Journal of
Water & Climate Change

IWA PUBLISHING

Journal of Water and Climate Change Vol 13 No 12, 4151 doi: 10.2166/wcc.2022.107

Performance evaluation of univariate time-series techniques for forecasting monthly rainfall data

P. Kabbilawsh *, D. Sathish Kumar and N. R. Chithra

Department of Civil Engineering, National Institute of Technology Calicut, Kozhikode, Kerala 673601, India
*Corresponding author. E-mail: kabbi.civil@gmail.com

PK, 0000-0003-3957-6237

ABSTRACT

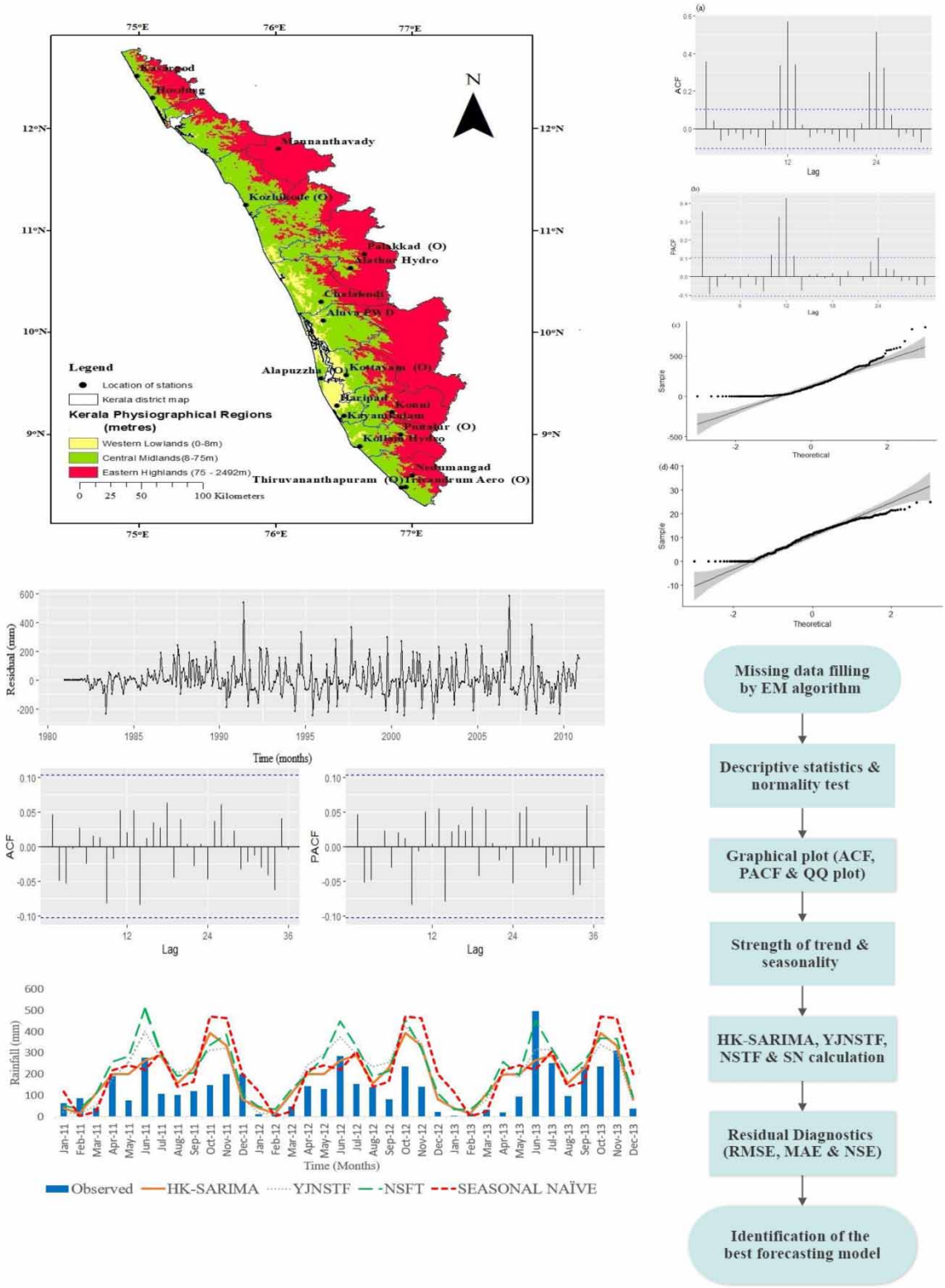
In this article, the performance evaluation of four univariate time-series forecasting techniques, namely Hyndman Khandakar-Seasonal Autoregressive Integrated Moving Average (HK-SARIMA), Non-Stationary Thomas-Fiering (NSTF), Yeo-Johnson Transformed Non-Stationary Thomas-Fiering (YJNSTF) and Seasonal Naïve (SN) method, is carried out. The techniques are applied to forecast the rainfall time series of the stations located in Kerala. It enables an assessment of the significant difference in the rainfall characteristics at various locations that influence the relative forecasting accuracies of the models. Along with this, the effectiveness of Yeo-Johnson transformation (YJT) in improving the forecast accuracy of the models is assessed. Rainfall time series of 18 stations in Kerala, India, starting from 1981 and ending in 2013, is used. A classification system based on root mean square error (RMSE), mean absolute error (MAE) and Nash–Sutcliffe model efficiency coefficient (NSE) is proposed and applied to find the best forecasting model. The models HK-SARIMA and YJNSTF performed well in the Western lowlands and Eastern highlands. In the Central midlands, out of 12 stations, the performance indices of 8 stations are in favour of the HK-SARIMA model. It can be concluded that HK-SARIMA models are more reliable for forecasting the monthly rainfall of the stations located in all geographic regions in the state of Kerala.

Key words: Hyndman Khandakar (HK) algorithm, mean absolute error (MAE), Nash–Sutcliffe model efficiency coefficient (NSE), root mean square error (RMSE), Thomas-Fiering (TF), Yeo-Johnson transformation (YJT)

HIGHLIGHTS

- YJT is applied to transform the non-normal rainfall time series more Gaussian-like distribution and simultaneously increases the TF model's forecasting ability.
- The HK algorithm used in this study connects the unit root test, minimising the corrected Akaike information criterion (AICc) and maximum likelihood estimator (MLE) to obtain the model order and parameter (coefficients) of a SARIMA model.
- The hyper-parameters of SARIMA models obtained from the HK algorithm identified that the seasonal autoregressive (AR) and moving average (MA) components are more prominent than the non-seasonal components.
- The concept of strength of seasonality and the strength of trend is utilised to explore the non-stationary nature of rainfall datasets, and it was inferred that all eighteen stations are indeed non-stationary.
- HK-SARIMA performs better than YJNSTF, NSTF and SN, which can be attributed to the fact that HK-SARIMA handles seasonality better than TF by creating an auto-regression equation on the time series dataset.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Time-series analysis is a straightforward approach for evaluating an ordered arrangement of data values. The data are collected at equally spaced time intervals instead of recording the values unsystematically or arbitrarily. It is not just assembling the data values in an orderly way but analysing how the data values vary over time. It delivers additional information on interdependencies among the data values. The time-series analysis techniques have the capability to forecast future data values from historical data. However, it necessitates that the data values confirm regularity and dependability. The application of time-series forecasting includes estimating the probable change in data values, like seasonality or cyclic behaviour, which deliver an improved understanding of data variables (Attah & Bankole 2012).

In climate change studies, time-series forecasting techniques are applied to generate various hydro-meteorological variables such as rainfall, streamflow, temperature and river water quality. In tropical regions, rainfall forecasts are carried out by distinctly considering dry and wet seasons. Several countries, which mainly depend on agricultural supplies as a source of income, have to forecast rainfall to determine the right time to start planting their crops to capitalise on their harvest. Besides this, there exists a necessity to calculate the amount of rainfall contributed to the water budget equation of any given watershed (Attah & Bankole 2012). Applications of time-series analysis in drought forecasting include drought analysis, environmental flows and sustaining reservoir levels (Sahoo *et al.* 2019). Forecasts are generally based on the theory of probability, stochastic processes and newly developed chaotic non-linear dynamical systems (Grimaldi *et al.* 2006).

There are two broad categories of time-series forecasting models commonly applied in the field of hydrology. The first category is based on the stochastic process, and the second is based on data mining techniques by applying artificial intelligence (AI) techniques. The stochastic forecasting models are classified as Thomas-Fiering (TF)-based models (Yurekli & Kurunc 2006; Teymouri & Fathzadeh 2015) and Autoregressive and Moving average (ARMA) family of models. The TF model is considered a characteristic stochastic methodology for generating synthetic datasets like streamflow in hydrology. The TF model calculates the variable at a particular time interval as a linear function of the same variable in the previous time interval. The TF model is applicable in several situations, easy to practice in spreadsheets and can be exercised for creating weekly, monthly, seasonal and annual streamflow (Kurunç *et al.* 2005). Stedinger & Taylor (1982) developed various monthly streamflow models, including the TF model, to generate synthetic data for the upper Delaware River basin in New York State. Their research corroborated that the TF model could accurately replicate the historical data statistics. Joshi & Gupta (2009) also developed a single-site TF model to generate monthly inflow series.

Similarly, many researchers have worked on ARIMA and seasonal-ARIMA (SARIMA) models to forecast meteorological data in different parts of India (Narayanan *et al.* 2013; Dabral & Murry 2017; Murthy *et al.* 2018; Saha *et al.* 2020; Ray *et al.* 2021). ARIMA models are used when seasonality is absent or forecasting is carried out for a specific season (Narayanan *et al.* 2013). When the forecasting is carried out on monthly datasets spanning multiple years, SARIMA models are developed and applied (Dabral & Murry 2017; Murthy *et al.* 2018; Ray *et al.* 2021). SARIMA models have the capability to describe time series that exhibit non-stationary behaviours both within and across seasons (Wang *et al.* 2013). Recently, many stochastic models have been developed using SARIMA models for forecasting hydrological time-series such as monthly rainfall, mean maximum and minimum temperature, evapotranspiration and monthly streamflows. Apart from the traditional classification of time-series forecasting methods as stochastic and data mining techniques, there are several least-square spectral analysis (LSSA) and their variants developed in the past 5 years. LSSA methods can examine and forecast hydrological data without any specific application for interpolation, gap-filling, de-spiking, seasonality and over/under-fitting (Ghaderpour Vujadinovic & Hassan 2021). With the rise in computing facilities, secondary hydrologic datasets can be clubbed with rainfall during analysis to increase the forecast accuracy. The temperature and streamflow data may aid precipitation forecasting, and climate data can be used for streamflow forecasting, etc. This coherency can be estimated using wavelet methods (Ghaderpour & Vujadinovic 2020). Even after developing newer techniques with many positive peculiarities, numerous journal articles are still getting published using stochastic techniques (Dabral & Murry 2017; Murthy *et al.* 2018; Ray *et al.* 2021).

Therefore, this study is focused on comparing the time-series forecasting models developed based on the stochastic approach. The fundamental assumptions in stochastic models are that rainfall time series must be stationary and follow a normal distribution (Unnikrishnan & Jothiprakash 2018). In real-world experience, rainfall is erratic, non-normal and non-stationary. Therefore, data pre-processing techniques such as normalisation and standardisation are carried out before applying the time-series model. The normalisation techniques commonly applied are Box-Cox transformation and Yeo-Johnson transformation (YJT). The Box-Cox transformation can be applied to datasets with positive values, whereas the YJT can

be applied to datasets with positive, negative and zero values. In many stations in the study area, the rainfall values for several months are found to be zero.

In several studies, Box-Cox transformation has been applied to rainfall datasets and reported a measurable enhancement in the rainfall estimations post-transformation (Cecinati *et al.* 2017; Dabral & Murry 2017; Pandey *et al.* 2019; Martínez-Acosta *et al.* 2020). However, rainfall forecasting studies that have applied the YJT are limited (Schepen *et al.* 2012; Zeynoddin & Bonakdari 2019) despite being able to deal with zero and negative numbers. The limited studies related to the application of YJT might create a belief among researchers that the application of YJT to the datasets may not significantly improve forecasts. Therefore, in this study, the effectiveness of YJT is analysed using the performance indices obtained from the rainfall forecasts carried out using the non-stationary Thomas-Fiering model (NSTF) with and without transformation.

The TF model was initially developed and applied for forecasting streamflows (Harms & Campbell 1967; Joshi & Gupta 2009; Sharma *et al.* 2018). Some studies have extended the application of the TF model to forecast monthly rainfall (Mallikarjuna & Vardhan 2002; Ünal *et al.* 2004). In such studies, the NSTF model and its improvisations are mainly compared with TF by considering TF as a parametric method. However, in most cases, the validation of the assumption regarding the normal distribution of datasets has not been established before applying a parametric method. In the present study, the nature of the distribution of the data is evaluated using two normality tests. The assumption of non-stationarity is validated by applying the concept of strength of trend and seasonality.

There are limited studies that have compared the performance of ARIMA and TF models for carrying out rainfall forecasts (Ahmad *et al.* 2001; Kurunç *et al.* 2005; Yurekli & Kurunc 2006; Yousif *et al.* 2016). However, overall results from these studies are contradictory and inconclusive. Some studies conclude that TF is a better model than ARIMA (Kurunç *et al.* 2005; Yousif *et al.* 2016) and few studies conclude otherwise (Ahmad *et al.* 2001; Yurekli & Kurunc 2006). The monthly rainfall datasets belonging to different geographical locations may follow different statistical distributions and relative magnitudes of seasonality and trend. Most of the earlier studies failed to examine such critical issues in the rainfall datasets before concluding their findings. Statistical analysis is carried out in this study to determine central tendency and dispersion measures. The relative dominance of the pattern (influenced by trend or seasonality or both) is quantified by applying seasonal adjustment, detrending and statistical tests.

