Int. J. Econ. Environ. Geol. Vol. 13 (4) 35-41, 2022
Journal home page: www.econ-environ-geol.org

Open Access
ISSN: 2223-957X

Rainfall forecast using SARIMA model along the coastal areas of Sindh Province

Noor Fatima, *Aamir Alamgir and Moazzam Ali Khan

Institute of Environmental Studies, University of Karachi, Karachi, Pakistan
* Email: aamirkhan.ku@gmail.com

Abstract: Rainfall forecasting is critical for economic activities such as agriculture, watershed management, and flood control. It requires mathematical modelling and simulation. This paper investigates the time series analysis and forecasting of the monthly rainfall for the Sindh coastline, Pakistan. The seasonal autoregressive integrated moving average (SARIMA) model was used for the last three decades (1991-2020) and forecasting was done for the next two years. The model is based on the Box Jenkins methodology. The decomposition of time series plots into trend, seasonal and random components showed a seasonal effect. The Augmented Dickey–Fuller (ADF) and Mann–Kendall (MK) tests showed the inherent stationarity of the rainfall data. The best SARIMA models for monthly rainfall were SARIMA $(1,0,1)(3,1,1)_{12}$ and SARIMA $(1,0,1)(1,1,1)_{12}$ with Akaike information criterion corrected (AICC) values of 1507 and 1387, respectively. The model predictions indicate that, in the years 2021/22, July will likely have the most rainfall, followed by August and June. The diagnostic statistical test values directed that the adequacy of the models is consistent for projected monthly rainfall forecasts.

Keywords: SARIMA, forecast, rainfall, seasonal, coastline

Introduction

The rainfall has a significant impact on a region, especially developing nations where agriculture, the condition of the soil's moisture, water supplies, and agricultural yield are crucial factors in their economies. Its amount fluctuates across time and space and is influenced by a number of intricate physical processes (Akrour et al., 2015). The fluctuation in rainfall pattern during last decades is also evident as a consequence of climate change (Udayanshankara et al., 2016). It is necessary to have precise information and projections about rainfall and other climate variables in order to manage floods, droughts, watershed, crop yield, flood protection, and civil works schedule. Precipitation projections and forecasts were necessary for the majority of water management programs and operations involving water resources.

Because rainfall is mostly erratic and composite, mathematical models and simulation were employed to forecast it. Climate time series forecasting also enables the detection and forecasting of climate change. For of modelling and predicting climatic data series, it is typically necessary to use time series that are essentially stochastic models, such as exponential smoothing, integrated seasonal and non-seasonal ARMA and GARCH (Khandelwal *et al.*, 2015; Chen *et al.*, 2018); Scholars have applied ARIMA and SARIMA models and extensions for varied scientific and technical applications of climatic data thence comprise research on precipitation (Abdul-Aziz *et al*.,2013; Bari *et al*., 2015; Dimri *et al*., 2020); A time series is a convenient way to express the most of the environmental data, including climate data, which are often recorded at regular intervals. These data are typically collected over extended periods of time and show specific patterns that can be recognized, modeled, and forecasted for short periods of time. The cyclic data series with seasonal components is subjected to the use of seasonal ARIMA (SARIMA) models (Wang et al. 2013). The SARIMA models are simple autoregressive moving average (ARMA) models applied to altered time-series, where trend differencing and seasonal differencing remove the time-series' seasonality and non-stationary behaviour.

In the present study, SARIMA models were created to provide long-term forecasts of the monthly rainfall time-series originating from two meteorological stations located along Sindh coastline. The area is regarded as a vulnerable hotspot for climate extremes that have occurred in the last two decades (Fatima et al., 2021). Candidate SARIMA models were created employing the notion of parsimony as opposed to the approach used by Hyndman and Khandakar (2008). Using AIC criteria, the best SARIMA was selected from the candidate models. The primary goal was to model rainfall time series data in order to build models that accurately depict the time series' fundamental structure and properties, as well as to collect and assess prior data. This forecasting predicts future values based on historical data to provide preventive measures for economic activities such as agriculture, watershed management, and flood control that are dependent on rain.

