) is ARIMA (p, d, q)×(P, D, Q)S. With "p" refers to non-seasonal AR order, "d" is the non-seasonal differencing, "q" means non-seasonal MA order, "P" is the seasonal AR order, "D" refers to seasonal differencing, "Q" is seasonal MA order, and "S" means the time-period of repeating seasonal pattern. However, a regular pattern of changes in a time series is called the seasonality over S periods (PennState Eberly College of Science, n.d.-b). The Seasonal ARIMA model equation (Mahmud et al., 2016) stands more formally as,

$$\phi_p(B)\Phi_P(B^S)\nabla^d \nabla_S^D \hat{y}_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (1.4)$$

Where "ε_t" is the random shock (i.e., white noise error) and "B" is the backward shift operator.

2.4 Coefficient of Determination (R^2)

Coefficient of determination (R^2), one of the performance indication parameters, well-known for the output of regression analysis, uses in this study to analyze the performance of the different models. The coefficient of determination (R^2) equation stands as one minus the ratio of residuals sum of squares to the total sum of squares (PennState Eberly College of Science, n.d.-a).

$$R^2 = 1 - \frac{RSS}{TSS} \quad (1.5)$$

Where RSS is the residuals of the sum of squares and TSS is the total sum of squares. RSS obtained from the data points, "y_i" vary around the estimated regression line, "$\hat{y}_t$" and how much the data points, "y_i" vary around their mean, "$\bar{y}$" indicates the TSS (PennState Eberly College of Science, n.d.-a).
The coefficient of determination helps in determining the variation in the dependent variables by variation of independent variables. The R^2 value ranges between 0 and 1, and the higher the value of R^2, the better the model's prediction and strength (EDUCBA, n.d.).

2.5 Standard Error of the Estimate (i.e., Root Means Square Error) (RMSE)

One of the techniques for assessing how well a regression model fits a dataset is the standard error of the estimate, which is a measure that informs us of the average difference between the projected values from the model and the actual values in the dataset. In reality, root means square error (RMSE) is the distance in a plotted sheet between the dependent variables (i.e., projected value) and the regression line itself, which predicts the average dependent variables linked with the independent variables (Stanford

University, 2000). The lower the RMSE, the greater the capacity of a model to "fit" a dataset (Zach, 2021). The equation of RMSE stands as,

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(\widehat{y_t} - y_i)^2}{n}} \tag{1.6}$$

Where, "y_i" is the observed value for the observation and "$\widehat{y_t}$" is the predicted value for the given time series dataset.

3. RESULTS AND DISCUSSION

3.1 Linear Regression (LR)

During the analysis of the monthly maximum and minimum temperature dataset, the linear regression (LR) interpret the Pearson's correlation coefficient (r) as 0.092 and 0.254, the probability value (p) as 0.025 and 0.000, which are less than the significance level (α) of 0.05 indicates the poor positive relationship, but it is significant. The R^2 value of the linear regression results as low as 0.009 and 0.065 close to 0 indicates that both temperature data cannot predict adequately with this model. The produced RMSE values of 2.674 and 5.551 indicate a significant difference between the predicted and observed datasets, which is higher.

In the normal P-P plot in figures 1 and 2, the maximum and minimum temperature residuals do not follow the normality line accurately and assume no normal distribution, though they have a few drastic deviations.

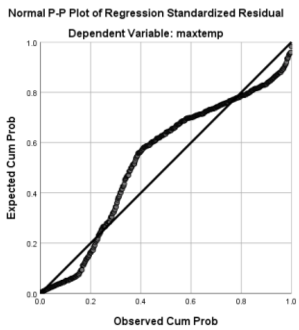

Figure 1: Normal P-P plot of maximum temperature datasets

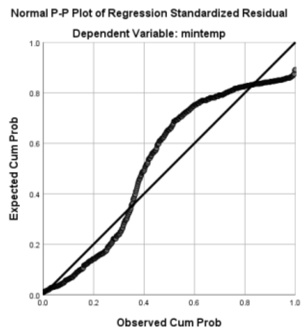

Figure 2: Normal P-P plot of minimum temperature datasets

The scatterplot of regression standardized predicted value in the x-axis and regression standardized residual in the y-axis in figures 3 and 4 indicates less homoscedasticity for maximum and minimum temperature in the linear regression models.

