SARIMA Modeling and Forecasting of Seasonal Rainfall Patterns in India

Subbaiah Naidu K.CH.V
Department of Mathematics, BT College, Madanapalle, India.

Abstract:
Rainfall is of critical importance for many people particularly those whose livelihoods are dependent on rain fed agriculture. Predicting the trend of rainfall is a difficult task in meteorology and environmental sciences. Statistical approaches from time series analysis provide an alternative way for predicting the patterns of rainfall. This paper describes an empirical study of modeling and forecasting seasonal rainfall patterns in India. The Box-Jenkins SARIMA Methodology has been adopted for forecasting, the diagnostic checking has shown that the seasonal model $(0,0,0)\times(0,1,1)_4$ *fitted to the series is appropriate, and forecast are obtained on the basis of the fitted model. Seasonal ARIMA model was a proper method for modeling and predicting the time series of seasonal rainfall patterns.*

Key Words — *Rainfall, Autocorrelation, Partial autocorrelation, SARIMA, Forecast.*

1. Introduction

The pattern and amount of rainfall are among the most important factors that affect agricultural systems. The analysis of rainfall records for long periods provides information about rainfall patterns and variability. Rain plays a major role in hydrology that finds its greatest applications in the design and operations of water resources, engineering works as well as agricultural systems. Stochastic models are used in operational hydrology to generate synthetic time series which exhibit similar statistical characteristics as the observed data. One of the crucial problems in stochastic modeling of hydrologic time series is to find a model which is capable of preserving the historical statistical characteristics that affect the variability of the data. Furthermore, the model should be capable of reproducing certain statistics that are related to the intended use of the model. Generally, the properties of a process include the mean, variance, skewness, and the correlation structure of the data.

Autoregressive Integrated Moving Average Model (ARIMA), is a widely used time series analysis model for analyzing chronological data in statistics. ARIMA model was firstly proposed by Box and Jenkins in the early 1970s, which is often termed as Box-Jenkins model or B-J model for simplicity (Stoffer and Dhumway, 2010)[1]. ARIMA is a type of short-term prediction model in time series analysis, because this method is relatively systematic, flexible and can grasp more original time series information. This is the most widely used method in meteorology, engineering technology, Marine, business statistics and prediction technology, (Kantz and Schreiber, 2004; Cryer and Chan, 2008)[3].

The general ARIMA model is also applicable for non-stationary time series that have some clearly identifiable trends (Stoffer and Dhumway, 2010)[1]. The general notation of Auto Regressive Integrated Moving Average model as ARIMA (p,d,q), where p is the order of autoregressive part, d is the order of differencing and q is the order of the moving average process and all are non-negative integers. General ARIMA model named as non-seasonal ARIMA (p,d,q) model, we should also consider some periodical time series. The periodicity of the time series is usually due to seasonal changes like quarterly, monthly and degree of weeks change or some other natural changes. If a time series has seasonal variation, there will be autocorrelation at lag S and possibly at multiples of that lag, depending on the persistence and nature of that autocorrelation. This seasonal dependence can be described by seasonal ARIMA (SARIMA) model of order (P,D,Q) if it is necessary to take seasonal difference. Weesakul and Lowanichchai (2005)[12] developed ARIMA models most appropriate to forecast annual rainfall in all regions of Thailand with acceptable accuracy, which are able to fulfill the requirement for agricultural water allocation planning.

Considering the data relation, we can build a multiplication $SARIMA(p,d,q)\times(P,D,Q)_s$, the term $(P,D,Q)_s$ gives the order of the seasonal part. The model has been successfully applied in many seasonal time

series. In practical applications, the order of model SARIMA is usually not too large (Guo, 2009)[9]. If the period of time series equals to 4, it can be denoted as $SARIMA(p,d,q)\times(P,D,Q)_4$. In the adjustment of the season, this is a very convenient, steady model.

2. Materials and Methods:

2.1 Seasonal ARIMA Model

The general form of multiplicative seasonal model $\text{SARIMA}(p,d,q)\times(P,D,Q)_s$ is given by

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d x_t = \mu + \Theta_Q(B^s)\theta_q(B)a_t \quad (1)$$

Where,

$\Phi_P(B^s) = 1 - \Phi B^s - \Phi_2 B^{2s} - \text{L L } \Phi_P B^{Ps}$ is the seasonal autoregressive operator of order P.

