Journal of Root Crops
Indian Society for Root Crops
ISSN 0378-2409, ISSN 2454-9053 (online)
Journal homepage: https://journal.isrc.in

Time series modelling of monthly rainfall in southern Kerala

R. S. Neethu* and Brigit Joseph

College of Agriculture, Vellayani, Thiruvananthapuram 695 522, Kerala, India

Abstract

This paper aimed to fit SARIMA model based on Box-Jenkins methodology to the time series data corresponds to monthly rainfall in three agro climatic regions *viz*., Regional Agricultural Research Station (RARS) Vellayani, RARS Kumarakom and Cardamom Research Station (CRS), Pampadumpara representing different regions of Southern part of Kerala. The empirical model gave a picture of climate change scenario happened in both temporal and regional wise. The SARIMA model was fitted to monthly rainfall for all the regions Vellayani, Kumarakom, and Pampadumpara using the data for 31 years from 1991 to 2021. The best identified SARIMA models for rainfall were ARIMA (1, 0, 0) (0, 1, $1)_{12}$, ARIMA (0, 0, 0) (0, 1, $1)_{12}$ and ARIMA (0, 0, 0) (0, 1, $1)_{12}$. The model parameters were obtained by using maximum likelihood method and the best model were selected using Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and Hannan-quinn coefficient. The adequacy of the check of the selected models confirmed that the selected models were free from autocorrelation and the residuals are normally distributed.

Keywords: Seasonal Autoregressive Integrated Moving Average, Regional Agricultural Research station, Akaike Information Criteria, Bayesian Information Criteria

Introduction

Indian agriculture has been historically explained as a gamble with monsoon because agricultural activity in most parts of the country depends mainly on monsoon. India is heavily dependent on South-West monsoon (June-September) for most of its annual rainfall. The Kerala state, known as "Gateway of monsoon in India" is one of the unique regions in the humid tropical monsoon climate which receives high solar radiation and warm temperature throughout the year since it is at a short distance away from the equator. Unimodal and bimodal distribution of rainfall with undulating topography, varied soil types and sharp changes in physiography (below msl to 2500 m above msl), together with 44 rivers, many freshwater lakes and estuarine backwaters give rise to contrasting ecological units congenial for high biological activity, contribute for its rich biodiversity in Kerala. The principal rainy seasons in Kerala are the South-west monsoon (June-September) and the North-East monsoon (October-November). The pre-monsoon months (March-May) are characterized by major thunderstorm activity in the state and winter months (December-January) are marked by low clouding and low rainfall season. Time series modeling of weather parameters especially rainfall will describethe overall variations noticed in the pattern and predict the future distributional behavior.

Materials and Methods

The time series approach used in this study is based on ARIMA - Box-Jenkins methodology. ARIMA uses the autocorrelation relationship exists in the data set for model development and forecasting.

* Corresponding author
E-mail: neethurs37@gmail.com; Tel: +91 9496790080

Received: 11 July 2021; Revised:09 August 2021; Accepted: 20 August 2021

Stationarity and differencing

Stationary time series data is characterized by its unique nature of time independency of its various properties like mean, variance. A time series $\{x_t\}$ is said to be strictly stationary, if the joint probability distribution of observations $(x_t, x_{t+1}, ..., x_{t+n})$ is exactly same as the joint probability distribution of observations $(x_{t+ +1}, x_{t+ +2}, ..., x_{t+ +n})$ for every point (t, t + 1, ..., t + n) where h is the time space. The process $\{x_t\}$ is said to be weakly stationary, if it has a constant mean, finite variance and its auto-covariance function $\gamma(t,s)$ depends only on the time lag $|t-s|$. There are many ways in which a time series fails to be stationary, and those are said to be non-stationary time series. Modelling of a non-stationary data will have no sense, so data should be stationary before fitting a model. By the method of differencing non-stationary data can be converted to stationary.

Differencing will stabilize the mean of the time series by eliminating or reducing trend and seasonality. Differenced series will be the change between consecutive observations. Ordinary differencing and seasonal differencing are the common ways to eliminate non-stationarity in the data.

