Chapter 3
Exploring the Behavior and Changing Trends of Rainfall and Temperature Using Statistical Computing Techniques

Abdus Salam Azad, Md. Kamrul Hasan, M. Arif Imtiazur Rahman, Md. Mustafizur Rahman, and Nashid Shahriar

Abstract The present study aimed at quantifying the change in surface air temperature and monthly total rainfall. The changing trend was detected using Mann–Kendall trend test, seasonal Mann–Kendall trend test, and Sen's slope estimator. *K*-means clustering algorithm was used to identify the rainfall distribution patterns over the years and also their changes with time. A comparative analysis was done among different time series prediction models to find out their suitability for forecasting daily temperature in climatic condition of Bangladesh. The analysis was performed using daily temperature and rainfall data of more than last 40 years (till 2009). The study found an increasing trend in maximum temperature during June to November and in minimum temperature during December to January in Bangladesh. There has been seen no significant change in rainfall over the years. However on the western side of the country, the amount of rain is significantly less than the eastern side. The study found that different prediction models were appropriate for different conditions.

Keywords Climate change • *K*-means clustering algorithm • Mann–Kendall trend test • Sen's slope estimator • Data mining • Pattern recognition • Time series prediction • Statistical analysis

3.1 Introduction

Bangladesh is likely to be one of the countries in the world which is most vulnerable to climate change. In recent times natural hazards are more frequent and intense compared to the similar kind of events occurred in one or two decades ago.

A.S. Azad (✉) • M.K. Hasan • M.A.I. Rahman • M.M. Rahman • N. Shahriar
Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Palashi, Dhaka 1000, Bangladesh
e-mail: azadsalam2611@gmail.com; kamrulhasan326@gmail.com; arif.imtiaz216@gmail.com; mustafiz_rahman@cse.buet.ac.bd; nshahriar@cse.buet.ac.bd

T. Islam et al. (eds.), *Computational Intelligence Techniques in Earth and Environmental Sciences*, DOI 10.1007/978-94-017-8642-3_3,

National governments and IPCC (Intergovernmental Panel on Climate Change) scientists accepted that these climate hazards are the result of climate change at the global and regional level. According to the Fourth Assessment Report (AR4) of IPCC in Climate Change 2007: Synthesis Report, during the last hundred years the global temperature increased by 0.74 ± 0.18 °C. The model results of AR4 for Bangladesh are appropriate for global scale. But they did not use the local data of 37 stations in Bangladesh operated by the Bangladesh Meteorological Department (2012). Various researchers have contributed to the study of rainfall and temperature of Bangladesh with local data.

Rana et al. (2007) attempted to construct linear relationship between monthly, seasonal, and annual rainfall over Bangladesh with the southern oscillation index (SOI). Ahasan et al. (2010) analyzed the variability and trends of summer monsoon (Jun–Sep) rainfall over Bangladesh and found that the annual profile of the monthly total country average rainfall shows a unimodal characteristic with highest in July followed by June and August and lowest in January followed by December and February.

Basak et al. (2013) tried to detect trends in the monthly average maximum and minimum temperature and rainfall based on linear regression method. Long-term changes of near surface air temperature over Bangladesh have also been studied by Islam (2009). Various studies like Warrick et al. (1994), Karmakar and Shrestha (2000), Nahrin et al. (1997), Chowdhury and Debsarma (1992), Mia (2003), and Debsarma (2003) also focused on trends of change in rainfall and temperature in the context of Bangladesh.

However, to the best of our knowledge, several effective and advanced methods like Mann–Kendall trend test, clustering, etc. have not been utilized to investigate the climatic conditions to a greater extent.

The Mann–Kendall trend test is a widely known method for finding trends in time series data. Jain et al. (2013) examined trends in monthly, seasonal, annual rainfall, and temperature for the northeast region of India. The magnitude of trend in a time series was determined using Sen's estimator and statistical significance of the trend was analyzed using the Mann–Kendall trend test. Tripathi and Govindaraju (2009) used the Mann–Kendall trend test to study the changes in rainfall and temperature patterns over India. The method was also used to analyze the changing trend of the nonuniformity in rainfall of the upper Yangtze basin by Huang and Wang (2011). A number of researchers like Xi-ting et al. (2011), Gaoliao et al. (2012), Yue and Hashino (2003), Singh et al. (2008a, b), and Kumar and Jain (2010) also employed MK test to find trend in climatic data.

Clustering is a process of partitioning a set of data in a set of meaningful subclasses called clusters, which can effectively be used to analyze the distribution pattern of rainfall. Ramos (2001) used clustering techniques including *K*-means clustering method to analyze the rainfall distribution patterns over the years and their changes over time in a Mediterranean region. The study found that the variations of the mean annual rainfall in the Alt Penedès area throughout a period of 111 years have not followed a consistent trend. Pelczer and Cisneros-Iturbe (2008) also

applied the *K*-means clustering technique to establish intensity classes to identify rainfall patterns over the Valley of Mexico. Major circulation patterns, associated with daily precipitation in Portugal, were classified by Corte-Real et al. (1998) based on the *K*-means clustering algorithm coupled with principal component analysis.

In our study inspection was done on more than 40 years climatic data of all the stations (37) of Bangladesh. Then five important places—Dhaka, Cox's Bazar, Khulna, Sylhet, and Rajshahi were analyzed with their geographic position in consideration, as these stations give quite a clear picture of the entire country. This study undertook the challenge of finding the trends in daily temperature changes and monthly total rainfall on those stations using the Mann–Kendall trend test and Sen's slope. For finding any trend in monthly total rainfall, seasonal Mann–Kendall trend test was also run to incorporate the seasonality. Monthly total rainfall of these selected regions was investigated to find the distribution of rainfall throughout the year. *K*-means clustering algorithm was used for this purpose.

As an agricultural country short-term temperature prediction is very important for Bangladesh. The study analyzed some of the well-known time series prediction models for finding their applicability to local data of Bangladesh. The models that were analyzed are autoregressive integrated moving average (ARIMA), Naive, random walk with drift (RWD), and the Theta model.

The later sections provide the details of our study. Section 3.2 provides an overview of the methods and materials that have been used in our study. In Sect. 3.3 the results of our study are discussed and in Sect. 3.4 we conclude the study with our key findings.

3.2 Materials and Methods

In this section we briefly discuss about the different methods and materials used in our study.

3.2.1 K-*Means Clustering*

Clustering is a main task of explorative data mining. It assigns a set of objects into groups (clusters) so that the objects in the same cluster are more similar. For clustering *K*-means clustering algorithm (MacQueen 1967) was used. It is an algorithm for putting N data points (x_1,x_2, ..., x_n) in an I-dimensional space into K clusters. The mean of each cluster is denoted by $m^{(k)}$. Each vector x has I components x_i. Distances between points in real space is as follows:

$$d(x, y) = \frac{1}{2}\sum_i (x_i - y_i)^2$$

The K-means $m^{(k)}$ is initialized to a random value. Then it follows two steps.

Assignment step: Each data point n is assigned to the nearest mean. For the cluster $k^{(n)}$, the point $x^{(n)}$ belongs to

$$\hat{k}^{(n)} = \arg\min_k d\left(m^{(k)}, x^{(n)}\right)$$

The indicator variable, $r_k^{(n)}$ is set to one if mean k is the closest mean to data point $x^{(n)}$; otherwise, $r_k^{(n)}$ is zero.

$$r_k^{(n)} = \begin{cases} 1 & \text{if } \hat{k}^{(n)} = k \\ 0 & \text{if } \hat{k}^{(n)} \neq k \end{cases}$$

Update step: The model parameters, the means, are adjusted to match the sample means of the data points that they are responsible for.

$$m^{(k)} = \begin{cases} \dfrac{\sum_n r_k^{(n)} x^{(n)}}{R^{(k)}} & \text{if } R^{(k)} > 0 \\ \text{Oldest } m^{(k)} & \text{if } R^{(k)} = 0 \end{cases}$$

where $R^{(k)}$ is the total responsibility of mean k,

$$R^{(k)} = \sum_n r_k^{(n)}$$

The assignment step and update step is repeated until the assignments do not change.