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \text{L L } - \phi_p B^p$ is the regular autoregressive operator of order p.

$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \text{L L } - \Theta_Q B^{Qs}$ is the seasonal moving average operator of order Q;

$\theta_q(B) = 1 - \theta B - \theta_2 B^2 - \text{L L } - \theta_q B^q$ is the regular moving average operator of order q;

Where μ is the intercept term or mean term $\nabla^d = (1-B)^d; \nabla_s^D = (1-B^s)^D;\ B^k x_t = x_{t-k}$

a_t ,the non-stationary time series a_t is the usual Gaussian white noise process; s is the period of the time series and B is the backshift operator.

In this study, we concentrate on southwest monsoon monthly precipitation time series. If the seasonal period of the series s = 4. It is clear that we may then rewrite Equation (1) as:

$$\Phi_P(B^4)\phi_p(B)\nabla_4^D\nabla^d x_t = \mu + \Theta_Q(B^4)\theta_q(B)a_t \quad (2)$$

2.2 Model Identification

In time series analysis, the most crucial steps are to identify and built a model based on the available data. At this stage it is necessary to identify the values of (p,d,q) and $(P,D,Q)_s$. The goal is to employ computationally simple techniques to narrow down the range of parsimonious models. The Box-Jenkins method is only suitable for stationary time series data. In such case, first, we should construct a time plot of the data and inspect the graph for any anomalies (Cryer and Chan, 2008) [2]. Through careful examination of the plot, usually we get an idea about whether the series contains a trend, seasonality, outliers; non-constant variances, and other non-normal and non-stationary phenomena. This will give us to choose proper data transformation. If the variance grows with time, we should use variance-stabilizing transformations and differencing. A series with non-constant variance often needs a logarithmic transformation. The next step is to identify preliminary values of autoregressive order p, the order of differencing d, the moving average order q and their corresponding seasonal parameters P, D and Q. Here, the autocorrelation function (ACF), the partial autocorrelation function (PACF) and inverse autocorrelation function (IACF) are the most important elements (Stoffer and Dhumway, 2010) [1]. The ACF measures the amount of linear dependence between observations in a time series that are separated by a lag q. The PACF helps to determine how many autoregressive terms p are necessary. The IACF is useful for detecting over-differencing [10] and if the data have been over-differenced [11] - [14], the IACF looks like an ACF from a non-stationary process. The parameter d is the order of difference frequency changing from non-stationary time series

to stationary time series. Furthermore, a time series plot and ACF of data will typically suggest whether any differencing is needed. If differencing is called for, the time plot will show some kind of linear trend.

When preliminary values of D and d have been fixed, the next step is to check the ACF and PACF of $\nabla_4^D \nabla^d x_t$ to determine the values of P, Q, p and q. We can further choose parameters using minimum Akaike's Information Criterion, called Bayesian Information Criterion (BIC) is computed as $-2\ln(L)+\ln(n)k$, where L is the likelihood function and k is the number of parameters (Stoffer and Dhumway, 2010) [1].

2.3 Parameters Estimation

Once the model is tentatively established, the parameters and the corresponding standard errors can be estimated using statistical techniques, such as Conditional Least Square Estimation (CLS) and Yule-Walker. The CLS estimates are conditional on the assumption that the past unobserved errors equal to 0. The series x_t can be represented in terms of the previous observations, as $x_t = a_t + \sum_{i=1}^{\infty} \pi_i B^i$. The weights (π) are computed from the ratio of the ϕ and θ polynomials as $\frac{\phi(B)}{\theta(B)} = 1 - \sum_{i=1}^{\infty} \pi_i B^i$. The CLS method produces estimates minimizing $\sum_{i=1}^{n} \hat{a}_t^2 = \sum_{i=1}^{n} (x_t - \sum_{i=1}^{\infty} \hat{\pi}_i x_t - i)^2$ where the unobserved past values of x_t are set to 0 and $\hat{\pi}_i$ are computed from the estimates of ϕ and θ at each iteration. The k -step forecast of x_{t+k} is computed as

$$\hat{x}_{t+k} = \sum_{i=1}^{k-1} \hat{\pi}_i \hat{x}_{t+k-i} + \sum_{i=k}^{\infty} \hat{\pi}_i \hat{x}_{t+k-i}$$

2.4 Diagnostic Checking

After parameter estimation, we have to assess model adequacy by checking whether the model assumptions are satisfied. The basic assumption is that they a_i are white noise. Generally, this step includes the analysis of the residuals as well as model comparisons. If the model fits well, the standardized residuals should behave as an independent and identically distributed sequence with zero mean and constant variance (Cryer and Chan, 2008) [2]. A standardized residuals plot or a Q-Q plot can help in identifying the normality (Stoffer and Dhumway, 2010) [1]. The model should pass the parametric test and diagnostic check.