Materials and Methods

The study was conducted on Sindh's coastline, located in Pakistan's southernmost section, between 23°43’ to 25°26’ N and 67°05’ to 68°45’E. The study area has been exposed to significant climatic tragedies during the last two decades in terms of climate variations, floods, cyclones and land degradation (Fig.1). Precipitation in the province of Sindh occurs in two seasons i.e., summer and winter. Generally, heavy

rainfall occurs during the summer season from June to September; and during the winter season (southwestern monsoon) from January to February (Awan, 2003). The rainfall data were gathered from two stations of the Pakistan Meteorological Department using a time series of mean monthly data from 1991 to 2020. The Badin meteorological station (MET-1) is situated between 24°38'N and 67°54'E and was used to cover the district Badin's coastline area. The Karachi Meteorological Station (MET-2) is situated between 24°54'N and 67°08'E and was used to cover the coastline region of the districts of Karachi, Thatta, and Sujawal. The data for the aforementioned stations ware accurate, reliable, continuous, and gap-free. The software XLSTAT was used for data analysis.

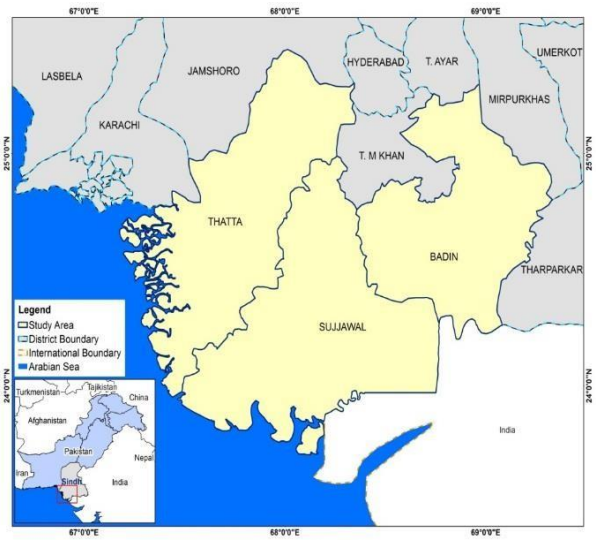


Fig. 1- Study district of Sindh coastline

In order to forecast monthly rainfall, Box and Jenkins (1976) autoregressive integrated moving average (ARIMA) model was utilized. The predictors in the linear (i.e., regression-type) equation for ARIMA model for a stationary time series data are the lags of the dependent variable. A moving average portion MA(q), an integrating part I and an autoregressive part AR(p) are used to depict it. ARIMA models are also used in modelling seasonal data to a large extent. SARIMA (p,d,q) (P,D,Q)m is a seasonal ARIMA model that is produced by inserting supplemental seasonal components (P for seasonal order (AR), D for seasonal differencing and Q for seasonal order (MR)) into ARIMA models. The seasonal components are very similar to the non-seasonal terms, but they also incorporate seasonal backshift operators.

The Augmented Dickey-Fuller (ADF) and Mann-Kendall tests were used (at critical probability p= 0.05) to identify the non-stationarity and trend in the series. The series depicting stationarity was subjected to SARIMA model. The suitable values of parameters p, and q, are determined by evaluating its autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots which are also used to provide a simple test for stationarity. By locating the values of P, D, and Q for m = 12 along the same lines as the p, d, and q were calculated, the seasonality in the datasets is eliminated. Once all the model parameters have been calculated, the best model is selected (p, d and q; P, D and Q). The computed SARIMA residuals are tested for white noise, and the model with the best residual behavior is chosen. The forecasting period is 2021–2022. (2 years). The model's projected results are confirmed for the years 2018 through 2020. The efficacy of the chosen SARIMA model for rainfall to assess the relative quality of statistical model for a given dataset is examined using AIC criterion. The residuals of the model are checked to find if they present within the range of Hessian standard error envelope. Moreover, the residuals are also evaluated for independence, homoscedasticity, normal distribution.

Result and Discussion

Initial data analysis shows that the mean monthly rainfall at MET-1 is 16.99mm, with the greatest record being in Aug-1994 (358.6 mm). Similar to that, the average monthly rainfall at MET-2 is 12.86mm with the largest quantity seen there in July 2003 (270.4 mm). Both stations' time series plots displayed the predictable up-and-down pattern indicative of seasonality. Decomposing time series plots (Figs. 2 and 3) into seasonal and random components allowed for a more thorough analysis.

