Theoretical and Applied Climatology (2019) 136:1115–1131
https://doi.org/10.1007/s00704-018-2537-y

ORIGINAL PAPER

Comparison of missing value estimation techniques in rainfall data of Bangladesh

Farzana Jahan[1] · Narayan Chandra Sinha[2] · Md. Mahfuzur Rahman[3] · Md. Morshadur Rahman[4] · Md. Sanaul Haque Mondal[5] · M. Ataharul Islam[6]

Received: 9 May 2017 / Accepted: 6 June 2018 / Published online: 22 June 2018
© Springer-Verlag GmbH Austria, part of Springer Nature 2018

Abstract

The presence of missing values in daily rainfall data may hamper the analyses to determine effective results for solving problems of hydrological, agricultural, and climatological issues. The study attempts to select an appropriate method for estimating the missing value of daily rainfall data of Bangladesh. For this purpose, eight methods and seven comparison techniques are employed. For imputation of missing values employing these methods, three sets of daily rainfall data (1, 5, and 10% missing values) with 1000 repetitions are considered randomly for five regions of the country. These samples are artificially created as missing and then imputation for these missing values is made applying the selected methods. The relative performance of the methods are examined using some comparison criteria. The following observations can be made from the study regarding the choice of the appropriate missing value estimation technique: for imputation of the missing values of daily rainfall data, the arithmetic average method for rainfall stations Chittagong and Rajshahi in the south-east region and the north-west region, respectively, is found as the best methods. Further, the single best estimator method for rainfall stations Sylhet and Dhaka in the north-east region and the mid-region, respectively, and the EM-MCMC method for rainfall station Khulna of the south-east region are also identified as the best methods in respect of Kolmogorov-Smirnov test, the lowest bias of estimate, the value of *S* index, etc.

1 Introduction

Rainfall is an important factor in the field of hydrological study. The occurrences of rainfall provide the input of crop growth and production models. It also indicates the situation of landfills, tailing dams, and land disposal of liquid waste materials which are environmentally sensitive to any region or overall country. Generally, the rainfall amount is measured in daily time scale method, and then, it may be converted into a monthly or annual series. Therefore, the analysis of rainfall plays a significant role in the field of agriculture, ecology, and climatology studies (Asati 2012; Williams 1998; Cong and Brady 2012; and Silva et al. 2007). Besides, it is a highly influential factor for flood formation. Rainfall data analysis is always hampered by the shortage of consecutive data (Silva et al. 2007; Simolo et al. 2010). The presence of missing values in the rainfall data of different countries in the world is a common problem for data analysis. Rainfall data may be missing for various reasons such as loss of yearbooks, human errors, wars, fire accidents, occasional interruptions of automatic stations, instrument malfunctions, and network reorganizations (Simolo et al. 2010). A similar situation may also be observed in Bangladesh.

For performing the effective analysis of rainfall, it is essential to estimate the missing value of daily rainfall data. For this purpose, different authors have suggested suitable methods for estimating the missing values for specific countries or regions using several comparison techniques to the missing data

✉ Farzana Jahan
kakon18@yahoo.com

[1] School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD, Australia

[2] Dhaka School of Economics, Dhaka 1000, Bangladesh

[3] Green Business School, Green University of Bangladesh, Dhaka 1207, Bangladesh

[4] Department of Statistics, University of Dhaka, Dhaka 1000, Bangladesh

[5] Tokyo Institute of Technology, Tokyo, Japan

[6] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

estimation methods. Because the performance of any method for estimating missing values generally depends on the nature of the missing mechanism, nature of consecutive occurrences of rainfall, nature of neighboring stations, other intrinsic characteristics of the climate variables, etc. (Little and Rubin 1987).

To estimate the missing value of daily rainfall data, Silva et al. (2007) and Suhalia et al. (2008) have compared different methods such as inverse distance, normal ratio, arithmetic mean and aerial precipitation ratio, inverse weighting distance, and correlation coefficient method for Sri Lanka and for Malaysia, respectively, following the suggestions of Simanton and Osborn (1980), Tabios and Salas (1985), Young (1992), Hubbard (1994), Lennon and Turner (1995), Tang et al. (1996), Xia et al. (1999), Eischeid et al. (2000), Teegavarapu and Chandramouli (2005), Ahrens (2006), Garcia et al. (2008), and Chen and Liu (2012). For comparing these methods, they used techniques such as similarity index (S index), mean absolute error (MAE), and coefficient of correlation (R).