2.5 Fitting and Prediction

Once a model has been identified and all the parameters have been estimated, we can predict future values of a time series with this model.

3. Data

In this study, the time series is the weighted average seasonal rainfall data of India, from 1951-2014, obtainable from Open Government Data (OGD) Platform India. Here there are 4 seasons; seasonal-I having months January & February, season-II having months March, April & May, season-III having months June, July, August & September and season-IV having the months October, November & December. The data processing tool is the SAS software. Time series plot is shown in **Fig. 1**. The descriptive statistics for our data are summarized in **Table 1.**

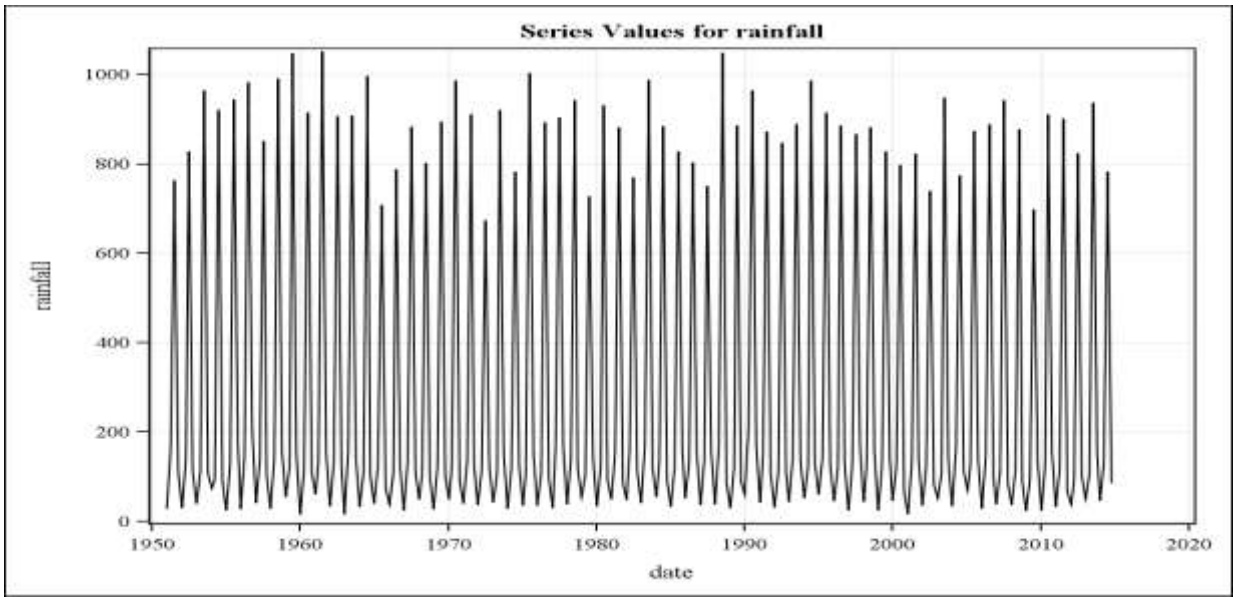

Fig. 1. Time series of seasonal rainfall data for All India (1951-2014)