First order differenced series:

$$y'_t = y_t - y_{t-1}$$

Second order differenced series:

$$y''_t = y'_t - y'_{t-1}$$

$$y''_t = y_t - 2\, y_{t-1} + y_{t-2}$$

Usually, ordinary second order differencing will be enough to make the data stationary. Sometimes seasonal differencing will also be found necessary to remove non-stationarity. This is nothing but difference between consecutive observations in the same season denoted as $y'_t = y_t - y_{t-m}$, where m is seasonal term.

Unit root test

The modern technique used to detect stationarity of the time series data is through unit root test. Several unit roots tests are available such as Augmented Dickey Fuller test (ADF), Kwiatkowski-Phillips-Schmidt-Shin(KPSS) test *etc.* In this study ADF test is used for detecting the stationarity.

The null hypothesis and the alternative hypothesis for ADF test was:

H_0: Presence of unit root indicating time series data is non stationary.

H_1: Absence of unit root indicating time series data is stationary.

The test statistic for ADF test is defined as follows:

$$DE_t = \frac{\gamma}{SE(\gamma)}$$

If DE_t was found greater than critical value or *p*-value less than 0.05, then H_0 was rejected.

Autocorrelation and Partial autocorrelation functions (ACF and PACF)

The classical method used to determine whether data is stationary or not is by analyzing the nature of ACF and PACF plots. These plots graphically summarize the strength of association betweenobservations in present time with its previous period.

Auto correlation is the correlation between observations of a variable taken at different time points. Auto Correlation Function (ACF) plots are widely for checking randomness in a data set. This randomness is ascertained by computing auto correlations for data values at varying time lags. Partial Auto Correlation Function (PACF) of $\{Z_t\}$ is a partial correlation coefficient between $\{Z_t\}$ and $\{Z_{t-k}\}$ by fixing the effect of others. PACF of order k is the correlation coefficient between $\{Z_t\}$ and a suitable linear combination of $Z_t, Z_{t-1}, ...,$. ACF and PACF plots are drawn by considering correlation coefficients on the y-axis with number of lags in the x-axis.

Autoregressive model (AR Model)

In an autoregressive (AR) Model, each value in a series should be a linear function of the preceding value or values. In a first-order autoregressive process, only the single preceding value is used or in a second-order process, the two preceding values are used, and so on. These processes are commonly indicated by the notation AR(p) or ARIMA(p-0-0), where the number in parentheses indicates the order.

AR model of order p can be written as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

Moving average model (MA Model)

In this model, instead of considering past values of forecast variables past values of forecast errors are considered in the regression equation. In a moving-average process, each value is determined by the weighted average of the current disturbanceand one or more previous disturbances. The order of the moving-average process specifies how many previous disturbances are averaged into the new value.

MA model of order q can be written as

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

Non-seasonal ARIMA Model

Combination of AR and MA models along with order of integration or difference will form an Autoregressive Integrated Moving Average (ARIMA) model.

The full model will be in the form:

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_2 y'_{t-p} + \theta 1\ \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where y'_t the differenced series and right-hand side contain predictors of lagged values of y_t and ε_t (residual term).

The form of ARIMA (p,d,q) can also be written as,

$$\phi+(B)\ (1\text{-}B)^d Z_t=\theta(B)\,\varepsilon_t$$

Where ϕ – Coefficient of non-seasonal AR component

B – Backshift operator

θ – Coefficient of non-seasonal MA component

This can be notated simply as ARIMA (p, d, q)

Where p – Order of autoregressive part

d – Order of integration

q – Order of moving average part

Seasonal ARIMA (SARIMA) Model

The SARIMA model is formed by including a seasonal component to the ARIMA model. It can be represented as ARIMA (p, d, q) (P, D, Q) in which p and q are non-seasonal autoregressive and moving average parameters, P and Q are the seasonal autoregressive and moving average parameters, respectively. The two other parameters, d and D, are non-seasonal and seasonal differencing respectively, used to make the series stationary.