As discussed earlier, the observations on the performance of different univariate time-series models are contradictory and mostly location-specific. Therefore, it is necessary to evaluate the models using rainfall datasets obtained from stations belonging to different physiographic regions and experiencing rainfall through different mechanisms. The rainfall stations in the state of Kerala are grouped into three main clusters based on distinct rainfall patterns as the percentage of rainfall received during the southwest monsoon, northeast monsoon and pre-monsoon thunderstorms vary significantly (Simon & Mohankumar 2004). Therefore, analysing the time-series rainfall datasets of Kerala can bring better insight into the performance of these models. In this study, three time-series forecasting models, namely NSTF, YJNSTF and HK-SARIMA, are compared with Seasonal Naïve (SN) (Hyndman & Athanasopoulos 2018) and benchmarked.

The overall objectives of this study are: (1) to develop the best-fit models by comparing four time-series forecasting approaches, namely HK-SARIMA, NSTF, YJNSTF and SN models using 33 years' (1981–2013) monthly rainfall time series; (2) to compare the forecasting efficiency of the four models using scale-dependent errors (root mean square error (RMSE) and MAE), normalised RMSE and NSE; (3) to assess the improvement in the forecasting efficiency of the TF models after applying the YJT technique.

2. STUDY AREA AND DATASETS

The physiography of Kerala has been broadly grouped into three regions (Figure 1), namely the Western lowlands (up to 8 m from mean sea level (MSL)), the Central midlands (between 8 and 75 m above MSL) and the Eastern highlands (75 m above from MSL) (Chattopadhyay & Franke 2006). Eastern highlands receive higher rainfall quantities due to the orographic effect. Similarly, the northern parts of the state receive a relatively large amount of rainfall compared with the southern region. The rainfall datasets from 130 stations covering all three physiographic regions of Kerala were obtained from the Indian Meteorological Department (IMD). However, several data gaps were found in the time-series datasets obtained. The rainfall time series from 18 stations having less than 20% missing data from 1981 to 2013 are chosen for the study. For Kasargod and Palakkad (O) stations, the rainfall datasets are available from the year 1977 to 2009 and 1969 to 2001, respectively. The first 30 (1981–2010) years' data are used to train the model, and the remaining 3 years' data (2011–2013) are used to test the forecasting models. Five hundred and eighteen

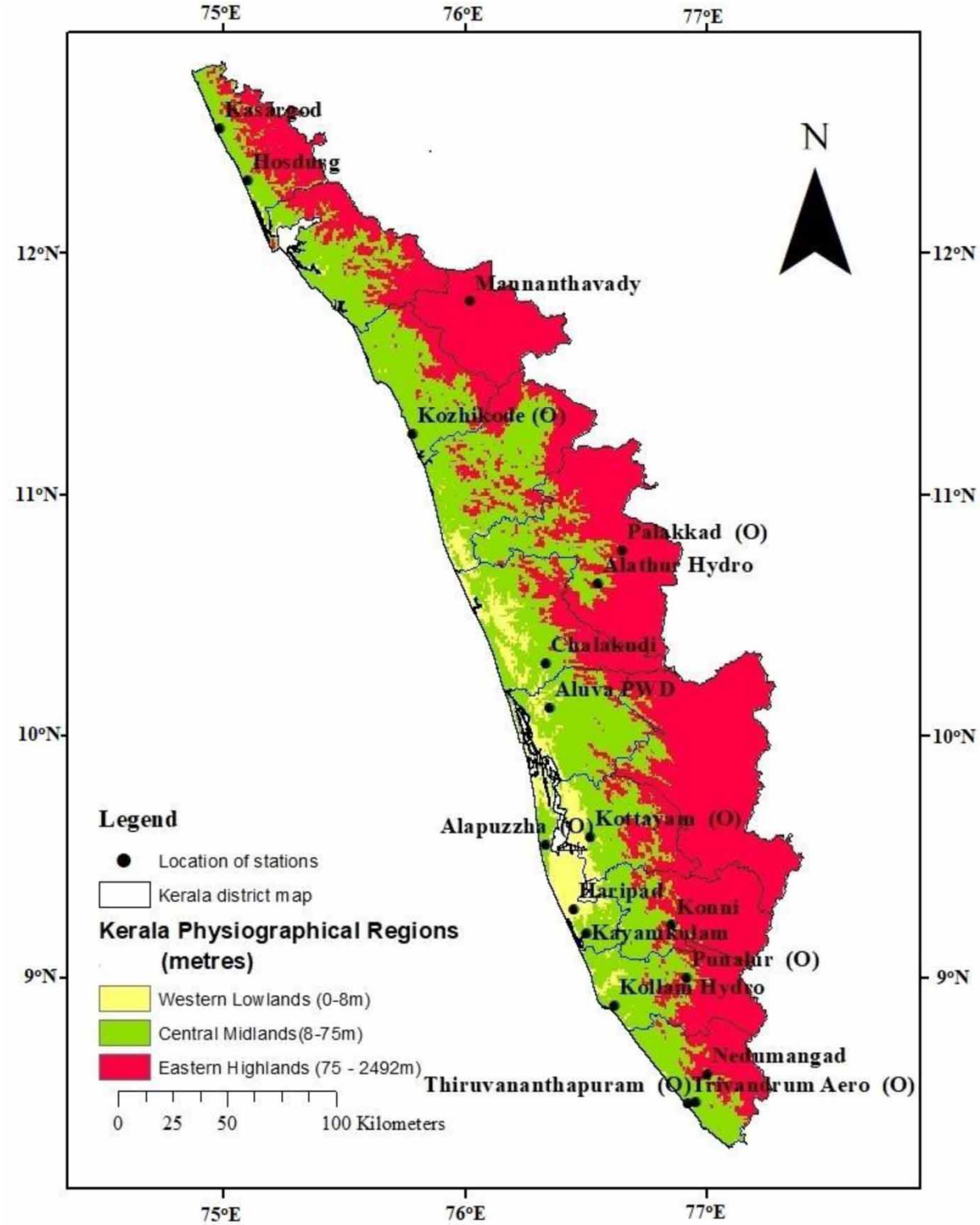


Figure 1 | The location map of the 18 rainfall stations in the study area.

intermittent monthly rainfall data points (about 7.27% of the data) were missing in the available dataset. The missing data reduces the characteristic features of the time series (Kang 2013). Thereby, the missing values are estimated first by application of the expectation–maximisation (EM) algorithm. The data points so generated are used to obtain continuous time-series datasets.

3. METHODOLOGY

3.1. Expectation–maximisation (EM) algorithm

The EM algorithm is a machine learning data-filling technique created by Dempster *et al.* in 1977. The purpose of the development of the algorithm was to overcome the limitations in the application of maximum likelihood methods (Dempster *et al.* 1977). The ideology is to use the existing observed data of the rainfall time series to estimate the missing rainfall values of latent variables and then apply the data to update the parameter value in the maximisation step (Firat *et al.* 2010). The procedure consists of four steps.

(a) The first step consists of introducing initial values to the parameters, which leads to delivering the incomplete observed rainfall time series to the system with the assumption that the observed rainfall data belong to a specific probability distribution.
(b) The second step involves employing observed rainfall values to estimate the missing data and, in turn, the variable's value gets updated.
(c) In the third step, the entire data consisting of observed and filled data from the expectation step is used to appraise the values of the parameter. The hypothesis gets updated.
(d) In the fourth step, the EM loop is carried out until there is a convergence in rainfall data.

3.2. Descriptive statistics

It consists of explanatory numbers which summarise a given rainfall time series. The rainfall time series analysed can either demonstrate the entire population or a sample from a population. It mainly consists of measures of central tendency and measures of variability. Mean, median and mode make up the measure of central tendency. Range, standard deviation (SD), coefficient of variation (CV), variance, interquartile range (IQR), standard error of the mean (SEM) and Skewness and Kurtosis are the measures of variability.

3.3. Autocorrelation function (ACF)

Autocorrelation function (ACF) systematically depicts the correlation between time series and their lagged form over sequential time intervals. It is practically analogous to the correlation coefficient between two diverse time series. Still, the ACF applies the same series twice in its initial format and the other in its lagged version (NIST/SEMATECH 2022).

3.4. Partial autocorrelation function (PACF)

The partial autocorrelation at lag k is the AC between Y_t and Y_{t-k} that is not accounted for by lags 1 through $k-1$. Explicitly, partial autocorrelations help recognise the order of an autoregressive model. There is detailed technical background on the ACF and PACF elsewhere (NIST/SEMATECH 2022).

3.5. Yeo-Johnson's transformation (YJT)

Power transformation consists of a group of 'power functions' that are applied on time series to establish a monotonic transformation. The power transformation used in this study is the YJT. The purpose of applying YJT is to stabilise variance, give the time-series data more Gaussian-like distribution and enhance the correlation coefficient between two variables. In this study, the rainfall time series is initially transformed, then the dataset is forecasted and back-transformation is carried out. On applying YJT, the rainfall time series closely represents a normal distribution (Yeo & Johnson 2000). The application of normality will help build confidence intervals and conduct hypothesis testing. YJT can lower the skewness value of the dataset and help maintain a linear relationship between rainfall and its lagged values. Let $y_1, y_2 \ldots y_n$ be the time series and λ be the power parameter, then transformation is defined as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{((y_i+1)^{\lambda}-1)}{\lambda}, & \text{if } \lambda \neq 0, y \geq 0 \\ \dfrac{-((-y_i+1)^{(2-\lambda)}-1)}{(2-\lambda)}, & \text{if } \lambda = 0, y \geq 0 \\ \log(y_i+1), & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_i+1), & \text{if } \lambda = 2, y < 0 \end{cases} \quad (1)$$

3.6. Normality test

3.6.1. Shapiro–Wilk (SW) test

The test estimates a W statistic that checks whether a random sample $x_1,\ x_2 \ldots x_n$ has been chosen from a normally distributed population. The smaller value of W indicates deviation from normality, and critical values of the W statistic are achieved from Monte Carlo simulations (Shapiro & Wilk 1965).

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\left(\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2\right)} \tag{2}$$

where $x_{(i)}$ are rainfall time-series data points and a_i are constants obtained from mean, variance and covariances of a sample of size n from a normal distribution. The assumptions, advantages and disadvantages of SW can be found in the e-Handbook (NIST/SEMATECH 2022).

3.6.2. Anderson–Darling (AD) test

The AD test is applied to check whether the rainfall time series belongs to a specific distribution (normal distribution in the current study). It is an improvement over the Kolmogorov–Smirnov test as it provides more weight at the tail of the distributions than the Kolmogorov–Smirnov test. The AD test is not a distribution-free test, meaning that the critical values obtained from the AD test are influenced by the particular distribution for which it is tested. The most significant advantage of this property is that the test becomes highly sensitive even for small perturbations in the time series. The limitation is that critical values must be computed for each distribution (Anderson & Darling 1952; NIST/SEMATECH 2022).

3.7. Graphical visualisation tools

3.7.1. Quantile-Quantile (QQ) plot

The QQ plot is a visualisation tool that assists in evaluating whether the rainfall time series chosen is a sample from a population of particular theoretical distribution (normal distribution in this study). It is a scatterplot produced by mapping two sets of quantiles against one another. In the present case, if the selected rainfall time-series dataset follows a perfectly normal distribution, then the points in the plot fall on a straight line (NIST/SEMATECH 2022).

3.8. Strength of trend and seasonality

Rainfall time series is highly erratic and varies by a large margin temporally and spatially. One of the best methods to study rainfall and its inherent characteristic is by splitting the entire time series into respective components. There are three main components, namely trend, seasonality and remainder. Each component of the time series epitomises an underlying pattern. Decomposition of time series helps understand the properties and improve forecast accuracy (Wang *et al.* 2006). Besides this, time-series decomposition can be used to measure the strength of trend and seasonality. The quantification of trend and seasonal components helps choose a specific time-series model for forecasting depending upon the relative magnitudes.

The most crucial step before applying any statistical models to the rainfall time series for forecasting is to study the property of stationarity. A time series is defined as stationary when the characteristics of the series do not vary with time (Ukkola *et al.* 2019). The existence of trend and seasonality will make the time series non-stationary. Conventional techniques are used mathematically to check whether the trend and seasonal component present are sufficiently high in the time series to declare the time series as non-stationary. The most widely used conventional method is the non-parametric Mann–Kendall trend test followed by Sen's slope. The Mann–Kendall trend test (Mann 1945; Kendall 1975; Gilbert 1987) on rainfall time series is to check whether there exists a statistically significant monotonic increasing or decreasing trend. Sen's slope is used to quantify the magnitude of the detected monotonic trend (Theil 1950; Helsel & Hirsch 1992). Sen's slope is unresponsive to outliers and considerably precise in comparison with non-robust simple linear regression (Ordinary least square). Sen's slope has the ability to detect trends in skewed and heteroskedastic time series (Sen 1968). There exist a few assumptions (Conover 1999) of Sen's slope which need to be fulfilled before applying it to a time series:

(a) The rainfall data points attained as a time series are not autocorrelated. The data points are illustrative of the true conditions of the environment when the sampling procedure is carried out.

(b) The methods of data collection, compilation, processing, treatment and quality of the measurement methods remain unbiased over time and contain the same characteristics as the original population.
(c) The trend which exists in the time series is linear. The best-fit trend line shows no variation for different time scales, e.g., months or calendar quarters.

