Acta Geophysica (2020) 68:1161–1174
https://doi.org/10.1007/s11600-020-00462-9

RESEARCH ARTICLE - HYDROLOGY

Trend analysis and SARIMA forecasting of mean maximum and mean minimum monthly temperature for the state of Kerala, India

P. Kabbilawsh[1] · D. Sathish Kumar[1] · N. R. Chithra[1]

Received: 5 February 2020 / Accepted: 3 July 2020 / Published online: 16 July 2020

Abstract

The development of temperature forecasting models for the state of Kerala using Seasonal Autoregressive Integrated Moving Average (SARIMA) method is presented in this article. Mean maximum and mean minimum monthly temperature data, for a period of 47 years, from seven stations, are studied and applied to develop the model. It is expected that the time-series datasets of temperature to display seasonality (and hence non-stationary), and a possible trend (due to the fact that the data spans 5 decades). Hence, the key step in the development of the models is the determination of the non-stationarity of the temperature time-series, and the transformation of the non-stationary time-series into a stationary time-series. This is carried out using the Seasonal and Trend decomposition using Loess technique and Kwiatkowski–Phillips–Schmidt–Shin test. Before carrying out this process, several preliminary tests are conducted for (1) finding and filling the missing values, (2) studying the characteristics of the data, and (3) investigating the presence of the trend and seasonality. The non-stationary temperature time-series are transformed to stationary temperature time-series, by one seasonal differencing and one first-order differencing. This information, along with the original time-series, is further utilized to develop the models using the SARIMA method. The parsimonious and best-fit SARIMA models are developed for each of the fourteen variables. The study revealed that SARIMA$(2, 1, 1)(1, 1, 1)_{12}$ as the ideal forecasting model for eight out of the fourteen time-series datasets.

Keywords Autocorrelation function (ACF) · Partial autocorrelation function (PACF) · Sen's slope estimator · Seasonal autoregressive integrated moving average (SARIMA) · Mann–Kendall (MK) trend test

Introduction

India, with a population of more than 1.3 billion, has more than 50% of its population dependent on agriculture (Arjun 2013). Most states in India still heavily rely on rainfall for various agricultural activities. It is well known that rainfall, a part of the hydrological cycle, is susceptible to changes in global temperature (Allen and Ingram 2002; Andronova and Schlesinger 2000; Trenberth 1999). Hence, an exclusive look into the long-term temperature variations would constitute a vital part in the analysis of agricultural output of any region of the country.

In this regard, many researchers have carried out studies in the last decade on global, continental and regional level long-term temperature variations (Hänsel et al. 2016; Jain and Kumar 2012; Kocsis et al. 2017). Also, many attempts have been undertaken by researchers to develop models for understanding and extrapolating the temperature variation (Hänsel et al. 2016; Mills 2014; Tiwari et al. 2016). In India, among all the studies focused on temporal temperature variation, the most noteworthy study is the one conducted by the Indian Network for Climate Change Assessment (INCCA) (2010). The projections of mean annual surface temperature for the 2030s (average of 2021–2050) were carried out on country level using PRECIS (Providing Regional Climates for Impact Studies), with the data obtained from 1970s (average of 1960–1990). In this study, it was predicted that the annual mean surface air temperature would rise by 1.7–2 °C over the entire Indian subcontinent. Though this study indicates that significant

✉ P. Kabbilawsh
kabbi.civil@gmail.com

D. Sathish Kumar
sathish@nitc.ac.in

N. R. Chithra
chithranr@nitc.ac.in

[1] Department of Civil Engineering, NIT Calicut, Calicut 673601, India

changes could be expected in the overall characteristics of rainfall, the projections are at a macroscopic level (i.e. for the entire Indian subcontinent), and not for each individual states. Regional studies focussing on individual states are necessary to get a better understanding of the local factors that influence these variations. A state-wise study is important because local policies and actions can be exclusively implemented by the state governments to combat any expected adverse changes in their respective states. In this study, the temporal variation of the monthly mean maximum (MMAX) and mean minimum temperature (MMIN) is analysed for the state of Kerala.

The analysis is carried out for a period of 47 years, starting from 1969 to 2015. The overall objective of the study is to develop a model for future forecasts of MMAX and MMIN for the state of Kerala. Prior to the time-series modelling, it is necessary to carry out preprocessing of the data to identify the missing values. The time-series data available for each station and the number of missing values are listed in Table 1. The data gaps are to be eliminated before any time-series modelling. The data infilling process is carried out using expectation–maximization algorithm. Further, for the construction of a suitable forecasting model, it is necessary to evaluate the time-series datasets to understand the existing pattern. This preliminary analysis provides a good insight regarding the available data. It comprises of (1) a descriptive statistical analysis of the monthly data, (2) performance of the normality test, (3) test to check for outliers, (4) Mann–Kendall trend analysis and (5) performance of the Sen's slope test. The results obtained from the preliminary analysis revealed the presence of non-stationarity in the datasets. In order to confirm the preliminary results obtained, the time-series datasets are decomposed using STL decomposition to get the time-series components. The obtained time-series components also revealed the presence of seasonality and the presence or absence of a trend. The value of parameters (seasonal and non-seasonal differencing, D and d, respectively) needed for converting the non-stationary time-series to a stationary series is obtained using the results of Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. These values, along with the original time-series datasets, are used for the SARIMA model building process. The next section ("Temperature data and research methodology" section) briefly describes each of the process (the preliminary tests, STL decomposition, Unit root test and SARIMA) applied in this study. Section "Temperature data and research methodology" describes the application of these tests to our data. Also, in "Result and discussions" section, the result of each test is analyzed and elaborated, and a final forecast is also delivered with the developed model. Lastly, section "Summary and conclusions" concludes the article with an overview of the entire study.

Temperature data and research methodology

The main reason for carrying out the temperature-related studies for the state of Kerala is that the state is the gateway of the summer monsoon (South-West) for India. Any disturbance to the South-West monsoon creates a cascading effect on the rainfall patterns in the entire country. As stated earlier, this section presents the techniques applied for (1) estimating the missing values, (2) conducting the preliminary analysis, (3) decomposing the time-series data, (4) converting the non-stationary data to stationary data, and (5) developing the model.

An account of Kerala and its temperature dataset

The state Kerala is a small strip of coastal land located in the southern part of India. It consists of an area of 38,850 km^2. It is located between 8° 18′ N–12° 48′ N latitudes and 74° 52′ E–77° 24′ E longitudes. Figure 1 shows the location map of the study area. The state has a shoreline of 580 km, and the width of the state varies between 30 and 120 km. Geologically, the state Kerala can be categorised into three climatically distinct regions: the eastern highlands (rugged and cool mountainous terrain), the central midlands (rolling hills), and the western lowlands (coastal plains). The lowlands and highlands bound the state of Kerala, where the lowlands comprise the regions which adjoin the shoreline, and highlands cover the region slopping down from the Western Ghats. The midlands spread between the highlands and lowlands. Area-wise, the highlands comprise of 18,650 km^2, while the midlands and lowlands comprise of 16,200 km^2 and 4000 km^2 respectively. Tea, coffee and rubber are major plantation crops grown in the highlands. It also houses several endemic flora and fauna. Wide variety of fruits, nuts and vegetables are grown in the midland region. Paddy and coconut are grown in the fertile lowlands.