Further, Lo Presti et al. (2010) identified the Theil method as the best among the regression-based methods (simple substitution, parametric regression, ranked regression, and Theil method) for estimating the missing value of daily rainfall data of Candelaro River Basin, Italy. Besides, Coulibaly and Evora (2007) suggested artificial neural network (ANN) algorithms for imputation of daily rainfall missing precipitation. This algorithm is adapted on the basis of weighted interpolation technique from adjacent stations. Yozgatligil et al. (2013) suggested the Monte Carlo Markov Chain based on expectation-maximization (EM-MCMC) algorithm as the best technique for estimating missing value for the Turkish meteorological data. These studies indicate that to estimate the missing value of daily rainfall data for different stations, different techniques are found appropriate for separate station or region. Therefore, to analyze the daily rainfall data of different rainfall stations of Bangladesh, a suitable missing value estimation technique is essential for separate stations or regions.

Bangladesh is an agro-based country. Around 50% of the country's labor forces are engaged in this sector. Its contribution on the gross domestic product (GDP) is 15.33% in the overall growth of 7.05% for the FY 2015–2016 (Bangladesh Economic Review 2016). It indicates that the analysis of daily rainfall data has a significant role in the development of agricultural sector. Therefore, to analyze the daily rainfall data of Bangladesh, several authors applied different simple techniques for replacing or handling the missing data problems, such as omission of the missing data, replacing the missing values in a month by average value of the same month from previous, and subsequent years (Kripalani et al. 1996). However, none of the works has been done till date to identify the best method for estimating the missing value of daily rainfall data for different stations in Bangladesh. Therefore, the study is an attempt to compare several missing value estimation methods and suggests a suitable method to estimate the missing value of daily rainfall data for different rainfall stations of Bangladesh.

Following this section, the study is organized as below. The daily rainfall data and the behavior of daily rainfall missing data are discussed in Sect. 2. The different methods and their comparison techniques to identify the best method for estimating the missing value of daily rainfall data for target stations are also discussed in the same section. The discussions regarding the results obtained by applying the selected methods and comparison techniques are depicted in Sect. 3. Finally, the conclusions of the study are drawn in Sect. 4.

2 Data and methods

2.1 Data

To perform the above objective, this study considers 27 out of 35 daily rainfall recording stations under Meteorological Department of Bangladesh. The metric unit millimeter is the measurement unit of daily rainfall data. These stations record daily rainfall data for consecutive days. We have considered five climatic sub-zones of Bangladesh according to the geographical condition such as south-east region, north-east region, mid-region, south-west region, and north-west region (Rashid 1991). From each climatic sub-zone, one station is considered as target station and the stations surrounding 100 km of it are considered as reference stations (Tronci et al. 1986). The climatic sub-zone-wise daily rainfall measuring stations, sub-zone-wise target and reference stations, the availability of rainfall data for corresponding stations, and station-wise geographical condition are shown in Table 2.

2.1.1 Overview of missing data for selected stations

To perform the study, it is identified that each considered rainfall station contains some missing data. The proportions of missing data in percentage for available years in the selected stations are displayed in Fig. 1. It indicates that the percentage of missing observations in the 27 stations varies from 1.9% in Ambagan to 6.6% in Hatiya. The study considers different methods for estimating the missing value of daily rainfall observations including different comparison techniques for these methods to identify the best method for each of the selected stations.

2.1.2 Missing data mechanism

The problems of missing data may arise due to different observational behaviors. Under probabilistic response, the patterns of missing data may be classified into three phenomena:

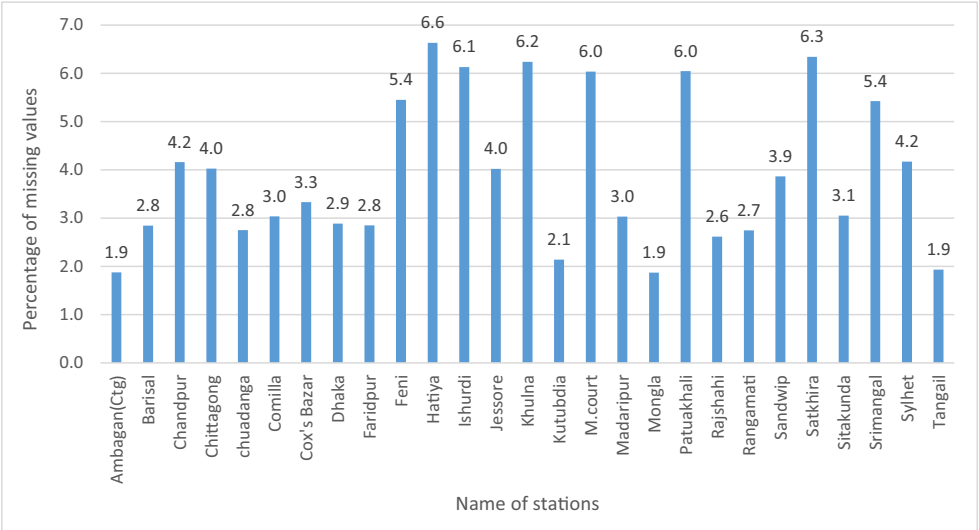

Fig. 1 Percentage of missing values (ratio of number of missing observations to total number of observations of rainfall in each station) of daily rainfall data for available years of each selected station for Bangladesh

missing at random (MAR), missing completely at random (MCAR), and not missing at random (NMAR) (Rubin 1976; Schafer 1997; Little and Rubin 1987). The Probability that missing data of daily rainfall observations depends on the observed responses but not on missing data itself indicates the MAR. The probability that the missing data of daily rainfall observations does not depend on its own values or observed data provides MCAR; i.e., MCAR is the special case of MAR. The probability that the missing data of daily rainfall depends on the value of missing observations itself can be termed as NMAR.

To measure the patterns of missing data different authors (Dempster et al. 1977; Little and Rubin 1987; Rubin 1987; Schafer 1997; Collins et al. 2001; Graham et al. 1997) have suggested different techniques such as maximum likelihood (ML) estimation method and multiple imputation (MI) method under expectation-maximization (EM) algorithm based on Bayesian framework, following the indication of Rubin (1976). Because the formulation of a statistical model using NMAR data creates different complexities, such as the missing data model may not be correctly specified, the estimated parameters may contain sizable bias, etc. Therefore, for testing the existence of NMAR mechanism in the daily rainfall missing observations, Lo Presti et al. (2010) suggested to verify the following statements:

(i) The existence of a positive correlation between the missing data (yearly percentage of days with missing data in each station) and the elevation of the stations and
(ii) The amount of missing data are affected by evident seasonal behavior; for instance, monsoon and autumn seasons are more rainy than summer, late autumn, and winter seasons.

To verify the statement (i), the study observed that the correlation coefficient between the missing data and the elevation of the corresponding station is found negative ($r = -0.133$ with p value 0.507). Although this is not significant, the value of the correlation coefficient appears to be negative implying non-positive correlation between elevation and proportion of missing data of corresponding stations. The result indicates that the daily rainfall missing observations for different stations of Bangladesh do not follow NMAR mechanism. Further, to verify the statement (ii), Lo Presti et al. (2010) suggested standardized entropy (H) which is stated as below:

$$H = -\frac{\sum_{k=1}^{K}\{[\ln p(k)] \times p(k)\}}{\ln k} \quad (1)$$

where $p(k)$ is the proportion of missing observations to the total number of rainfall observations for a station at the k^{th} month during the study period and lnk indicates the upper boundary of the measurement months. The value of standardized entropy (H) close to 1 indicates that the missing data distribution for study period is not affected by the seasonal behavior; i.e., the hypothesis NMAR may be rejected. For instance, the Table 1 shows the measurement result of standardized entropy for south-east region's target station and its reference stations of Bangladesh. The standardized entropy is found near to 1 for the selected stations of south-east region of the country, which indicates that the distribution of missing observations of rainfall data does not follow NMAR (Table 1). Similar results are also observed for other regions of the country.