The magnitude of the linear trend can be determined using Sen's slope. However, the technique is inefficient for calculating the magnitude of a non-linear trend in the data series. Therefore, an alternative method is proposed to find the nature of the time series (stationary or not) and the relative amount of that nature (trend and seasonality) in a single run. The trend or seasonality can be rejected from a time series when the magnitude of the trend or seasonal component is less than the remainder component. Thereby to determine the property of stationarities and quantify the enormity of components, the concept of strength of trend and strength of seasonality is introduced (Hyndman & Athanasopoulos 2018).

The strength of the trend F_T is defined as one minus ratio of the variance of the remainder component to the sum of the variance of the trend and remainder component in a seasonally adjusted time series (Hyndman & Athanasopoulos 2018).

$$Y_t = T_t + S_t + R_t \tag{3}$$

where Y_t is the time series, T_t is the trend component, S_t is the seasonal component and R_t is the reminder component.

The seasonally adjusted time series $(R_t + T_t)$ must have a higher variance than the remainder component in a rainfall time series with a strong trend pattern. In the case of a shallow trend in a time series, both variances must be approximately the same. The strength of the trend (F_T) is defined as

$$F_T = \max\left(\begin{matrix}0\\ 1 - \left[\dfrac{\mathrm{Var}(R_t)}{\mathrm{Var}(R_t + T_t)}\right]\end{matrix}\right) \tag{4}$$

Similarly, the strength of seasonality (F_S) is defined as the one minus ratio of the variance of the remainder component to the sum of seasonal and reminder components in a detrended $(S_t + R_t)$ time series.

$$F_S = \max\left(\begin{matrix}0\\ 1 - \left[\dfrac{\mathrm{Var}(R_t)}{\mathrm{Var}(R_t + S_t)}\right]\end{matrix}\right) \tag{5}$$

3.9. TF method

Thomas & Fiering (1962) pioneered their work by applying stochastic principles in water resource systems scheduling and forecasting. The novel idea of projecting synthetic time series based on similar correlation performance of the original streamflow time series was initiated. The TF method used the Markov chain model to generate monthly streamflow by considering the serial correlations of monthly flows (Thomas & Fiering 1962). The method can model the seasonal components and variability of the time series by considering the month-to-month coefficient of correlation (Clarke 1973; Kurunç *et al.* 2005). Harms & Campbell (1967) improved TF by reinforcing with the following assumptions to the data series: (1) annual time series are normally distributed; (2) monthly time series are log-normally distributed; (3) correlation exists between annual time series and (4) correlation exists between monthly time series (Harms & Campbell 1967). Two variants of Thomas-Fiering's models are applied based on the stationarity of the time series. The first one is stationary and the second one is non-stationary. Since the rainfall time series exhibit seasonality, the NSTF model is used in the current study.

3.9.1. NSTF model

The NSTF model is given as follows:

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j}(X_{i,j} - \mu_j) + t_{i,j+1}\sigma_{j+1}\sqrt{1 - \rho_j^2} \tag{6}$$

where $X_{i,j+1}, X_{i,j}$ is the rainfall value in $(j+1)$th and jth time step belonging to ith year, μ_{j+1}, μ_j is the long-term mean in $(j+1)$th and jth time step belonging to ith year, σ_{j+1}, σ_j is the standard deviation belonging to $(j+1)$th and jth time step, ρ_j is the lag-one correlation between $(j+1)$th and jth time step and $t_{i,j+1}$ is the standard normal deviate.

3.9.2. Assumptions of the NSTF model

(a) The first-order Markov model assumes that the process is non-stationary in its first three moments.
(b) It is possible to generalise the model so that the periodicity in hydrologic data is accounted for to a certain extent. The main application of this generalisation has been in generating monthly rainfall data where pronounced seasonality in the monthly rainfall exists.

3.10. Hyndman Khandakar-Seasonal Autoregressive Integrated Moving Average (HK-SARIMA)

The seasonal autoregressive and moving average model is abbreviated as SARIMA $(p, d, q)(P, D, Q)_m$, where seven model orders p, d, q, P, D, Q and m need to be determined. Hyndman & Khandakar (2008) proposed an automated algorithm in the R programming language to find the model order using the forecast package. A summary consisting of three steps is presented subsequently, and detailed information is provided in Hyndman & Khandakar (2008). The value of m is declared by the user depending on the type of time-series object. When the time series is annual, the value of m is 1, $m = 12$ for monthly rainfall and $m = 365$ for daily data (Hyndman & Khandakar 2008).

3.10.1. Step 1: application of unit root test

The first step consists of determining the seasonal and non-seasonal difference terms (D, d) by extended Canova–Hansen (CH) (Canova & Hansen 1995) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski *et al.* 1992), respectively. The motivation of applying the KPSS test along with CH test is that both have similar null hypotheses. The null hypothesis of KPSS test is 'The process is trend stationary' and the alternative hypothesis is the series has a unit root (series is not stationary). The null and alternative hypothesis of the CH test is defined as no unit root against the presence of a seasonal unit root. Canova & Hansen (1995) only determined the critical values (C_m) compared with the test statistic for $2 < m < 13$. Later, to make the test applicable for time series whose seasonal value (m) is more than 12, Hyndman & Khandakar (2008) plotted critical value (C_m) versus m up to 365. It was inferred that a straight line fit was suitable to describe the variation between (C_m) and m. The equation they arrived at is $C_m = 0.269m^{0.928}$. Thereby, the CH test can be applied for any value of $m > 1$. In order to simplify things, HK proposed a constraint $D < 2$ as more than one seasonal difference leads to poor forecasting ability. First, $D = 0$ is applied, if the CH test-statistic is more than the critical value C_m at a chosen level of significance, the null hypothesis is rejected and the alternative hypothesis of the presence of seasonal root is accepted, then $D = 1$ is applied to the time series to remove the seasonal part. Suppose the CH test statistic is less than C_m at $D = 0$ unit root. Thereby, no seasonal difference is applied. After the determination of D, the KPSS test is successively applied on D-times seasonally differenced data. KPSS is a statistical test applied to classify the rainfall time series as stationary or non-stationary. The null hypothesis of the KPSS test is that the time-series data is stationary, and the alternative hypothesis is that the time series is non-stationary. In the presence of sufficient statistical indication, the null hypothesis may be rejected. The complete information about KPSS is found in Kwiatkowski *et al.* (1992). Applying the KPSS test, the initial time series is tested for stationarity. If the p-value is less than 0.05, then the time series is non-stationary, and the null hypothesis is rejected at a 95% confidence level. First-order differencing is applied to the time series, and differenced series (∇y_t) is once again tested using KPSS for the presence of any more non-stationarity. This process is followed until stationary series is obtained (Meer 2019).

3.10.2. Step 2: fitting of current model and its perturbations with constraints

The following mentioned four models are fitted to the stationary time series after D-seasonal and d-non-seasonal differenced data.

- SARIMA $(2, d, 2)(1, D, 1)$
- SARIMA $(0, d, 0)(0, D, 0)$
- SARIMA $(1, d, 0)(1, D, 0)$
- SARIMA $(0, d, 1)(0, D, 1)$.

Thirteen variations of the current model, along with its constraints, are determined by applying the following guidelines:

- Any one of the model orders $p, q,\ P$ and Q are allowed to vary ± 1 from the current model subjected to the constraint that the model orders $p, q < 5$ and $P,\ Q < 2$. Thereby, the maximum of eight models can be formed.
- p, q; both are allowed to vary ± 1 from the current model subjected to the constraint where the model orders $p, q < 5$. Thereby, the maximum of two models can be formed.
- P and Q are allowed to vary ± 1 from the current model subjected to the constraint where the model orders $P,\ Q < 2$. Thereby, the maximum of two models can be formed.
- c is incorporated when the current model has $c = 0$ or omitted if the current model has $c \neq 0$. Thereby, one model is formed.

Out of the 13 models, the one with the lowest AIC is selected and considered the current model. Once again, 13 models are formed based on the new current model, and this procedure is repeated until the chosen current model has the lowest AIC value among its variants. Akaike's information criterion (AIC) is the measure of the goodness of fit of the model and is given as

$$\mathrm{AIC} = -2\mathrm{Log}(L) + 2k \tag{7}$$

where L is the Gaussian likelihood of data and k is the count of independent parameters in the model. Out of 13 models, the one with the least AIC is chosen as the initial model. The concept applied in AIC is that likelihood extends monotonically as the number of parameters keeps on adding to the model. Thereby, maximising the likelihood will support a model that overfits the data. AIC averts such overfitting by correcting the likelihood with a term that is proportional to the number of parameters used in the model (Meer 2019).

3.10.3. Step 3: parameter estimation by maximum likelihood estimation

The autoregressive and moving average of seasonal and non-seasonal parameters are predicted by applying the maximum likelihood estimation (MLE) method (Hyndman & Athanasopoulos 2018; Meer 2019). The likelihood is defined as the probability of procuring the observed time series when the SARIMA model and its parameters are known. Those factors which maximise the likelihood are called MLEs, and a comprehensive mathematical explanation of MLE for SARIMA is given by Brockwell & Davis (2002).

The HK algorithm is certain to find a suitable model since the model space is finite. Therefore, at least one of the initial 13 models will be established either with no AR or MA parameters. Steps 1 and 2, which consist of model order determination, are automated by the HK algorithm. Step 3, which consists of model parameter determination by MLE, is also automated. However, it is not a part of the HK algorithm. The forecast package's auto.arima() function is used to run all three steps in a single run in the R programming language.

3.10.4. Ljung-Box test

The Ljung-Box test is an analytical measure applied to test the lack of fit of a rainfall time series. The measure examines whether the residuals of a rainfall time series after fitting a SARIMA model are white noise. The test inspects the AC of the residuals. If the value of autocorrelations is small, then it can be concluded that the SARIMA model developed does not reveal a significant lack of fit. The null hypothesis (H_0) is defined as: The SARIMA model does not demonstrate a lack of fit. There is no AC between the residual time series and its lagged version. The alternative hypothesis (H_a), defined as the developed SARIMA model, demonstrates a lack of fit. There is significant AC between the residual time series and its lagged version. Given a time series Y of length n, the test statistic is defined as:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\widehat{r_k}}{n-k} \tag{8}$$

where $\widehat{r_k}$ is the estimated AC of the series at lag k, m is the number of lags being tested and n is the number of data points. A value of p greater than 0.05 signifies a failure to reject the null hypothesis, the residual time series is white noise and there is no significant AC. The time-series model is reasonably fit. A p-value less than 0.05 signifies rejecting the null hypothesis and concluding that the time series is not white noise (NIST/SEMATECH 2022).

3.11. Assumptions of the SARIMA model

(a) The rainfall dataset should be stationary.
(b) If the rainfall datasets are non-stationary, then SARIMA models have the internal ability to convert into a stationary time series by applying seasonal and first-order differencing.
(c) SARIMA is applicable only when there is a presence of strong seasonality or data is autocorrelated to itself; else, ARIMA is applied.
(d) Rainfall time series should be univariate. The autoregressive and moving average part is about regression with previous values and errors.
(e) Residual obtained from a SARIMA fitted time series should be white noise or a series with cyclic behaviour as they are deemed to be stationary series.

3.12. SN method

In the case of highly seasonal data, SN is a better choice than the naïve model. A forecast of a particular season is equal to the value of the past specific season's value from the previous year (Hyndman & Athanasopoulos 2018). The monthly rainfall values are forecasted using the following equation:

$$\widehat{Y_{T+h|T}} = Y_{T+h} - m(K+1) \tag{9}$$

where $\widehat{Y_{T+h|T}}$ is the rainfall estimate of time series consisting of T data points (Y_1, Y_2, Y_3, ... Y_T) for a forecast horizon h, m is the seasonal period and K is the integer part of $(h-1)/m$.

The forecasts obtained using the SN are used as a benchmark to assess the performance of HK-SARIMA, YJNSTF and NSTF techniques. SN is applied to improve the inference made from results and act as a standard reference to measure the relative outcomes of HK-SARIMA, NSTF and YJNSTF.

3.13. Scale-dependent errors

RMSE is the square root of the average of the square of all of the errors between forecasted rainfall and observed rainfall time series. It calculates the accuracy; it can only evaluate errors of different models for a specific variable and not among variables (Agboola *et al.* 2013). Mean absolute error (MAE) calculates the mean magnitude of the errors in forecasts set without considering their direction. It quantifies the amount of accuracy for continuous variables. It has a linear score which denotes that all the individual differences are weighted equally in the average (Agboola *et al.* 2013)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)} \tag{10}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|S_i - O_i|}{n} \tag{11}$$

where S_i is the forecasted value from HK-SARIMA and YJNSTF models, O_i is the observed rainfall value and n is the number of observations available for analysis.

3.14. Nash–Sutcliffe model efficiency coefficient (NSE)

Conventionally, NSE is applied as a performance indicator (Nash & Sutcliffe 1970; Hu *et al.* 2020) to evaluate the goodness of fit between the modelled and observed time series in the field of hydrology. Mathematically, it is expressed as one minus the fraction of the error variance of the modelled time series divided by the variance of the observed time series. NSE value in the vicinity of one (NSE = 1) designates a higher predictive skill of the considered model. NSE value equal to or near zero signifies the same predictive skill as that of the mean of the time series. If the value is negative (NSE < 1), it specifies that the

observed mean is a better predictor than the simulated results.