3.2.2 Mann–Kendall Trend Test

The Mann–Kendall test (Mann 1945; Kendall 1975) is a nonparametric test for identifying trends in time series data. The test compares the relative magnitudes of sample data rather than the data values themselves (Gilbert 1987). Here it is assumed that that there exists only one data value per time period. Let $(x_1, x_2, \ldots, x_n)$ represent n data points where x_j represents the data point at time j. Then the Mann–Kendall statistics (S) is given by

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sign}\left(x_j - x_k\right)$$

where

$$\operatorname{sign}(x_j - x_k) = \begin{cases} +1 & \text{if } x_j - x_k > 0 \\ 0 & \text{if } x_j - x_k = 0 \\ -1 & \text{if } x_j - x_k < 0 \end{cases}$$

A very high positive value of S is an indicator of an increasing trend, and a very low negative value indicates a decreasing trend. However, it is necessary to compute the VAR(S), Sen's slope associated with S and the sample size n, to statistically quantify the significance of the trend. When $n \geq 8$, the S is approximately normally distributed with the mean. The variance of S, VAR(S), by the following equation (Helsel and Hirsch 1992):

$$\mathrm{VAR}(S) = \frac{n*(n-1)*(2n+5) - \sum_{i=1}^{m} t_i(i)(i-1)(2i+5)}{18}$$

where t_i is considered as the number of ties up to sample i. VAR(S) and Sen's slope estimator both are used to estimate the trend in time series data. A positive S value indicates a positive trend and a negative value indicates a negative trend in time series data.

3.2.3 Seasonal Mann–Kendall Trend Test

Hirsch et al. (1982) developed this test that is used to find the monotonic trend in time series data with seasonal variation. Mann–Kendall statistics S is computed separately for each month and then summed to obtain the overall test statistic. Mann–Kendall statistics S_k for each season k ($k = 1, 2, \ldots, p$):

$$S_k = \sum_{i=1}^{n_k - 1} \sum_{j=i+1}^{n_k} \operatorname{sign}(x_{jk} - x_{ik})$$

where

x_{jk} = Observation from season k in year j
x_{ik} = Observation from season k in year i

Then these statistics are summed to form overall statistics S_n:

$$S_n = \sum_{k=1}^{p} S_k$$

Here the overall variance of the test statistics VAR(S_n) is obtained by summing the variances of the Kendall score statistics for each month.

Test interpretation for Mann–Kendall trend test and seasonal Mann–Kendall trend test:

The null and alternative hypotheses for both the Mann–Kendall trend test and seasonal Mann–Kendall trend test:

H_0: There is no trend in the series
H_a: There is a positive/negative trend in the series

Here the significance level alpha (α) is 5 %. If the computed p-value is greater than alpha, then the null hypothesis H_0 cannot be rejected.

3.2.4 Sen's Slope Estimator

In nonparametric statistics, Sen's slope estimator (1968) is a method for robust linear regression that chooses the median slope among all lines through pairs of two-dimensional sample points. The magnitude of linear trend is predicted by the Sen's estimator. The slope (Q) of all data pair is

$$Q = \frac{x_i^{'} - x_i}{i^{'} - i} \quad \text{for } i = 1, 2, 3, \ldots, N$$

where

Q = slope between data points X_i and $X_i^{'}$
$X_i^{'}$ = data measurement at time i'
X_i = data measurement at time i
i' = time after time i

Sen's estimator of slope is simply given by the median slope (Q'), shown below as:

$$Q^{'} = \begin{cases} Q\left[\dfrac{N+1}{2}\right] & \text{if } N \text{ is odd} \\ \dfrac{Q[N+1] + Q[N+2]}{2} & \text{if } N \text{ is even} \end{cases}$$

where N is the number of calculated slopes.

Then, Q'med is computed by a two-sided test at 100 $(1 - \alpha)$ % confidence interval and then a true slope can be obtained by the nonparametric test. Positive value of Sen's slope indicates increasing trend and a negative value indicates a decreasing trend.

3.2.5 Naive Model

Naive approach is the simplest but an efficient forecasting model of time series data. If the data is stable according to this model, forecast of any periods are equal to the actual value of previous period. When the sequence of observations begins at time $t = 0$, the simplest form of Naive is given by the formulae:

$$y_t = y_{t-1}, \quad t > 0$$

Naive forecast can be used as a benchmark against which other forecasting models can be compared.

3.2.6 Autoregressive Integrated Moving Average Model

ARIMA model is a generalization of an autoregressive moving average (ARMA) model (Box and Jenkins 1970). This model is used to predict future points in the series. ARIMA models aim to describe the autocorrelations in the data. It combines differencing with autoregression and moving average (MA) model.

The full model can be written as:

$$y_t^{'} = c + \sum_{i=1}^{p} \phi_i y_{t-i}^{'} + \sum_{i=1}^{q} \theta_i e_{t-i}$$

ARIMA models are defined for stationary time series. Therefore, if the data shows non-stationary time series, then it is needed to *difference* the time series until a stationary time series is obtained.

Differencing method makes time series stationary by computing the differences between consecutive observations. First-differenced series $y_t^{'}$ can be written as:

$$y_t^{'} = y_t - y_{t-1}$$

When the differenced series is white noise with nonzero mean, the original series can be written as

$$y_t - y_{t-1} = c + e_t$$

where c is the average of the change between consecutive observations.

$$e_t = \text{White noise}$$

Sometimes it is needed to difference the data second time to obtain stationary series:

$$\begin{aligned} y_t^{''} &= y_t^{'} - y_{t-1}^{'} \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

The backward shift operator B shifts the data y_t back to one period.

$$By_t = y_{t-1} \tag{3.1}$$

So first difference can be written as:

$$y_t^{'} = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t$$

Similarly second difference can be written as:

$$y_t^{''} = y_t^{'} - y_{t-1}^{'} = (1 - B)^2 y_t$$

In general, a dth order difference can be written as:

$$y_t^d = (1 - B)^d y_t$$

Autoregressive model (*AR*) is a multiple regression model that specifies that the output variable can be computed from linear combination of its own previous values. This model forecasts the value at time t by the weighted average of past few observations. The notation AR(p) indicates an autoregressive model of order p. The AR(p) model is defined as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + e_t \tag{3.2}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of this model. Using backshift operator B from Eq. 3.1 the AR(p) model can be written as:

$$y_t = c + \sum_{i=1}^{p} \phi_i B^i y_{t-i} + e_t$$

Changing the values of $\phi_1, \phi_2, \ldots, \phi_p$ results in different time series patterns. Autoregressive model can be restricted to stationary data by putting some constraints on the values of the parameters. This equation only gives one step ahead forecast. For n ahead step forecast, past forecast values are used for weighted average on the right side of the equation (Eq. 3.2). There are some problems with forecasting using the AR model. It will not tell very much about future. Forecast will not reach the peaks of the data, because it smoothes out the data by taking the mean of observed values.

The moving average (MA) model is another model to forecast time series data. While the AR model uses past values of forecast variable in regression, the MA model uses past forecast errors. The MA(q) model can be written as:

$$y_t = c + \sum_{i=1}^{q} \theta_i e_{t-q} + e_t$$

where e_t = white noise and $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model.

Any value at time t is the weighted average of past few forecast errors. Some constraints are needed to put on parameters for restricting the model to stationary data. Time series pattern can be changed by changing the parameters. But changing the value of e_t will just change the scale not the pattern.