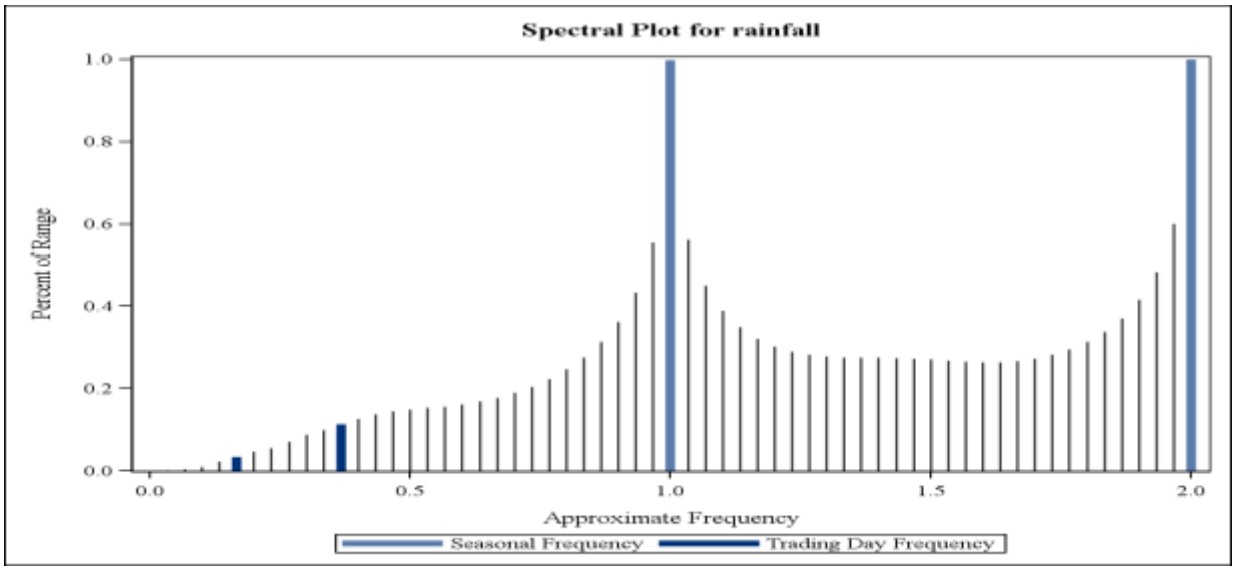

Fig. 2: Spectral Plot for weighted average seasonal rainfall series for All India (1951-2014)

Basic Statistical Measures and Moments			
No of observations	256	Range	1036
Mean	292.53	Inter quartile Range	374.35
Std Deviation	345.09	Skewness	1.1687
Coefficient of Variation	117.97	Kurtosis	-0.4859

Table 1. Basic Statistics for weighted average seasonal rainfall (in mm) of India