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T} (Q_0^t - Q_m^t)^2}{\sum_{t=1}^{T} (Q_0^t - \overline{Q_0})^2} \tag{12}$$

where Q_0^t is the observed rainfall time series, $\overline{Q_0}$ is the mean of the observed rainfall time series and Q_m^t is the modelled rainfall time series.

4. RESULTS AND DISCUSSION

All the statistical tests and forecasting methods explained in the earlier section are applied to the monthly rainfall time-series datasets of 18 stations located in Kerala. The outcomes of these statistical tests and their importance are discussed in this section. Initially, the missing rainfall data were filled by the iterative EM algorithm. The percentage of missing rainfall data at each rainfall station varies between 0.76% (at Palakkad (O) and Trivandrum Aero (O)) and 17.42% (at Chalakudi).

The initial 90% of the data, i.e., 1981–2010, is considered for training (30 years), and the remaining 10% of data, i.e., 2011–2013 (3 years), is retained for testing purposes. The forecasting methodology is explained using the rainfall time series of the Nedumangad station in the following sections. A similar methodology is followed for the remaining 17 stations.

4.1. Descriptive statistics

4.1.1. Measures of central tendency

The descriptive statistics of monthly rainfall data of the 18 stations are listed in Table 1. The mean of monthly rainfall data ranged between 141.29 mm (at Trivandrum Aero (O)) and 295.6 mm (at Hosdurg) (Table 1). The mode of all 18 rainfall time series are equal, and the value is zero. The mode frequency varies between 17 (at Konni) and 88 (at Hosdurg) (Table 1). The inference is that classical Box-Cox transformation and log transformation cannot be applied since all the 18 rainfall series have at least 17 data points whose value is equal to zero. The median of the 18 rainfall varies from 85.50 mm (at Mannanthavady) to 235.80 mm (at Konni) (Table 1). The mean value of 18 stations is more than the respective median values, which indicates the rainfall dataset is right-skewed and has a longer tail in the high end than the low end.

4.1.2. Measures of dispersion

The SEM varies from 6.95 mm (at Trivandrum Aero (O)) to 21.20 mm (at Kasargod). There is not much difference in the SEM value of the three regions. The average SEM obtained (including all 18 stations) was 12.25 mm (Table 1). The amount of dispersion of the sample mean from the population mean is very low (negligible) compared with the range of rainfall. It can be concluded that all 18 stations' sample mean is the best estimator of their population mean, respectively. The SD varies from 131.79 mm (at Trivandrum Aero (O)) to 401.40 mm (at Kasargod). SD is the measure of the dispersion of the data around the mean. It was noticed there was a drop of 42.13 mm (rainfall) in the average value of SD in the Eastern highlands compared with the Western lowlands and Central midlands (Table 1). The variance of the rainfall dataset varies from 17,369.67 mm^2 (at Trivandrum Aero (O)) to 161,086.80 mm^2 (at Kasargod). There is a decrease in the amount of variance in the Eastern highlands compared with that of the Western lowlands and Central midlands. The variance of rainfall was unequal between the stations, leading to biased and skewed test outcomes. Thereby any statistical technique applied to rainfall datasets needs to be non-parametric since parametric tests are susceptible to changes in the variance. The CV is helpful in assessing the degree of variation between two time-series datasets, even when there is a significant difference in mean values. CV varies between 83.94 mm (at Konni) and 139.19 mm (at Kasargod). The range of the time series varies between 761.20 mm (at Trivandrum Aero (O)) and 1,776.90 mm (at Kasargod) (Table 1). The IQR of the time series varies between 186.2 mm (at Trivandrum Aero (O)) and 467.8 mm Hosdurg (at Hosdurg).

The skewness value of 17 stations is more than 1. The skewness value ranged between 0.82 (at Konni) and 1.71 (at Mannanthavady). Except for Konni, all other rainfall time series are highly positively right-skewed. Thereby mean (157.23) > median (114.65) > mode (0) for Nedumangad station (Table 1). A similar pattern exists in the remaining 17 stations. The excess Kurtosis of all 18 stations is more than 0. The excess Kurtosis ranged between 0.26 (at Chalakudi) and 5.14 (at Punalur (O)). Five stations have excess Kurtosis values between 0 and 1, implying mesokurtic distribution where the tail part of the

Table 1 | Descriptive statistics computed for monthly rainfall time-series data of 18 stations

	Measures of central tendency				Measures of dispersion							
Station name	**Mean (mm)**	**Median (mm)**	**Mode (mm)**	**Frequency of the mode**	**SEM (mm)**	**SD (mm)**	**Variance (mm^2)**	**CV**	**Range (mm)**	**IQR (mm)**	**Skewness**	**Kurtosis**
Alapuzzha (O)	234.10	166.90	0.00	20.00	11.70	222.30	49,414.20	94.97	1,190.80	340.30	1.02	0.60
Alathur Hydro	161.29	86.95	0.00	66.00	9.74	184.79	34,147.92	114.57	889.50	260.00	1.35	1.59
Aluva PWD	238.90	140.60	0.00	43.00	13.60	258.90	67,039.40	108.39	1,241.40	392.30	1.13	0.69
Chalakudi	275.00	171.30	0.00	52.00	15.50	294.70	86,875.70	107.16	1,293.20	452.80	1.03	0.26
Haripad	237.00	168.60	0.00	49.00	12.20	230.90	53,294.50	97.40	1,140.70	320.40	1.00	0.34
Hosdurg	295.60	100.50	0.00	88.00	21.00	398.30	158,658.00	134.77	1,758.80	467.90	1.43	1.14
Kasargod	288.40	89.40	0.00	84.00	21.20	401.40	161,086.80	139.19	1,776.90	431.60	1.58	1.69
Kayamkulam	199.80	154.90	0.00	32.00	10.10	192.20	36,941.00	96.19	1,189.50	278.60	1.18	1.74
Kollam Hydro	164.64	122.20	0.00	32.00	8.62	163.51	26,736.00	99.32	1,113.50	229.32	1.38	3.00
Konni	260.90	235.80	0.00	17.00	11.50	219.00	47,953.90	83.94	1,178.50	330.70	0.82	0.32
Kottayam (O)	238.30	178.90	0.00	41.00	12.20	231.10	53,428.60	97.00	1,416.30	362.10	1.15	1.55
Kozhikode (O)	252.50	121.30	0.00	55.00	16.20	307.10	94,317.80	121.62	1,467.10	399.10	1.45	1.73
Nedumangad	157.23	114.65	0.00	25.00	8.08	153.30	23,501.39	97.50	859.20	215.88	1.31	2.04
Mannanthavady	208.20	85.50	0.00	59.00	14.40	273.00	74,541.80	131.11	1,367.20	317.20	1.71	2.83
Palakkad (O)	164.58	88.60	0.00	62.00	9.89	187.56	35,178.80	113.96	878.80	245.00	1.32	1.22
Punalur (O)	225.30	195.00	0.00	29.00	10.50	198.30	39,338.80	88.05	1,537.00	284.20	1.47	5.14
Thiruvananthapuram (O)	145.49	115.80	0.00	19.00	7.15	135.61	18,389.64	93.21	872.00	193.85	1.21	2.06
Trivandrum Aero (O)	141.29	114.95	0.00	27.00	6.95	131.79	17,369.67	93.28	761.20	186.28	1.25	1.92

Note: SEM, standard error of the mean; SD, standard deviation; CV, coefficient of variation; IQR, interquartile range.

distribution is similar to that of normal distribution (Table 1). The mesokurtic distribution has a Kurtosis value close to zero or negligible probability of having extreme outcomes. The remaining 13 stations have excess Kurtosis with values exceeding 1, thereby classifying them as leptokurtic distributions. A leptokurtic distribution has heavier tails than the normal distribution, implying a high probability of extreme outcomes. The overall inference derived from the descriptive statistics is that all 18 rainfall time series have a non-normal distribution. Two normality tests and graphical plots, namely the QQ plots, are used to ascertain the inference of non-normal distribution.

4.2. ACF and PACF plot

The ACF and PACF plot results are used to graphically examine the presence of AC and non-stationary behaviour in the time series. The presence of AC in a time-series results in underestimating the standard error of independent variables. Underestimation leads to endorsing the coefficients of the variables as statistically significant values (when they are not). Thereby, the pattern of spikes in ACF and PACF plots in the time series before and after the modelling process needs to be compared statistically to make accurate inferences.

The rule of thumb is that non-stationary time series display a slow and gradual decay of the spikes (vertical bars) in AC value as the time lag increases. In contrast, the value of spikes in stationary time series drops to zero comparatively faster. In other words, in the nonappearance of AC in a time series, the succeeding spikes would drop rapidly to almost zero or at least between the dashed lines.

In Figure 2(a) and 2(b), there is neither a gradual decay nor a sudden drop in the spike. Instead, a sinusoidal pattern is observed, which is classified as AC and the reason for this particular pattern is due to the presence of strong seasonality. Thereby, the time series is classified as non-stationary. The residuals' pattern of ACF and PACF is discussed in the following sections.

4.3. Empirical normality test, graphical plots and YJT

SW and AD tests are carried out to ascertain whether the rainfall dataset follows a normal distribution. The result consists of the test statistic and its p-value at α value of 0.05 as the desired significance level. Suppose the p-value is less than 0.05

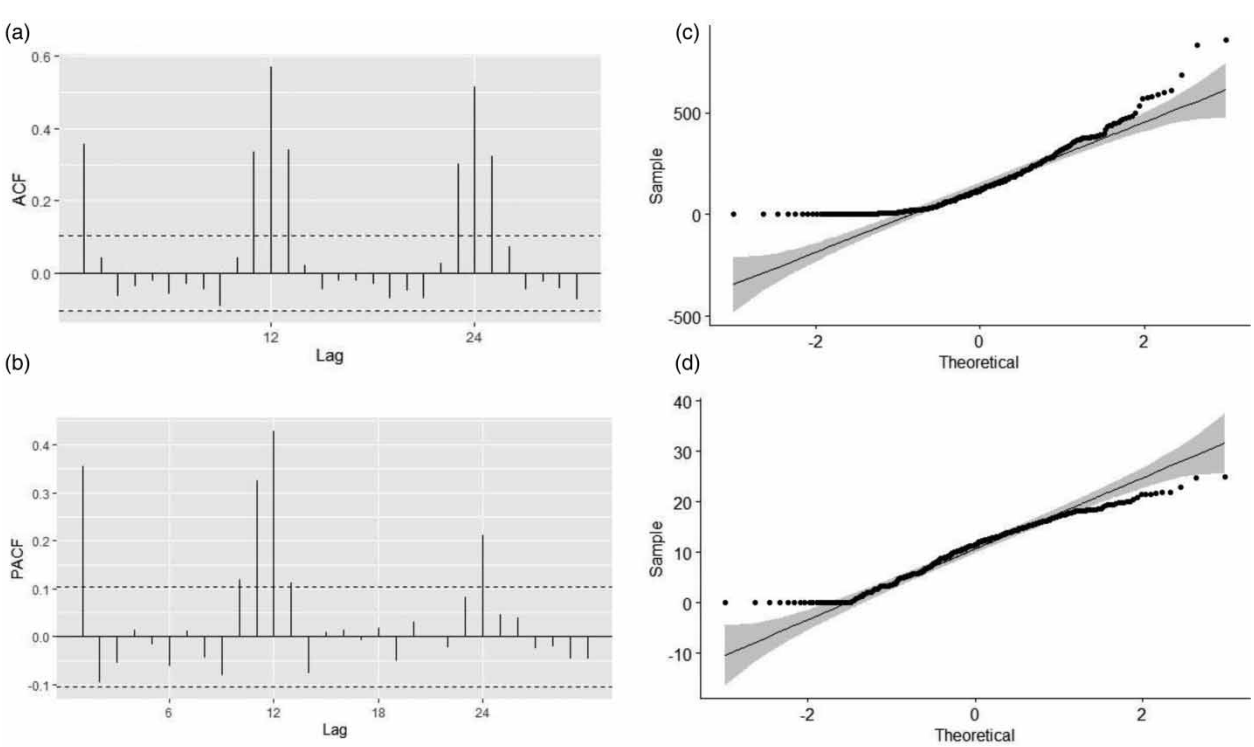


Figure 2 | (a) The ACF and (b) PACF plot of Nedumangad time series, QQ plot (c) before and (d) after Yoe-Johnson Transformation for Nedumangad time series.

($p \leq 0.05$), the null hypothesis is rejected, and if the p-value is more significant than 0.05, then the null hypothesis is not rejected. The p-value of the SW test varies between 2.26×10^{-16} (at Alathur Hydro) and 1.88×10^{-12} (at Konni) from Table 2. Since all p-values are less than 0.05, the null hypothesis is rejected by accepting the alternative hypothesis, implying that the datasets are non-normally distributed. After applying YJT, the p-value increased for all 18 stations (Table 2). The p-value of the SW test varies between 1.79×10^{-15} (at Hosdurg) and 7.64×10^{-7} (at Nedumangad). Though the result still indicates the distribution is non-normal. It can be inferred that Yeo-Johnson's transformation converted the time series to more Gaussian-like distribution but not 100% perfectly normal. Similarly, on applying the AD test p-value varies between 2.20×10^{-16} (at Alapuzzha (O)) and 1.5×10^{-15} (at Punalur(O)) (Table 2). After the transformation, the p-value increased for all the 18 stations. It varies between 2.20×10^{-16} (at Chalakudi) and 3.55×10^{-7} (at Alapuzzha (O)).