The temperature datasets from 13 observatories that cover the entire state of Kerala are obtained from the Indian Meteorological Department (IMD). The observatories spread across all the three climatic regions. The observatories are located at Palghat, Fort Cochin, Kovalam, Karipur, Trichur, Ernakulum, Kozhikode, Kannur, Alappuzha, Punalur, Kottayam, Thiruvananthapuram and Trivandrum Airport. Out of 13 stations, three observatories (Palghat, Fort Cochin and Kovalam) are not properly functioning for the past 15 years, and the datasets for recent years are not available. For three observatories (Karipur, Trichur and Ernakulum), the datasets are available starting from the year 1996 and

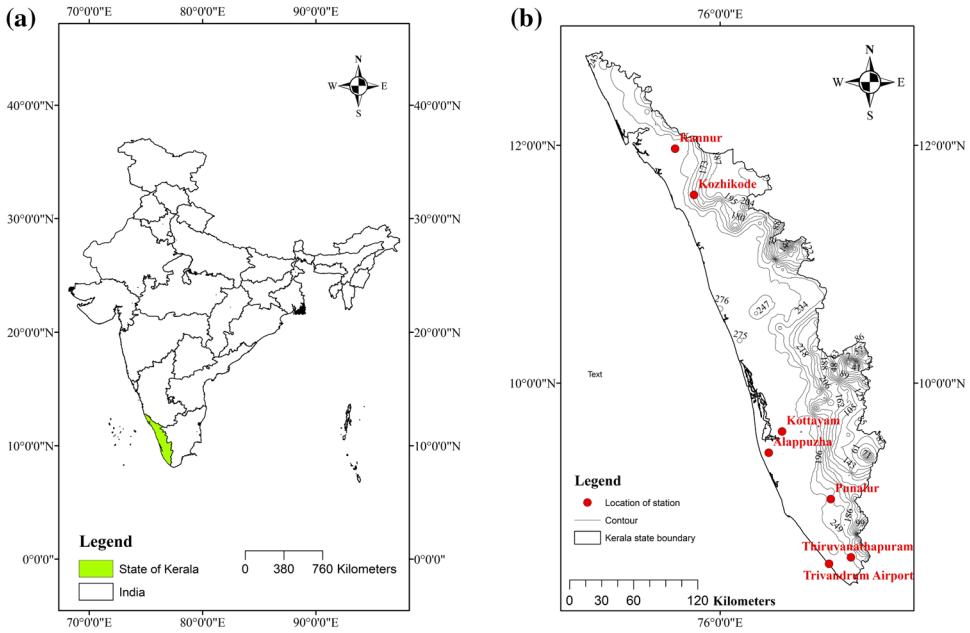

Fig. 1 **a** Location map of the state Kerala, **b** the location map of the seven stations for which study is conducted

Table 1 The amount of missing data present in the meteorological observatories

Station name	Starting year of time series	Ending year of time series	Total length of data (years)	MMAX			MMIN		
				Number of monthly values present	Number of monthly values missing	% of missing values	Number of monthly values present	Number of monthly values missing	% of missing values
Kozhikode	1969	2015	47	564	0	0	564	0	0
Kannur	1981	2015	34	404	16	3.81	404	16	3.81
Alappuzha	1969	2015	47	549	15	2.66	548	16	2.84
Punalur	1969	2015	47	532	32	5.67	500	64	11.34
Kottayam	1973	2011	43	498	18	3.49	496	20	3.88
Thiruvananthapuram	1969	2015	47	564	0	0	564	0	0
Trivandrum Airport	1969	2015	47	545	19	3.37	545	19	3.37

later. Adequate datasets for the analysis are available only for seven meteorological stations. The spatial locations of the observatories are shown in Fig. 1b.

The data from the rest of the seven stations are found ideal for the study. Table 1 shows the data availability for the selected seven stations. A total of 235 intermittent monthly values (about 3.13% of the data) are found to be missing in the available dataset. Table 1 lists the number and types of missing values for each station. Datasets with missing values present several problems in the representativeness of the samples (Kang 2013). Hence, the missing values are to be determined first. For this purpose, the expectation–maximization algorithm is used. The missing values estimated through this method is used to fill the data gaps in order to obtain continuous time-series datasets. Preliminary statistical tests are conducted using these datasets. The results indicated the presence of skewness and kurtosis. Further, a test for normality is carried out using the Shapiro-Wilk normality test and the outliers are identified using Grubb's test. The results indicated that datasets followed a non-normal distribution without any outliers. Therefore, a nonparametric Mann–Kendall trend test (Gocic and Trajkovic 2013; Kocsis et al. 2020) and Sen's slope test are used to determine the direction and magnitude of monotonic trends in the time-series.

Mann–Kendall trend test

The Mann–Kendall trend test (Mann 1945; Kendall 1975; Gilbert 1987) is widely used test in the field of

Hydro-meteorology, dealing with variables like temperature, rainfall and streamflow. The Mann–Kendall test is used to statistically assess the presence of an increasing or decreasing trend in the series. The Mann–Kendall test operates by checking whether to reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1). The null hypothesis (H_0) means there is no trend in the temperature over time, and the alternative hypothesis (H_1) implies the presence of either an increasing or decreasing trend in the temperature data. The sign of the computed Mann–Kendall test statistic Z_{MK} reveals the direction of the trend. The positive value of Z_{MK} indicates that the temperature tends to increase with time, while the negative value of Z_{MK} denotes the decrease in temperature over time. The null Hypothesis (H_0) is rejected, and the alternative hypothesis (H_1) is accepted if $|Z_{MK}| \geq Z_{1-\alpha/2}$ at the Type I error rate α.

Sen's slope estimation

All the available statistical techniques may not be equally good in detecting the magnitude of the trend in the time-series data (Radziejewski and Kundzewicz 2004). A simple parametric least-square regression technique is not suitable to calculate the magnitude of the trend for non-normal time-series. In such cases, a test which is nonparametric, robust against outliers would be an appropriate choice. Sen's Slope estimation, a nonparametric test is selected to detect the magnitude of trends in the temperature time-series. It is impartially resistant to outliers, with a breakdown point of 0.29 (Sen 1968). It was initially proposed in 1968 to account for the non-normality of precipitation data. A mathematical explanation of the scheme is not detailed here, as it has been already presented in detail by various authors (Gocic and Trajkovic 2013; Kocsis et al. 2020).

STL decomposition

The Mann–Kendall trend test and Sen's slope estimation are carried out as a part of the initial investigation. Further to provide a better understanding of the datasets, the time-series data is decomposed as the trend component, the seasonal component and the remainder component. It is carried out using STL (Seasonal and Trend decomposition using Loess) decomposition method (Cleveland et al. 1990). The decomposed components are plotted for graphical visualisation of the data. It allows us to visualise the presence of trend and seasonality in the data. Compared to the other classical decomposition methods, STL has several advantages like the ability to handle any type of seasonality (daily, monthly, quarterly, annual, etc.), being robust to outliers, facilitating the user to control the smoothness of trend cycle, and allowing the user to control the rate of change of seasonal component.

Unit root test

The trend and seasonal components obtained from the STL decomposition will reveal the presence of non-stationarity in the temperature time-series. The non-stationarity is only inferred from the graphs of the decomposed components (only visual inference). To mathematically confirm the presence of non-stationarity in the time-series, the unit root tests are performed. In the present study, KPSS (Kwiatkowski–Phillips–Schmid–Shin) unit root test is performed to confirm the presence of non-stationarity (Kwiatkowski et al. 1992). The original temperature time-series and the decomposed components are used for this purpose. The KPSS method proceeds with the null hypothesis (i.e. the data are stationary) and tries to find evidence to show that the null hypothesis is false for the selected time-series. If the non-stationarity is confirmed, then the next step is the conversion of the non-stationary data to stationary data. The p values determined from the KPSS test provides information about the differencing; small p values typically, less than 0.05 points the necessity of differencing for the conversion of the time-series.