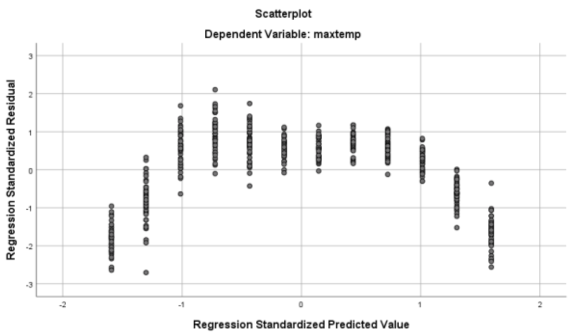

Figure 3: Scatterplot of maximum temperature datasets

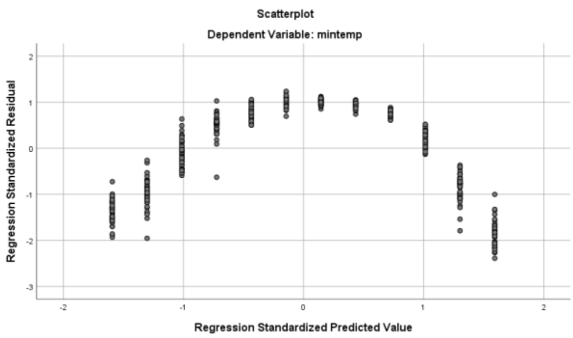

Figure 4: Scatterplot of maximum temperature datasets

The Durbin-Watson test statistic is 0.642 and 0.404 outside the range of 1.5 to 2.5 (Marshall, n.d.) for the monthly maximum and minimum temperature dataset, which requires the higher statistical order interpretation of the datasets.

3.2 Seasonal Autoregressive Integrated Moving Average (SARIMA)

3.2.1 Model Identification

The identification phase of the SARIMA model initiates with the stationarity test. The autocorrelation function (ACF) and partial autocorrelation function (PACF) are responsible for drawing the temporal correlation structure to identify the stationarity in the dataset. For the maximum temperature time series, ACF and PACF in figure 5 show the seasonal effect.