The form of ARIMA (p,d,q)×(P,D,Q) has the following form,

$$\phi_p(B)\Phi_p(B^s)\nabla^d\nabla^D{}_S Z_t=\theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Where ϕ – Coefficient of seasonal AR component

Φ – Coefficient of seasonal MA component

To obtain the ARIMA model by the Box-Jenkins methodology, there are three steps that must be considered which are identification, parameter estimation, and diagnostic checking (goodness of fit test).

Identification

In this step, three integers p, d, and q and P, D, Q representing respectively the number of autoregressive orders, the number of differencing orders, and the number of moving-average orders of both non-seasonal and seasonal part of ARIMA model are determined. Stationarity check of the data set reveals the nature of order of integration included in the model. It can be done by using classical methods involving autocorrelation functions (ACF) and partial autocorrelation functions (PACF) plots and modern methods such as Augmented Duckey Fuller test (ADF) (Saha et al., 2016).

Estimation of parameters

After estimating order of the model next step is to determine the parameters such as c, ϕ_1,...., ϕ_p, θ_1, ..., θ_q etc. The parameters can be estimated using a function minimization algorithm, either minimize the sums of squared residuals or maximize the likelihood (probability) of the observed series. To compute the sums of squares (SS) of the residuals, the approximate maximum likelihood method (MLE) is chosen, as this method is the fastest and can be used for very large data sets. For ARIMA model, MLE was similar to least square estimate which is based on minimizing the function $\Sigma^T_t=_1\varepsilon^2_t$. Since the ARIMA model is much complicated to estimate the regression models, certain model selection criteria were used by most of the software including open-source software Gretl, which is used in this study.

Information criteria

Model selection was done based on Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC) and Hannan-Quinn Criteria (HQIC).

AIC is useful in selecting predictors for regression as well as determining order of an ARIMA model. It can be written as

$$AIC=-2\log(L)+2\ (P+Q+K+1)$$

Where, L was the maximum likelihood function and last term represent the number of estimated parameters, in which K=0 if c=0 and K=1 if c≠0. (Akaike, 1974).

For ARIMA model, corrected AIC denoted as AIC_C can be written as:

$$AIC_C=(\frac{2n}{n-k-1})\,K-2ln[L]$$

Where, n was the number of observations

BIC or Schwarz information criteria (SIC)

$$SIC=\ln(n)K-2ln(L)\ \text{(Schwartz, 1978)}$$

$$HQIC=2In[\ln(n)]K-2ln(L)\ \text{(Hannan and Quin, 1979)}$$

Validation of the model

Once the preferred model is identified, standardized residuals should be analyzed. According to our model assumption, observations are normally distributed and thus, the standardized residuals should be standard normally distributed. Now, if a model was found to be not good enough, then errors will no longer remain uncorrelated and like a time series depends on its past values, the errors will remain uncorrelated as well. So, model validation can be made by analyzing the nature of residuals in terms of autocorrelation and normality.

Residual Analysis

When a model has been identified as best fit to a time series, it is inevitable to check that whether the selected model provides an adequate representation of the data. This is usually done by looking at the residuals. For a good model, residuals are stationary and uncorrelated, and a model validation usually consists of plotting the

residuals in various methods. Another way is by detecting whether residuals follow a normal distribution, and if so, the model selected will be good.

Ljung-Box Test

The test was used to determine whether the autocorrelations for the errors or residuals are non-zero (Modified Box-Pierce statistic) (Sallehuddin et al*.,* 2007; Kane and Yusof, 2013)

The null and alternate hypothesis of the test are given below:

H_0: The errors are uncorrelated

H_1: The errors are correlated.

The test statistic was:

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{Y_k^2}{n-k}$$

Where n was the number of observations, Y_k was the autocorrelation between residuals with lag k and m total number of lags. The statistic Q_m had a finite sample distribution that was much closer to that of x^2 (m-p-q). The procedure was to reject the null hypothesis of uncorrelated residuals, if the computed value of Q_m is larger than the chi-square table value for a specified significance level.

Normality plot of residuals

Graphical tool used for comparing data set with normal distribution. From the nature of histogram one can easily identify whether it is normally distributed or not.