Seasonal autoregressive integrated moving average (SARIMA) model

After the non-stationary time-series is converted to a stationary time-series (i.e. after the determination of d and D), the next step is to develop a model for future predictions. Forecasting models developed from the historical records are generally used to predict the future changes in the climate variables. Several authors have proposed temperature models using a number of forecasting techniques (Aguado-Rodríguez et al. 2016; Tiwari et al. 2016; Wang et al. 2019; Lai and Dzombak 2020; Wanishsakpong and Owusu 2020). Several climate variables are generally influenced by seasonality, and one of the best forecasting models for such variables is the SARIMA model. It combines the advantage of the autoregressive model and the moving average model.

In an autoregressive model, a linear combination of the past values of the variable is used to predict the future of the variable. Mathematically, Eq. 1 represents an autoregressive model of order p, i.e. $AR(p)$ model.

$$y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + \varepsilon_t \quad (1)$$

The equation shows that the observation y at time t (y_t) is estimated from p previous observations (y_{t-i}, $i = 1, 2, 3, \ldots, p$). θ_k, with $k = 1, 2, 3, \ldots, p$ are the parameters, and ε_t is the white noise.

In a moving average model, the forecast is done using the past forecast errors in a regression-like model.

$$y_t = \phi_0 + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \cdots + \phi_q\varepsilon_{t-q} \quad (2)$$

Equation 2 describes the moving average model of order q, i.e. $MA(q)$ model, where y_t is the observation at time t; ϕ_i, with $i = 1, 2, 3, \ldots, q$, are the parameters and ε_{t-k}, with $k = 1, 2, 3, \ldots, q$ are the error terms, respectively.

By combining differencing with autoregression and a moving average model, the non-seasonal Autoregressive Integrated Moving Average (ARIMA) model is obtained. Mathematically, the ARIMA model is represented by Eq. 3.

$$y'_t = \theta_0 + \theta_1 y'_{t-1} + \theta_2 y'_{t-2} + \cdots + \theta_p y'_{t-p} + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \cdots + \phi_q\varepsilon_{t-q} + \varepsilon_t \quad (3)$$

y'_t denotes the differenced series. It has to be noted that the series may have been differenced more than once, and the degree of the differencing involved is denoted by d. This series is represented as the $ARIMA(p, d, q)$ model, where p denotes the order of autoregressive part, d denotes the degree of differencing, and q denotes the moving average part.

If seasonality is observed in a time-series, then a seasonal-*ARIMA* model or SARIMA model (Hyndman and Athanasopoulos 2018) has to be applied. The seasonal-*ARIMA* model is obtained by including the additional seasonal terms to the *ARIMA* models. The seasonal-*ARIMA* model is represented as $\text{SARIMA}(p,d,q)(P,D,Q)_m$. The non-seasonal part of the model is represented as (p, d, q), and the seasonal part of the model is given by $(P,D,Q)_m$. The terms P, D and Q represents the order of the seasonal autoregressive term, degree of the seasonal differencing and order of the seasonal moving average part, respectively. The term m represents the number of observations per year. The terms of the seasonal part of the model are similar to the non-seasonal part expect that they involve backshifts of the seasonal period.

Result and discussions

All the tests discussed in the previous section are applied parallelly or sequentially based on the requirements. The results of these tests, and their significance, are discussed in this section.

Results from descriptive statistics

Descriptive statistics of the temperature datasets are obtained after filling the missing values using the expectation–maximization algorithm. The analysis is carried out for seven stations for two variables (MMAX and MMIN) in each station. Therefore, altogether fourteen time-series datasets are analysed. The average MMAX and MMIN temperature covering all the seven stations are 31.84 °C and 23.48 °C respectively. The MMAX varies between 31.11 °C (at Trivandrum Airport), and 33.06 °C (at Punalur), and MMIN varies between 22.34 °C (at Punalur) and 24.22 °C (at Kozhikode). The standard error of the mean of all fourteen variables ranges between 0.04 °C and 0.1 °C. The maximum deviation of the sample-mean from the population-mean is 0.2 °C, at a confidence level of 95%. The difference between the sample-mean and the population-mean is negligible. Therefore, it can be concluded that a sample mean is a genuine representation of the population mean. The descriptive statistics of the variables are listed in Table 2. In the tabulation, SEM,

Table 2 Descriptive statistics of the variables

Station name	Variable type	Mean	SEM	SD	Variance	CV	Q_1	Q_3	Range	IQR	Skewness	Excess kurtosis
Kozhikode	MMAX	31.50	0.08	1.79	3.22	5.69	30.20	32.80	8.40	2.60	– 0.23	– 0.64
Kozhikode	MMIN	24.22	0.05	1.24	1.54	5.12	23.50	24.80	8.10	1.30	0.22	0.29
Kannur	MMAX	32.13	0.10	2.04	4.16	6.35	30.50	33.70	8.90	3.20	– 0.14	– 0.89
Kannur	MMIN	23.47	0.07	1.33	1.78	5.68	22.70	24.20	7.20	1.50	0.12	– 0.12
Alappuzha	MMAX	31.48	0.07	1.64	2.70	5.22	30.10	32.80	7.70	2.70	– 0.25	– 0.90
Alappuzha	MMIN	23.92	0.05	1.17	1.36	4.88	23.20	24.60	6.50	1.40	0.07	0.03
Punalur	MMAX	33.06	0.09	2.14	4.58	6.47	31.40	34.70	10.30	3.30	0.36	– 0.57
Punalur	MMIN	22.34	0.05	1.19	1.42	5.33	21.70	23.10	7.10	1.40	– 0.25	0.14
Kottayam	MMAX	32.03	0.08	1.76	3.11	5.51	30.70	33.40	9.50	2.70	– 0.09	– 0.65
Kottayam	MMIN	23.08	0.04	0.98	0.95	4.23	22.70	23.70	6.20	1.00	– 0.91	1.52
Thiruvananthapuram	MMAX	31.56	0.06	1.38	1.90	4.37	30.50	32.68	6.20	2.18	0.04	– 0.83
Thiruvananthapuram	MMIN	23.56	0.04	0.97	0.94	4.12	23.00	24.10	5.90	1.10	0.24	0.13
Trivandrum Airport	MMAX	31.11	0.05	1.18	1.39	3.79	30.20	31.90	5.90	1.70	0.21	– 0.66
Trivandrum Airport	MMIN	23.76	0.05	1.10	1.20	4.61	23.20	24.32	6.60	1.12	– 0.23	0.60

SD, CV, *Q* and IQR stand for standard error of the mean, standard deviation, coefficient of variation, quartile, and inter-quartile range, respectively.

The MMIN variable at Kottayam has a skewness of −0.91, and since this value is within the range of −0.5 and −1, it implies that data is moderately skewed. Moreover, the excess kurtosis values of all fourteen variables are nonzero, which implies that all fourteen temperature time-series are non-mesokurtic. Although the excess kurtosis values (nonzero) indicate the non-normal distribution of all the variables, it has to be noted that the values are small. Therefore, it necessitates a dedicated normality test. Consequently, a Shapiro-Wilk test is conducted to validate the nature of the distribution.

Test for normality

The test for normality indicated that all fourteen variables are indeed non-normally distributed. The results of the Shapiro-Wilk test are presented in Table 3.

Additionally, the Grubb's test is also conducted to determine the presence of outliers in the data. The results of the Grubb's test are presented in Table 4. The G-statistic values of all fourteen variables are found to be less than their corresponding critical values indicating that there are no outliers in any of the fourteen temperature datasets.