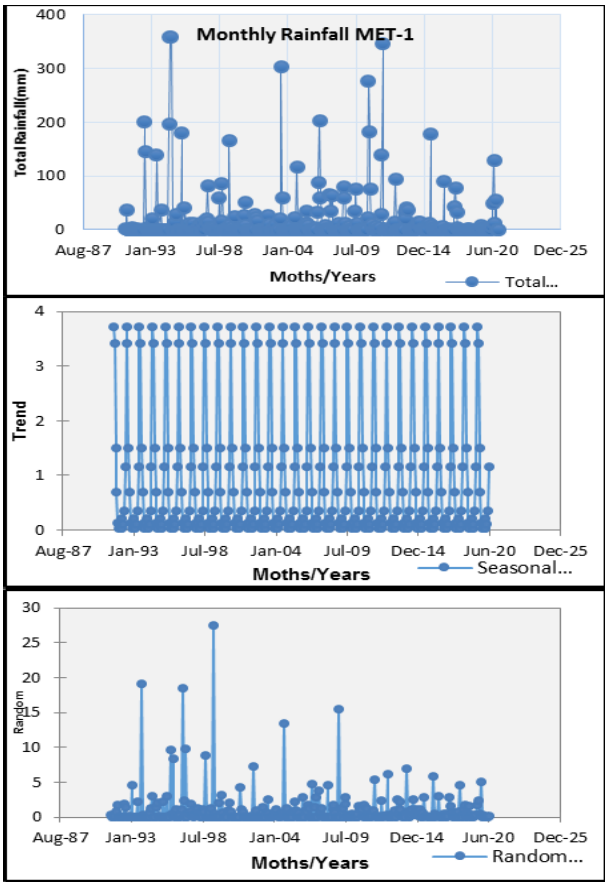


Fig.2 Time series plot of total, seasonal and random of monthly rainfall MET-1

As shown in Figs. 2 and 3, the data had seasonality, with an annual up and down trend. This suggests that the seasonal component's pattern had an impact on the annual average of monthly rainfall data. While the random component remained constant over time, the trend appeared to be extremely consistent across time. The Box-Cox Transformation function was used to transform the series of both stations in order to get around the issue of non-normality of the time series data. The best value of was found to be between -0.077 (MET-1) and -0.210 (MET-2).

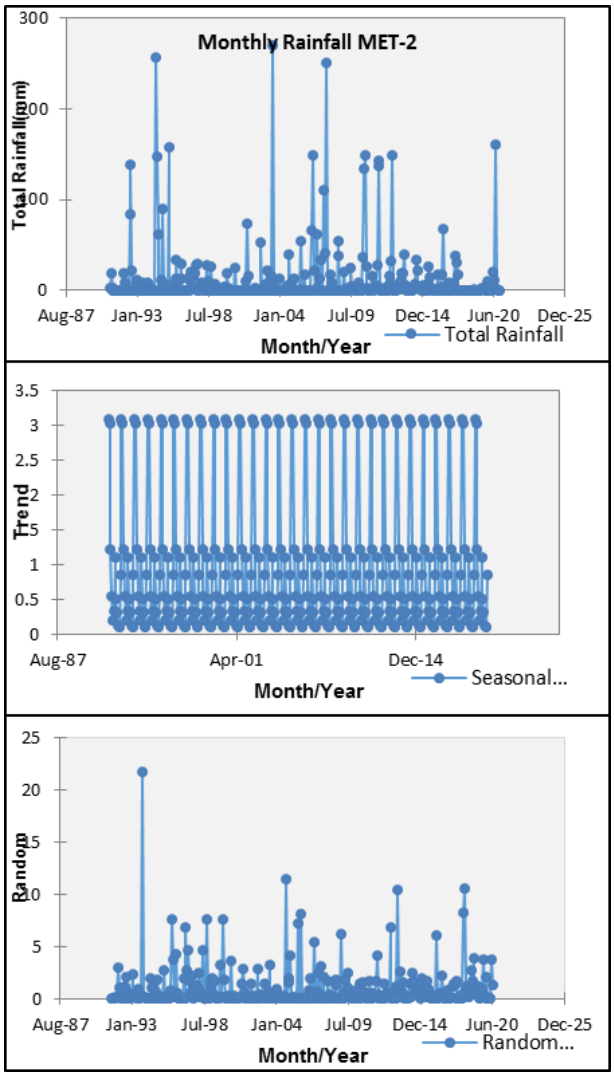


Fig.3 Time series plot of total, seasonal and component random component of monthly rainfall MET-2