If the data points in the rainfall time series are approximately normally distributed, then the QQ plot would be a straight line with most observations in its vicinity. Then, rainfall data points are plotted against appropriate quantiles from the standard normal distribution. The ith order value is plotted against $i\text{th}/n + 1$ quantile of the standard normal distribution. The straight line drawn goes through the first and third quartiles of the distribution. The QQ plot before transformation (Figure 2(c)) displays that relatively fewer points fall close to the 45° line. After transformation (Figure 2(d)), more points are in the vicinity of the straight line.

4.4. Strength of trend and strength of seasonality

The strength of the trend varied between 0.047 (at Thiruvananthapuram (O)) and 0.224 (at Kayakulam). The strength of seasonality varied between 0.480 (at Thiruvananthapuram (O)) and 0.856 (at Hosdurg) (Table 3). There exists no literature to classify and quantify the magnitude. Thereby, a new classification system is proposed. The value of strength between 0 and 0.3 is considered negligible, between 0.3 and 0.7 as moderate and between 0.7 and 1 as high. If the value of strength is more than 0.3, it can be stated as non-stationary due to a particular component (either trend or seasonality). All 18 time series had their $F_T < 0.3$ (Table 3). The trend component is considered negligible, and the trend component is considered

Table 2 | Results of normality test before and after applying YJT on rainfall time-series data of all the 18 stations

	Shapiro–Wilk test				Anderson–Darling test			
	Before transformation		After transformation		Before transformation		After transformation	
Station name	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Alapuzzha (O)	0.89	2×10^{-15}	0.97	2.11×10^{-7}	11.87	$<2.20 \times 10^{-16}$	2.85	3.55×10^{-7}
Alathur Hydro	0.83	$<2.20 \times 10^{-16}$	0.92	3.08×10^{-13}	18.77	$<2.20 \times 10^{-16}$	9.50	$<2.20 \times 10^{-16}$
Aluva PWD	0.85	$<2.20 \times 10^{-16}$	0.94	5.93×10^{-11}	17.78	$<2.20 \times 10^{-16}$	6.04	7.03×10^{-15}
Chalakudi	0.86	$<2.20 \times 10^{-16}$	0.93	3.91×10^{-12}	17.59	$<2.20 \times 10^{-16}$	7.38	$<2.20 \times 10^{-16}$
Haripad	0.89	1.02×10^{-15}	0.95	3.95×10^{-10}	12.41	$<2.20 \times 10^{-16}$	4.96	2.71×10^{-12}
Hosdurg	0.76	$<2.20 \times 10^{-16}$	0.89	1.79×10^{-15}	33.54	$<2.20 \times 10^{-16}$	12.81	$<2.20 \times 10^{-16}$
Kasargod	0.75	$<2.20 \times 10^{-16}$	0.90	1.90×10^{-14}	34.99	$<2.20 \times 10^{-16}$	10.69	$<2.20 \times 10^{-16}$
Kayamkulam	0.88	7.65×10^{-16}	0.96	3.99×10^{-8}	10.97	$<2.20 \times 10^{-16}$	3.68	3.34×10^{-9}
Kollam Hydro	0.87	$<2.20 \times 10^{-16}$	0.96	5.68×10^{-8}	12.04	$<2.20 \times 10^{-16}$	3.53	7.90×10^{-9}
Konni	0.92	1.88×10^{-12}	0.97	6.43×10^{-7}	6.80	$<2.20 \times 10^{-16}$	3.16	6.22×10^{-8}
Kottayam (O)	0.88	5.94×10^{-16}	0.95	1.14×10^{-9}	11.69	$<2.20 \times 10^{-16}$	4.77	7.66×10^{-12}
Kozhikode (O)	0.81	$<2.20 \times 10^{-16}$	0.92	1.87×10^{-12}	23.57	$<2.20 \times 10^{-16}$	7.83	$<2.20 \times 10^{-16}$
Nedumangad	0.88	$<2.20 \times 10^{-16}$	0.97	7.64×10^{-7}	11.47	$<2.20 \times 10^{-16}$	2.86	3.26×10^{-7}
Mannanthavady	0.77	$<2.20 \times 10^{-16}$	0.94	1.52×10^{-10}	29.14	$<2.20 \times 10^{-16}$	5.07	1.48×10^{-12}
Palakkad (O)	0.83	$<2.20 \times 10^{-16}$	0.92	1.87×10^{-12}	19.55	$<2.20 \times 10^{-16}$	7.99	$<2.20 \times 10^{-16}$
Punalur (O)	0.89	$<2.44 \times 10^{-15}$	0.96	6.17×10^{-8}	6.32	1.51×10^{-15}	3.88	1.12×10^{-9}
Thiruvananthapuram (O)	0.89	$<2.42 \times 10^{-15}$	0.97	3.96×10^{-7}	9.52	$<2.20 \times 10^{-16}$	3.17	6.01×10^{-8}
Trivandrum Aero (O)	0.89	2.11×10^{-15}	0.96	4.96×10^{-8}	9.15	$<2.20 \times 10^{-16}$	3.97	6.62×10^{-10}

Table 3 | Strength of trend and strength of seasonality obtained for monthly rainfall time-series datasets

Station name	Strength of trend (F_T)	Strength of seasonality (F_S)
Alapuzzha (O)	0.052	0.681
Alathur Hydro	0.151	0.758
Aluva PWD	0.069	0.774
Chalakudi	0.187	0.805
Haripad	0.145	0.697
Hosdurg	0.123	0.856
Kasargod	0.121	0.850
Kayamkulam	0.224	0.710
Kollam Hydro	0.208	0.558
Konni	0.161	0.678
Kottayam (O)	0.101	0.713
Kozhikode (O)	0.066	0.770
Nedumangad	0.211	0.493
Mannanthavady	0.067	0.813
Palakkad (O)	0.082	0.751
Punalur (O)	0.124	0.586
Thiruvananthapuram (O)	0.047	0.480
Trivandrum Aero (O)	0.104	0.482

statistically stationary. Compared with the trend, the magnitude of the seasonal component is more, and the values are close to one for 10 stations ($F_s > 0.7$), which is classified as high. In the remaining eight rainfall stations, the values were between 0.3 and 0.7, classified as moderate. All 18 stations had their $F_s > 0.3$ (Table 3); thereby, it can be concluded that the seasonal component is more dominant than the trend component in the datasets.

4.5. TF method

Several statistical parameters are to be derived from monthly rainfall time-series data for developing the TF model (Ghanbarpour *et al.* 2009). To construct a TF model for a chosen month, the mean monthly rainfall, the correlation coefficient of the chosen month with its preceding month and the SD of monthly rainfall values of that specific month must be determined. A distinct model is developed for each month to forecast the monthly rainfall. Table 4 provides a list of TF models fitted to monthly rainfall data after applying YJT for each month for Nedumangad station.

4.6. Hyndman Khandakar-Seasonal Autoregressive Integrated Moving Average (HK-SARIMA)

There are five steps in HK-SARIMA modelling and the auto.arima() function in the R programming language automates Steps 1–3, which are explained in detail in Section 3.10. Initially, 90% of data (from 1981 to 2010) is used for training the data, and the remaining 10% of data (from 2011 to 2013) is used to test the data (Hyndman & Athanasopoulos 2018). Here, a brief explanation using the Nedumangad rainfall station is provided. In application of Step 1, it was observed that one seasonal difference is sufficient to make rainfall time series stationary. In Step 2, a number of unique SARIMA models with different model parameters and orders were formed by applying the constraints. The model with the least AICc and Bayesian Information Criterion (BIC) is chosen as the best. The output obtained from the HK algorithm is HK-SARIMA$(1,0,1)(0,1,1)_{12}$ for the rainfall data at the Nedumangad station. In Step 3, the magnitude and significance of the model parameters were determined using MLE, and coefficients were tested for their significance using the coeftest() function. For the monthly rainfall datasets of Nedumangad station, all the coefficients had a p-value > 0.05, failing to reject the null hypothesis. Thereby, it can be concluded that the HK-SARIMA model developed is statistically significant. In the subsequent step, ACF and PACF plots for the residuals from HK-SARIMA$(1,0,1)(0,1,1)_{12}$ indicated that all the autocorrelations are within the threshold limit. All 36 spikes in the residual ACF and PACF are insignificant (Figure 3). Finally, the p-value of the Ljung-Box statistic is

Table 4 | List of Thomas-Fiering models fitted to the monthly rainfall data of Nedumangad station for each month

$X_{JAN} = 3.833 + 0.143(X_{DEC} - 6.742) + t_{1,JAN}\sqrt{1 - 0.162^2}$
$X_{FEB} = 2.929 + 0.240(X_{JAN} - 3.833) + t_{1,FEB}\sqrt{1 - 0.275^2}$
$X_{MAR} = 6.607 + 0.308(X_{FEB} - 2.929) + t_{1,MAR}\sqrt{1 - 0.215^2}$
$X_{APR} = 12.178 + 0.415(X_{MAR} - 6.607) + t_{1,APR}\sqrt{1 - 0.409^2}$
$X_{MAY} = 12.431 + 0.155(X_{APR} - 12.178) + t_{1,MAY}\sqrt{1 - 0.153^2}$
$X_{JUN} = 16.415 + 0.200(X_{MAY} - 12.431) + t_{1,JUN}\sqrt{1 - 0.259^2}$
$X_{JUL} = 14.322 + 0.275(X_{JUN} - 16.415) + t_{1,JUL}\sqrt{1 - 0.291^2}$
$X_{AUG} = 11.442 + 0.265(X_{JUL} - 14.322) + t_{1,AUG}\sqrt{1 - 0.244^2}$
$X_{SEP} = 12.107 + 0.748(X_{AUG} - 11.442) + t_{1,SEP}\sqrt{1 - 0.523^2}$
$X_{OCT} = 15.819 + 0.160(X_{SEP} - 12.107) + t_{1,OCT}\sqrt{1 - 0.151^2}$
$X_{NOV} = 15.258 + 0.286(X_{OCT} - 15.819) + t_{1,NOV}\sqrt{1 - 0.461^2}$
$X_{DEC} = 6.742 + 0.518(X_{NOV} - 15.258) + t_{1,DEC}\sqrt{1 - 0.436^2}$

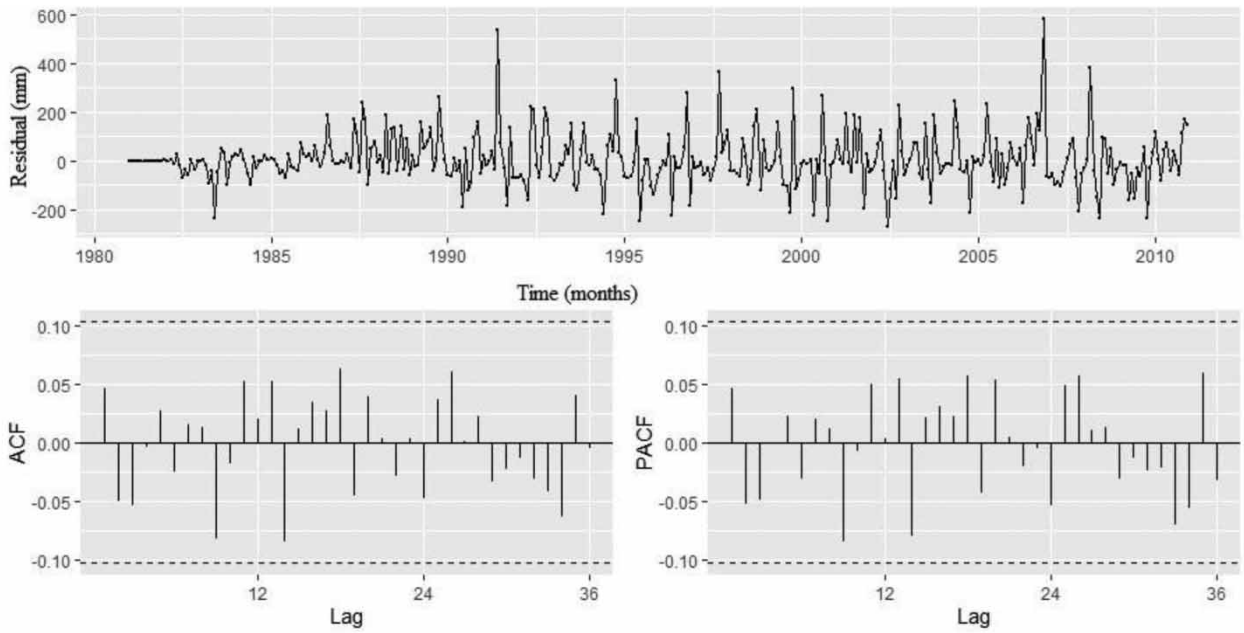


Figure 3 | The residual time series of the fitted HK-SARIMA(1,0,1)(0,1,1)$_{12}$ model for the Nedumangad station, ACF and PACF plot of the residuals.

found to be 0.771 (Table 5). The *p*-value is more significant than 0.05, implying failure to reject the null hypothesis and the residuals are white noise. The forecast for the testing period is carried out and R^2 is calculated to be 0.654 (Figure 4).

Similar studies are carried out in the remaining 17 stations, and respective models have arrived. Some interesting statistical results were explicitly obtained for the Ljung-box test when the above procedure was applied to 17 stations. Table 5 gives the result of the Ljung-Box test. Out of the 18 stations, 16 stations have a *p*-value of more than 0.05 and two stations have a *p*-value of less than 0.05. Four spikes are significant in the ACF residual plot (Figure 5(a)), and four spikes are significant in the PACF plot (Figure 5(b)) for Thiruvananthapuram (O). Four spikes are significant in the ACF residual plot (Figure 5(c)), and three spikes are significant in the PACF plot (Figure 5(d)) for Trivandrum Aero (O). Therefore, it is concluded that the HK-SARIMA model can capture only 88.89% of information from the data in ACF and PACF plots, which is less than the threshold limit of 95%.