Results and Discussion

Box-Jenkins (1970) methodology was applied to model the rainfall data and it includes identification of the model, estimation of the model parameters and validation of the model (Hipel et al., 1977). The time series data should be stationary which means that it should have aconstant mean, variance, and covariance which dependent only on time before fitting ARIMA models. The most used method to transform non-stationary data to stationary is differencing the data points, which replaces each value in the series by the difference between two consecutive values as t^{th} and $t\text{-}1^{th}$ periods for a first order differenced series.

Identification of the model

Stationarity was checkedusing unit root test (ADF test) and examining the autocorrelation function (ACE) and partial autocorrelation function (PACF) to identify the potential models. Null hypothesis for the ADF test was the presence of unit rootindicating non-stationary and the alternate hypothesis as no unit root indicating a stationary time series. ADF test results of rainfall data in all the three stations were found to be in rejection zone indicating stationarity and the order of integration is zero. Based on the significant value of the ADF test the order for integration for both non seasonal and seasonal component wasdetected and it isshown in Table 1.

Table 1. Order of integration based on unit root test result of rain fall data

Stations	ADF test *P*-value	Regular difference order	Seasonal difference order
Vellayani	-9.89	0	1
Kumarakom	-13.63	0	1
Pampadumpara	-13.28	0	1

Classical methods based on ACF and PACF were also performed to identify AR and MA components for both non-seasonal and seasonal parts. Fig.1, 2 and 3 shows

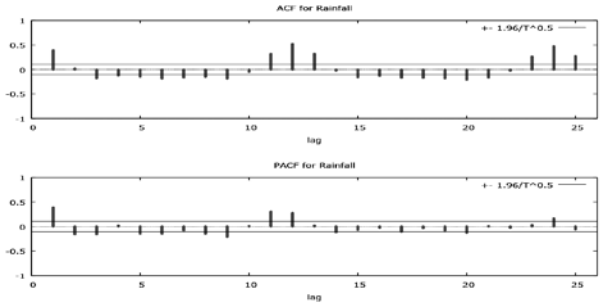

Fig.1. ACF and PACF plot for rainfall at Vellayani, Thiruvananthapuram, Kerala

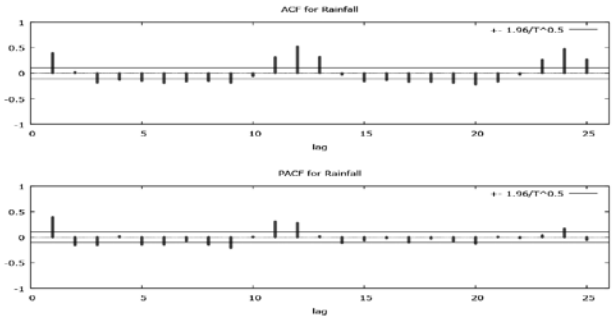

Fig. 2. ACF and PACF plot for rainfall at Kumarakom, Kottayam, Kerala

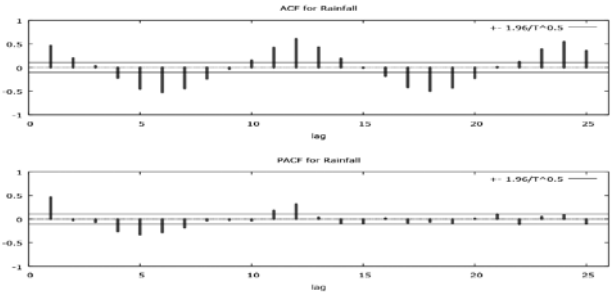

Fig. 3. ACF and PACF plot for rainfall at Pampadumpara, Idukki, Kerala

the correlogram corresponds to the rainfall data of the stations with lag length of 25 in X-axis and autocorrelation values in the Y-axis.

The seasonal autocorrelation relationship was observed and quite prominent from ACF and PACF and the gradual decay observed in the plot again indicate stationarity nature of data set. Based on the nature of the correlogram and the result of the unit root test we can choose a temporary model for rainfall and the model could be ARIMA (p, 0, q) (P, 1, Q).