The ACF and PACF were examined up to 26-lag delay. The ACF and PACF were examined up to 26-lag delay. The ACF and PACF of both stations exhibit a steady deterioration at several delays, which is important. The data series was not a white noise process because there were substantial peaks in the ACF and PACF plots, which allowed modelling to move forward (Fig. 4). Alternating positive and negative values diminishing at zero with increasing lag was the seasonal pattern that was observed. This confirms the earlier claim that the data were seasonal, necessitating the use of seasonal differencing with a 12-month period.

For the non-seasonal component, after lag 1 the ACF plots of both stations tails off and the PACF plot ends. As a result, the auto-regression (AR) and moving averages (MA) components of the model may start with one lag back and one term in the data series, respectively.

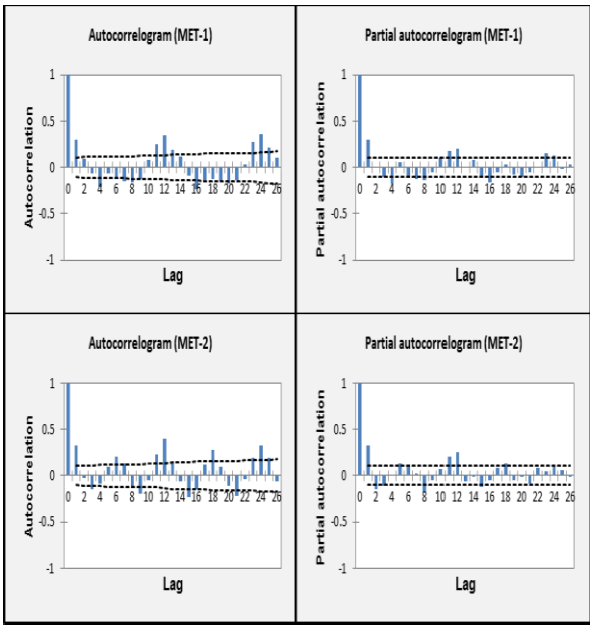


Fig. 4- The ACF and PACF plots of monthly rainfall (mm) MET-1 and MET-2

The ADF test (Table 1) established that there is trend stationarity and that neither of the series has a unit root. There was no trend in either station's rainfall series was detected in Mann-Kendall test. The results supported the conclusions of the ACF and PACF plots, showing that both series have an inherent stationarity. As a result, the model execution requires no differencing, keeping d=0.

Table 1- Stationarity test (α=0.05 at 95% confidence interval)

Test	Observations	MET-1	MET-2
ADF	p-value (one-tailed)	< 0.0001	0.0003
Mann-Kendall	p-value (Two-tailed)	0.12	0.564

The model non-seasonal part showed that the MR and AR process for both stations would be of order 1 (p=1; q=1). Since the series does not require non-seasonal differencing (d=0), 1, 0, and 1 would be the proper values for p, d, and q to fit the time series data of both stations, respectively.

The ACF plot of both stations revealed substantial peaks at delays multiples of 12 for the seasonal

component of the model (12 and 24). The seasonal delays in the PACF plots descended exponentially (12 and 24). As a result, the seasonal component of both models has an autoregressive term of order 1 (P=1) and a moving average term of order 1 (Q=1). Since the series must include seasonal differencing (D=1) for both stations. A preliminary SARIMA $(1,0,1)\ (1,1,1)_{12}$ model is suggested based on the characteristics indicated by the plots.