Further, Lo Presti et al. (2010) indicated that the measurement of MCAR for the missing data of rainfall observations always depends on the efficient measurement of rainfall amount. To measure the rainfall amount, Rubel and Hantel (1999) identified three leading sources of error: (i) wind-

Table 1 Proportion of missing observations per month and standardized entropy for study period (2011–2014) of south-east region taking Chittagong as the target station and its reference stations

Month	Chittagong	Cox's Bazar	Feni	Hatiya	M. Court	Sitakunda	Sandwip	Rangamati
January	0.24	0.23	0.17	0.27	0.31	0.32	0.23	0.31
February	0.00	0.22	0.16	0.20	0.21	0.31	0.21	0.33
March	0.19	0.24	0.26	0.21	0.17	0.32	0.22	0.31
April	0.29	0.23	0.17	0.21	0.19	0.31	0.32	0.30
May	0.23	0.02	0.17	0.15	0.14	0.18	0.13	0.32
June	0.15	0.11	0.18	0.07	0.20	0.22	0.09	0.14
July	0.19	0.22	0.33	0.19	0.22	0.03	0.02	0.03
August	0.15	0.04	0.25	0.12	0.22	0.00	0.22	0.05
September	0.18	0.21	0.25	0.19	0.21	0.03	0.22	0.03
October	0.02	0.24	0.25	0.23	0.15	0.03	0.31	0.00
November	0.20	0.22	0.10	0.26	0.20	0.18	0.22	0.00
December	0.34	0.31	0.00	0.27	0.18	0.03	0.00	0.00
Standardized entropy (H)	0.88	0.92	0.91	0.96	0.97	0.79	0.88	0.73

Proportion of missing values in each month is calculated as a ratio of number of missing observations to total number of observations in the whole data set for a specific month

induced losses, (ii) wetting of the walls and evaporation from the tipping bucket, and (iii) instrumental accuracy and precision, which lead to underestimation of the actual rainfall amount. Except these, several secondary sources of error affect the measurement of rainfall amount such as splash in, splash out, wind shield, and temperature (Lo Presti et al. 2010; Goodison et al. 1998).

Bangladesh Meteorological Department (BMD) measures the rainfall observation in each station using natural siphon rainfall recorders and Snowdon rain gauge (Chowdhury 2013). Recently, this technique is highly popular for the efficient measurement of rainfall observation; however, there may also arise some reasonable errors, such as influence of other variables, instrumental failure, weak efficiency, and precision of technician. Considering these arguments, MCAR mechanism may not be appropriate for the missing data distribution of rainfall of the country. Besides, Rubin (1976) and Scheffer (2002) stated that the missing data of rainfall observations are very rare to follow MCAR. That is, the rejection of MCAR hypothesis leads us to consider the MAR mechanism for missing data distribution of rainfall observations in Bangladesh.

2.2 Methods

To estimate the missing value of daily rainfall observations, several authors employed different methods which are already discussed in Sect 1. The present study employed eight methods for estimating the missing values and made their comparison following some comparison measures. For performing the study, daily rainfall data from the year 2011 to 2014 (total number of days, $n = 1461$) are considered for each of the five target stations. From each target stations, 1% (sample size, $n = 14$), 5% (sample size, $n = 73$), and 10% (sample size, $n = 146$) non-missing observations are chosen randomly, and these are artificially created missing values. The actual values of those days are considered as observed values. Thereafter, different methods for estimating missing values are employed and their comparisons are made to identify the suitable method for each target station. This random process for sample selection, estimation process, and comparison techniques are repeated 1000 times. In the end, the arithmetic mean of the comparison measures of those 1000 repetitions is considered for the final decision for choosing the best missing value estimation technique.

2.2.1 Methods of missing value estimation for daily rainfall data

The methods employed in the study for estimating the missing values of daily rainfall data are discussed in this section. Let Y_{mi} indicates the missing value of m^{th} day of i^{th} target station in the study period (2011–2014) which is to be estimated, and Y_{mj} indicates the rainfall amount of m^{th} day of j^{th} reference station, where $i = 1,2,3, \ldots, n$ and $j = 1,2,3, \ldots, k$.

Arithmetic average (AA) method To estimate the missing value of daily rainfall observations, this method is used generally (Silva et al. 2007; Xia et al. 1999; Yozgatligil et al. 2013). In this method, missing values are estimated by the arithmetic average of concurrent observations of the reference stations

which have similar features with the target station (Paulhus and Kohler 1952). The arithmetic average for estimating the missing value of m^{th} day of i^{th} target station is given by

$$\hat{Y}_{mi} = \frac{\sum Y_{mj}}{k} = \frac{1}{k}(Y_{m1} + Y_{m2} + \ldots + Y_{mk}) \quad (2)$$