Table 5 | HK-SARIMA models developed for the monthly rainfall time-series datasets of the stations

Station name	Best-fit HK-SARIMA model	Information criterion based on the training dataset			Ljung-Box test		
		AIC	AICc	BIC	*Q**	df	*p*-value
Alapuzzha (O)	HK-SARIMA$(0,0,3)(1,1,1)_{12}$	4,409.21	4,409.46	4,432.33	20.01	19	0.39
Alathur Hydro	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	4,271.79	4,271.86	4,283.35	16.96	22	0.77
Aluva PWD	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	4,479.87	4,479.94	4,491.43	24.40	22	0.33
Chalakudi	HK-SARIMA$(1,0,1)(2,1,0)_{12}$	4,553.61	4,553.79	4,572.88	23.61	20	0.26
Haripad	HK-SARIMA$(0,0,2)(2,1,0)_{12}$	4,492.96	4,493.13	4,512.22	14.09	20	0.83
Hosdurg	HK-SARIMA$(1,0,0)(2,1,0)_{12}$	4,598.52	4,598.64	4,613.93	23.10	21	0.34
Kasargod	HK-SARIMA$(0,0,1)(0,1,2)_{12}$	4,550.97	4,551.09	4,566.38	15.72	21	0.79
Kayamkulam	HK-SARIMA$(2,0,0)(2,1,0)_{12}$	4,349.16	4,349.33	4,368.42	20.19	20	0.45
Kollam Hydro	HK-SARIMA$(4,0,0)(0,1,1)_{12}$	4,311.96	4,312.21	4,335.07	9.78	19	0.96
Konni	HK-SARIMA$(1,0,1)(2,1,0)_{12}$	4,500.29	4,500.46	4,519.55	28.07	20	0.11
Kottayam (O)	HK-SARIMA$(0,0,0)(0,1,1)_{12}$	4,402.36	4,402.40	4,410.07	26.28	23	0.29
Kozhikode (O)	HK-SARIMA$(0,0,0)(1,1,1)_{12}$	4,530.07	4,530.14	4,541.62	17.28	22	0.75
Nedumangad	HK-SARIMA$(1,0,1)(0,1,1)_{12}$	4,284.54	4,284.66	4,299.95	15.97	21	0.77
Mannanthavady	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	4,442.56	4,442.56	4,442.56	29.92	22	0.12
Palakkad (O)	HK-SARIMA$(0,0,0)(1,1,2)_{12}$	4,227.22	4,227.34	4,242.63	22.25	21	0.39
Punalur (O)	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	4,484.37	4,484.44	4,495.92	19.87	22	0.59
Thiruvananthapuram (O)	HK-SARIMA$(0,0,0)(2,0,0)_{12}$	4,454.50	4,454.62	4,470.05	50.55	21	0.00
Trivandrum Aero (O)	HK-SARIMA$(0,0,3)(2,0,0)_{12}$	4,433.03	4,433.35	4,460.23	44.46	18	0.00

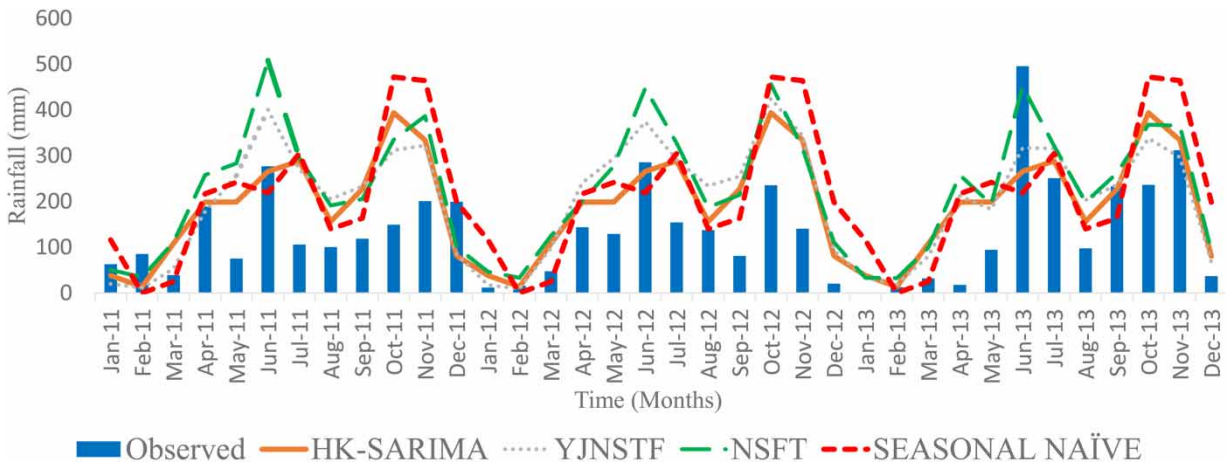


Figure 4 | The observed time series, HK-SARIMA forecast, YJNSTF forecast, NSTF forecast and Seasonal Naive (January 2011–December 2013) for the Nedumangad station.

4.7. Model performance

4.7.1. RMSE

The benefit of using RMSE is that it sums the magnitude of the deviations between the forecasted rainfall and observed rainfall data into a distinct amount of predictive power. RMSE is the measure of accuracy to associate forecasting deviations of different models for a specific dataset (rainfall test data in our study). The RMSE values range between 97.82 mm (at

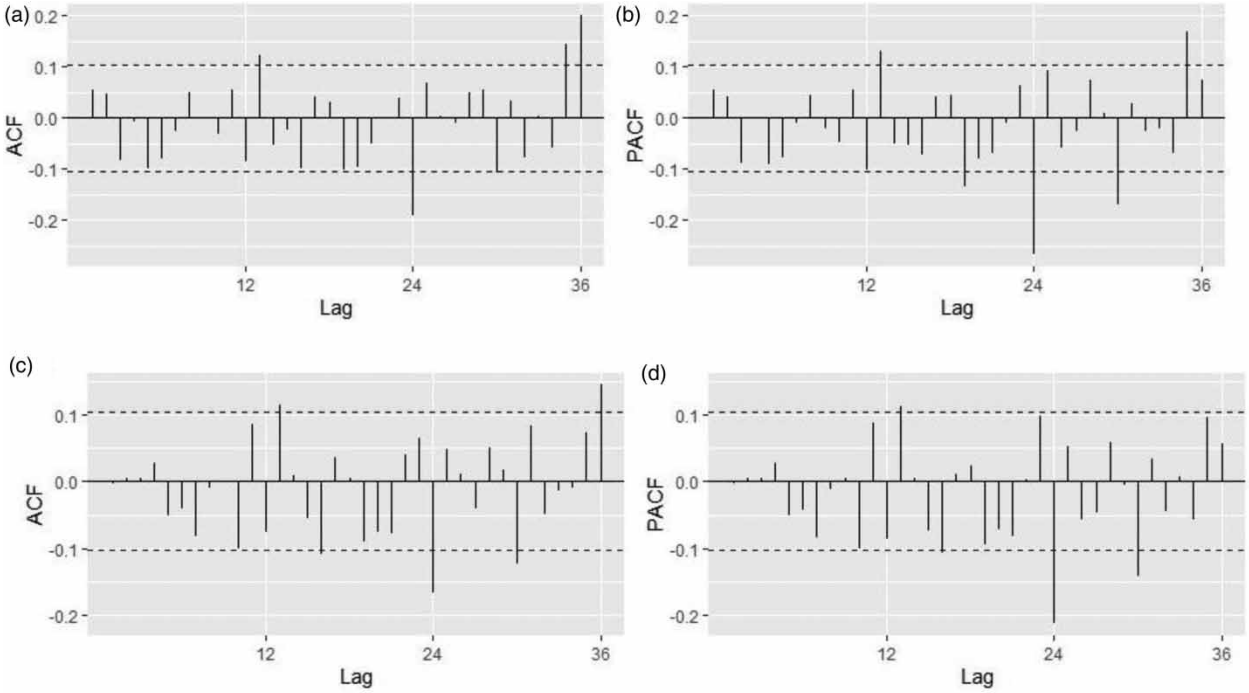


Figure 5 | (a) The ACF and (b) PACF plot of residuals of Thiruvananthapuram (O) time series and (c) ACF and (d) PACF plot of residuals of Trivandrum Aero (O) time series.

Thiruvananthapuram) and 195.92 mm (at Mannanthavady) for HK-SARIMA models. The RMSE ranges between 111.49 mm (at Alathur Hydro) and 195.26 mm (at Palakkad (O)) for NSTF models (Table 6). The RMSE ranges between 105.36 mm (at Alathur Hydro) and 190.75 mm (at Hosdurg) for YJSNTF. The RMSE values for SN range between 78.25 mm (at Alathur Hydro) and 258.92 mm (at Kasargod) (Table 6). Out of 18 monthly rainfall datasets, except one, the other 17 datasets have shown a reduction in RMSE value after applying YJT. Konni is the only station where the RMSE value marginally increased from 153.34 to 159.72 mm on applying the YJT. The average reduction of RMSE values between YJNSTF and NSTF is about 6.46% (Table 6).

The amount of deviations expressed in the form of RMSE is relative to the magnitude of squared errors; thus, larger errors have a significant effect on the computed RMSE. Kerala receives more than 70% of rainfall during the monsoon season, spanning June to September. The delayed setting of the monsoon in some years has resulted in significant variations in the monthly precipitation. The CV in the monthly rainfall at various locations in the state ranges between 24 and 53 mm in the month of June, 28–48 mm in the month of July, 30–66 mm in the month of August and 50–74 mm in the month of September (Guhathakurta *et al.* 2020). Similarly, pre-monsoon thundershowers are erratic, leading to significant variations in the monthly rainfall values for different years. These large deviations in the monthly rainfall values from its mean can be attributed to the higher RMSE values in the model forecasts.

The percentage reduction of RMSE by three models (HK-SARIMA, NSTF and YJNSTF) is compared with the SN model. It was inferred that the average reduction of RMSE per station was equal to 15.43% for HK-SARIMA, 7.34% for NSTF and 13.23% for YJNSTF (Table 6). The amount of reduction in RMSE of the forecasts by HK-SARIMA is not that high compared with YJNSTF. The results indicate that both HK-SARIMA and YJNSTF models are suitable for forecasting monthly rainfall data (Table 6).

4.7.2. MAE

RMSE is an average deviation statistic that depends on squared deviations, distributions of deviation magnitudes and $\sqrt{n}$ (where n is the number of data points). The limitation of RMSE is that it becomes more significant in magnitude when the distribution of deviation magnitude becomes more variable (Willmott & Matsuura 2005). Therefore, MAE, which is described as a natural measure of average deviations, unambiguous for dimensioned assessment, is computed as an additional

Table 6 | Comparison of the goodness-of-fit statistics for HK-SARIMA, YJNSTF, NSTF and SN model for the monthly rainfall time-series datasets of all stations

Station name	Models	RMSE (mm)	% Reduction in RMSE when compared with SN	MAE (mm)	% Reduction in MAE when compared with SN	NSE	NRMSE (mean)	NRMSE (SD)	Best model
Western lowlands									
Haripad	HK-SARIMA$(0,0,2)(2,1,0)_{12}$	141.50	18.50%	109.44	18.16%	0.55	0.62	0.66	YJNSTF
	NSTF	156.23	10.02%	113.65	15.01%	0.45	0.68	0.73	
	YJNSTF	**137.56**	**20.77%**	**102.36**	**23.45%**	**0.58**	**0.60**	**0.64**	
	Seasonal Naïve	173.62		133.72		0.32	0.76	0.81	
Aluva PWD	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	**131.60**	**16.81%**	92.18	22.44%	**0.77**	**0.53**	**0.47**	HK-SARIMA and YJNSTF
	NSTF	142.58	9.87%	94.47	20.52%	0.73	0.57	0.51	
	YJNSTF	132.44	16.28%	**84.97**	**28.51%**	0.77	0.53	0.48	
	Seasonal Naïve	158.20		118.86		0.67	0.64	0.57	
Central midlands									
Alapuzzha (O)	HK-SARIMA$(0,0,3)(1,1,1)_{12}$	**149.97**	**20.31%**	**106.86**	**16.88%**	**0.46**	**0.75**	**0.72**	HK-SARIMA
	NSTF	168.87	10.27%	128.52	0.03%	0.32	0.85	0.82	
	YJNSTF	157.92	16.09%	118.79	7.60%	0.40	0.79	0.76	
	Seasonal Naïve	188.19		128.56		0.15	0.94	0.91	
Chalakudi	HK-SARIMA$(1,0,1)(2,1,0)_{12}$	**154.75**	**17.22%**	110.96	21.75%	**0.68**	**0.63**	**0.54**	HK-SARIMA
	NSTF	167.38	10.47%	118.89	16.15%	0.63	0.68	0.60	
	YJNSTF	159.85	14.50%	**108.51**	**23.47%**	0.66	0.65	0.57	
	Seasonal Naïve	186.95		141.80		0.54	0.76	0.67	
Hosdurg	HK-SARIMA$(1,0,0)(2,1,0)_{12}$	186.96	−3.16%	106.71	3.80%	0.76	**0.63**	**0.48**	Seasonal Naïve
	NSTF	192.76	−6.36%	124.58	−12.31%	0.75	0.65	0.50	
	YJNSTF	190.75	−5.25%	111.13	−0.19%	0.75	0.64	0.49	
	Seasonal Naïve	**181.24**		**110.92**		**0.78**	**0.61**	**0.47**	
Kasargod	HK-SARIMA$(0,0,1)(0,1,2)_{12}$	172.32	33.45%	115.88	33.41%	0.78	0.57	0.46	YJNSTF
	NSTF	177.53	31.43%	110.85	36.30%	0.77	0.58	0.47	
	YJNSTF	**155.12**	**40.09%**	**99.29**	**42.95%**	**0.82**	**0.51**	**0.41**	
	Seasonal Naïve	258.92		174.03		0.51	0.85	0.69	
Kayamkulam	HK-SARIMA$(2,0,0)(2,1,0)_{12}$	155.78	23.17%	117.58	20.06%	0.33	0.79	0.81	YJNSTF
	NSTF	142.81	29.56%	107.72	26.76%	0.44	0.73	0.74	
	YJNSTF	**135.41**	**33.21%**	**102.67**	**30.20%**	**0.49**	**0.69**	**0.70**	
	Seasonal Naïve	202.75		147.09		−0.14	1.03	1.05	
Kollam Hydro	HK-SARIMA$(4,0,0)(0,1,1)_{12}$	**118.71**	**20.75%**	**73.47**	**28.77%**	**0.43**	**0.73**	**0.74**	HK-SARIMA
	NSTF	153.36	−2.39%	106.74	−3.48%	0.05	0.93	0.96	
	YJNSTF	146.93	1.91%	107.77	−4.48%	0.13	0.90	0.92	
	Seasonal Naïve	149.78		103.15		0.10	0.92	0.94	
Konni	HK-SARIMA$(1,0,1)(2,1,0)_{12}$	**138.63**	**25.56%**	**108.78**	**19.63%**	**0.20**	**0.70**	**0.88**	HK-SARIMA
	NSTF	153.34	17.66%	120.91	10.67%	0.02	0.77	0.98	
	YJNSTF	159.72	14.23%	129.53	4.30%	−0.61	0.80	1.02	
	Seasonal Naïve	186.23		135.36		−0.44	0.93	1.18	