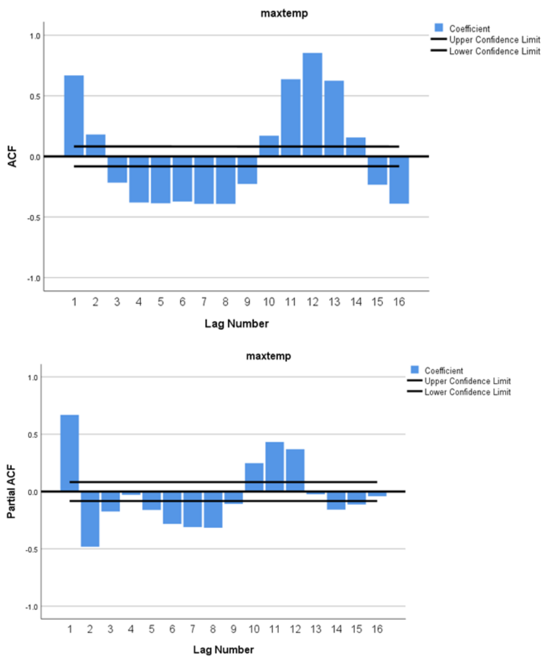

Figure 5: ACF and PACF of maximum temperature datasets

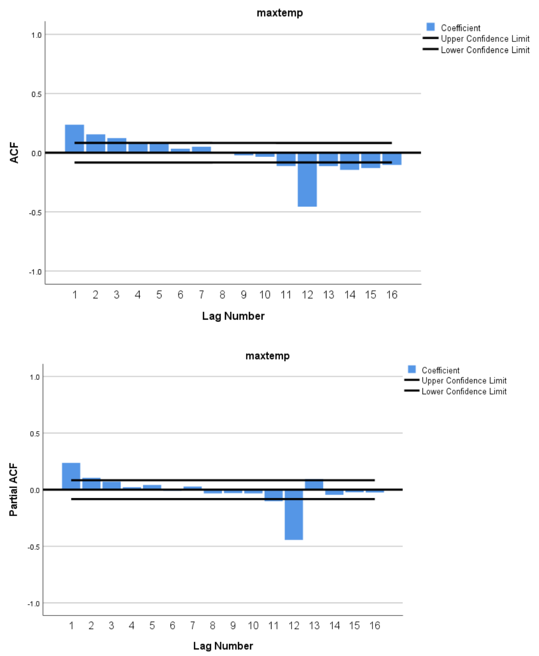

Figure 6: ACF and PACF of maximum temperature datasets

The seasonal differencing then takes place to remove the seasonality from the maximum temperature time series. After differencing in figure 6, the first and second spikes of both ACF and PACF take into account. However, after testing all feasible options to the entire degree of possible combinations, the model selected for maximum temperature is SARIMA(2,0,1)(1,1,2) and SARIMA(1,0,1)(1,1,1) for the minimum temperature time series.

3.2.2 Parameter Estimation

The parameter estimation also takes place during the model selection process for a better-fitted model. It is vice versa. The parameter estimation process cannot proceed without model identification, and getting the better-fitted model parameter estimation is very important. The output of the estimation of the parameters, the value of the coefficient of determination (R^2), and standard error of the estimate (i.e., root means square error) (RMSE) showing in the below table 1 for the maximum and minimum temperature.

Table 1: Output of the estimation of the parameters

Dataset	Model	Parameter	
		R^2	RMSE
Maximum Temperature	**SARIMA(2,0,1)(1,1,2)**	**0.869**	**0.979**
	SARIMA(2,0,1)(1,1,1)	0.867	0.983
	SARIMA(1,0,1)(1,1,2)	0.869	0.978
Minimum Temperature	**SARIMA(1,0,1)(1,1,1)**	**0.964**	**1.092**
	SARIMA(1,0,1)(1,1,2)	0.964	1.093
	SARIMA(2,0,1)(1,1,1)	0.964	1.093

The coefficient of determination (R^2) value 0.869 means that 86.9 percent of the maximum temperature data points are close to the regression line, and the year can predict 86.9 percent variance in maximum temperature. This percentage indicates the strong association between the dependent variable (i.e., temperature) and the independent variable (i.e., year). The variance in minimum temperature also predicted accurately around 96.4 percent with the year variable- stated by the R^2 value of 0.964. The selected SARIMA algorithms produced RMSE as just 0.979 and 1.092 values that are extremely near to one. For the dataset, the predicted maximum temperature has only a distance of 0.979 from the observed one, and the minimum one has a 1.092 difference between the observation and prediction.

3.2.3 Diagnostics Check

3.2.3.1 ACF and PACF of residuals

The below figures 7 and 8 of ACF and PACF of residuals show no significant correlation with having all the spikes within the confidence limit for both maximum and minimum temperature time series.