Estimation of Parameters of the model

Even though the order of integration was identified the parameters and the best ARIMA model was identified by trial-and-error method based on the value of AIC, BIC and Hannan Quinncriteria. The different models estimated with different criteria usingthe open-sources of tware Gretl are shown in Table 2, 3 and 4.

For Vellayani, the model having p=1, d=0, q=0, P=0, D=1, Q=1has lower values for AIC, BIC and Hanann

Table 2. ARIMA models for rainfall at Vellayani

ARIMA Model		Coefficient	*P*-value	AIC	BIC	Hannan Quinn
(001) (111)	phi-1	0.07	0.33	4037.84	4060.74	4046.97
	theta-1	0.12	0.02**			
	theta-1	-0.89	3.39e-052***			
(003)(010)	theta-1	0.12	0.02**	4185.58	4208.49	4194.71
	theta-2	0.12	0.02**			
	theta-3	-0.03	0.53			
	phi-1	-0.44				
	phi-2	-0.70	0.003***0.00			
	phi-1	0.06	07***0.43			
(202)(111)	theta-1	0.56	2.49e-06***	4037.90	4072.25	4051.60
	theta-2	0.803460	1.44e-05***			
	theta-1	-0.893021	1.37e-05***			
(100)(011)	phi-1	0.14	0.009***7.1	4035.98	4055.07	4043.59
	theta-1	-0.85	5e-06***			

Table 3. ARIMA models for rainfall at Kumarakom

ARIMA Model		Coefficient	*P*-value	AIC	BIC	Hannan Quinn
(000)(112)	theta-1	−1.01	0.0009***	4172.91	4191.99	4180.52
	theta-2	0.01	0.81			
	phi- 1	0.07	0.19			
(100)(111)	phi-1	−0.02	0.77	4173.18	4196.08	4182.31
	theta-1	−0.99	0.003***			
	phi-1	−0.66	0.0004***			
	phi-1	−0.005	0.93			
(101)(111)	theta-1	0.74	7.75e-06***	4173.22	4199.95	4183.88
	theta-1	−1.00	5.66e-09***			
(100)(110)	phi-1	0.12	0.02**	4250.59	4269.68	4258.19
	phi-1	−0.55	1.20e-32***			
(000)(011)	theta-1	-1	$2.14*10^{-12}$	4170.97	4186.23	4177.05

Table 4. ARIMA Models for Rainfall in Pampadumpara

ARIMA Model		Coefficient	*P*-value	AIC	BIC	Hannan Quinn
(001)(112)	phi-1	0.708	0.0002***	3994.501	4021.221	4005.152
	theta-1	0.046	0.3788			
	theta-1	−1.66	5.76e-022***			
	theta-2	0.711	1.82e-06***			
(101)(012)	phi-1	0.669	0.0084***	3996.582	4023.302	4007.233
	theta-1	−0.608	0.0239**			
	theta-1	−0.907	4.61e-048***			
	theta-2	0.029	0.6301			
(100)(110)	phi-1	0.080	0.1404	4088.251	4107.336	4095.859
	phi-2	−0.499	1.93e-024***			
(000)(011)	theta-1	0.88	1.46*10-144***	3993.09	3999.17	4008.36

Quinn criterion which revealed the best model for rainfall was ARIMA (1, 0, 0) (0, 1, 1)12. The model having p=0, d=0, q=0, P=0, D=1, Q=1 has lower values for AIC, BIC and Hanann Quinn criterion for Kumarakom data which revealed that the best model was ARIMA (0, 0, 0) (0, 1, 1)12. Pampadumpara model having p=0, d=0, q=0, P=0, D=1, Q=1 has lower AIC, BIC and Hanann Quinn criterion and so the best model for rainfall was ARIMA (0, 0, 0) $(0, 1, 1)_{12}$. All the coefficients of the estimatedmodels were highly significantsince *p*-values were less than 0.05. Fig. 4, 5 and 6 shows the plot for actual and fitted values, where there dlinere presents the actual values and blue line represent the fitted values