Normal ratio (NR) method Paulhus and Kohler (1952) proposed the method for spatial interpolation using weights, W_i. Afterwards, several authors used the method for imputing the missing value of daily rainfall data. The weights are estimated by the ratio of total annual rainfall amount for target station, T_i, to the total annual rainfall amount for each reference station, T_j. Then, the NR method is explained as (Yozgatligil et al. 2013)

$$\hat{Y}_{mi} = \frac{1}{k}\sum_j^k \frac{T_i}{T_j} Y_{mj} = \frac{1}{k}\left(\frac{T_i}{T_1}Y_{m1} + \frac{T_i}{T_2}Y_{m2} + \ldots\ldots + \frac{T_i}{T_k}Y_{mk}\right) \quad (3)$$

Normal ratio method considering the weight as correlation function (NRWC) Young (1992) modified the NR method considering the weight as correlation function instead of proportion to annual rainfall amount of target station to the reference station for the selected period in which missing value exists. To formulate the method, the weight is defined as

$$w_{ij} = \left[r_{ij}^2\left(\frac{n_{ij}-2}{1-r_{ij}^2}\right)\right] \quad (4)$$

where r_{ij} is the correlation coefficient between the i^{th} target station and j^{th} reference station and n_{ij} is the number of rainfall observations for measuring correlation coefficient. Then, the NRWC is defined as

$$\hat{Y}_{mi} = \frac{1}{\sum_j^k w_{ij}}\sum_j^k w_{ij} Y_{mj} = \frac{1}{\sum_j^k w_{ij}}(w_{i1}Y_{m1} + w_{i2}Y_{m2} + \ldots\ldots + w_{ik}Y_{mk}) \quad (5)$$

Inverse distance (ID) method Shepard (1968) suggested the method for analyzing two-dimensional interpolation functions for irregularly spaced data. Then, various authors used this method for estimating the missing value of daily rainfall observations (Lam 1983; Tronci et al. 1986; Hubbard 1994; Xia et al. 1999; Eischeid et al. 2000). The method is explained as the weighted interpolation technique which is defined as

$$\hat{Y}_{mi} = \sum_{j=1}^k w_{ij} Y_{mj} \quad (6)$$

where weight, w_{ij} is explained as:

$$w_{ij} = \frac{d_{ij}^{-p}}{\sum_j^k d_{ij}^{-p}} \quad and \quad \sum_{j=1}^k w_{ij} = 1 \quad (7)$$

Here, p indicates the exponent of inverse distance and d_{ij} indicates the distance of proximity neighboring j^{th} reference station from i^{th} target station. To calculate the distance, d_{ij} from i^{th} to j^{th} station, the latitude and longitude values of the respective stations are used. Latitude and longitude values of each station are converted into decimal degrees. Then, the distance from i^{th} to j^{th} station is computed using *Great Circle Calculator of National Hurricane Center of USA* (National Hurricane Center of USA n.d).

The method is used to estimate the missing observations of meteorological or hydrological variables under interest for assigning more weight to closer points. That is, weight is decreased as the distance from the interpolated points increase. The higher value of exponent p indicates a high influence of closer values to the interpolated point (Suhalia et al. 2008). Xia et al. (1999) indicated that usual value of p ranges from 1.0 to 6.0, and this value is generally considered as 2. Thus, the study considers the value of p as 2.

Multiple imputation using EM-MCMC method To estimate missing value of the data set, the multiple imputation method is developed by Rubin (1976, 1978) to overcome the uncertainty of the missing value estimates which rises due to the insufficient measurement of sampling variability. The method demonstrates that the missing values are imputed by estimating the parameters of the appropriate model to incorporate the random variation of multiple times and the average of multiple values. Then, to interpolate the missing data, the Monte Carlo Markov chain method-based expectation-maximization (EM-MCMC) algorithm is employed on the basis of Bayesian sampling procedure as the multiple imputation method (Tanner and Wong 1987; Schafer 1997). The method considers missing data according to proportional information of the sample to estimate the parameter of interest through conditional expectations. Therefore, the EM algorithm provides an estimation of parameters and imputations using MCMC procedure under iteration method (Yozgatligil et al. 2013).