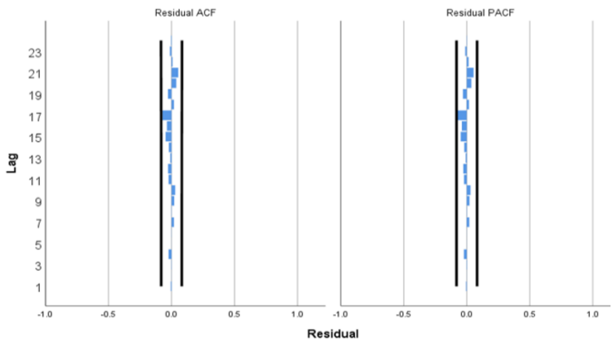

Figure 7: ACF and PACF of residuals of maximum temperature datasets

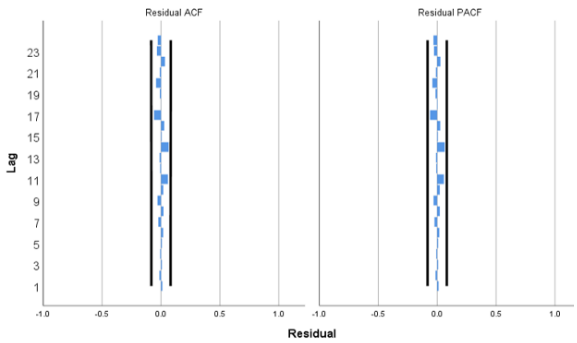

Figure 8: ACF and PACF of residuals of minimum temperature datasets

3.2.3.2 Lack of Fit-Test

In the Ljung-Box test for the maximum and minimum temperature dataset, the probability value p is 0.791 and 0.832, and both are greater than the significance level α of 0.05. Based on these probability values, the null hypothesis is white noise and that the test models are adequate with the time series.

3.2.3.3 Normal probability of residuals

The quantile-quantile (Q-Q) plot, a normal probability plot, demonstrates how the observed distribution of data compares to the expected normal distribution. If the observed data points approximately appear on a straight line, then the data is normally distributed. If the data is non-normal, the observed data points form curvature rather than a straight line (Analyse-it, n.d.). The below figures 9 and 10 show the almost linear representation of the maximum and minimum temperature dataset residuals, i.e., the normally distributed dataset.

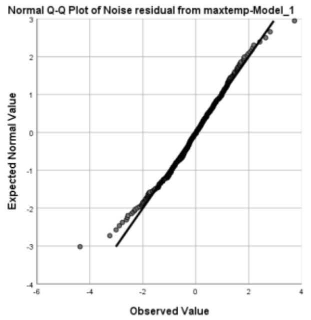

Figure 9: Normal Q-Q plot of maximum temperature datasets residuals

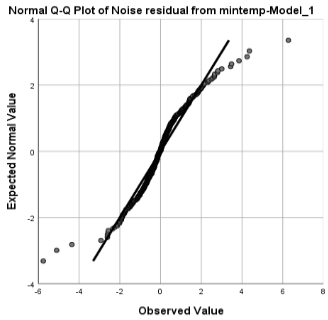

Figure 10: Normal Q-Q plot of minimum temperature datasets residuals

3.2.3.4 Residuals versus prediction plot

The evenly distributed (i.e., spread equally) residuals around the mean indicate the best fit model while plotting prediction in the X-axis and residuals in the Y-axis (Mahmud et al., 2016). The below figures 11 and 12 represent the equally distributed datapoint of both the maximum and minimum temperature datasets residuals, meaning SARIMA(2,0,1)(1,1,2) and SARIMA(1,0,1)(1,1,1) are the best fit model.

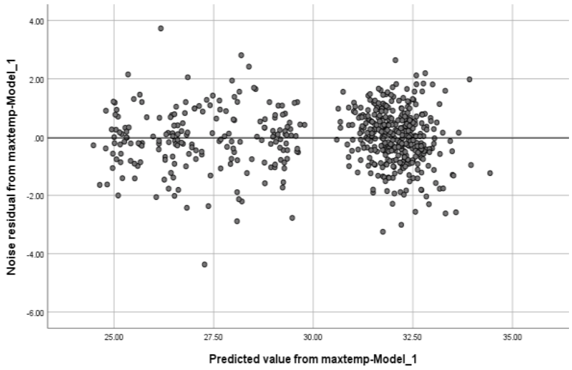

Figure 11: Scatterplot of maximum temperature datasets residuals

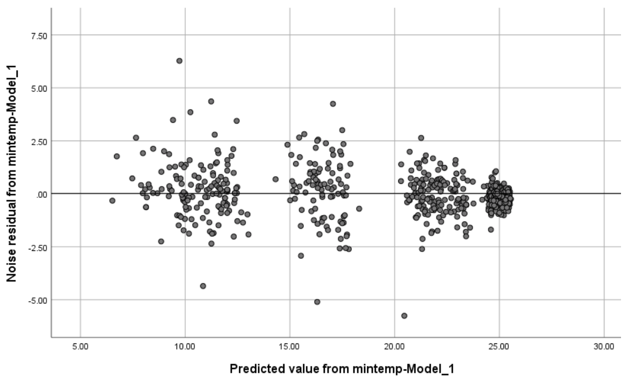

Figure 12: Scatterplot of minimum temperature datasets residuals

3.2.4 Temperature Forecasting

The forecasting of maximum and minimum temperature of Sreemangal has been accomplished with 12 months lead time, using SARIMA(2,0,1)(1,1,2) and SARIMA(1,0,1)(1,1,1) models. The predicted values of the temperature time series follow the observed data closely enough in figures 13 and 14. The more significant coefficient of determination (R^2) and the most inferior value of the standard error of the estimate (i.e., root means square error) (RMSE) indicates the well enough performance of the forecasting model.