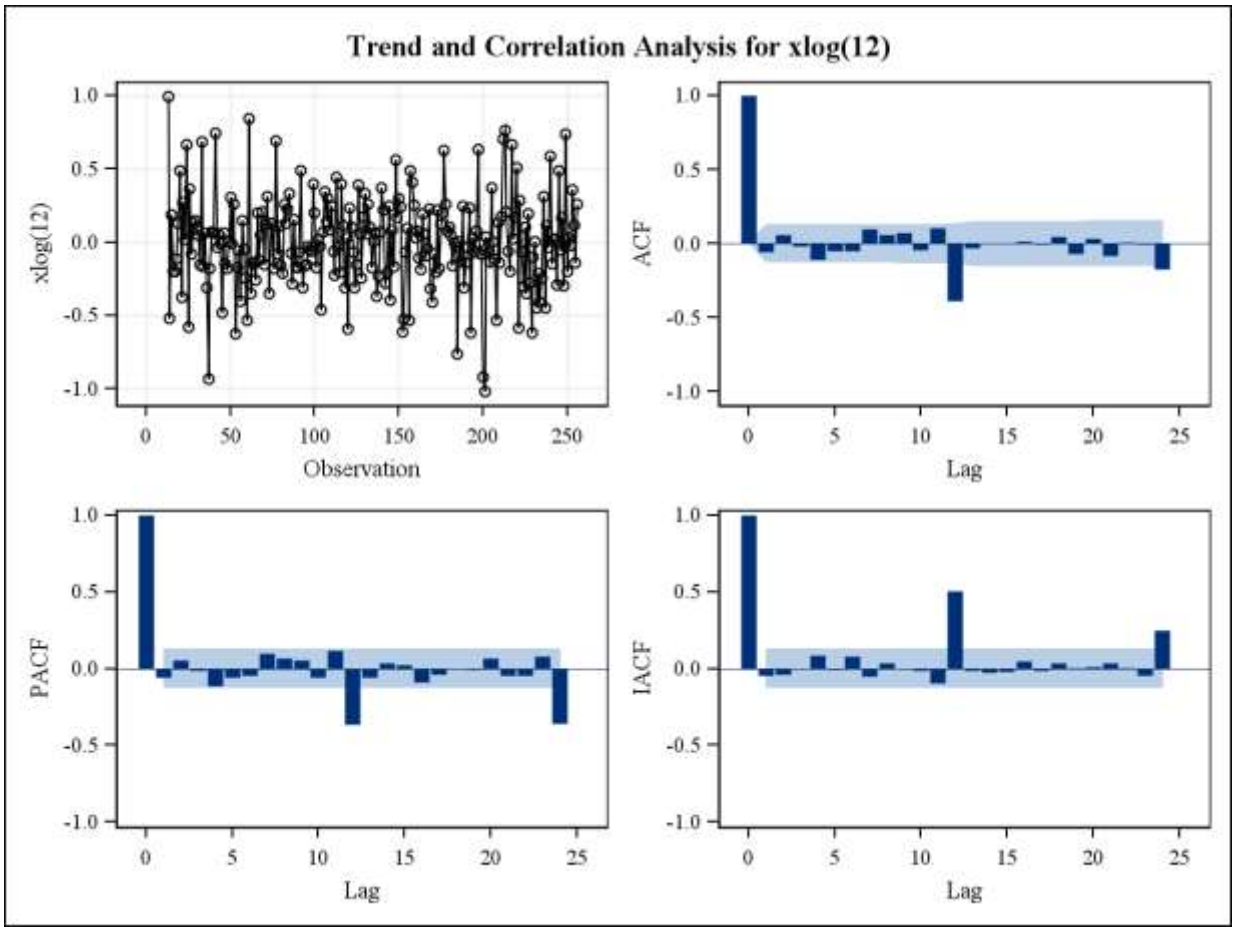

Fig. 3: Trend and Correlation analysis, Autocorrelation (ACF), Partial Autocorrelation (PACF) and Inverse Autocorrelation (IACF) for transformed time series of weighted rainfall data in India.

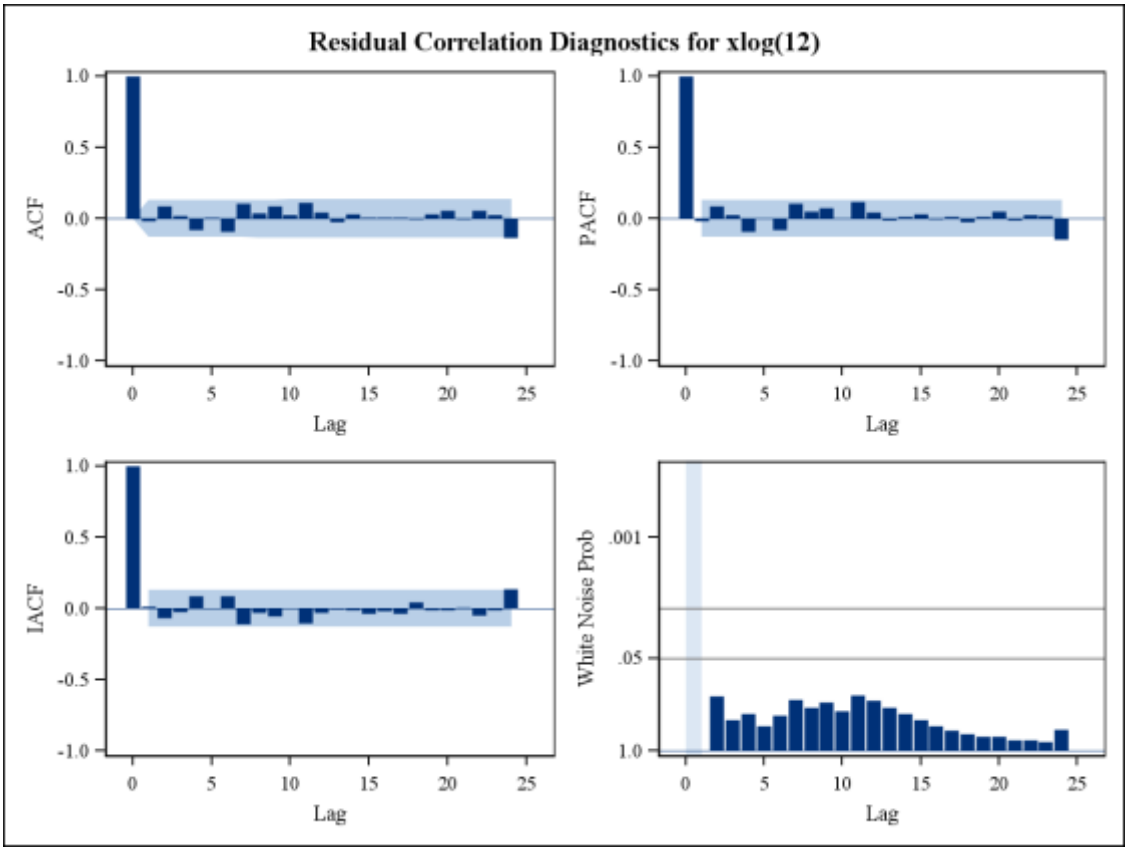

Fig.4: Autocorrelation (ACF), Partial Autocorrelation (PACF), Inverse Autocorrelation (IACF) and White Noise Probability for Residual and Correlation diagnostics for xlog (12).