Table 3 The results of the test for normality of the variables

Station name	Variable type	Degrees of freedom	Shapiro–Wilk		
			Statistic	*p* value	Decision at level (5%)
Kozhikode	MMAX	564	0.983	3E−06	Reject normality
Kozhikode	MMIN	564	0.977	9E−08	Reject normality
Kannur	MMAX	420	0.978	6E−06	Reject normality
Kannur	MMIN	420	0.987	1E−03	Reject normality
Alappuzha	MMAX	564	0.972	6E−09	Reject normality
Alappuzha	MMIN	564	0.991	2E−03	Reject normality
Punalur	MMAX	564	0.979	3E−07	Reject normality
Punalur	MMIN	564	0.989	4E−04	Reject normality
Kottayam	MMAX	516	0.988	2E−04	Reject normality
Kottayam	MMIN	516	0.953	1E−11	Reject normality
Thiruvananthapuram	MMAX	564	0.984	7E−06	Reject normality
Thiruvananthapuram	MMIN	564	0.983	5E−06	Reject normality
Trivandrum Airport	MMAX	564	0.983	4E−06	Reject normality
Trivandrum Airport	MMIN	564	0.981	1E−06	Reject normality

Table 4 The results obtained from the Grubb's test for the variables

Station name	Variable type	G-statistic	Critical value	Approximate *p* value (%)	Decision
Kozhikode	MMAX	2.4	3.9	9.18	No outliers
Kozhikode	MMIN	3.49	3.9	0.26	No outliers
Kannur	MMAX	2.22	3.82	10.87	No outliers
Kannur	MMIN	2.95	3.82	1.3	No outliers
Alappuzha	MMAX	2.61	3.9	5.05	No outliers
Alappuzha	MMIN	2.93	3.9	1.85	No outliers
Punalur	MMAX	2.82	3.9	2.61	No outliers
Punalur	MMIN	1.83	2.89	1.78	No outliers
Kottayam	MMAX	2.97	3.87	1.49	No outliers
Kottayam	MMIN	3.78	3.87	0.01	No outliers
Thiruvananthapuram	MMAX	2.35	3.9	10.4	No outliers
Thiruvananthapuram	MMIN	3.26	3.9	0.6	No outliers
Trivandrum Airport	MMAX	2.88	3.9	2.22	No outliers
Trivandrum Airport	MMIN	3.34	3.9	0.45	No outliers

Trend analysis using Mann–Kendall test and Sen's slope estimation

Since the datasets are not normally distributed, the Mann–Kendall test is applied to check the presence or absence of the trend in the datasets. The results of the trend analysis are presented in Table 5. As mentioned earlier, the results of the Mann–Kendall indicate only the presence or the absence of a trend in the series and its direction. However, it fails to quantify the magnitude of the trend.

In this test, a p value greater than α (i.e. 0.05), indicates the absence of the trend. The sign of MK-statistic indicates the direction of the trend. The test results indicate that a certain amount of trend is present in ten variables. The magnitude of trend determined using Sen's slope estimation is presented in Table 5. The β-slope represents the magnitude of the trend. This is consistent with the findings (p value) from the Mann–Kendall's test. The four stations that indicated the absence of a trend in the Mann–Kendall test resulted in very low values of β-slope. It may be noted that for non-stationary series with small slopes (< 0.0002), even at $p < 0.01$, Mann–Kendall trend test rejects null-hypothesis, resulting in Type-I error. The other specific inferences that could be made from this test is that the MMAX series of Kozhikode, Kannur and Thiruvanathapuram has a significant trend (β-slope exceeding 0.2%), and MMIN series of Alapuzzha station is the only one with a decreasing trend, confirming the result obtained from the Mann–Kendall test.

Analysis through STL decomposition

Though the previous two tests reveal the presence or absence of a trend, the presence of non-stationarity resulting from seasonality cannot be directly inferred from them. A time-series decomposition technique is applied to obtain the components. The fourteen time-series variables are decomposed to get each one's trend, seasonal and remainder components using the STL decomposition. Figure 2 shows the original data, trend, seasonal, and remainder components of the MMAX variable of Kozhikode station.

For STL decomposition, only the first 80% of the time-series datasets are utilized. The remaining 20% of the data is retained for the validation of the forecasting model. The increasing trend, which was predicted by both the Mann–Kendall test and Sen's slope estimation, can be visualised from the figure. It also indicates the existence of a strong seasonal pattern. Similar to Kozhikode MMAX variable, the other thirteen variables also exhibited seasonal patterns. On this basis, it is possible to conclude that the datasets are non-stationary. Before developing a forecasting model, the non-stationary datasets must be converted to stationary datasets, and subsequently the parameters d and D must be determined. The non-stationarity of the datasets are validated by applying the unit root test.

Table 5 The results of Mann–Kendall trend test ($\alpha = 0.05$) and Sen's slope test ($\alpha = 0.05$)

Station name	Variable type	MK S-statistic	Standard error	z statistic	p value	Presence of trend	Sen's slope	Sen's-slope (lower 95 % Confidence Interval)	Sen's-slope (Upper 95 % Confidence Interval)
Kozhikode	MMAX	35,826	4469.78	8.02	0	Yes	0.0038	0.0029	0.0046
Kozhikode	MMIN	26,011	4467.88	5.82	0	Yes	0.0015	0.001	0.0021
Kannur	MMAX	16,583	2873.8	5.77	0	Yes	0.0048	0.0032	0.0063
Kannur	MMIN	9118	2873	3.17	0	Yes	0.0016	0.0006	0.0025
Alappuzha	MMAX	9169	4469.78	2.05	0.04	Yes	0.0009	0	0.0017
Alappuzha	MMIN	– 14,724	4468.65	– 3.3	0	Yes	– 0.001	– 0.0016	– 0.0004
Punalur	MMAX	2670	4470.12	0.6	0.55	No	0.0003	– 0.0008	0.0015
Punalur	MMIN	5220	4468.91	1.17	0.24	No	0.0003	– 0.0001	0.001
Kottayam	MMAX	5403	3912.08	1.38	0.17	No	0.0007	– 0.0003	0.0018
Kottayam	MMIN	– 4099	3909.71	– 1.05	0.3	No	0	– 0.0008	0
Thiruvananthapuram	MMAX	33,767	4469.45	7.56	0	Yes	0.0028	0.0021	0.0035
Thiruvananthapuram	MMIN	20,561	4467.24	4.6	0	Yes	0.001	0.0006	0.0015
Trivandrum airport	MMAX	18,572	4469.01	4.16	0	Yes	0.0013	0.0007	0.0019
Trivandrum airport	MMIN	21,046	4467.87	4.71	0	Yes	0.0011	0.0006	0.0016