The most suitable model with the appropriate parameter values (P, Q) has been identified through several models taking P and Q values from 0. For each station, the model with the lowermost AIC criterion value was selected as the preferred one. According to Table 2, the SARIMA $(1,0,1)\ (3,1,1)_{12}$ model provided the greatest match for Badin meteorological station (MET-1), with an AIC criterion score of 1507.19. The putative SARIMA $(1,0,1)\ (1,1,1)_{12}$ was determined to be the best fit model for Karachi meteorological station (MET-2) with an AIC criterion score of 1389.73. The parameter values for both models were within the estimated Hessian standard errors' confidence interval (Table 2). The residuals' homoscedasticity was confirmed by the homoscedasticity plots (Fig. 5) and the outcomes of the Breusch Pagan test and White test (for both stations Table 3). The calculated p-values were higher than the 0.05 level of significance. The residuals' distribution plot verified that they were normal given the rainfall data. The histograms (Fig. 6) demonstrated that, for the rainfall series for MET-1 and MET-2, the residual of the best-selected SARIMA model basically followed normal distribution, a reasonable confidence interval for the future forecast must be produced.

Table 2- Parameters of the Best SARIMA model

Model	AICC Value	Parameter	Value	Hessian standard Error	Lower Bound (95%)	Upper Bound (95%)
SARIMA (1,0,1) (3,1,1)	1507.087	AR(1)	0.022	0.056	-0.088	0.133
		SAR(1)	0.000	0.059	-0.115	0.115
		SAR(2)	0.000	0.058	-0.114	0.114
		SAR(3)	0.000	0.060	-0.117	0.117
		MA(1)	0.000	0.057	-0.112	0.111
		SMA(1)	0.000	0.201	-0.394	0.394
SARIMA (1,0,1) (1,1,1)	1389.731	AR(1)	0.628	0.205	0.226	1.031
		SAR(1)	0.000	0.070	-0.138	0.138
		MA(1)	-0.431	0.236	-0.892	0.031
		SMA(1)	0.000	0.051	-0.101	0.101

Table 3- Diagnostic test (α=0.05 at 95% confidence interval)

Stations	Breusch-Pagan test	White test	Interpretation
MET-1	0.215	0.01	Homoscedastic
MET-2	0.707	0.178	Homoscedastic

Additionally, standardized residuals were shown on P-P and Q-Q plots of normal probability (Fig. 7) essentially followed a straight-line trend, proving that the predicted and observed data series had quantile distributions that were quite comparable. Some of the points that stray from the straight line can be as a result of a mean deviation brought on by the time series data' variability. Because their standardized residuals have the same gamma distribution, the best fitted models should be accurate.

According to (Fig. 8) the Residual Autocorrelogram Function (RACF) and Residual Partial Autocorrelogram Function (RPACF) plots of MET-1 and MET-2 were originate to be inside the 95% confidence limit. They were insignificant, demonstrating the residuals' independence and continuous variance. The residuals' ACF and PACF revealed no significant violations of the model suppositions, demonstrating that they will deliver the necessary level of forecast accuracy.

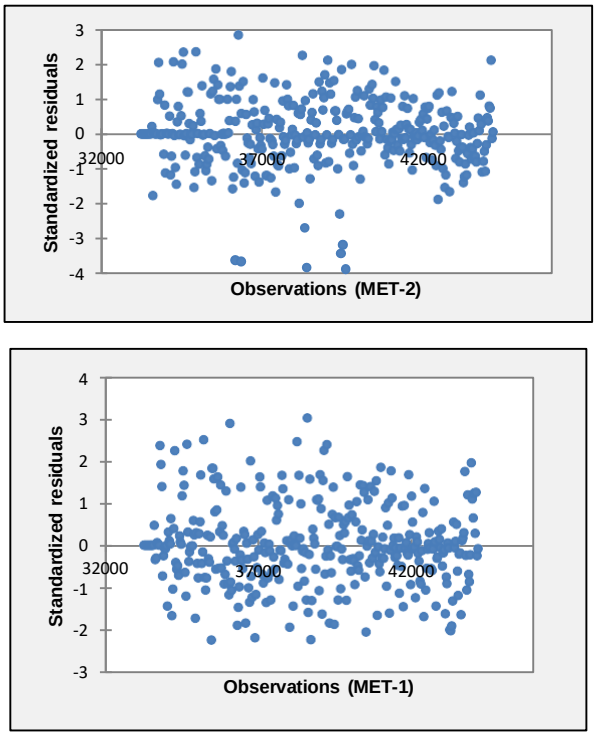
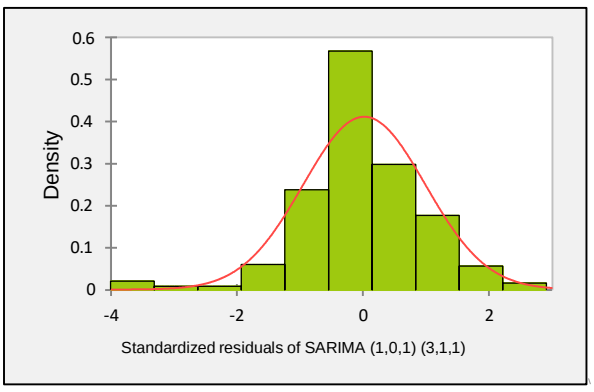