ARIMA model:

ARIMA model is the combination of all these methods. First the data must be *differenced d* times to obtain stationary series if needed. Then full ARIMA model is applied on stationary data to forecast. The full model is:

$$y_t^{'} = c + \sum_{i=1}^{p} \phi_i y_{t-i}^{'} + \sum_{i=1}^{q} \theta_i e_{t-i} \tag{3.3}$$

This is called ARIMA(p, d, q) model, where

p = order of the autoregressive part
d = degree of first differencing involved
q = order of the moving average part

Equation 3.3 can be written using backshift notation:

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)(1 - B)^d y_t = c + \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) e_t$$

Determining the appropriate value of p, d, and q for the data is difficult. The appropriate value of p, d, and q cannot be found from a time plot. From autocorrelations and partial autocorrelations plot, the value of q and p can be determined. ARIMA model selection becomes problematic with missing observations and other data irregularities.

3.2.7 Random Walk with Drift

When a time series shows irregular pattern, it is better to try to predict the change that occurs from one period to the next. Random walk model can be written as:

$$Y(t) = Y(t-1) + \alpha$$

where

$Y(t)$ = Predicted value at time t
$Y(t-1)$ = Previous value
α = Drift: Average change between periods

Time series value at any period will be equal to the last period's value plus the average change between periods. Average change between two periods is called drift (α) which acts like a trend. This model is known as "random walk" model: it assumes that from one period to the next, the original time series merely takes a random step away from its last position. If the constant term (α) in the random walk model is zero, then it is called random walk without drift. This is similar to the Naive model.

If the series being fitted by a random walk model has an average upward trend ($\alpha > 0$) or downward trend ($\alpha < 0$), then a nonzero constant (α) drift term must be added. This drift along with previous frequency determines new forecast (Pesaran and Pick 2009). It is known as random walk with drift.

3.2.8 Theta Model

Theta method is used for obtaining forecast from series of data. Forecasts obtained by theta method are equivalent to simple exponential smoothing (SES) with drift. Theta method is simply a special case of SES with drift, where the drift parameter is half the slope of the linear trend fitted to the data. The method performed well, particularly for monthly series and for microeconomic data. The detail of the Theta model is described by Assimakopoulos and Nikolopoulos (2000).

3.2.9 Wilcoxon Test

The Wilcoxon signed rank test is a nonparametric method of testing whether two populations X and Y have the same continuous distribution without assuming them to follow the normal distribution (Wilcoxon 1945). Two data samples are matched if they come from repeated observations of the same subject. The null and alternative hypotheses for the Wilcoxon signed rank test are:

H_0: median difference between the pairs is zero
H_a: median difference is not zero

The output of Wilcoxon signed rank test is p-value. It is used to determine which hypothesis will be accepted. Ninety-five percent confidence level is used. If the value is less than 0.05 significance level, null hypothesis can be rejected which yields that two samples are not identical. Otherwise data samples follow the same distribution.

3.2.10 RMS Error

Root mean square (RMS) error is a measure of the average error, weighted according to the square of the error. It is a good measure of accuracy, but only to compare forecast errors of different models. RMSE can be defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

where

F_i = the forecast values
O_t = the corresponding verifying value
N = the number of observation points

3.2.11 SMAPE

Symmetric mean absolute percentage error (SMAPE) is an accuracy measure based on percentage error. For n forecast, the SMAPE can be defined as:

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - f_i|}{(y_i + f_i)/2}$$

where

y_i = actual value
f_i = forecast value

The absolute value of y_i and f_i is divided by their average.

3.2.12 Materials Used During the Study

The Mann–Kendall trend test and seasonal Mann–Kendall trend test are implemented in Java. *K*-means clustering algorithm is implemented using WEKA (Hall et al. 2009). The time series prediction models are implemented using the forecast package of R (R Core Team 2013).

The analysis was done during 2012–2013 in BUET, Bangladesh.

3.2.13 Data Used in the Study

The daily minimum temperature, maximum temperature, and rainfall data in the period 1947–2009 from 37 weather stations of Bangladesh were collected from Bangladesh Meteorological Department (BMD). Daily minimum, maximum temperature and rainfall data were cleaned by filling in missing values with mean values of adjacent days.

With geographic position in consideration, we selected Dhaka, Rajshahi, Khulna, Sylhet, and Cox's Bazar stations for our study.

3.3 Results and Discussion

In this section results of different analysis are presented and investigated. In Sects. 3.3.1 and 3.3.2 the changing trends of temperature and rainfall are inspected by means of the Mann–Kendall trend test and Sen's slope estimator. The distribution of rainfall in Bangladesh is discussed in Sect. 3.3.3. In Sect. 3.3.4 a comparative study is presented on applicability of some of the well-known time series prediction models for providing temperature forecast.

3.3.1 Change in Temperature

Regression analysis and then the Mann–Kendall trend test were run on average maximum and minimum monthly temperature on the selected stations. The results show that the maximum temperature (T_{max}) of the months June to November has increased and the minimum temperature (T_{min}) of winter has increased in Dhaka, Cox's Bazar, and Sylhet. On the contrary T_{min} shows negative trend in Khulna and Rajshahi.

The result of the Mann–Kendall trend test and Sen's slope estimator on average maximum and minimum temperature are shown in Tables 3.1 and 3.2, respectively. The change in temperature per decade is shown in Fig. 3.1.

In Dhaka an increasing trend in maximum temperature is observed during the months June to November (Table 3.1). On average the maximum temperature has increased 0.19 °C/decade. Figure 3.1c clearly shows the increasing trend during this time span. On the other hand the T_{min} has significantly increased during November to March (Table 3.2). A very high increase, more than 0.55 °C/decade is observed (Fig. 3.1c).

At Cox's Bazar positive trend is detected throughout the year. Only the month of May has not experienced any trend for T_{min}. Specially from the start of the Monsoon (Jun–Sep) till the month of November T_{max} showed a significant rise, more than 0.29 °C/decade (Fig. 3.1d). While T_{min} is increased significantly in

Table 3.1 Mann–Kendall statistics (S) and Sen's slope (SS) of maximum temperature of different regions

Month	Rajshahi	Khulna	Dhaka	Sylhet	Cox's Bazar
January	$S = -227$	$S = -567$		$S = +221$	$S = +427$
	SS $= -0.02$	SS $= -0.026$		SS $= +0.015$	SS $= +0.015$
February				$S = +231$	$S = +639$
				SS $= +0.032$	SS $= +0.029$
March		$S = -315$			$S = +679$
		SS $= -0.016$			SS $= +0.029$
April					$S = +752$
					SS $= +0.031$
May				$S = +286$	$S = +680$
				SS $= +0.026$	SS $= +0.027$
June	$S = +219$	$S = +505$	$S = +625$	$S = +525$	$S = +672$
	SS $= +0.03$	SS $= +0.019$	SS $= +0.034$	SS $= +0.036$	SS $= +0.029$
July	$S = +398$	$S = +773$	$S = +732$	$S = +460$	$S = +724$
	SS $= +0.03$	SS $= +0.022$	SS $= +0.024$	SS $= +0.033$	SS $= +0.28$
August	$S = +523$	$S = +653$	$S = +836$	$S = +580$	$S = +887$
	SS $= +0.037$	SS $= +0.022$	SS $= +0.032$	SS $= +0.039$	SS $= +0.034$
September		$S = +330$	$S = +521$	$S = +431$	$S = +929$
		SS $= +0.012$	SS $= +0.022$	SS $= +0.031$	SS $= +0.032$
October		$S = +435$	$S = +557$	$S = +677$	$S = +955$
		SS $= +0.014$	SS $= +0.028$	SS $= +0.036$	SS $= +0.037$
November	$S = +268$	$S = +438$	$S = +548$	$S = +634$	$S = +892$
	SS $= +0.018$	SS $= +0.017$	SS $= +0.03$	SS $= +0.033$	SS $= +0.041$
December				$S = +448$	$S = +905$
				SS $= +0.025$	SS $= +0.038$

Positive value of SS and S signify positive trend and vice versa
Empty cell denotes no trend

winter (Dec–Feb) by 0.35 °C/decade. These changes are verified by the Mann–Kendall statistics and Sen's slope (Tables 3.1 and 3.2).