Figure 13: Comparison of observed and predicted maximum temperature datasets

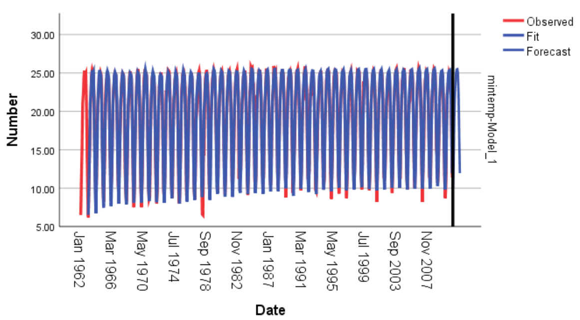

Figure 14: Comparison of observed and predicted minimum temperature datasets

The forecasted 12-months maximum and minimum temperatures of 2011 are present in below table 2 compared with BMD's Sreemangal station dataset. With the most miniature root means square error as 0.78 and 0.79 for both time series, the forecasted and observed data shows almost similarity.

Table 2: Forecasted maximum and minimum temperatures with BMD's recorded dataset

Month	Maximum Temperature			Minimum Temperature		
	BMD	Forecast	RMSE	BMD	Forecast	RMSE
Jan 2011	23.70	25.09	0.78	8.60	9.94	0.79
Feb 2011	28.80	27.93		12.40	12.54	
Mar 2011	31.30	32.05		17.60	17.68	
Apr 2011	33.20	33.04		20.60	21.51	
May 2011	32.20	32.38		22.80	23.25	
Jun 2011	32.30	31.97		25.00	24.97	
Jul 2011	32.50	32.24		25.50	25.45	
Aug 2011	32.40	32.71		25.30	25.49	
Sep 2011	33.30	32.40		25.10	24.93	
Oct 2011	33.00	31.72		22.20	22.39	
Nov 2011	30.10	29.59		14.80	16.94	
Dec 2011	25.90	26.93		12.00	11.95	

4. CONCLUSION

Due to global warming, the temperature is in the more or less increasing trend with the significant seasonality effect. SARIMA(2,0,1)(1,1,2)$_{12}$ and SARIMA(1,0,1)(1,1,1)$_{12}$ are the most fitted model for analyzing the concerned station's monthly temperature data and forecasting the future temperature. Most of the residuals of the maximum and minimum temperature time series follow the straight-line of normality in the Q-Q plot, and having all spikes under the confidence line in the residual ACF and PACF validate the model performance, and R^2 close to 1 and negligible RMSE also make chorus to stand boldly with these SARIMA models. As a significant economically important tea-capital with naturally enriched bio-diversity, for Sreemangal to medicate global warming and mitigate the challenges of unstable climate changes, the roleplay of SARIMA models in forecasting with reasonable precision is very significant for further decision-making.

REFERENCES

Analyse-it. (n.d.). Normal probability (q-q) plot. Retrieved november 9, 2021, from https://analyse-it.com/docs/user-guide/distribution/continuous/normal-probability-plot

Banglapedia. (n.d.). Sreemangal upazila. Retrieved november 7, 2021, from https://en.banglapedia.org/index.php?title=sreemangal_upazila

Barc. (n.d.). Climate information management system. Retrieved november 9, 2020, from http://climate.barcapps.gov.bd/

Bibi, u. M., kaduk, j., & balzter, h. (2014). Spatial-temporal variation and prediction of rainfall in northeastern nigeria. Climate.