Fig. 5- Distribution of standardized residuals SARIMA(1,0,1) (3,1,1) and SARIMA(1,0,1)(1,1,1)



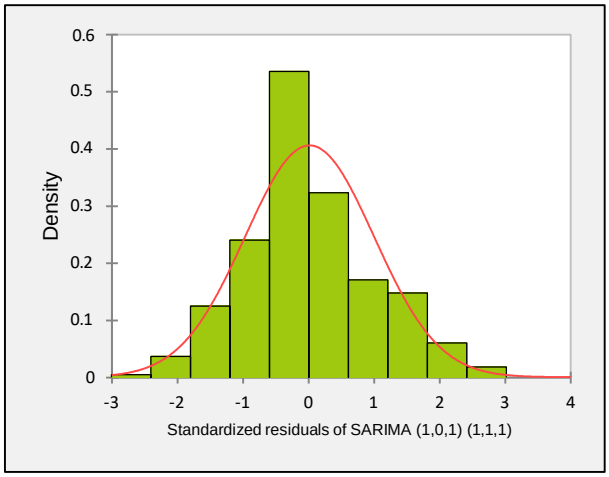


Fig. 6- Normality distribution of the residuals of SARIMA (1,0,1) (3,1,1) and SARIMA (1,0,1) (1,1,1)