Also in Sylhet the temperature has seen a positive trend almost throughout the year. During June to November, the positive trend in T_{max} is about 0.29 °C/decade. In winter T_{min} has increased by about 0.31 °C/decade (Fig. 3.1e). The Mann–Kendall statistics and Sen's slope also suggests that.

In Rajshahi T_{max} has shown a positive trend in June, July, August, and November (Table 3.1). T_{min} increased in July, August (Table 3.2). While January has shown a negative trend both in minimum and maximum temperature. In monsoon T_{max} has shown increasing trend more than 0.2 °C/decade (Fig. 3.1a). In winter (Dec–Feb) T_{min} is found stable (Fig. 3.1a). The Mann–Kendall trend test has shown no trend.

In Khulna T_{max} has shown a positive trend in June to November and negative trend in January and March (Table 3.1). T_{max} has increased about 0.16 °C/decade (Fig. 3.1b). Only in Khulna it is found that in winter T_{min} has shown negative trend. T_{min} has decreased in December to February and May (Table 3.2). A slightly decreasing trend in T_{min}, about 0.11 °C/decade, is found in winter (Fig. 3.1b).

Table 3.2 Mann–Kendall statistics (S) and Sen's slope (SS) of minimum temperature of different regions

Month	Rajshahi	Khulna	Dhaka	Sylhet	Cox's Bazar
January	$S = -266$	$S = -355$	$S = +727$	$S = +339$	$S = +569$
	SS $= -0.034$	SS $= -0.04$	SS $= +0.05$	SS $= +0.025$	SS $= +0.025$
February	$S = -226$	$S = +786$	$S = +393$	$S = +759$	
	SS $= -0.021$	SS $= +0.055$	SS $= +0.025$	SS $= +0.04$	
March			$S = +0.479$	$S = +358$	$S = 558$
			SS $= +0.04$	SS $= +0.029$	SS $= +0.033$
April					$S = +347$
					SS $= +0.017$
May		$S = -239$			
		SS $= -0.015$			
June			$S = +319$		$S = +506$
			SS $= +0.011$		SS $= +0.016$
July	$S = +169$		$S = +312$	$S = +244$	$S = +581$
	SS $= +0.007$		SS $= +0.008$	SS $= +0.009$	SS $= +0.012$
August	$S = +185$		$S = +431$	$S = +372$	$S = +667$
	SS $= +0.009$		SS $= +0.009$	SS $= +0.012$	SS $= +0.015$
September				$S = +234$	$S = +631$
				SS $= +0.011$	SS $= +0.012$
October			$S = +336$	$S = +373$	$S = +660$
			SS $= +0.013$	SS $= +0.025$	SS $= +0.012$
November			$S = +726$	$S = +495$	$S = +439$
			SS $= +0.048$	SS $= +0.043$	SS $= +0.034$
December		$S = -263$	$S = +747$	$S = +464$	$S = +559$
		SS $= -0.022$	SS $= +0.056$	SS $= +0.038$	SS $= +0.033$

Positive value of SS and S signify positive trend and vice versa
Empty cell denotes no trend

3.3.2 Changes in Monthly Total Rainfall

The Mann–Kendall trend test and seasonal Mann–Kendall trend test were run on monthly total rainfall on the selected stations with intent to find any trend in rainfall over the years. The significance level, α was set to 5 % for both the tests. The period value was set to 12 for the seasonal Mann–Kendall trend test.

Table 3.3 summarizes the tests results. It tabulates the statistics (S), Sen's slope (SS), and p-values for the Mann–Kendall trend test. For the seasonal Mann–Kendall trend test, the statistics (S_k) and p-values are tabulated. p-Values less than 0.05 (significance level, α) indicate the presence of trend in the time series data. If there is a trend, positive values of statistics (S) and Sen's slope (SS) (for the Mann–Kendall trend test) denote positive trend in the series and vice versa.

From the results of the Mann–Kendall trend test, it can be seen that all the p-values are greater than 0.05. So, the null hypothesis cannot be rejected. It indicates that monthly total rainfall has seen no trend over the years in the selected regions, if seasonality is not taken into account.

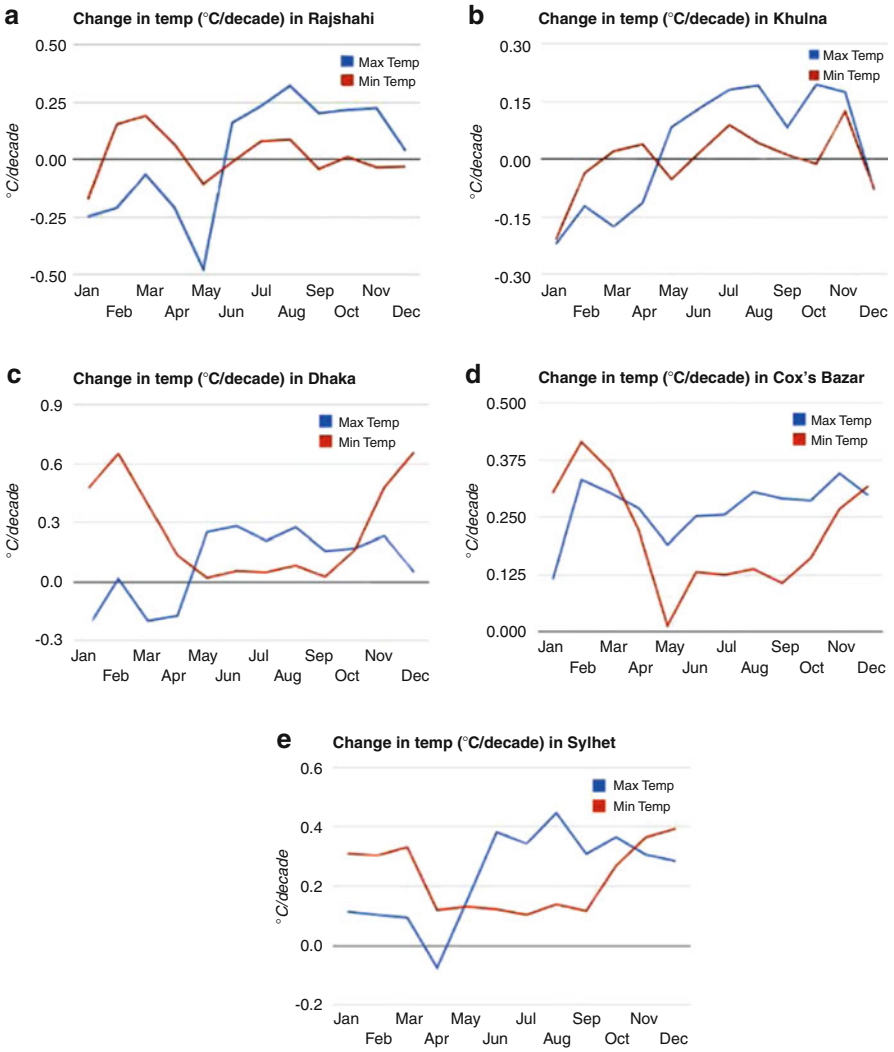

Fig. 3.1 Change in minimum and maximum temperature per decade (°C/decade). (**a**) Rajshahi, (**b**) Khulna, (**c**) Dhaka, (**d**) Cox's Bazar, and (**e**) Sylhet

The seasonal Mann–Kendall trend test shows the same result too, except in Cox's Bazar, where there is seen a positive change over the years.