The daily rainfall data always contains incomplete data with two types of observations (non-missing and missing value); these observations are explained as $Y = (Y_{oi}, Y_{mi})$. Here, Y_{oi} and Y_{mi} indicate the non-missing value and missing value of rainfall data, respectively, of i^{th} day. To perform the multiple

imputation techniques using EM-MCMC algorithm based on the Bayesian framework, the unknown θ and Y_{mi} are considered as random variables for the performance of statistical inference on the parameter θ (Schafer 1997). Then, the posterior predictive distribution is stated as

$$f(Y_{m\,i}|Y_{o\,i}) = \int f\left(Y_{m\,i},\ \theta|Y_{o\,i}\right)d\theta = \int f(Y_{m\,i}|Y_{o\,i},\theta)\, f(\theta|Y_{o\,i})d\theta \quad (8)$$

where the functions $f(Y_{m\,i}|\,Y_{o\,i},\,\theta)$ and $f(\theta|\,Y_{o\,i})$ indicate the conditional predictive distribution of Y_{mi} and the posterior distribution of θ in respect of the non-missing value of rainfall observations, respectively. The posterior distribution, $f(\theta|\,Y_{oi})$, is determined through the intensification of Y_{oi} providing the assumed value of Y_{mi}, which is measured by two-step procedure (Yozgatligil et al. 2013). The first step is to impute the missing value, Y_{mi}, from the conditional predictive distribution, $f(Y_{mi}|\,Y_{oi},\,\theta)$in the k^{th} step, i.e.,

$$Y_{mi}^{(k+1)} \sim f\left(Y_{mi}|Y_{oi},\ \ \theta^{k}\right) \quad (9)$$

The second step provides the new value of θ from the posterior distribution of non-missing data given the missing data.

$$\left(Y_{oi}, Y_{mi}^{(k+1)}\right),\ \text{ i.e., }\ \theta^{(k+1)} \sim f\left(\theta|Y_{oi}, Y_{mi}^{(k+1)}\right) \quad (10)$$

These two steps are repeated through the iteration process starting with initial value as $\theta^{(0)}$, and the process yields a Markov chain, i.e., $\left(Y_{mi}^{(1)},\ \theta^{(1)}\right)$, $\left(Y_{mi}^{(2)},\ \theta^{(2)}\right)$, $\left(Y_{mi}^{(3)},\ \theta^{(3)}\right)$ and so on.

The distribution of these transition counts of the Markov chain provides the joint conditional distribution, $f(Y_{mi},\ \theta|\,Y_{oi})$. If the value of parameter $\theta^{(k)}$ satisfies the convergence of distribution, then the posterior distribution, $f(\theta|\,Y_{oi})$, is drawn from non-missing data using this value of the parameter. Then, from the posterior predictive distribution, $f(Y_{mi}|\,Y_{oi})$, the $Y_{mi}^{(k)}$ is considered as an appropriate selection. This method is perfectly valid, provided that the missing data of rainfall observations do not follow the NMAR mechanism (Scheffer 2002).

The whole process of multiple imputations using EM-MCMC method can be done by using PROC MI in the University Edition of SAS (Yim 2015). This study used PROC MI to make the multiple imputations of daily rainfall missing data for the target stations using the concurrent rainfall data of reference stations as covariates. The underlying distribution of the data is considered to be multivariate normal in this study.

Single best estimator (SBE) method To estimate the missing value of daily rainfall data, various authors employed this method (Wallis et al. 1991; Xia et al. 1999; Eischeid et al. 2000). For performing this method, the daily rainfall data of proximity neighboring station corresponding to the missing data of target station is considered as the estimated missing value, provided that the data of neighboring and target station would have the highest positive correlation. This is analogous to the simple substitution or closest neighboring station method (Lo Presti et al. 2010; Garcia et al. 2006). To select proximity neighboring station to the target station, minimum distance with the target station is considered, because the rainfall amount of closest neighboring station and the target station always provide highest positive correlation compared to the other neighboring stations. For instance, in mid region of the study, Faridpur is found to be the closest station to target station Dhaka (distance 57 km), and in the south-east region, Ambagan is found to be the closest station to the target station Chittagong (distance 15 km) (Table 2). The distance measurement procedure is discussed in the inverse distance method.

Linear regression (LR) method To formulate the linear regression method for estimating the missing data of daily rainfall occurrences, the study considers the following estimated form (Dumedah and Coulibaly 2011; Xia et al. 1999):

$$\hat{Y}_{mi} = \hat{\alpha} + \hat{\beta}X_{mj}, \qquad i = 1, 2, \ldots, n \quad (11)$$

where $\hat{Y}_{mi}$ indicates the estimated value of missing rainfall observation of m^{th} day for i^{th} target station and X_{mj} indicates the observation of m^{th} day rainfall of the closest reference station j. The closest reference station is selected by considering the minimum distance to the target station within the neighboring stations. Here, $\hat{\alpha}$ and $\hat{\beta}$ are the parameters which are estimated by using least squares method from the simple linear regression model. To estimate the parameters (α and β), the daily rainfall observations of i^{th} target station and proximity neighboring j^{th} reference station are considered as dependent and independent variables, respectively.