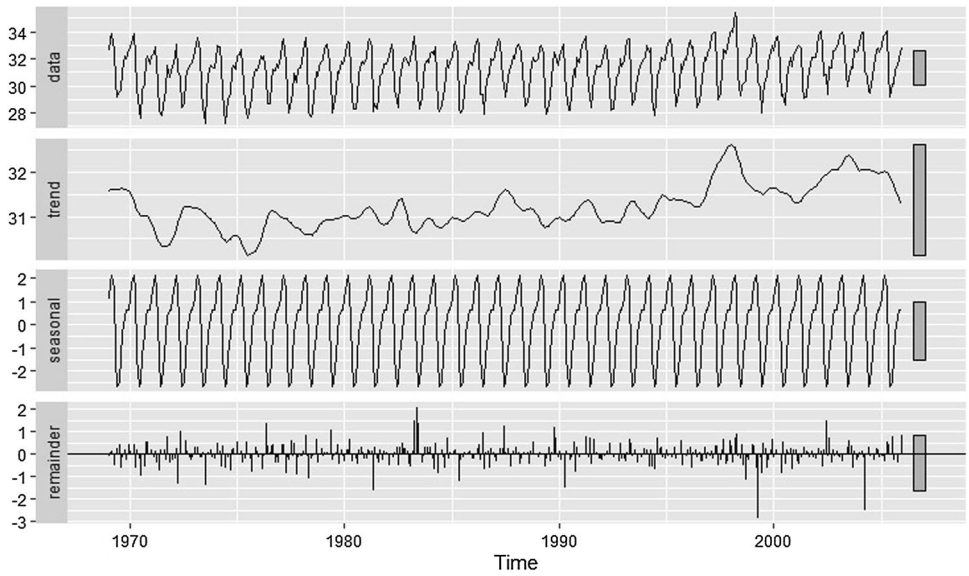

Fig. 2 The decomposed components of the Kozhikode MMAX time-series

Unit root test and the conversion to a stationary series

The Kwiatkowski–Phillips–Schmid–Shin (KPSS) test is applied to categorise the datasets as stationary or non-stationary. The results of the unit root test are presented in Table 6. The test results for the original temperature time-series are listed in the third column of the table. A p value of less than 0.05 implies that the series is non-stationary. The results indicate that time series datasets of Alappuzha, Punalur and Kottayam corresponding to MMAX variable, and MMIN variable of Punalur and Kottayam are stationary. This is contrary to what was inferred from the STL seasonal plots. To resolve this paradox, the autocorrelation (ACF) plots and partial autocorrelation (PACF) plots of those five variables are analyzed. The ACF and PACF plots of the Punalur MMAX are shown in Fig. 3. The ACF and PACF values in the plots follow a decaying sinusoidal pattern, which indicates seasonality and the dataset is non-stationary.

Similar trends are observed in the ACF and PACF plots for the other four variables. Therefore, it is conclusive that all fourteen variables are indeed non-stationary. As the seasonality is confirmed, at least one seasonal differencing is necessary to convert the non-stationary time-series to a stationary time-series. As this technique ascertains non-stationarity by means of seasonality, there are possibilities that the non-stationarity may exist exclusive of the seasonal component. In other words, it is possible that there could

Table 6 The result of KPSS test for level stationarity

Station name	Variable type	Original series (p value)	Seasonally adjusted series (p value)	Series after seasonal adjustment and first-order difference (p value)
Kozhikode	MMAX	0.01	0.01	0.1
Kozhikode	MMIN	0.01	0.01	0.1
Kannur	MMAX	0.01	0.01	0.1
Kannur	MMIN	0.049	0.01	0.1
Alappuzha	MMAX	0.1	0.01	0.1
Alappuzha	MMIN	0.01	0.01	0.1
Punalur	MMAX	0.1	0.1	0.1
Punalur	MMIN	0.1	0.051	0.1
Kottayam	MMAX	0.1	0.01	0.1
Kottayam	MMIN	0.06	0.01	0.1
Thiruvananthapuram	MMAX	0.01	0.01	0.1
Thiruvananthapuram	MMIN	0.01	0.01	0.1
Trivandrum Airport	MMAX	0.01	0.01	0.1
Trivandrum Airport	MMIN	0.01	0.01	0.1

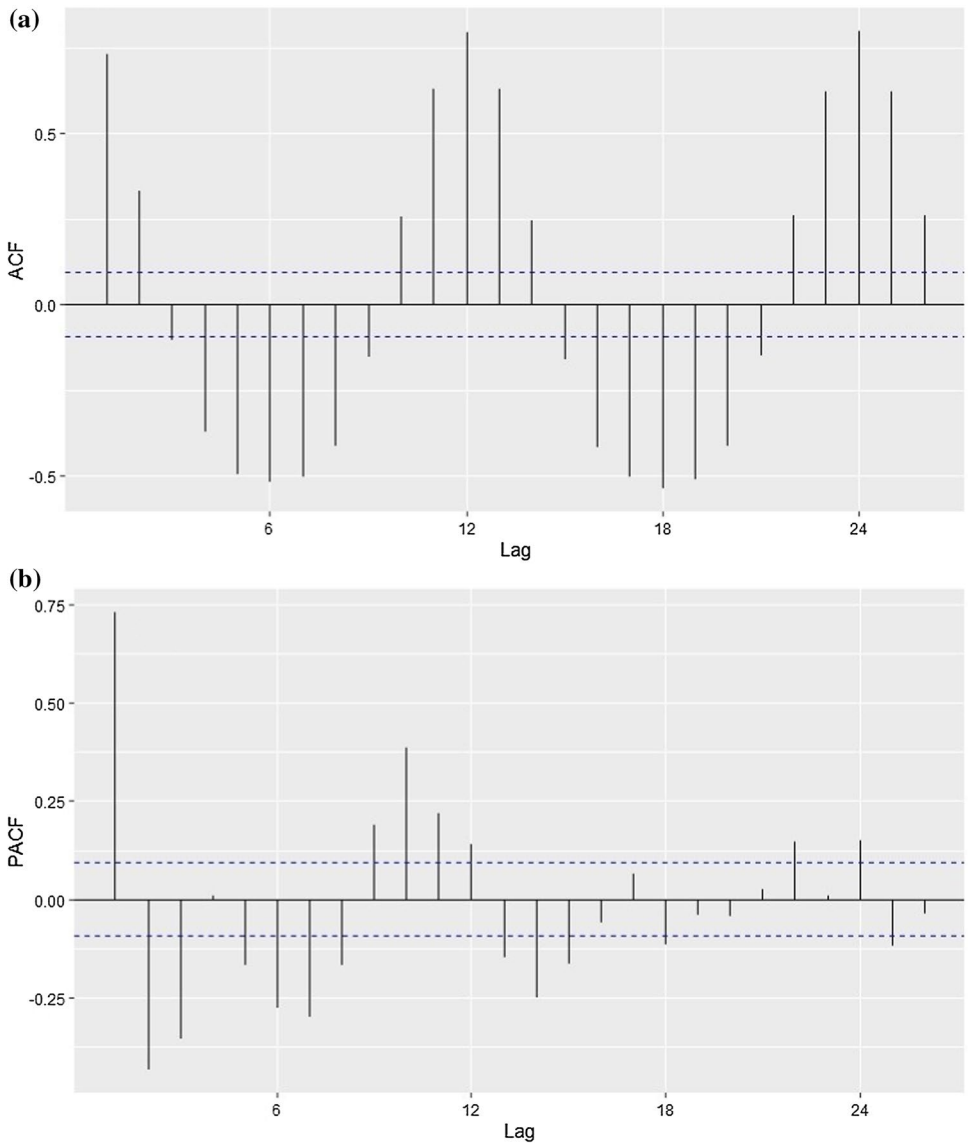

Fig. 3 **a** ACF plot of MMAX variable at Punalur station, **b** PACF plot of MMAX variable at Punalur station

be a non-stationarity in the non-seasonal component of the time-series. In order to determine this possibility, the KPSS test is conducted on the seasonally adjusted time-series.

The seasonally adjusted series is obtained by subtracting the seasonal component from the original time-series datasets (seasonal differencing). The test results for the seasonally adjusted time-series are listed in Table 6. The results indicate that most of the seasonally adjusted series are non-stationary. Hence, it is evident that the non-stationarity of the datasets is not just due to the presence of seasonality alone. It indicates that, in addition to seasonal differencing, performing the first-order difference would be prudent in conversion of non-stationary time-series to stationary time-series. Subsequently, all fourteen time-series datasets are subjected to one seasonal differencing and a first-order difference. The KPSS test is performed on the resulting time-series datasets. The test results are presented in the last column of Table 6, where it can be observed that all the differenced time-series datasets are stationary.

Modelling by seasonal autoregressive integrated moving average (SARIMA) method

Forecasting models are developed by applying the SARIMA method using the original time-series datasets. The model has an inherent ability to transform non-stationary data into stationary data using the parameters d and D determined earlier. SARIMA models for each variable are developed with a different combination of parameters, and the best-fitting model is selected based on statistical evaluation. The procedure followed to develop the SARIMA model for MMAX variable of Kozhikode station is detailed. A similar procedure is adopted for the other thirteen variables.