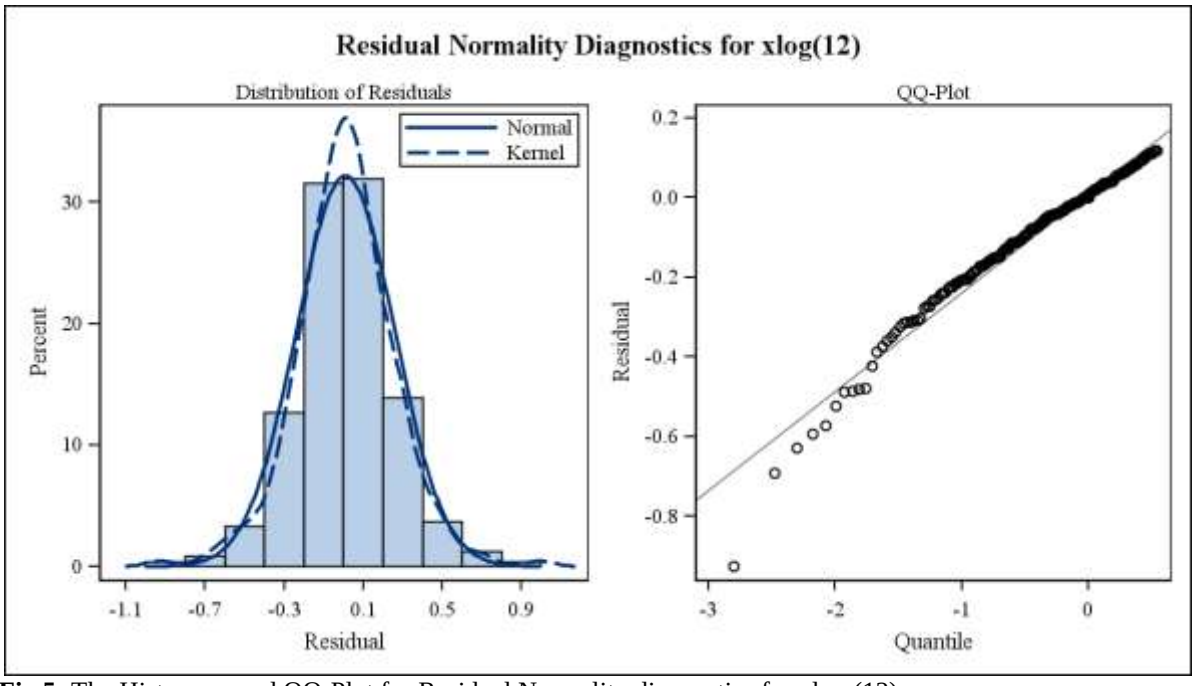

Fig.5: The Histogram and QQ-Plot for Residual Normality diagnostics for xlog (12).