Multiple regression (MR) method Kemp et al. (1983), Tabony (1983), Young (1992), and Eischeid et al. (1995) explained different facilities of the regression model for data interpolation and missing data estimation. Following their suggestions, Xia et al. (1999) indicated multiple regression method for estimating the missing value of daily rainfall occurrences. Therefore, for estimating the missing value of daily

Table 2 Classification of 27 selected stations according to climatic sub-zones with geographic position and data availability

Climatic sub-zones	Station Name	Years of data availability	Elevation (Meter)	Lat. (N)		Long. (E)		Distance from target to reference station (km)	Correlation between target and reference station's for rainfall amount
				Deg.	Mts.	Deg.	Mts.		
South east region	Ambagan	1999-2015	5.5	22	13	91	48	15	0.9155934**
	Chittagong*	1949-2015	33.2	22	21	91	49	0	-
	Cox's Bazar	1948-2015	2.1	21	27	91	58	101	0.5692109**
	Feni	1973-2015	6.4	23	02	91	25	86	0.5057946**
	Hatiya	1966-2015	2.44	22	27	91	06	75	0.546802**
	Kutubdia	1977-2015	2.74	21	49	91	51	59	0.7158667**
	M.Court	1951-2015	4.87	22	52	91	06	94	0.5001792**
	Rangamati	1957-2015	68.89	22	38	92	09	51	0.5505097**
	Sandwip	1966-2015	2.1	22	29	91	26	43	0.6450881**
	Sitakunda	1977-2015	7.3	22	38	91	42	33	0.6464453**
North east region	Srimangal	1948-2015	21.95	24	18	91	44	68	0.3093841**
	Sylhet*	1956-2015	33.53	24	54	91	53	0	-
Mid region	Dhaka*	1953-2015	8.45	23	46	90	23	0	-
	Madaripur	1977-2015	7	23	10	90	11	70	0.4890244**
	Faridpur	1948-2015	8.1	23	36	89	51	57	0.6027658**
	Chandpur	1964-2015	4.88	23	14	90	42	68	0.4340542**
	Tangail	1987-2015	10.2	24	15	89	56	70	0.4459477**
	Comilla	1948-2015	7.5	23	26	49	51	90	0.4401898**
South west region	Jessore	1948-2015	6.1	23	12	89	20	53	0.4562205**
	Khulna*	1948-2015	2.1	22	47	89	34	0	-
	Barisal	1949-2015	2.1	22	43	90	22	82	0.4979408**
	Patuakhali	1973-2015	1.5	22	20	90	20	93	0.4495002**
	Mongla	1991-2015	1.8	22	28	89	36	35	0.6060664**
	Satkhira	1948-2015	3.96	22	43	89	05	51	0.5904685**
North west region	Chuadanga	1989-2015	11.58	23	39	88	49	81	0.4504169**
	Ishwardi	1961-2015	12.9	24	09	89	02	41	0.5079663**
	Rajshahi*	1964-2015	19.5	24	22	88	42	0	-

Data Source: Bangladesh Meteorological Department

*Indicates the target station, and remaining stations indicate the reference stations for each region

**Indicates p value <2.2e-16

rainfall occurrences, the study considers the following estimated multiple regression model as an interpolation method:

$$\hat{Y}_{mi} = \hat{\alpha} + \sum_{j=1}^{k} \hat{\beta}_j X_{mj}, \quad i = 1, 2, \ldots, n \tag{12}$$

where $\hat{Y}_{mi}$ indicates the estimated value of rainfall observation of the m^{th} day in the i^{th} target station and X_{mj} indicates the observation of m^{th} day of the j^{th} reference station (j= 1,2,3,...,k; where k is the number of reference stations of station i. Here, $\hat{\alpha}$ and $\hat{\beta}_j$ are the parameters which are estimated by using least squares method from the multiple regression model. To estimate the parameters (α and β), the daily rainfall observations of i^{th} target station and j^{th} reference stations are considered as dependent and independent variables, respectively.