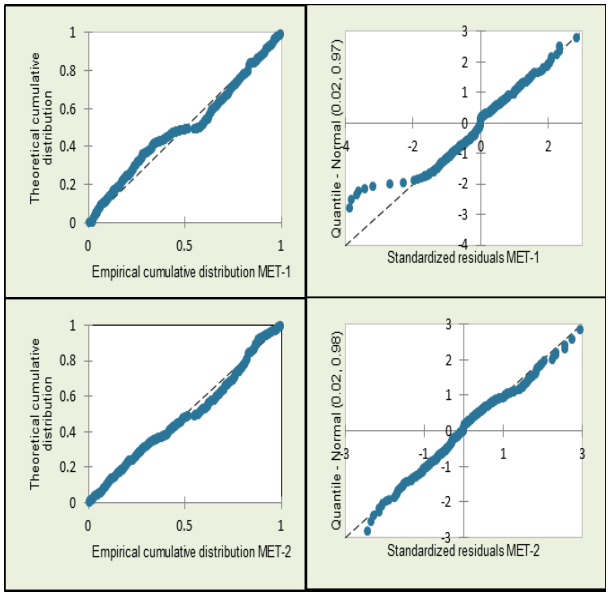


Fig. 7- P-P and Q-Q Plots of the residuals of MET-1 and MET-2

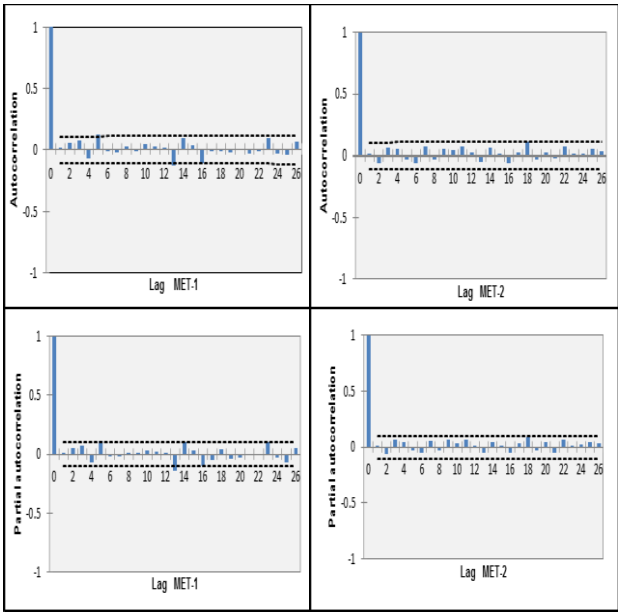


Fig. 8- RACF and Residual RPACF of MET-1and MET-2

The accuracy of the SARIMA model was assessed by comparing the observed data to the produced projected series. To validate the model, the data were split into training data set (Jan, 1991- Dec, 2018) and testing data set (Jan, 2019- Dec, 2020). Additionally, the forecast series were built with 95 percent confidence intervals for a lead period of two years. The original series and its variable seasonal pattern were found to be ahead of the anticipated data series and its residual (Fig.9 and 10).

With a few overestimations of severe rainfall occurrences, particularly during the monsoon, the rainfall forecast is in good agreement with the observed data. The predicted numbers indicate that, for example, in the years 2021/22, there won't be any precipitation in the month of October and that the month of July, trailed by the months of August and June, will have the most precipitation. For both stations from 2021 to 2022, the predicted series confirmed the modelling strategy. The figures of the predicted precipitation for both sites indicated slightly less precipitation during the following two years.

The sector which is highly prone to high or low rainfall is the agriculture. Since Sindh is located at extreme south of Pakistan, it is already at risk due to coastal dynamics, water shortage and lack of natural resources. Climate changes induced irregular rainfall patterns further aggravate these conditions and make the environment further harsh to sustain livelihood opportunities for the local community (Lohano & Mari, 2020). The damages due to heavy rains affecting the crop production has direct and indirect impacts on the agro-based major and minor industries located nearby coastal areas (Dahri *et al.*, 2020). The revenue generated for the farmers after the sale of raw agricultural crops left the farmers with riskier returns. These patterns of uncertainty are causing long term impacts on the productivity of food which upset the economies of agriculture based countries (Wilkinson and Peters 2015; FAO, 2000).

The SARIMA (1,0,1) $(3,1,1)_{12}$ and SARIMA(1,0,1) $(1,1,1)_{12}$ for MET-1 and MET-2 was identified as the best suitable models respectively. These model seemed to have the least adequacy values of sum of mean absolute percentage error (MAPE; MET-1=105.8; MET-2=105.7), squares error (SSE; MET-1=1685.9; MET-2= 1250.5), mean square error (MSE; MET-1=5.2; MET-2=3.8), root mean square error (RMSE; MET-1=2.3; MET-2=1.9). The SARIMA (1,0,1) $(3,1,1)_{12}$ appears to accomplish a little lower than that of the SARIMA(1,0,1) $(1,1,1)_{12}$.

For the coastal region of Sindh, two years' worth of monthly rainfall data could be predicted using the SARIMA model. These residuals of the models consistently predicted future rainfall since they lagged behind the normal distribution and other performance efficiency criteria.