In the SARIMA model, where the model is represented by SARIMA$(p, d, q)(P, D, Q)_m$, the determination of the parameters p, d, q, P, D, Q and m, for a particular time-series data, completes the development of the model for that dataset. In the present case, from the previous analysis, it was found that $d = 1$ and $D = 1$, and for monthly data $m = 12$. Therefore, it is necessary to determine the values for parameters p, q, P and Q alone. These parameters are determined from the ACF and PACF plots of the stationary series (the seasonally and first-order differenced series) (Hyndman and Athanasopoulos 2018). The ACF and PACF plots of Kozhikode MMAX which are seasonally and first-order differenced are shown in Fig. 4, where the first 30 lags are considered for determining the parameters.

The value of non-seasonal autoregressive term (p) and seasonal autoregressive term (P) are determined from the PACF plot (Fig. 4b). In the first span of seasonality, there are significant spikes at lag 1, lag 2 and lag 3, and this indicates that a non-seasonal autoregressive component up to $AR(3)$ (i.e. $p \leq 3$) would be appropriate. The spikes at lags 1, 2 and 3 are considered, while the spikes at lags 5, 6 and 11 are ignored, because lags 1, 2 and 3 serially lie outside the bounds and lag 4 lies within the bound, and thus break the continuity. All the out of bound lags, in the first span of seasonality, after lag 4 are ignored for this reason. Both the second (lags 12 to 23) and third span (lags 24 to 35) of seasonality have out of bound lags. Therefore, a seasonal autoregressive component $AR(2)$ (i.e. $P \leq 2$) would be appropriate. Similarly, the moving average components are determined from the ACF plot Fig. 4a. The appropriate values of moving average components are $q \leq 1$ and $Q \leq 2$ (seasonal). Thus, the candidate model is SARIMA$(3, 1, 1)(2, 1, 2)_{12}$.

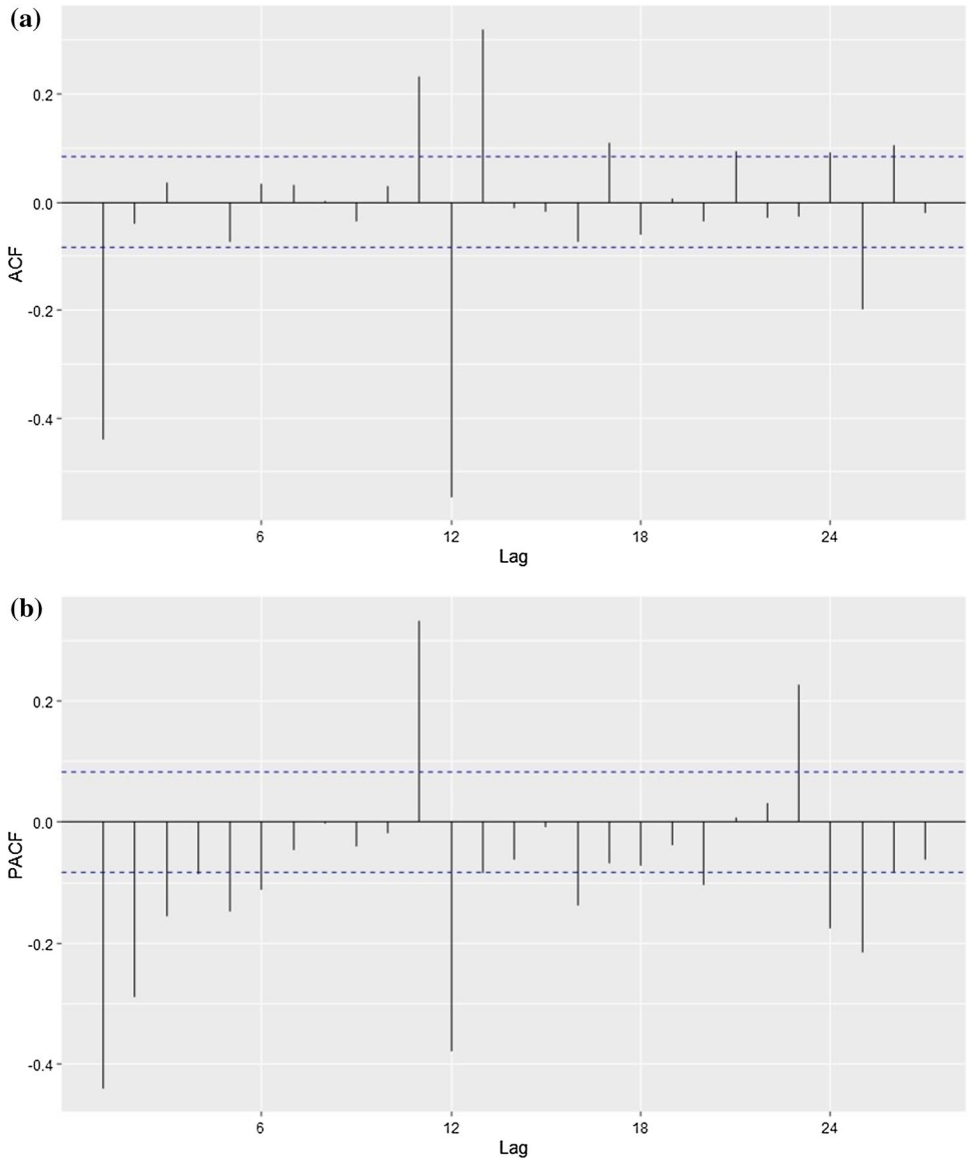

Fig. 4 **a** ACF plots for the stationary series of MMAX variable at Kozhikode station, **b** PACF plot for the stationary series of MMAX variable at Kozhikode station

It may be noted that the nature of the fourth span (lags 36 to 47) in the PACF and ACF plots is unknown. Therefore, due consideration should also be given for the seasonal autoregressive component $AR(3)$ (i.e. $P = 3$) and the seasonal moving average component $MA(3)$ (i.e. $Q = 3$). In the model development phase, it is necessary for the developer to ensure that the model is parsimonious. In order to satisfy the parsimony principle, the sum of the parameters p, q, P and Q of the SARIMA model should be less than or equal to six. Therefore, these four parameters of the candidate model are perturbed in the range of −1 and + 1. It resulted in 15 possible combinations to build the SARIMA model. Out of these, the best model is the one which minimises AICc (corrected Akaike information criteria) and BIC (Bayesian information criteria). The AICc and BIC values for the 15 models are presented in Table 7. The most suitable model that corresponds to the lowest AICc and BIC value is SARIMA$(2, 1, 1)(1, 1, 2)_{12}$. However, it may be noted that the AICc and BIC values of the other four models SARIMA$(2, 1, 2)(1, 1, 1)_{12}$, SARIMA$(2, 1, 1)(1, 1, 1)_{12}$, SARIMA$(2, 1, 1)(2, 1, 1)_{12}$ and SARIMA$(3, 1, 1)(1, 1, 1)_{12}$ are also closer to the selected model.

Statistical evaluation of the developed models is carried out using the validation dataset. The computed statistical measures are root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and mean percentage error (MPE). The results of the statistical evaluation of the models are presented in Table 8. In the statistical evaluation, the SARIMA$(2, 1, 1)(2, 1, 2)_{12}$ model did not produce the best results; nevertheless, the results are very close to the models that produced the best results. Therefore, the SARIMA$(2, 1, 1)(1, 1, 2)_{12}$ is considered to be an appropriate model for forecasting the MMAX variable of the Kozhikode station.