3.3.3 Clustering of Rainfall

Applying *K*-means clustering algorithm, monthly total rainfall were partitioned into five different clusters. The clusters give five categories of monthly rainfall

Table 3.3 Outcome of Mann–Kendall trend test and seasonal Mann–Kendall trend test on monthly total rainfall data

	Mann–Kendall trend test				Seasonal Mann–Kendall trend test		
Stations	S	SS	p-Value	Trend	S_k	p-Value	Trend
Dhaka	−941	0	0.781	No	−279	0.324	No
Cox's Bazar	4,221	1,464	0.114	No	714	0.007	Positive
Sylhet	−529	0	0.438	No	−364	0.1	No
Khulna	−1,806	1,159	0.303	No	−13	0.484	No
Rajshahi	−1,649	0	0.313	No	−372	0.094	No

Table 3.4 Range of rainfall corresponding to different cluster number

Cluster	Monthly rainfall (mm)	Class
1	0–120	Negligible
2	121–322	Less
3	323–578	Moderate
4	579–966	High
5	967–3,017	Very high

based on its amount. Cluster 1 depicts the months with negligible or no rainfall where Cluster 5 represents the highest (Table 3.4). Thus over the years, the distribution of months in different clusters depicts a clear picture of rainfall of that region on the time period (Fig. 3.2).

In Cox's Bazar there is seen a negligible amount of rain during December to March, falling in Cluster 1. Then it increases from April, reaching its peak rainfall during June to August. During this period, huge amount of rainfall is observed, mostly falling in Clusters 4 and 5. Then it starts decreasing from September (Fig. 3.2a–c).

Study for Sylhet shows slightly less rainfall than Cox's Bazar, negligible rain during November to February, an increase from March. The peak rainfall (Cluster 3–5) is seen during May to August. Then starts to decrease from September (Fig. 3.2m–o).

In Dhaka negligible rainfall is experienced during November to March (Cluster 1). Then it increases from April, reaches peak during May to September (Clusters 2 and 3) and then starts decreasing (Fig. 3.2d–f).

In Khulna during November to February very less rainfall is seen (Cluster 1). Then it increases from March, remains constant in its peak rainfall during May to September (Clusters 2–4) and starts decreasing from October (Fig. 3.2j–l).

The study shows the least amount of rainfall in Rajshahi. Here, the rainfall constantly remains in Cluster 1 from November to April spanning half of the year. The peak rainfall is seen during June to September mostly residing in Clusters 2 and 3 (Fig. 3.2g–i).

Higher cluster number denotes higher amount of rainfall, the actual range can be found in Table 3.4.

The study also shows that over the years the distribution of rainfall has no significant change. Only the month of September shows some increase in Khulna from the year 1996.

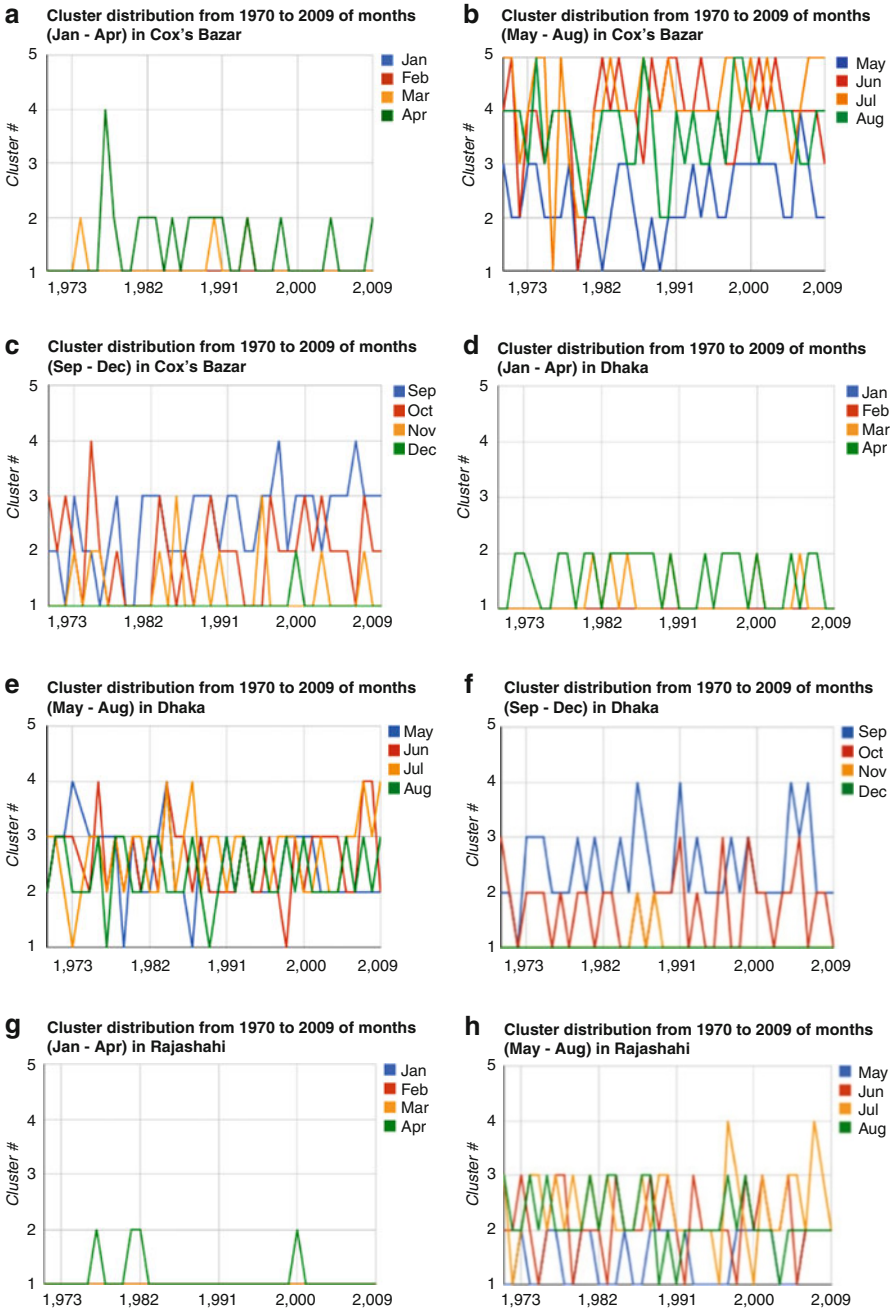

Fig. 3.2 Cluster distribution of rainfall of different regions throughout the year. (**a**) Cox's Bazar (Jan–Apr), (**b**) Cox's Bazar (May–Aug), (**c**) Cox's Bazar (Sep–Dec), (**d**) Dhaka (Jan–Apr), (**e**) Dhaka (May–Aug), (**f**) Dhaka (Sep–Dec), (**g**) Rajshahi (Jan–Apr), (**h**) Rajshahi (May–Aug), (**i**) Rajshahi (Sep–Dec), (**j**) Khulna (Jan–Apr), (**k**) Khulna (May–Aug), (**l**) Khulna (Sep–Dec), (**m**) Sylhet (Jan–Apr), (**n**) Sylhet (May–Aug), and (**o**) Sylhet (Sep–Dec)