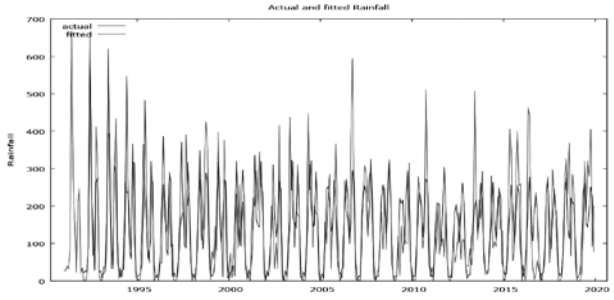

Fig. 4. Actual versusfitte dplot for rainfall in Vellayani

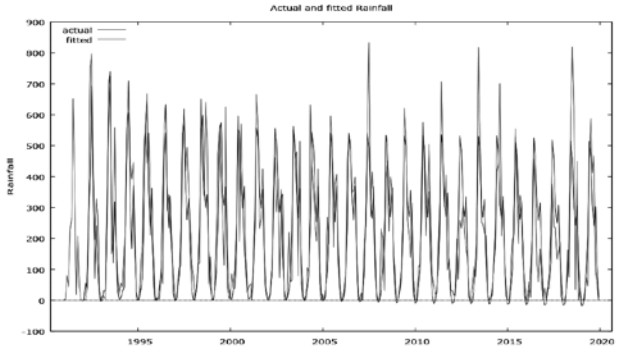

Fig. 5. Actual versus fitted plot for rainfall in Kumrakom

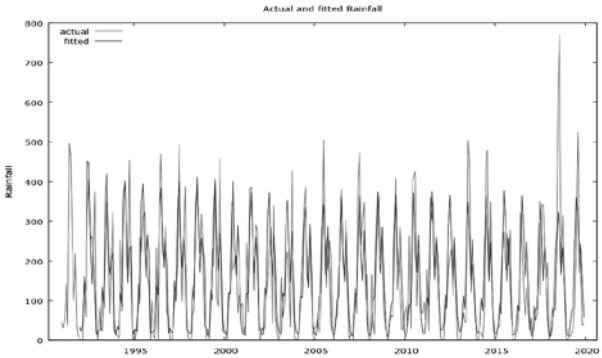

Fig. 6. Actual versus fitted plot for rainfall in Pampadumpara

and from the graph it is obvious that fitted and actual were closer.

Model validation

The best fitted model for rainfall in Vellayani was found to be ARIMA $(1,0, 0) (0, 1,1)_{12}$.

The functional form of the model is:

$$y_t - y_{t-12} = \phi_1 y_{t-1} - \phi_1 y_{t-13} + \Theta_1 e_{t-12} + e_t$$

Let $y_t - y_{t-12} = Z_t$ then,

$$Z_t = \phi_1 Z_{t-1} + \Theta_1 e_{t-12} + e_t$$

Here ϕ_1=0.14 and Θ_1=-0.85

$$Z_t = 0.14 Z_{t-1} - 0.85 e_{t-12} + e_t$$

The best fit model for rainfall in Kumarakom was ARIMA $(0, 0, 0) (0, 1, 1)_{12}$. The functional form of the model is:

$$y_t - y_{t-12} = \Theta_1 e_{t-12} + e_t$$

Let $y_t - y_{t-12} = Z_t$ then,

$$Z_t = \Theta_1 e_{t-12} + e_t$$

Here, Θ_1=- 1

$$Z_t = -e_{t-12} + e_t$$

The best fit model for rainfall in Pampadumpara was ARIMA (0, 0, 0) (0, 1, 1)$_{12}$. The functional form of the model is:

$$y_t - y_{t-12} = \Theta_1 e_{t-12} + e_t$$

Let $y_t - y_{t-12} = Z_t$ then,

$$Z_t = \Theta_1 e_{t-12} + e_t$$

Here $\Theta_1 = 0.8$

$$Z_t = 0.88\, e_{t-12} + e_t$$

Two methods are commonly used to test the adequacy of the selected model one method is by checking the autocorrelation of the residuals using Ljung-Box Q test (Box et al., 1995) and secondis by checking the normality of the residuals. It has been found to measure the over all adequacy of the chosen model by examining a quantity Q known as L jung-Box statistic (Yurekli et al., 2005; Sallehuddin et al., 2007), which is a function of auto correlations of residuals and its approximate distribution was Chi-square. If Ljung-Box statistic value is found non-significant then residuals are uncorrelated and hence the model selected was good enough for the prediction. The estimated Ljung-box test statistic of rainfall at three stations are shown in Table 5, result indicated that the residuals are not correlated. Fig. 7 displays the normality plot of residuals for rainfall and it clearly shows that residuals are normally distributed.

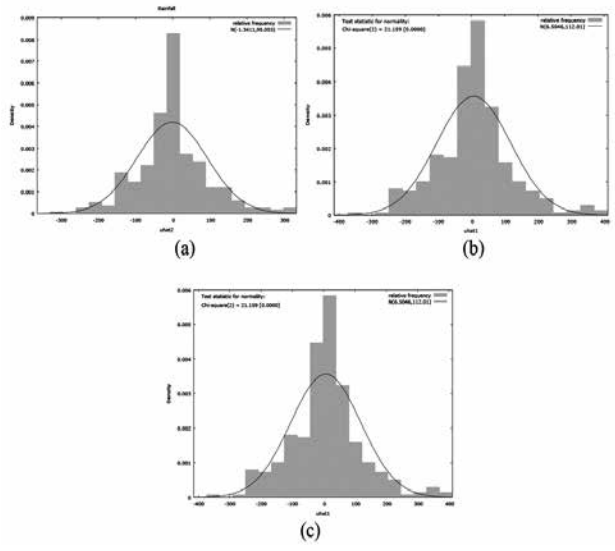

Fig. 7. Normality plot of residuals for rainfall at (a) Vellayani, (b) Kumarakom and (c) Pampadumpara

Tab. 5. Result of Ljung-Box test

Station	Ljung-box test statistics	*p* value
Vellayani	8.08	0.62
Kumarakom	7.57	0.757
Pampadumpara	9.74	0.55

Conclusions

The SARIMA model was fitted to monthly rainfall for all the regions Vellayani, Kumarakom, and Pampadumpara using the monthly data for the period from 1991 to 2019. The model parameters were obtained by using maximum likelihood method and the best model were selected using Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and Hannan-quinn coefficient. ARIMA (1, 0, 0)×(0, 1, 1)$_{12}$ was found best fit for rainfall for Vellayani, ARIMA (0, 0, 0)×(0, 1, 1)$_{12}$ for Kumarakom and Pampadumpara. The adequacy of the check of the selected models confirmed that the selected models were free from autocorrelation and the residuals are normal.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE. Trans. Automatic Control,* **19**:716-723.

Box GEP, Jenkins GM, Reinsel GC. 1995. Time series analysis: Forecasting and control. Prentice-Hall. 592.

Hipel, K.W., McLeod, A.I. and Lennox, W. C. 1977. Advances in Box Jenkins modeling: 1. Model construction. *Water Resour. Res.,* **13**(3):567-575.

Kane, I. L. and Yusof, F. 2013. Assessment of risk of rainfall events with a hybrid of ARFIMA-GARCH. *Mod. Appl. Sci.,* **7**(12):78-89.

Saha, E., Hazra, A., and Banik, P. 2016. SARIMA modelling of the monthly average maximum and minimum temperatures in the eastern plateau region of India. *Mausam.* **67**(4):841-848.

Sallehuddin, R., Shamsuddin, S.M.H., Hashim, S.Z.M. and Abraham, A., 2007. Forecasting time series data using hybrid grey relational artificial neural network and auto regressive integrated moving. *Neural Network World*, **17**(6), p.573.

Schwartz, G. 1978. Estimating the dimension of a model. *The An. Statist*. **6**:461-464.

Hannan, E.J. and Quinn, B.G., 1979. The determination of the order of an autoregression. *J. R. Statis. Soc*., **41**(2):190-195.