(*Continued.*)

Table 6 | Continued

Station name	Models	RMSE (mm)	% Reduction in RMSE when compared with SN	MAE (mm)	% Reduction in MAE when compared with SN	NSE	NRMSE (mean)	NRMSE (SD)	Best model
Kozhikode (O)	HK-SARIMA$(0,0,0)(1,1,1)_{12}$	**130.74**	**12.46%**	**89.41**	**16.51%**	**0.81**	**0.52**	**0.43**	HK-SARIMA
	NSTF	160.79	−7.66%	131.81	−23.08%	0.71	0.68	0.53	
	YJNSTF	148.93	0.28%	96.77	9.64%	0.75	0.63	0.49	
	Seasonal Naïve	149.35		107.09		0.75	0.63	0.49	
Punalur (O)	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	**128.31**	**28.85%**	**100.46**	25.12%	**0.25**	**0.66**	**0.85**	HK-SARIMA
	NSTF	148.31	17.76%	114.04	15.00%	0.00	0.75	0.98	
	YJNSTF	137.40	23.81%	102.68	23.47%	0.15	0.69	0.91	
	Seasonal Naïve	180.33		134.16		−0.47	0.91	1.20	
Thiruvananthapuram (O)	HK-SARIMA$(0,0,0)(2,0,0)_{12}$	**97.82**	**19.66%**	79.21	17.35%	**0.13**	**0.77**	**0.92**	HK-SARIMA
	NSTF	112.12	7.92%	88.11	8.06%	−0.14	0.88	1.05	
	YJNSTF	105.50	13.36%	**77.99**	**18.62%**	−0.01	0.82	0.99	
	Seasonal Naïve	121.77		95.83		−0.34	0.95	1.14	
Trivandrum Aero (O)	HK-SARIMA$(0,0,3)(2,0,0)_{12}$	**104.21**	**23.11%**	81.44	21.91%	**0.08**	**0.84**	**0.95**	HK-SARIMA
	NSTF	115.74	14.61%	87.02	16.55%	−0.13	0.93	1.05	
	YJNSTF	109.97	18.86%	**79.20**	**24.05%**	−0.02	0.89	1.00	
	Seasonal Naïve	135.54		104.28		−0.56	1.09	1.23	
Kottayam (O)	HK-SARIMA$(0,0,0)(0,1,1)_{12}$	**123.62**	**27.24%**	90.65	30.57%	**0.68**	0.54	0.56	HK-SARIMA and YJNSTF
	NSTF	137.11	19.30%	92.79	28.93%	0.60	0.59	0.62	
	YJNSTF	123.70	27.19%	**89.73**	**31.28%**	0.67	0.54	0.56	
	Seasonal Naïve	169.90		130.57		0.39	0.74	0.77	
Eastern highlands									
Mannanthavady	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	195.92	−3.38%	115.03	−5.98%	0.59	0.88	0.64	YJNSTF
	NSTF	188.13	0.73%	110.77	−2.06%	0.62	0.85	0.61	
	YJNSTF	**165.33**	**12.76%**	**96.74**	**10.87%**	**0.70**	**0.74**	**0.54**	
	Seasonal Naïve	189.51		108.54		0.61	0.85	0.62	
Palakkad (O)	HK-SARIMA$(0,0,0)(1,1,2)_{12}$	**158.56**	**15.12%**	**101.91**	**13.47%**	**−0.67**	**2.07**	**1.28**	HK-SARIMA
	NSTF	195.26	−4.52%	134.57	−14.25%	−1.53	2.55	1.57	
	YJNSTF	184.67	1.14%	119.66	−1.60%	−1.27	2.41	1.49	
	Seasonal Naïve	186.81		117.78		−1.32	2.44	1.50	
Alathur Hydro	HK-SARIMA$(0,0,0)(2,1,0)_{12}$	113.08	−44.50%	71.99	−26.11%	0.39	0.76	0.77	Seasonal Naïve
	NSTF	111.49	−42.47%	73.07	−28.00%	0.41	0.75	0.76	
	YJNSTF	105.36	−34.64%	63.17	−10.66%	0.47	0.71	0.72	
	Seasonal Naïve	**78.25**		**57.08**		**0.71**	**0.52**	**0.53**	
Nedumangad	HK-SARIMA$(1,0,1)(0,1,1)_{12}$	**108.11**	**26.58%**	**84.87**	**25.78%**	**−0.04**	0.81	1.00	HK-SARIMA
	NSTF	123.76	15.95%	104.54	8.57%	**−0.36**	0.92	1.15	
	YJNSTF	112.67	23.48%	93.99	17.80%	**−0.13**	0.84	1.05	
	Seasonal Naïve	147.24		114.34		−0.92	1.10	1.37	

Note: All the bold numerals correspond to the least RMSE, MAE, NRMSE (mean and standard deviation) and maximum NSE for particular rainfall stations.
HK, Hyndman Khandakar; YJNSTF, Yeo-Johnson transformed non-stationary Thomas-Fiering; NSTF, non-stationary Thomas-Fiering.

error statistic. MAE varies between 71.99 mm (at Alathur Hydro) and 117.58 mm (at Kayakulam) for HK-SARIMA models (Table 6). MAE varies between 63.17 mm (at Alathur Hydro) and 129.53 mm (at Konni) for YJNSTF models. MAE varies between 73.07 mm (at Alathur Hydro) and 134.57 mm (at Palakkad(O)) for NSTF. Among all 18 time-series datasets, the application of YJT to rainfall datasets of Konni and Kollam Hydro stations has resulted in increased MAE values. In Kollam Hydro, the increase is marginal (106.74–107.77 mm), whereas in Konni, the increase is a little significant (120.91–129.53 mm). The average reduction percentage of MAE value between YJNSTF and NSTF is 8.99%, which is a statistically significant quantity (Table 6). The MAE value of SN varied between 57.08 mm (at Alathur Hydro) and 174.03 mm (at Kasargod). The percentage reduction of MAE by three models (HK-SARIMA, NSTF and YJNSTF) is compared with the SN model. It was inferred that the average reduction of RMSE per station was equal to 16.86% for HK-SARIMA, 6.63% for NSTF and 15.52% for YJNSTF. Similar to RMSE, it can be resolved that both HK-SARIMA and YJNSTF are better forecasting models (Table 6).

4.7.3. NSE

NSE values vary from $-\infty$ to 1.0. The criteria used to make meaningful inferences from NSE values are as follows (Moriasi *et al.* 2007):

(a) The value NSE = 1 is obtained when the modelled rainfall time series matches the observed rainfall time series.
(b) The value of NSE between 0 and 1 means the modelling procedure followed is better than the mean of the observed time series and is usually regarded as the acceptable level of performance.
(c) The value of NSE < 0 (negative) means the mean of the observed time series is better than the forecasting model, and the resulting forecasts are deemed to be unacceptable.

The NSE values of the forecasts are calculated to understand whether the rainfall forecasts obtained using various modelling techniques are within acceptable limits. It also provides an idea about the relative performance of the models. The forecasts obtained using HK-SARIMA models resulted in negative NSE values for two rainfall stations (Palakkad and Nedumangad). The NSE statistics computed for the rainfall forecasts obtained using NSTF and YJNSTF models have resulted in zero or negative values for five stations. Thiruvananthapuram (O), Trivandrum Aero (O), Palakkad (O) and Nedumangad are the four stations where the NSE statistics for the forecasts from both NSTF and YJNSTF yielded negative values. For the Punalur station, forecasts from NSTF produced an NSE value of zero (Table 6). Similarly, NSE computed for the forecasts from YJNSTF for the Konni station resulted in a negative value. Except for the forecasts obtained for the Konni station, the forecasts from other stations have shown an improvement in NSE values on applying the YJT before developing the NSTF model (Table 6).

For 11 stations, the NSE values of the forecasts given by HK-SARIMA are significantly better than YJNSTF. In four stations, namely Haripad, Kasargod, Kayakulam and Mannanthavady, the NSE values for the forecasts of YJNSTF are marginally better than HK-SARIMA. In the case of Hosdurg and Alathur Hydro stations, forecasts of SN gave better NSE values than NSTF and YJNSTF. For Kottayam (O), forecasts from both HK-SARIMA and YJNSTF yielded similar NSE values. Overall, the NSE statistics indicate that HK-SARIMA performs better than other models (Table 6).

4.7.4. Identification of the best model using performance indicators

The time-series rainfall datasets are split into training and test data. The training dataset is used to determine the parameters of the modelling procedures, and test data is used to cross-check the ability of each modelling approach. Three error statistics, namely RMSE, MAE and NSE, were calculated to check each model's performance. The model with lower RMSE, MAE and higher NSE (probably closer to 1) is chosen as the best model. There is no classical procedure to find the best method from three error statistics. Therefore, a new classification procedure is proposed and followed to determine the best model. A specific forecasting method is considered the best when at least two out of the three statistics are in its favour (Table 6).

Only two stations are located in the Western lowlands, namely Haripad and Aluva public works department (PWD). The reliable method for Haripad is YJNSTF which has lower RMSE, MAE and higher NSE, compared with other methods. In the case of Aluva PWD, both YJNSTF and HK-SARIMA models are good since they gave similar results. The value for RMSE is marginally lower for the forecasts of HK-SARIMA, and the MAE value is lower for YJNSTF than other metrics. The forecasts of both models resulted in a similar NSE value (0.77). In Haripad and Aluva PWD, the difference in the error statistics (RMSE and MAE) between YJNSTF and HK-SARIMA is negligible (<10 mm). Therefore, both YJNSTF and HK-SARIMA methods are equally suitable for forecasting monthly rainfall in the Western lowlands (Table 6).

Central midlands comprise 12 stations and HK-SARIMA models are found to be better forecasting the monthly rainfall of eight stations. For Kasargod and Kayakulam stations in the Central lowlands, YJNSTF models provide better forecasts than HK-SARIMA (Table 6). In Hosdurg, all three methods failed to give better results compared with SN. In the remaining Kottayam (O) station, both YJNSTF and HK-SARIMA models yielded reasonably similar results, and therefore, both methods are considered equally good for carrying out rainfall forecasts (Table 6).

Out of four stations located in the Eastern highlands, YJNSTF models performed well for one station (Mannanthavady). The HK-SARIMA technique proved to be reliable for rainfall forecasting of Palakkad (O) and Nedumangad stations (Table 6). Similar to the Hosdurg station in the Central midlands, all the error statistics are less for the forecasts of the SN model for the Alathur Hydro rainfall station (Table 6).

From the above discussions, it is inferred that both HK-SARIMA and YJNSTF performed well for the stations located in the Western lowlands and Eastern highlands. In the Central highlands, HK-SARIMA outperformed YJNSTF by a significant margin. The HK-SARIMA model that gives more enhanced forecast results than NSTF and YJNSTF is attributed to the ability of the HK-SARIMA model to handle and model seasonality in a far better way than TF-based models. The TF model uses only one lag-one correlation (previous month), mean and SD of past values of the particular month. In contrast, SARIMA builds an autoregressive model which can be a function of more than one previous time step. The Ljung-Box test is applied to measure the magnitude of residuals in the HK-SARIMA model. However, the TF-based model does not apply any test to quantify the residuals after modelling the dataset.