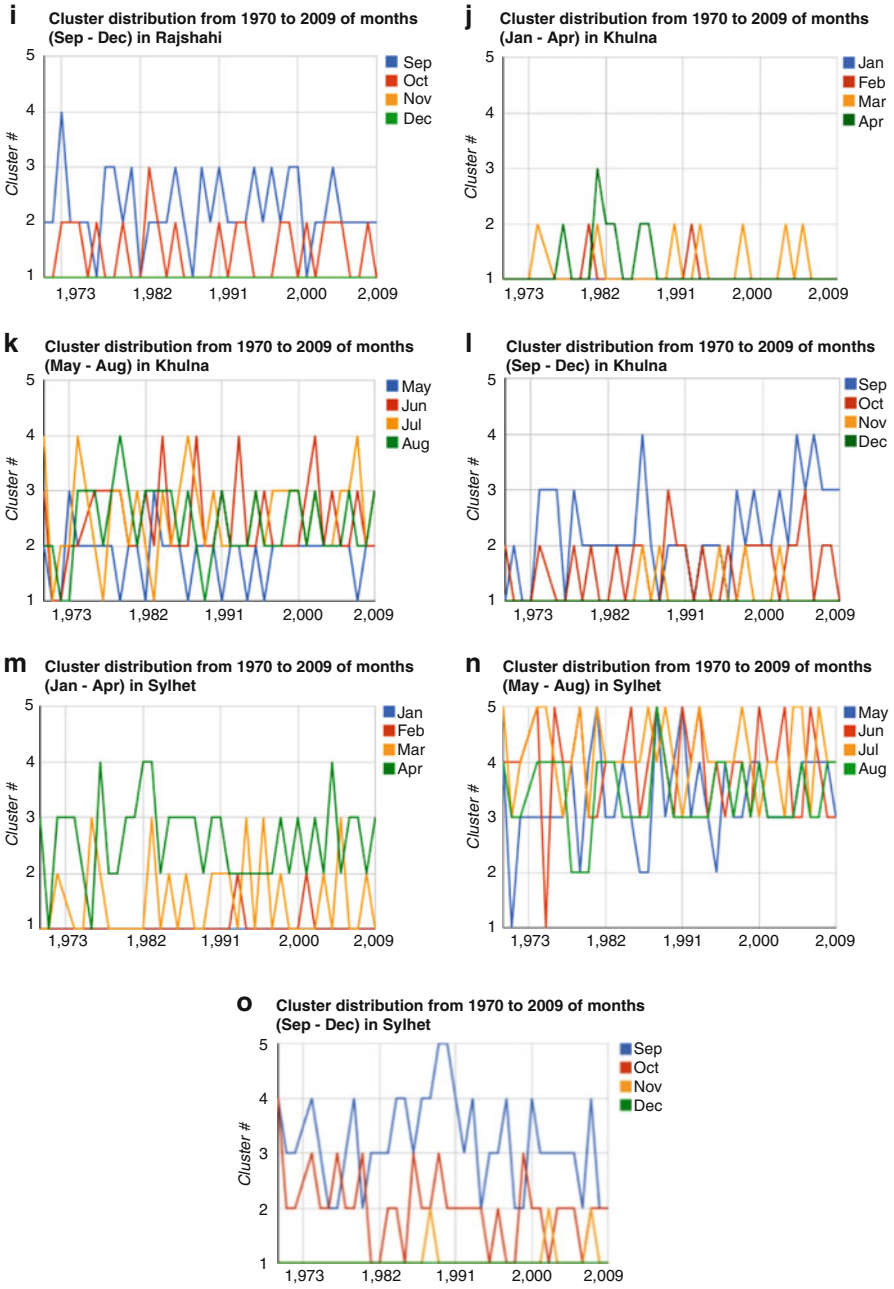

Fig. 3.2 (continued)

It is also evident from the distributions of clusters that in the eastern regions (Cox's Bazar and Sylhet) it rains remarkably more than the western regions (Rajshahi and Khulna).

3.3.4 Comparison of Different Time Series Forecasting Models

Different time series forecasting models were used to predict daily temperature. The models used are the Naive model, RWD model, Theta model, and ARIMA (1,1,0) model. For a particular month, the forecasting models were fed with the first 20 days' data (minimum and maximum temperature) and prediction of next 7 days were obtained. Then the RMS error and SMAPE were computed using the original values. This process was repeated for all the months from 1953 to 2009 for different stations.[1] Then the mean and standard deviation of the RMS error and SMAPE were computed to compare the accuracy of different models. In the cases, where the SMAPE and RMS error provides different ordering of the models, the ordering obtained from the SMAPE is considered for the discussion.

For ARIMA(p, d, q) model, the data is transformed to stationary time series by differencing with degree 1 ($d = 1$). For best fitting, the AR order (p) and MA order (q) were set to 1 and 0, respectively. ARIMA(1,1,0) is called autoregressive model with first difference.

Tables 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, and 3.14 tabulate the mean and standard deviation of the RMS error and SMAPE of minimum and maximum temperature for the selected stations. The models are sorted in descending order according to their mean of the SMAPE.

From the SMAPE and RMS error, it is evident that for all the selected regions, the RWD model shows the greatest SMAPE and RMS error for both minimum and maximum temperature. So it is the least applicable one among the selected models.

No simple ordering among the other three models can be found. Firstly, the scenario for minimum temperature is discussed. The Theta model provides best prediction for Dhaka, Sylhet, and Cox's Bazar while ARIMA and Naive model suits most for Rajshahi and Khulna region, respectively.

On the other hand, for prediction of maximum temperature, ARIMA model is most accurate for Rajshahi, Khulna, and Cox's Bazar. But for Dhaka and Sylhet, the Naive model, beating all the other models gives the best prediction.

The best models for different regions, for both maximum and minimum temperature, are listed in Table 3.15. In most of the cases the best model has the least SMAPE and RMS error simultaneously. In some cases where the SMAPE

[1] For time series prediction, missing data were replaced by the data of the previous day. Months, where more than 50 % data were missing, were not included in calculation.

Table 3.5 Mean RMS error and SMAPE of different models for Dhaka region (minimum temperature)

	RMS error		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	1.997	1.167	8.929	6.904
ARIMA	2.019	1.25	9.064	7.499
Naive	2.041	1.22	9.134	7.291
Random walk with drift	2.299	1.429	10.429	8.869

Table 3.6 Mean RMS error and SMAPE of different models for Rajshahi region (minimum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
ARIMA	2.005	1.204	9.936	8.04
Naive	2.016	1.167	9.946	7.841
Theta model	1.991	1.182	10.017	8.298
Random walk with drift	2.228	1.378	11.357	9.938

Table 3.7 Mean RMS error and SMAPE of different models for Sylhet region (minimum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	1.624	0.96	7.533	5.572
Naive	1.675	0.945	7.705	5.411
ARIMA	1.667	0.982	7.718	5.776
Random walk with drift	1.882	1.147	8.806	6.634

Table 3.8 Mean RMS error and SMAPE of different models for Khulna region (minimum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Naive	2.004	1.199	8.822	6.857
ARIMA	1.993	1.234	8.862	7.283
Theta model	1.982	1.227	8.929	7.357
Random walk with drift	2.226	1.381	9.906	8.071

Table 3.9 Mean RMS error and SMAPE of different models for Cox's Bazar region (minimum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	1.577	0.974	6.541	4.74
ARIMA	1.595	0.984	6.636	4.863
Naive	1.629	1.032	6.714	4.97
Random walk with drift	1.792	1.21	7.462	5.732

Table 3.10 Mean RMS error and SMAPE of different models for Dhaka region (maximum temperature)

	RMS error		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	2.163	1.331	6.17	4.114
ARIMA	2.145	1.426	6.144	4.879
Naive	2.126	1.329	6.05	4.18
Random walk with drift	2.391	1.56	6.871	4.949

Table 3.11 Mean RMS error and SMAPE of different models for Rajshahi region (maximum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
ARIMA	2.398	1.63	6.656	4.669
Naive	2.406	1.637	6.68	4.743
Theta model	2.41	1.621	6.674	4.542
Random walk with drift	2.713	1.884	7.585	5.49

Table 3.12 Mean RMS error and SMAPE of different models for Sylhet region (maximum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	2.655	1.535	7.809	4.924
Naive	2.617	1.542	7.699	5.053
ARIMA	2.618	1.527	7.711	5.114
Random walk with drift	2.934	1.884	8.721	6.32

Table 3.13 Mean RMS error and SMAPE of different models for Khulna region (maximum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Naive	2.229	1.348	6.162	4.135
ARIMA	2.22	1.378	6.14	4.267
Theta model	2.246	1.353	6.195	4.106
Random walk with drift	2.493	1.592	6.955	4.957

Table 3.14 Mean RMS error and SMAPE of different models for Cox's Bazar region (maximum temperature)

	RMS		SMAPE	
Model name	Mean	SD	Mean	SD
Theta model	1.861	1.054	5.387	3.38
ARIMA	1.807	1.052	5.192	3.396
Naive	1.821	1.093	5.222	3.572
Random walk with drift	2.041	1.289	5.946	4.304

Table 3.15 Listing of the most accurate models of different regions according to SMAPE

Region	Min temp	Max temp
Dhaka	Theta model	Naive
Rajshahi	ARIMA (Theta model)	ARIMA
Sylhet	ARIMA	Naive
Khulna	Theta model (Naive)	ARIMA
Cox's Bazar	Theta model	ARIMA

In cases where orderings obtained from SMAPE and RMS error differ, the best model according to RMS error is written in parentheses

and RMS error disagree about the best model, the SMAPE is given higher priority. The best model according to the RMS error is also listed in parentheses.