Table 7 The AICc and BIC values of the SARIMA models for MMAX variable of the Kozhikode station

SARIMA Model	AICc	BIC
SARIMA$(3,1,1)(1,1,1)_{12}$	– 2208.86	– 2180.66
SARIMA$(3,1,0)(2,1,1)_{12}$	– 2176.63	– 2148.43
SARIMA$(3,1,0)(1,1,2)_{12}$	– 2180.75	– 2152.56
SARIMA$(3,1,0)(1,1,1)_{12}$	– 2176.52	– 2152.33
SARIMA$(2,1,1)(2,1,1)_{12}$	– 2208.06	– 2179.86
SARIMA$(2,1,1)(1,1,2)_{12}$	– 2212.19	– 2183.99
SARIMA$(2,1,1)(1,1,1)_{12}$	– 2208.52	– 2184.33
SARIMA$(2,1,0)(2,1,2)_{12}$	– 2173.92	– 2145.72
SARIMA$(2,1,0)(2,1,1)_{12}$	– 2168.88	– 2144.68
SARIMA$(2,1,0)(1,1,2)_{12}$	– 2174.36	– 2150.16
SARIMA$(2,1,0)(1,1,1)_{12}$	– 2169.02	– 2148.83
SARIMA$(2,1,0)(1,1,3)_{12}$	– 2173.91	– 2145.72
SARIMA$(2,1,0)(3,1,1)_{12}$	– 2168.35	– 2140.16
SARIMA$(2,1,2)(1,1,1)_{12}$	– 2210.84	– 2182.64

Table 8 The statistical evaluation results of the SARIMA models developed for the Kozhikode MMAX

SARIMA Model	RMSE	MAE	MPE	MAPE
SARIMA$(3,1,1)(1,1,1)_{12}$	0.846	0.656	– 1.36	2.046
SARIMA$(3,1,0)(2,1,1)_{12}$	1.092	0.891	– 2.41	2.8
SARIMA$(3,1,0)(1,1,2)_{12}$	1.121	0.92	– 2.541	2.9
SARIMA$(3,1,0)(1,1,1)_{12}$	1.117	0.918	– 2.551	2.891
SARIMA$(2,1,1)(2,1,1)_{12}$	0.817	0.627	– 1.18	1.951
SARIMA$(2,1,1)(1,1,2)_{12}$	0.878	0.675	– 1.454	2.11
SARIMA$(2,1,1)(1,1,1)_{12}$	0.868	0.674	– 1.451	2.107
SARIMA$(2,1,0)(2,1,2)_{12}$	1.024	0.825	– 2.136	2.588
SARIMA$(2,1,0)(2,1,1)_{12}$	1.056	0.856	– 2.27	2.685
SARIMA$(2,1,0)(1,1,2)_{12}$	1.089	0.886	– 2.402	2.79
SARIMA$(2,1,0)(1,1,1)_{12}$	1.082	0.882	– 2.411	2.776
SARIMA$(2,1,0)(1,1,3)_{12}$	1.022	0.823	– 2.127	2.581
SARIMA$(2,1,0)(3,1,1)_{12}$	1.13	0.93	– 2.582	2.933
SARIMA$(2,1,2)(1,1,1)_{12}$	0.869	0.679	– 1.5	2.12
SARIMA$(4,1,0)(1,1,1)_{12}$	1.162	0.962	– 2.738	3.037

The selected model is also validated using the ACF of the residuals obtained from the fitted SARIMA$(2, 1, 1)(1, 1, 2)_{12}$ model to the complete time-series data. The residual plot and the ACF plot are shown in Fig. 5. Ideally, for a model to be absolutely perfect, it is expected to have autocorrelation of residuals close to zero.

However, if 95% of the spikes lie within the bounds ($\pm 2\sqrt{T}$, where T is the length of the time-series), that would confirm that the series is white noise without autocorrelation. Here, there are two significant spikes (at lag 25 and lag 35), and this translates to 94.4% of the spikes remaining within the bounds. The percentage of spikes lying within the bounds is close to 95% indicates that the selected model has the ability to provide good forecasting results.

Model building and validation for other thirteen variables are carried out using a similar procedure. The data length used for model building and validation is presented in Table 9. Finally, Table 10 shows the apt models for all of the fourteen temperature time-series.

Summary and conclusions

The development of temperature forecasting models for the state of Kerala, India, is presented in this article. Monthly mean maximum (MMAX) and mean minimum (MMIN) temperature time-series, obtained from seven stations of Kerala is used for the development of the model. The time-series temperature data observed over a period of 47 years, spanning from 1969 to 2015 is utilised in this study.

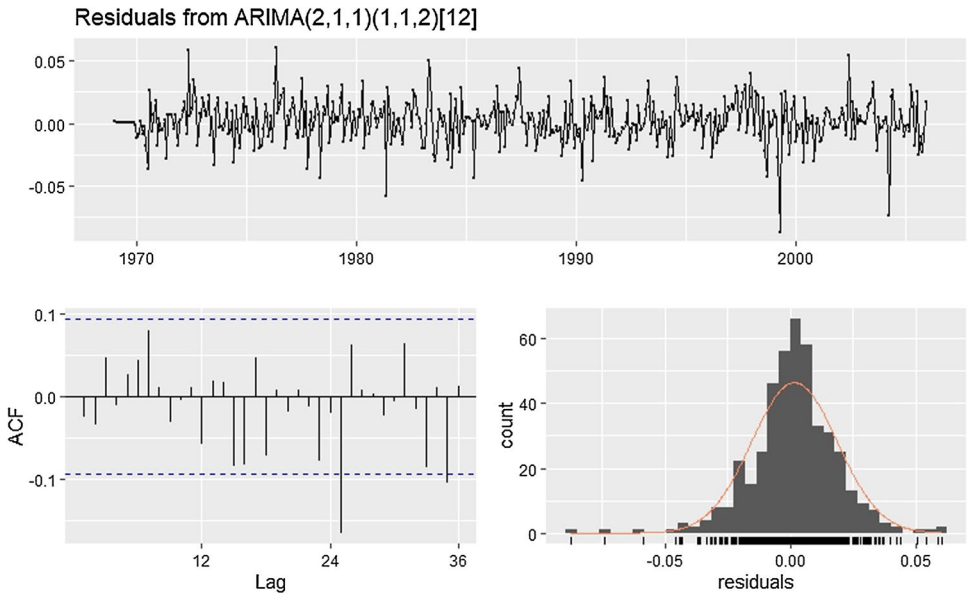

Fig. 5 The residual time-series of the fitted SARIMA$(2,1,1)(1,1,2)_{12}$ model for the Kozhikode MMAX variable, ACF plot of the residuals and the distribution of the residuals

Table 9 The training and the validation data length utilised

Station name	Variable type	Training data	Validation data
Kozhikode	MMAX	1969–2005 (37 years)	2006–2015 (10 years)
Kozhikode	MMIN	1969–2005 (37 years)	2006–2015 (10 years)
Kannur	MMAX	1981–2005 (28 years)	2006–2015 (7 years)
Kannur	MMIN	1981–2005 (28 years)	2006–2015 (7 years)
Alappuzha	MMAX	1969–2005 (37 years)	2006–2015 (10 years)
Alappuzha	MMIN	1969–2005 (37 years)	2006–2015 (10 years)
Punalur	MMAX	1969–2005 (37 years)	2006–2015 (10 years)
Punalur	MMIN	1969–2005 (37 years)	2006–2015 (10 years)
Kottayam	MMAX	1969–2003 (35 years)	2005–2011 (8 years)
Kottayam	MMIN	1969–2003 (35 years)	2005–2011 (8 years)
Thiruvananthapuram	MMAX	1969–2005 (37 years)	2006–2015 (10 years)
Thiruvananthapuram	MMIN	1969–2005 (37 years)	2006–2015 (10 years)
Trivandrum Airport	MMAX	1969–2005 (37 years)	2006–2015 (10 years)
Trivandrum Airport	MMIN	1969–2005 (37 years)	2006–2015 (10 years)

Some data gaps are identified in the datasets obtained from IMD. The missing values are estimated using the expectation–maximisation algorithm. It is natural for a long term time-series dataset of a meteorological variable to possess a trend. Moreover, the monthly mean of the meteorological variable is bound to have seasonal variations. The inherent seasonality in the variable induces a non-stationarity in the time-series datasets. Statistical analysis is carried out on the time-series datasets to understand the nature of the data.