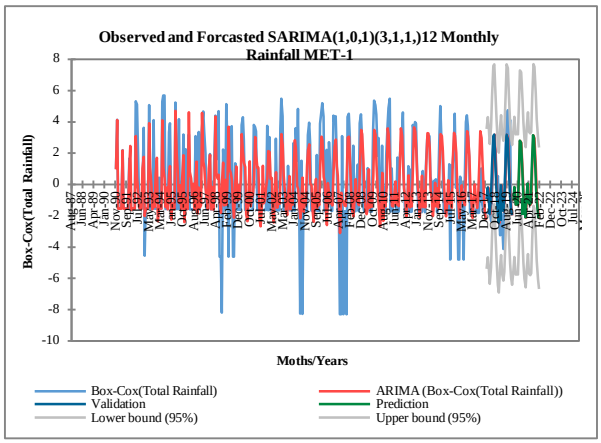


Fig. 9- Observed, synthetic and forecasted series for MET-1

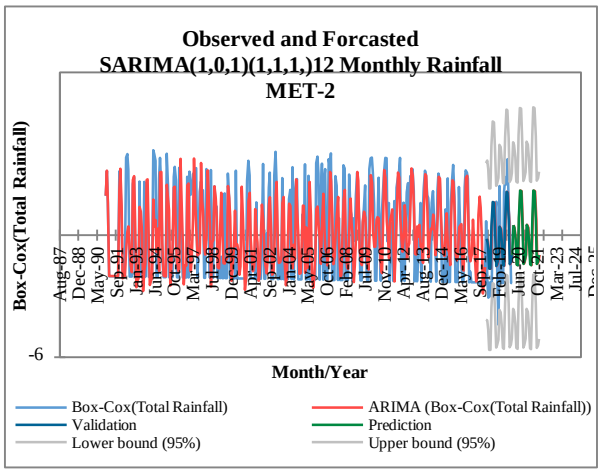


Fig. 10- Observed, synthetic and forecasted series for MET-2

Conclusion

The annual monthly rainfall over the three decades was examined using the SARIMA model. The models were created and tested using time series rainfall data from the Badin and Karachi meteorological stations along the coast-line of Sindh. With some overestimations of high rainfall, particularly in the wet season, the monthly rainfall forecast generally agrees with the observed data. According to the forecasted monthly rainfall amounts, for instance, the wettest months in 2022 and 2023 will be July, August, and June. No rain will fall during the month of October. The SARIMA models were reviewed and evaluated for validation after initial identification and diagnostic checks, and the chosen models were determined to be sufficient for predicting monthly rainfall and temperature on the coastal Sindh region.

References

Abdul-Aziz, A. R., Anokye, M., Kwame, A., Munyakazi, L., & Nsowah-Nuamah, N. N. N. (2013). Modelling and forecasting rainfall pattern in Ghana as a seasonal ARIMA process: The case of Ashanti region. International Journal of Humanities and Social Science, 3(3), 224-233.

Akrour, N., Chazottes, A., Verrier, S., Mallet, C., & Barthes, L. (2015). Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution. Water Resources Research, 51(9), 7417-7435.

Awan, S. A. (2003). Flood forecasting and management in Pakistan. In: Proceedings of Symposium on Water Resources Systems, IAHJS Publ. 281(90–98).

Bari, S. H., Rahman, M. T., Hussain, M. M., & Ray, S. (2015). Forecasting monthly precipitation in Sylhet city using ARIMA model. Civil and Environmental Research, 7(1), 69-77.

Box, G. E., & Jenkins, G. M. (1976). Timiie Series Analysis. For-ecastinig and Conitr-ol (rev. ed.).

Chen, J., Zeng, G. Q., Zhou, W., Du, W., & Lu, K. D. (2018). Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. Energy conversion and management, 165, 681-695.

Dahri, G. N., Talpur, B. A., Nangraj, G. M., Mangan, T., Channa, M. H., Jarwar, I. A., & Sial, M. (2020). Impact of Climate Change on Banana Based Cropping Pattern in District Thatta, Sindh Province of Pakistan. Journal of Economic Impact, 2(3), 103-109.

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.

Fatima, N., Alamgir, A., Khan, M., A., & Mehmood, K. (2021). Conceptual Framework for Climate Vulnerability and Conflicts in the Coastal Districts of Thatta and Sujawal, Sindh, Pakistan. Journal of Biosciences, 18(4), 60-76.

FAO (2000). The State of Food and Agriculture. Lessons from the Past 50 Years Food and Agriculture Organization of The United Nations. ISBN 92-5-104400-7

Hyndman, R.J and Khandakar, Y.(2008). Automatic time series forecasting: The forecast package for R; J. Stat. Softw.27(3) 1–42, https://doi.org/10.18637/jss.v027.i03.

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. Procedia Computer Science, 48, 173-179.

Lohano, H. Das, & Mari, F. M. (2020). Climate Change and Implications for Agriculture Sector

in Sindh Province of Pakistan. Mehran University Research Journal of Engineering and Technology, 39(3), 668–677.

Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An ARIMA-ANN hybrid model for time series forecasting. Systems Research and Behavioral Science, 30(3), 244-259.

Wilkinson, E., & Peters, K. (2015). Climate extremes and resilient poverty reduction: development designed with uncertainty in mind. Overseas Development Institute, London.

Udayashankara, T. H., Murthy, B. S., & Madhukar, M. (2016). Impact of climate change on rainfall pattern and reservoir level. Journal of Water Resource Engineering and Management, 3(1), 10-14.