The distributions of prediction obtained from the models are then compared by means of their RMS error and SMAPE against the Naive model. The Naive model is an efficient forecasting model and often gives better results than other more

Table 3.16 Wilcoxon test results (*p*-value) on minimum temperature of different models against Naive model

Region		Arima	Theta	Random Walk
Dhaka	RMS Error	0.002	0.512	2.22E-23
	SMAPE	0.016	0.364	2.56E-22
Cox's Bazar	RMS Error	0.188	0.986	2.22E-23
	SMAPE	0.394	0.627	2.56E-22
Sylhet	RMS Error	0.037	0.578	4.66E-16
	SMAPE	0.051	0.771	4.69E-15
Rajshahi	RMS Error	0.177	0.858	4.07E-09
	SMAPE	0.22	0.539	6.02E-09
Khulna	RMS Error	0.677	0.624	2.79E-16
	SMAPE	0.613	0.804	6.24E-14

Table 3.17 Wilcoxon test results (*p*-value) on maximum temperature of different models against Naive model

Region		Arima	Theta	Random Walk
Dhaka	RMS Error	0.186	0.232	9.07E-17
	SMAPE	0.423	0.232	1.17E-17
Cox's Bazar	RMS Error	0.154	0.003	7.09E-19
	SMAPE	0.6	0	1.58E-19
Sylhet	RMS Error	0.555	0.01	2.13E-16
	SMAPE	0.215	0.005	2.61E-15
Rajshahi	RMS Error	0.4	0.429	4.05E-13
	SMAPE	0.726	0.325	7.80E-13
Khulna	RMS Error	0.021	0.519	1.03E-14
	SMAPE	0.083	0.611	6.21E-15

sophisticated models, providing a benchmark to compare. Wilcoxon signed rank test was used to determine whether a model follows the same distribution as Naive forecast or not. If the *p*-value obtained from Wilcoxon test is less than the significance level (0.05), it indicates that the model is different from Naive forecast and vice versa.

Wilcoxon tests were run for both minimum and maximum temperature of the selected stations with the SMAPE vector of different models against the SMAPE vector of the Naive model. The same tests were also run for the RMS errors. Tables 3.16 and 3.17 shows the *p*-values resulted from the tests. Cells, having *p*-values greater than 0.05, are colored red, indicating same distribution as the Naive model.

From the results it can be seen that, for time series prediction of temperature value in Bangladesh, the RWD model is significantly different from the Naive model. On the contrary, forecasts given by the Theta model and ARIMA model follow quite the same distribution as the Naive model as expected.

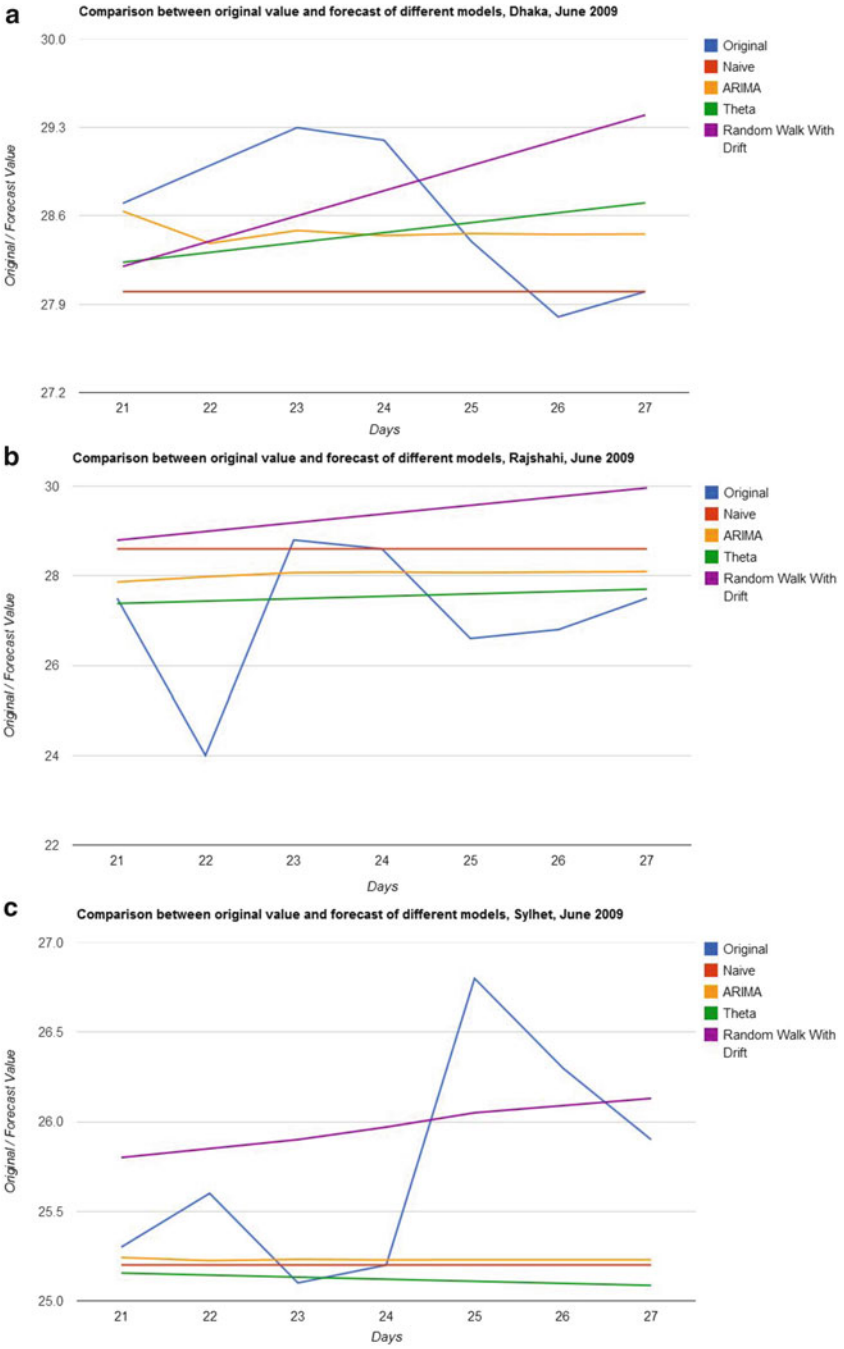

Fig. 3.3 Comparison between original value of temperature and forecast given by different models. (**a**) Dhaka, (**b**) Rajshahi, (**c**) Sylhet, (**d**) Khulna, and (**e**) Cox's Bazar