The results from the descriptive statistics indicated that most of the temperature time-series are kurtotic. A preliminary analysis is carried out to test the normality of the data and to check the presence of outliers. The Shapiro-Wilk test and the Grubb's test are conducted to test the normality and to check the outliers. The results indicated that the time-series datasets are non-normal and outliers are absent.

The trend analysis is carried out by applying Mann–Kendall's trend test and Sen's Slope estimation. The results indicated the presence of trend in at least ten of the fourteen time-series datasets. This served as the first indication for the non-stationary nature of the datasets. In order to confirm the presence of seasonality, with absolute confidence, STL decomposition and KPSS test are conducted. In STL decomposition, the time-series is decomposed into trend, seasonal, and remainder components. The results obtained from these tests clearly indicated the presence of seasonality and thereby, confirmed the non-stationarity of the all the fourteen time-series datasets. Subsequently, one seasonal difference and one first-order difference are applied to

Table 10 The best-fit SARIMA models developed for forecasting the variables

Station name	Variable type	Number of valid models	Best-fit SARIMA model	AICc	BIC
Kozhikode	MMAX	15	SARIMA(2,1,1)(1,1,2)$_{12}$	– 2212.19	– 2183.99
Kozhikode	MMIN	31	SARIMA(2,1,1)(1,1,1)$_{12}$	– 2026.97	– 2002.77
Kannur	MMAX	31	SARIMA(2,1,1)(1,1,1)$_{12}$	– 1545.43	– 1523.03
Kannur	MMIN	15	SARIMA(3,1,1)(1,1,1)$_{12}$	– 1446.43	– 1420.34
Alappuzha	MMAX	15	SARIMA(2,1,1)(1,1,1)$_{12}$	– 2154.94	– 2130.74
Alappuzha	MMIN	31	SARIMA(2,1,1)(1,1,1)$_{12}$	– 2045.26	– 2021.06
Punalur	MMAX	15	SARIMA(2,1,1)(1,1,1)$_{12}$	– 1917.4	– 1893.2
Punalur	MMIN	31	SARIMA(2,1,1)(2,1,1)$_{12}$	– 1822.05	– 1793.86
Kottayam	MMAX	15	SARIMA(2,1,1)(1,1,1)$_{12}$	– 1736.01	– 1712.17
Kottayam	MMIN	31	SARIMA(3,1,1)(1,1,1)$_{12}$	– 1822.03	– 1794.25
Thiruvananthapuram	MMAX	31	SARIMA(2,1,2)(1,1,1)$_{12}$	– 2154.22	– 2126.03
Thiruvananthapuram	MMIN	31	SARIMA(2,1,1)(1,1,1)$_{12}$	– 2165.97	– 2141.77
TrivandrumAirport	MMAX	15	SARIMA(2,1,2)(1,1,1)$_{12}$	– 2341.1	– 2312.9
TrivandrumAirport	MMIN	31	SARIMA(2,1,1)(1,1,1)$_{12}$	– 2054.48	– 2030.29

transform the non-stationary datasets into stationary datasets. The results assist in identifying the values of the differencing parameters necessary for building the SARIMA model.

The SARIMA models are developed individually for each of the fourteen variables using the original time-series datasets. The SARIMA models are developed individually for each of the fourteen time-series datasets. The results indicated that the SARIMA$(2, 1, 1)(1, 1, 1)_{12}$ model the ideal one to forecast eight out of the fourteen time-series variables. In order to have a better understanding of local influences, the studies must be carried out on a better spatial and temporal scales.

Acknowledgements The first author would like to thank the Ministry of Human Resource Development (MHRD) for the financial support, which was provided in the form of research stipend. The authors also acknowledge the Indian Meteorological Department (IMD) for the temperature datasets.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aguado-Rodríguez GJ, Quevedo-Nolasco A, Castro-Popoca M, Arteaga-Ramírez R, Vázquez-Peña MA, Zamora-Morales BP (2016) Predicción de variables meteorológicas por medio de modelos arima. Agrociencia 50(1):1–13

Allen MR, Ingram WJ (2002) Constraints on future changes in climate and the hydrologic cycle. Nature 419(6903):228–232

Andronova NG, Schlesinger ME (2000) Causes of global temperature changes during the 19th and 20th centuries. Geophys Res Lett 27(14):2137–2140

Arjun KM (2013) Indian agriculture-status, importance and role in Indian economy. Int J Agric Food Sci Technol 4(4):343–346

Cleveland RB et al (1990) Stl: a seasonal-trend decomposition procedure based on loess. citeulike-article-id:1435502

Gilbert RO (1987) Statistical methods for environmental pollution monitoring. Wiley, New York

Gocic M, Trajkovic S (2013) Analysis of changes in meteorological variables using Mann–Kendall and Sen's slope estimator statistical tests in Serbia. Glob Planet Change 100:172–182

Hänsel S, Medeiros DM, Matschullat J, Petta RA, de Mendonça Silva I (2016) Assessing homogeneity and climate variability of temperature and precipitation series in the capitals of North-Eastern Brazil. Front Earth Sci 4:29

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts, Melbourne

Indian Network for Climate Change Assessment and India Ministry of Environment (2010) Climate Change and India: a 4 × 4 assessment, a sectoral and regional analysis for 2030s, vol 2. Ministry of Environment & Forests, Government of India, New Delhi

Jain SK, Kumar V (2012) Trend analysis of rainfall and temperature data for India. Curr Sci 102:37–49

Kang H (2013) The prevention and handling of the missing data. Korean J Anesthesiol 64(5):402

Kendall M (1975) Rank correlation methods. Charles Griffin, London (There is no corresponding record for this reference)

Kocsis T, Kovács-Székely I, Anda A (2017) Comparison of parametric and non-parametric time-series analysis methods on a long-term meteorological data set. Cent Eur Geol 60(3):316–332

Kocsis T, Kovács-Székely I, Anda A (2020) Homogeneity tests and non-parametric analyses of tendencies in precipitation time series in Keszthely, Western Hungary. Theor Appl Climatol 139(3–4):849–859

Kwiatkowski D, Phillips PC, Schmidt P, Shin Y et al (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. J Econom 54(1–3):159–178

Lai Y, Dzombak DA (2020) Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation. Weather Forecast 35:959–976

Mann HB (1945) Nonparametric tests against trend. Econom J Econom Soc 13:245–259

Mills TC (2014) Time series modelling of temperatures: an example from k efalonia. Meteorol Appl 21(3):578–584

Radziejewski M, Kundzewicz ZW (2004) Detectability of changes in hydrological records/possibilité de détecter les changements dans les chroniques hydrologiques. Hydrol Sci J 49(1):39–51

Sen PK (1968) Estimates of the regression coefficient based on Kendall's tau. J Am Stat Assoc 63(324):1379–1389

Tiwari P, Kar S, Mohanty U, Dey S, Kumari S, Sinha P (2016) Seasonal prediction skill of winter temperature over North India. Theor Appl Climatol 124(1–2):15–29

Trenberth KE (1999) Conceptual framework for changes of extremes of the hydrological cycle with climate change. In: Weather and climate extremes. Springer, Dordrecht, pp 327–339

Wang H, Huang J, Zhou H, Zhao L, Yuan Y (2019) An integrated variational mode decomposition and arima model to forecast air temperature. Sustainability 11(15):4018

Wanishsakpong W, Owusu BE (2020) Optimal time series model for forecasting monthly temperature in the southwestern region of Thailand. Model Earth Syst Environ 6(1):525–532