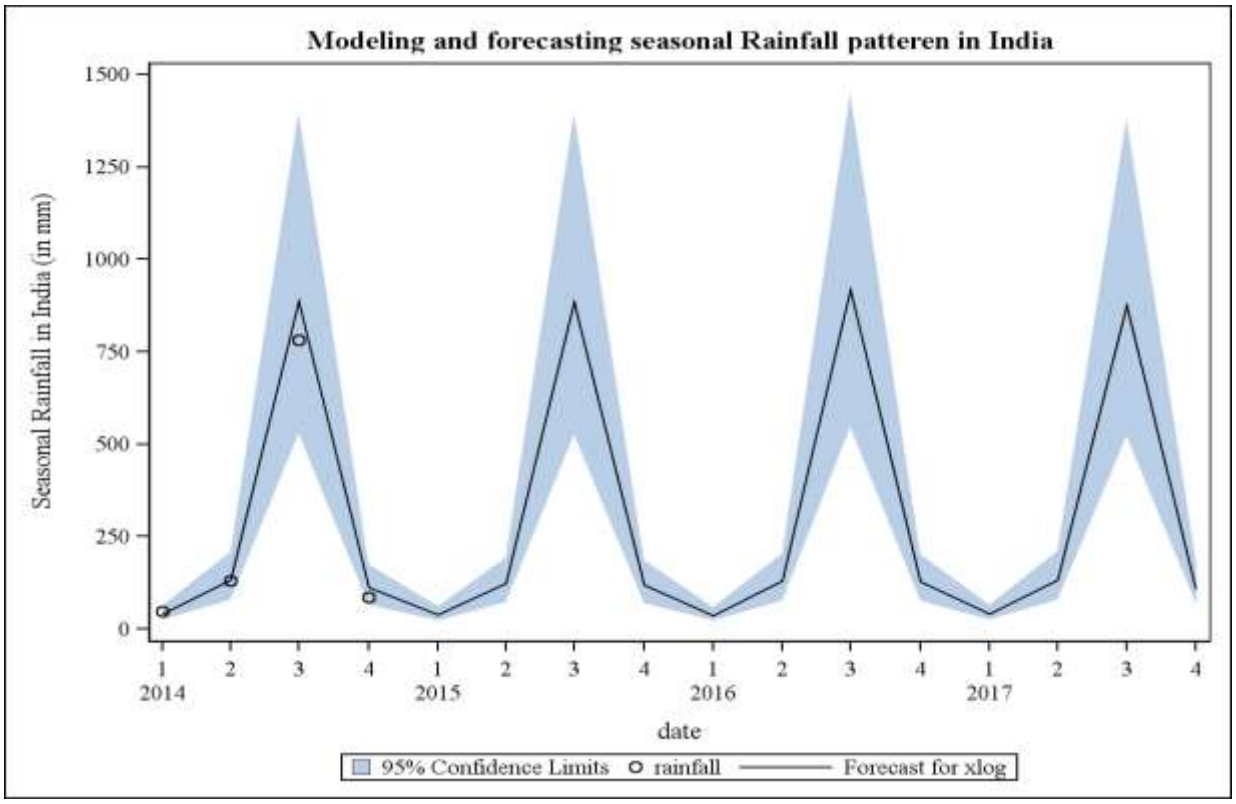

Fig.6: Forecasting the seasonal Rainfall patterns in India based on the Seasonal ARIMA $(0,0,0)\times(0,1,1)_4$ model

4. Results and Discussions

Fig. 1 show a seasonal fluctuation occur every 4 seasons, namely seasonal-I having months January & February, season-II having months March, April & May, season-III having months June, July, August & September and season-IV having the months October, November & December resulting in s = 4. For satisfying the stationary condition, we take logarithmic transformation. The ACF, PACF and IACF of the transformed data, $\log(x_t)$, t = 1, 2,…, 256, are shown in **Fig. 3**. The model parameters are estimated using Conditional Least Square Estimation (CLS). It can be observed that the parameters of model Seasonal ARIMA $(0,0,0)\times(0,1,1)_4$ are all significant. The result indicates that we consider the model and choose this is the best model based on minimum AIC known as BIC criteria. From Fig.5, the Histogram and QQ-Plot for Residual Normality diagnostics for $\log x(12)$ shows us the model fit is well although a small amount of autocorrelations still remains. Then from equations (1) & (2) the fitted form of the model in this case is

$$(1-B)(1-B^4)\log(x_t) = \mu + (1-B)(1-\Theta B^4)a_t \qquad (3)$$

This implies $\log x_t = \log x_{t-1} + \log x_{t-4} - \log x_{t-5} + \mu + a_t - a_{t-1} - \Theta a_{t-4} + \Theta a_{t-5}$

Therefore the fitted Seasonal ARIMA $(0,0,0)\times(0,1,1)_4$ model for this region is

$$(1-B)(1-B^4)\log(x_t) = 0.0004 + (1-B)(1-0.8660B^4)a_t \qquad (4)$$

i.e., $\log x_t = \log x_{t-1} + \log x_{t-4} - \log x_{t-5} + 0.0004 + a_t - a_{t-1} - 0.8660a_{t-4} + 0.8660a_{t-5}$

Therefore,

$$\hat{x}_t = Exp(\log x_{t-1} + \log x_{t-4} - \log x_{t-5} + 0.0004 + a_t - a_{t-1} - 0.8660a_{t-4} + 0.8660a_{t-5}) \quad (5)$$

With the white noise variance $\hat{\sigma}_e^2 = 0.0618$ and with AIC=15.0201and BIC=22.01443.