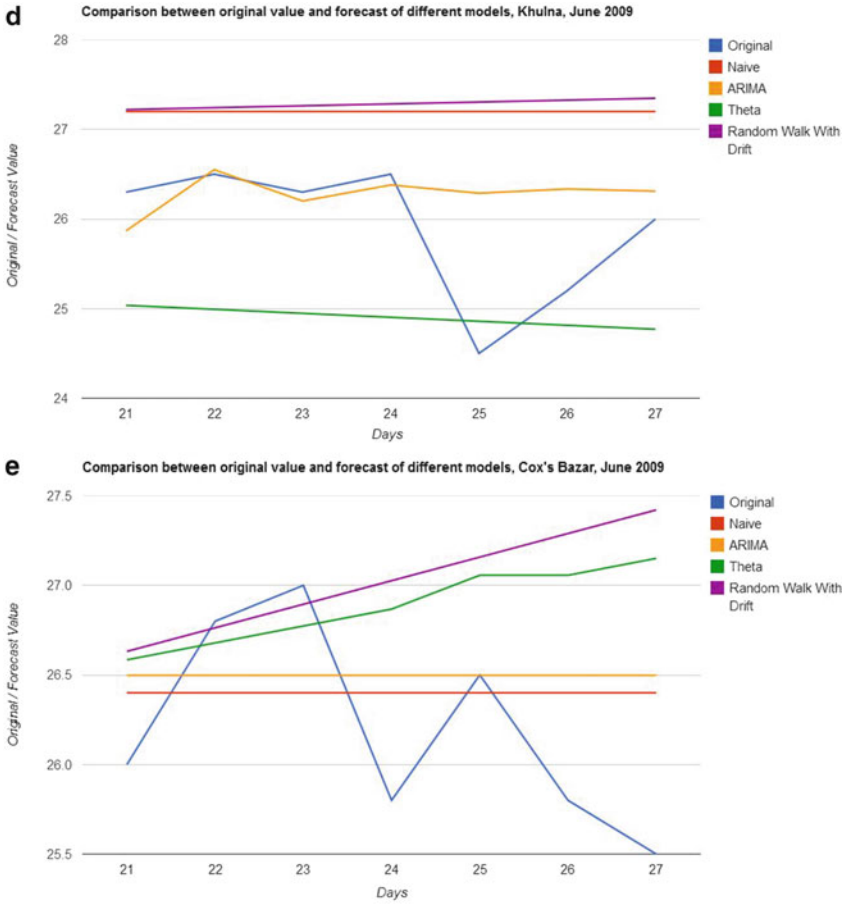

Fig. 3.3 (continued)

The original values and forecast corresponding to different models can be plotted to compare them visually. Figure 3.3 shows the comparison for June 2009 (arbitrarily chosen) for different regions.

3.4 Concluding Remarks

In this study the Mann–Kendall trend test and Sen's slope estimator were used to find the trends in temperature in Bangladesh. The study found that maximum temperature has shown remarkable positive trend during June to November in Bangladesh. On the other hand minimum temperature has increased during

December to January. The results also indicated that the eastern side has faced more change in temperature than the western side. Cox's Bazar and Sylhet exhibit an increasing trend almost throughout the year.

For analyzing the behavior of rainfall, the Mann–Kendall trend test, Sen's slope estimator, and seasonal Mann–Kendall trend test were used. *K*-means clustering algorithm was also employed to identify the rainfall distribution patterns over the years and their changes with time. The peak rainfall throughout the country is experienced during June to August and there has been no significant change in rainfall over the years. The study also reflects that over the years the western side of the country has experienced significantly less rainfall than the eastern side.

Performance of four time series prediction models (ARIMA(1,1,0), Theta, RWD, and Naive) were analyzed with respect to the climate condition of Bangladesh for forecasting daily minimum and maximum temperature. No obvious ordering could be found among ARIMA, Naive, and Theta models for prediction of daily minimum and maximum temperature, each one providing best prediction for different conditions. On the contrary, the RWD model is the least applicable one among the employed models.

References

Ahasan MN, Chowdhary Md AM, Quadir DA (2010) Variability and trends of summer monsoon rainfall over Bangladesh. J Hydrol Meteorol 7(1):1–17

Assimakopoulos V, Nikolopoulos K (2000) The theta model: a decomposition approach to forecasting. Int J Forecast 16:521–530

Bangladesh Meteorological Department (2012) http://www.bmd.gov.bd/

Basak JK, Titumir RAM, Dey NC (2013) Climate change in Bangladesh: a historical analysis of temperature and rainfall data. J Environ 2(2):41–46

Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco

Chowdhury MHK, Debsarma SK (1992) Climate change in Bangladesh—a statistical review. In: Report of IOC-UNEP workshop on impacts of sea level rise due to global warming, NOAMI, Intergovernmental Oceanographic Commission, Dhaka, 16–19 Nov 1992

Corte-Real J, Qian B, Xu H (1998) Regional climate change in Portugal: precipitation variability associated with large-scale atmospheric circulation. Int J Climatol 18:619–635

Debsarma SK (2003) Intra-annual and inter-annual variation of rainfall over different regions of Bangladesh. In: Proceedings of SAARC seminar on climate variability in the south Asian region and its impacts, SAARC Meteorological Research Centre, Dhaka, 20–24 Dec 2002

Gaoliao J, Zhiwei Z, Hailin Z (2012) Trend analysis of air temperature between 1979–2000 in Hubei Province. Paper presented at the World Automation Congress (WAC), Puerto Vallarta, 24–28 June 2012

Gilbert RO (1987) Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, New York

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18

Helsel DR, Hirsch RM (1992) Statistical methods in water resources. Elsevier, New York

Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. Water Resour Res 18(1):107–121

Huang F, Wang X (2011) Spatial and temporal variation of monthly rainfall nonuniformity of the upper Yangtze basin. Paper presented at international symposium on water resource and environmental protection, Xi'an, Shaanxi Province, 20–22 May 2011

Islam AS (2009) Analyzing changes of temperature over Bangladesh due to global warming using historic data. In: Proceedings of the young scientists of Asia conclave: pressing problems of humankind: energy & climate, Bangalore, 15–17 Jan 2009

Jain SK, Kumar V, Saharia M (2013) Analysis of rainfall and temperature trends in North East India. Int J Climatol 33(4):968–978

Karmakar S, Shrestha ML (2000) Recent climate changes in Bangladesh. In: SAARC Meteorological Research Centre (SMRC), SMRC-No. 4, SMRC Publication, Dhaka

Kendall M (1975) Rank correlation methods. Charles Griffin & Company, London, England

Kumar V, Jain SK (2010) Trends in seasonal and annual rainfall and rainy days in Kashmir valley in the last century. Quatern Int 212:64–69

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of fifth Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, pp 81–297

Mann HB (1945) Nonparametric tests against trend. Econometrica 13:245–259

Mia NM (2003) Variations of temperature in Bangladesh. In: Proceedings of SAARC seminar on climate variability in the south Asian region and its impacts, SAARC Meteorological Research Centre, Dhaka, 20–24 Dec 2002

Nahrin Z, Munim AA, Begum QN, Quadir DA (1997) Studies of periodicities of rainfall over Bangladesh. J Remote Sens Environ 1:43–54

Pelczer IJ, Cisneros-Iturbe HL (2008) Identification of rainfall patterns over the Valley of Mexico. Paper presented at the 11th international conference on urban drainage, Edinburgh, 31 Aug–5 Sept 2008

Pesaran MH, Pick A (2009) Forecasting random walks under drift instability. Cambridge working papers in economics, Faculty of Economics, University of Cambridge, Cambridge

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, http://www.R-project.org/

Ramos MC (2001) Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region. Atmos Res 57(2):123–138

Rana M, Shafee S, Karmakar S (2007) Estimation of rainfall in Bangladesh by using southern oscillation index. Pak J Meteorol 4(7):7–23

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

Singh P, Kumar V, Thomas T, Arora M (2008a) Basinwise assessment of temperature variability and trends in the northwest and central India. Hydrol Sci J 53:421–433

Singh P, Kumar V, Thomas T, Arora M (2008b) Changes in rainfall and relative humidity in different river basins in the northwest and central India. Hydrol Process 22:2982–2992

Tripathi S, Govindaraju RS (2009) Change detection in rainfall and temperature patterns over India. In: Proceedings of the third international workshop on knowledge discovery from sensor data, ACM, New York, pp 133–141

Warrick RA, Bhuiya AH, Mirza MQ (1994) The greenhouse effect and climate change: briefing document no. 1. Dhaka, Bangladesh Unnayan Parishad, pp 17–20

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull 1(6):80–83

Xi-ting L, Chun-qing G, Xiao Y (2011) Evolvement analysis about rainfall-runoff in the upper stream of Li River under the changeable environment. Paper presented at the international conference on remote sensing, environment and transportation engineering, Nanjing, 24–26 June 2011

Yue S, Hashino M (2003) Temperature trends in Japan: 1900–1990. Theor Appl Climatol 75: 15–27