Bmd. (n.d.). Bmd. Retrieved november 9, 2021, from http://live.bmd.gov.bd/

Climate-data.org. Retrieved november 7, 2021, from https://en.climate-data.org/asia/bangladesh/sylhet-division/sreemangal-970045/

Doulah, m. S.-u.-. (2018). Forecasting temperatures in bangladesh: an application of sarima models. Environmental science. Https://www.semanticscholar.org/paper/forecasting-temperatures-in-bangladesh%3a-an-of-doulah/1e5c582fb704dbf0c173420641b5cb29abd3b2ee

Educba. (n.d.). Coefficient of determination formula. Retrieved november 9, 2021, from https://www.educba.com/coefficient-of-determination-formula/

Google maps. (n.d.). Sreemangal upazila. Retrieved november 7, 2021, from https://www.google.com/maps/place/sreemangal+upazila/@24.3058311,91.5574841,11z/data=!3m1!4b1!4m5!3m4!1s0x37517a618ce21b2f:0x6fe8a619a6c2bc3d!8m2!3d24.3064554!4d91.7296439

Intergovernmental panel on climate change. (2007). Ar4 climate change 2007: the physical science basis. Https://www.ipcc.ch/report/ar4/wg1/

Mahmud, i., bari, s. H., & rahman, m. T. U. (2016). Monthly rainfall forecast of bangladesh using autoregressive integrated moving average method. Environmental engineering research. Https://doi.org/https://doi.org/10.4491/eer.2016.075

Mannan, m. A., chowdhury, m. A. M., karmakar, s., ahmed, s., & mason, s. J. (2015). Rainfall prediction over northeastern part of bangladesh during monsoon season. Dew-drop, 01, 14–25. Https://www.researchgate.net/profile/samarendra_karmakar/publication/287207846_rainfall_prediction_over_northeastern_part_of_bangladesh_during_monsoon_season/links/5673a69308aee7a42744b405/rainfall-prediction-over-northeastern-part-of-bangladesh-during-mo

Marshall, e. (n.d.). Simple linear regression in spss. Https://www.sheffield.ac.uk/polopoly_fs/1.885152!/file/75_simplelinearregression.pdf

Nau, r. (n.d.). Notes on linear regression analysis. Retrieved november 9, 2021, from https://people.duke.edu/~rnau/regintro.htm

Nury, a. H., hasan, k., & alam, m. J. Bin. (2017). Analysis of time series variations of temperature and its forecast in the northeastern bangladesh. Int. J. Global warming, 13. Https://www.inderscience.com/info/inarticle.php?artid=86283

Paul, s., & roy, s. (2020). Forecasting the average temperature rise in bangladesh: a time series analysisforecasting the average temperature rise in bangladesh: a time series analysis. Journal of engineering science, 11, 83–91. Https://www2.kuet.ac.bd/jes/images/files/v11_1/08_jes_270_15-05-2020.pdf

Pennstate eberly college of science. (n.d.-a). 2.5 - the coefficient of determination, r-squared. Retrieved november 9, 2021, from https://online.stat.psu.edu/stat462/node/95/

Pennstate eberly college of science. (n.d.-b). 4.1 seasonal arima models. Retrieved november 9, 2021, from https://online.stat.psu.edu/stat510/lesson/4/4.1#:~:text=ar%3a,%2b θ q b q

Rahman, m. R., & lateh, h. (2015). Climate change in bangladesh: a spatio-temporal analysis and simulation of recent temperature and rainfall data using gis and time series analysis model. Theoretical and applied climatology, 128. Https://doi.org/https://doi.org/10.1007/s00704-015-1688-3

Stanford university. (2000). Rms error. Https://statweb.stanford.edu/~susan/courses/s60/split/node60.html

The world bank. (n.d.). Climate change knowledge portal. Retrieved november 7, 2021, from https://climateknowledgeportal.worldbank.org/country/bangladesh/climate-data-historical

World climate guide. (n.d.). Climate - bangladesh. Retrieved november 7, 2021, from https://www.climatestotravel.com/climate/bangladesh

Zach. (2021). How to interpret root mean square error (rmse). Https://www.statology.org/how-to-interpret-rmse/