Finally, forecasting's based on the fitted model for the next three years and also the comparison between actual values and the forecast values are shown in Fig. 6. Fig. 2 shows the Spectral Plot for weighted average of seasonal rainfall series for All India (1951-2014). Practically from the stochastic environmental factors, such as geographic location, climate and temperature, the model state of rainfall is a complicated dynamical system. The time series model in study does not model the extreme values well. Further extensions of study may be undertaken by considering X-12 ARIMA seasonal adjustment method. This procedure makes additive or multiplicative adjustments and creates an out data set that contains the adjusted time series and intermediate calculations.

5. Conclusions

In this study, an ARIMA model incorporates the seasonality of time series. Using the time series of seasonal rainfall data in India, we build a Seasonal ARIMA $(0,0,0)\times(0,1,1)_4$ model. The model fitted the data well and the stochastic seasonal fluctuation was successfully modeled except some extreme values. The predictions based on this model indicate that the rainfall patterns in India almost same in the next three years. The increasing trend is consistent and this changing trend reminds us to make proper strategies for planning the agro sector and drinking water purposes.

6. Summary

We proposed a suitable Seasonal ARIMA model for forecasting average monthly rainfall in the Indian region by using logarithmic transformation. All the parameters have been estimated by Seasonal ARIMA $(0,0,0)\times(0,1,1)_4$ model and also we can predict future values of a time series with this model.

7. References

[1] Stoffer, D.S. and R.H. Dhumway, Time Series Analysis and its Application. 3rd Edn, Springer, New York, ISBN-10: 1441978658, pp: 596, 2010.

[2] Cryer, J. D. and K.S. Chan, Time Series Analysis with Application in R. 2nd Edn., Springer, New York, ISBN-10: 0387759581,2008, pp: 491.

[3] Kantz, H. and T. Schreiber. Nonlinear Time Series Analysis. 2nd Edn, Cambridge University Press, Cambridge, ISBN-10: 0521529026, 2004, pp: 369.

[4] He S.Y., 2004, Applied time series analysis 1st Edn., Peking University press, Beijing.

[5] Eni, D. and Adeyeye, F.J. (2015) Seasonal ARIMA Modeling and Forecasting of Rainfall in Warri Town, Nigeria. Journal of Geoscience and Environment Protection, **3**, 91-98. http://dx.doi.org/10.4236/gep.2015.36015

[6] Metrine Chonge, Kennedy Nyongesa , Omukoba Mulati, Lucy Makokha, Frankline Tireito (2015) A Time Series Model of Rainfall Pattern of Uasin Gishu County, IOSR Journal of Mathematics, Volume 11, Issue 5, pp. 77-84

[7] H. R. Wang1, C. Wang1, X. Lin2, and J. Kang2 An improved ARIMA model for precipitation simulations, Nonlin. Processes Geophys., 21, 1159–1168, 2014

[8] Wang, J., Y.H. Du and X.T. Zhang, Theory and Application with Seasonal Time Series. 1st Edn, Nankai University Press, Chinese, 2008.

[9] Guo, Z.W., The adjustment method and research progress based on the ARIMA model. Chinese J. Hosp. Stat., 161:2009, 65-69.

[10] Chartfield,C, "Inverse Autocorrelations" Journal of Royal Statistical Society, A142,1980, 363-377.

[11] Brocklebank,J.C. Dickey D.A., SAS System for Forecasting Time Series 2nd edition, Cary North Carolina: SAS Institute Inc., 2003

[12] Weesakul, U. and Lowanichchai, S. "Rainfall Forecast for Agricultural Water Allocation Planning in Thailand" Thammasat International Journal of Science and Technology, 10(3), 2005, pp. 18-27.

[13] Xinghua Chang, Meng Gao, Yan Wang and Xiyong Hou, Seasonal Autoregressive Integrated Moving Average Model for Precipitation Time Series. Journal of Mathematics and Statistics 8(4): 2012, pp. 500-505.

[14] Momani, M. and P.E. Naill, Time series analysis model for rainfall data in Jordan: Case study for using time series analysis. Am. J. Environ. Sci., 5: 599-604. 2009, DOI: 10.3844/ajessp.2009.599.604