Irrespective of the physiographical regions, the HK-SARIMA technique provides reasonably good forecasts compared to other techniques when RMSE, MAE and NSE were used as criteria. The limitations of all four techniques are that they cannot predict the exact amount of rainfall at peaks. For instance, both HK-SARIMA and YJNSTF forecasting models underestimated the maximum rainfall for the year 2013 (June 2013) in the forecasted time series (Figure 4). HK-SARIMA, in comparison with YJNSTF, was marginally better in capturing the peaks in June 2011 and 2012. Thus, graphically and statistically, it is inferred that the HK-SARIMA is the most suitable model for forecasting monthly rainfall.

5. CONCLUSION

Time-series analysis delivers a collection of scientific tools for modelling hydrological systems. So synthetically produced rainfall time series is of great prominence as the historic rainfall to review, investigate and analyse any possible replacements for planning, design and functioning of water resource projects. The significant findings and conclusions of this study are stated subsequently.

- The mode of the rainfall series of all 18 stations is zero. The frequency of the mode varies between 17 and 88; thereby, both Box-Cox and log transformation cannot be applied since both transformations are incapable of handling zero value.
- Two normality tests are applied before and after the YJT; it is inferred that the SW and AD tests display a significant increase in p-value; thereby, the transformation of the time series occurs.
- Unlike the TF model, SARIMA has inherent seasonal and first-order differencing capabilities for converting a non-stationary time series to stationary and also helps in stabilising the variance. Therefore, YJT is only applied on TF and not in SARIMA.
- The SARIMA models obtained from the HK algorithm identified that the seasonal autoregressive (P) and moving average components (Q) are more prominent than non-seasonal components (p,q). Eight out of 18 stations had their entire non-seasonal component zero (p,q,$d = 0$). Sixteen out of 18 stations have undergone one seasonal differencing to convert non-stationary series into stationary series. The first-order differencing parameter (d) is zero for all 18 stations, showing the seasonal component's prominence compared to the trend component. The obtained results are in coherence with the outcome of the statistical test (strength of seasonality and strength of trend).
- For 17 stations, the values of RMSE and MAE were less for YJNSTF than NSTF. It can be concluded that YJT is a capable methodology for converting non-normal datasets into a more Gaussian-like probability distribution, which enhances the forecasting performance of the time-series model. The result is coherent with the results presented by Pandey *et al.* (2019), where it was found that the power transformation (Box-Cox) could alleviate the variability and increase the forecast accuracy.

- Although these univariate techniques do not predict the exact amount of rainfall, they provide reasonable forecasts to enable planners to devise comprehensive water management plans for domestic water supply and agricultural water management.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Indian Meteorological Department (IMD) for the rainfall datasets.

AUTHORS CONTRIBUTION

The authors confirm their contribution to the paper as follows. P.K.: conceptualisation, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review and editing, visualisation, supervision and project administration. D.S.K.: formal analysis, investigation, writing – original draft, writing – review and editing, visualisation and supervision. N.R.C.: formal analysis, investigation, writing – original draft, writing – review and editing, visualisation and supervision.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Agboola, A. H., Gabriel, A. J., Aliyu, E. O. & Alese, B. K. 2013 Development of a fuzzy logic based rainfall prediction model. *International Journal of Engineering & Technology* **3** (4), 427–435.

Ahmad, S., Khan, I. H. & Parida, B. P. 2001 Performance of stochastic approaches for forecasting river water quality. *Water Research* **35** (18), 4261–4266.

Anderson, T. W. & Darling, D. A. 1952 Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *The Annals of Mathematical Statistics* **23** (2), 193–212.

Attah, D. A. & Bankole, G. M. 2012 Time series analysis model for annual rainfall data in lower Kaduna catchment Kaduna, Nigeria. *International Journal of Research in Chemistry and Environment (IJRCE)* **2** (1), 82–87.

Brockwell, P. J. & Davis, R. A. 2002 *Introduction to Time Series and Forecasting*. Springer New York, New York, NY.

Canova, F. & Hansen, B. E. 1995 Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business & Economic Statistics* **13** (3), 237–252.

Cecinati, F., Wani, O. & Rico-Ramirez, M. A. 2017 Comparing approaches to deal with non-Gaussianity of rainfall data in Kriging-based radar-gauge rainfall merging. *Water Resources Research* **53** (11), 8999–9018.

Chattopadhyay, S. & Franke, R. W. 2006 *Striving for Sustainability: Environmental Stress and Democratic Initiatives in Kerala*. Concept Publishing Company, New Delhi.

Clarke, R. T. 1973 *Mathematical Models in Hydrology, Irrigation and Drainage Paper*, Vol. 19. Food and Agriculture Organization of the United Nations, Rome, Italy.

Conover, W. J. 1999 *Practical Nonparametric Statistics*, 3rd edn. John Wiley & Sons, New York.

Dabral, P. P. & Murry, M. Z. 2017 Modelling and forecasting of rainfall time series using SARIMA. *Environmental Processes* **4** (2), 399–419.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39** (1), 1–22.

Firat, M., Dikbas, F., Koç, A. C. & Gungor, M. 2010 Missing data analysis and homogeneity test for Turkish precipitation series. *Sadhana* **35** (6), 707–720.

Ghaderpour, E. & Vujadinovic, T. 2020 The potential of the least-squares spectral and cross-wavelet analyses for near-real-time disturbance detection within unequally spaced satellite image time series. *Remote Sensing* **12** (15), 1–23.

Ghaderpour, E., Vujadinovic, T. & Hassan, Q. K. 2021 Application of the least-squares wavelet software in hydrology: Athabasca River Basin. *Journal of Hydrology: Regional Studies* **36** (June), 100847.

Ghanbarpour, M. R., Abbaspour, K. C. & Hipel, K. W. 2009 A comparative study in long-term river flow forecasting models. *International Journal of River Basin Management* **7** (4), 403–413.

Gilbert, R. O. 1987 *Statistical Methods for Environmental Pollution Monitoring*. John Wiley & Sons, New York.

Grimaldi, S., Koutsoyiannis, D., Piccolo, D. & Napolitano, F. 2006 Time series analysis in hydrology. *Physics and Chemistry of the Earth* **31** (18), 1097–1098.

Guhathakurta, P., Sudeepkumar, B. L., Menon, P., Prasad, A. K., Sable, S. T. & Advani, S. C. 2020 *Observed rainfall variability and changes over Kerala State*. India Meteorological Department Ministry of Earth Sciences, Pune.

Harms, A. A. & Campbell, T. H. 1967 An extension to the Thomas-Fiering model for the sequential generation of streamflow. *Water Resources Research* **3** (3), 653–661.

Helsel, D. R. & Hirsch, R. M. 1992 *Statistical Methods in Water Resources*, Vol. 49. Elsevier, Amsterdam, Netherlands.

Hu, C., Liu, C., Yao, Y., Wu, Q., Ma, B. & Jian, S. 2020 Evaluation of the impact of rainfall inputs on urban rainfall models: a systematic review. *Water (Switzerland)* **12** (9), 1–17.

Hyndman, R. J. & Athanasopoulos, G. 2018 *Forecasting: Principles and Practice*. OTexts, Melbourne.

Hyndman, R. J. & Khandakar, Y. 2008 Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* **27** (3), 1–22.

Joshi, G. S. & Gupta, K. 2009 A simulation model for the operation of multipurpose multireservoir system for River Narmada, India. *Journal of Hydro-Environment Research* **3** (2), 96–108.

Kang, H. 2013 The prevention and handling of the missing data. *Korean Journal of Anesthesiology* **64** (5), 402–406.

Kendall, M. G. 1975 *Rank Correlation Methods*, 4th edn. Charles Griffin, London.

Kurunç, A., Yürekli, K. & Çevik, O. 2005 Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilirmak River, Turkey. *Environmental Modelling & Software* **20** (9), 1195–1200.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. & Shin, Y. 1992 Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics* **54** (1–3), 159–178.

Mallikarjuna, P. & Vardhan, G. V. 2002 Stochastic modelling of monthly rainfall – a case study. *ISH Journal of Hydraulic Engineering* **8** (2), 60–72.

Mann, H. B. 1945 Non-parametric tests against trend'. *Econometrica: Journal of the Econometric Society* **13** (3), 245–259.

Martínez-Acosta, L., Medrano-Barboza, J. P., López-Ramos, Á., López, J. F. R. & López-Lambraño, Á. A. 2020 SARIMA approach to generating synthetic monthly rainfall in the Sinú river watershed in Colombia. *Atmosphere* **11** (6), 602.

Meer, L. V. D. 2019 *Spatio-Temporal Forecasts for Bike Availability in Dockless Bike Sharing Systems*. *Master Dissertation*, Institute for Geoinformatics, University of Münste.

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D. & Veith, T. L. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50** (3), 885–900.

Murthy, K. V. N., Saravana, R. & Vijaya Kumar, K. 2018 Modeling and forecasting rainfall patterns of southwest monsoons in North-East India as a SARIMA process. *Meteorology and Atmospheric Physics* **130** (1), 99–106.

Narayanan, P., Basistha, A., Sarkar, S. & Kamna, S. 2013 Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. *Comptes Rendus – Geoscience* **345** (1), 22–27.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10** (3), 282–290.

NIST/SEMATECH. 2022 *e-Handbook of Statistical Methods*. Available from: https://www.itl.nist.gov/div898/handbook/index.htm. https://doi.org/10.18434/M32189

Pandey, P. K., Tripura, H. & Pandey, V. 2019 Improving prediction accuracy of rainfall time series by hybrid SARIMA–GARCH modeling. *Natural Resources Research* **28** (3), 1125–1138.

Ray, S., Das, S. S., Mishra, P. & Al Khatib, A. M. G. 2021 Time series SARIMA modelling and forecasting of monthly rainfall and temperature in the South Asian Countries. *Earth Systems and Environment* **5** (3), 531–546.

Saha, A., Singh, K. N., Ray, M. & Rathod, S. 2020 A hybrid spatio-temporal modelling: an application to space-time rainfall forecasting. *Theoretical and Applied Climatology* **142** (3), 1271–1282.

Sahoo, B. B., Jha, R., Singh, A. & Kumar, D. 2019 Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica* **67** (5), 1471–1481.

Schepen, A., Wang, Q. J. & Robertson, D. E. 2012 Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *Journal of Geophysical Research-Atmospheres* **117** (20), D20107.

Sen, P. K. 1968 Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association* **63** (324), 1379–1389.

Shapiro, S. S. & Wilk, M. B. 1965 An analysis of variance test for normality (complete samples). *Biometrika* **52** (3/4), 591–611.

Sharma, P., Bhakar, S. R., Ali, S., Jain, H. K., Singh, P. K. & Kothari, M. 2018 Generation of synthetic streamflow of Jakham River, Rajasthan using Thomas-Fiering model. *Journal of Agricultural Engineering* **55** (4), 47–56.

Simon, A. & Mohankumar, K. 2004 Spatial variability and rainfall characteristics of Kerala. *Journal of Earth System Science* **113** (2), 211–221.

Stedinger, J. R. & Taylor, M. R. 1982 Synthetic streamflow generation: 1. Model verification and validation. *Water Resources Research* **18** (4), 909–918.

Teymouri, M. & Fathzadeh, A. 2015 Stochastic modeling of monthly river flow forecasting (Case study: Atrak River Basin, Iran). *Journal of Selçuk University Natural and Applied Science* **4** (2), 38–48.

Theil, H. 1950 A rank-invariant method of linear and polynomial regression analysis (Parts 1–3). *Indagationes Mathematicae* **12** (85), 173.

Thomas, H. A. & Fiering, M. B. 1962 Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In: *Design of Water Resource Systems* (A. Maass, M. M. Hufschmidt, R. Dorfman, H. A. Thomas, Jr., S. A. Marglin & G. Maskew Fair, eds.). Harvard University Press, Cambridge, MA, pp. 459–493.

Ukkola, A. M., Roderick, M. L., Barker, A. & Pitman, A. J. 2019 Exploring the stationarity of Australian temperature, precipitation and pan evaporation records over the last century. *Environmental Research Letters* **14** (12), 124035.

Ünal, N. E., Aksoy, H. & Akar, T. 2004 Annual and monthly rainfall data generation schemes. *Stochastic Environmental Research and Risk Assessment* **18** (4), 245–257.

Unnikrishnan, P. & Jothiprakash, V. 2018 Daily rainfall forecasting for one year in a single run using singular spectrum analysis. *Journal of Hydrology* **561**, 609–621.

Wang, X., Smith, K. & Hyndman, R. 2006 Characteristic-based clustering for time series data'. *Data Mining and Knowledge Discovery* **13** (3), 335–364.

Wang, S., Feng, J. & Liu, G. 2013 Application of seasonal time series model in the precipitation forecast. *Mathematical and Computer Modelling* **58** (3–4), 677–683.

Willmott, C. J. & Matsuura, K. 2005 Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30** (1), 79–82.

Yeo, I. K. & Johnson, R. A. 2000 A new family of power transformations to improve normality or symmetry. *Biometrika* **87** (4), 954–959.

Yousif, A. A., Aswad, F. K. & Ibrahim, S. A. 2016 Performance of ARIMA model and modified Thomas-Fiering model for predicting the monthly rainfall data for Tallafar Station. *Journal of Polytechnic* **6** (1), 293–309.

Yurekli, K. & Kurunc, A. 2006 Performances of stochastic approaches in generating low streamflow data for drought analysis. *Journal of Spatial Hydrology* **5** (1), 20–32.

Zeynoddin, M. & Bonakdari, H. 2019 Investigating methods in data preparation for stochastic rainfall modeling: a case study for Kermanshah synoptic station rainfall data, Iran. *Journal of Applied Research in Water and Wastewater* **6** (1), 32–38.

First received 14 March 2022; accepted in revised form 13 October 2022. Available online 25 October 2022