2.2.2 Techniques of comparison for the missing value estimation methods

To identify the appropriate matching between observed and expected observations, the following comparison criteria are considered in the study. For calculating the value of each comparison criterion, firstly, the study considers randomly selected portion of data as missing although there exist observed observations for target station of daily rainfall data, and then, these values are estimated by using different missing value estimation techniques. These estimated values of daily rainfall missing data are considered as the expected values $\left(Y_i^{\text{est}}\right)$, and these are compared with the observed amount of observations $\left(Y_i^{\text{obs}}\right)$. Here, $i (i = 1, 2, \ldots, n)$ indicates the number of sample observations.

Kolmogorov-Smirnov (K-S) test Kolmogorov-Smirnov test for goodness of fit would be used to determine whether a method provides good estimates of missing values or not (Massey 1951; Wilks 1995; Simolo et al. 2010). It uses the cumulative frequency distribution function, say $F_n(x)$-based non-parametric test. Here, x indicates any specific value of daily rainfall data and $F_n(x)$ indicates the proportion of cumulative frequency of individuals for the daily rainfall distribution. Further, $S_n(x)$ indicates the proportion of cumulative frequency of individuals for the estimated daily rainfall distribution. Then, the Kolmogorov-Smirnov test statistic for goodness of fit is defined as

$$D_n(x) = \max_x |F_n(x) - S_n(x)| \tag{13}$$

If the p value of above statistic is large, then the estimated daily rainfall observations provide a good fit to the observed rainfall observations.

Bias or mean of error (ME) In the concepts of statistics, bias indicates the difference between the estimator's expected value and the true value of the parameter. If this result is 0 (zero), it indicates unbiased estimation (Walther and Moore 2005). Therefore, the study considers differences between the observed value of daily rainfall amount $\left(Y_i^{\text{obs}}\right)$ and the estimated value of daily rainfall missing observation $\left(Y_i^{\text{est}}\right)$ for the corresponding observed value indicate the errors. Then, the mean of errors indicates the bias of estimate which is stated as (Simolo et al. 2010)

$$ME = n^{-1}\sum_{i=1}^{n}\varepsilon_i, \quad where \ \varepsilon_i = Y_i^{obs.} - Y_i^{est.} \tag{14}$$

The bias is calculated for all estimation methods and the method with the minimum bias is considered as the best.

MAE Mean absolute error is computed as the mean of the absolute differences of observed values and the estimated missing values of daily rainfall data. The estimation method having the lowest MAE value is considered as the best (Suhalia et al. 2008). Therefore, the method is defined as

$$MAE = n^{-1}\sum_{i=1}^{n}|\varepsilon_i| \ , \ where \ \ \varepsilon_i = Y_i^{obs.} - Y_i^{est.} \tag{15}$$

Root-mean-square error (RMSE) The RMSE is frequently used to measure the difference between the values (sample and population values) predicted by a model or an estimator and the values actually observed (Li and Zhao 2001; Chai and Draxler 2014). This measure is also used to compare the different estimating techniques or methods for identification of the best method. The method with the lowest value of RMSE indicates the best method. The study considers RMSE to measure the best technique or method using the difference between the observed values $\left(Y_i^{\text{obs}}\right)$ of daily rainfall data and estimated values $\left(Y_i^{\text{est}}\right)$ of daily rainfall missing data (Simolo et al. 2010). The measurement formula for RMSE is given below:

$$RMSE = \sqrt{n^{-1}\sum_{i=1}^{n}\varepsilon_i^2}, \quad where \quad \varepsilon_i = Y_i^{obs} - Y_i^{est.} \tag{16}$$

Coefficient of variation of root-mean-square error (CVRMSE) To identify the forecasting performances for time series data, RMSE is commonly used as a measure of accuracy under scale measurement. However, to eliminate scale dependencies of comparison criterion, Yozgatligil et al. (2013) suggested CVRMSE measurement. The measurement RMSE is divided by the mean of actual (observed) values gives the CVRMSE. To compare missing value estimation techniques, the RMSE divided by the mean of observed daily rainfall data for the artificially created missing period provides CVRMSE,

$$\text{CVRMSE} = \frac{\text{RMSE}}{\overline{\text{Y}}^{\text{obs.}}} \ , \quad \text{where} \ \ \overline{\text{Y}}^{\text{obs.}} = \text{n}^{-1}\sum_{\text{i}}^{\text{n}}\text{Y}_{\text{i}}^{\text{obs.}} \tag{17}$$

Minimum CVRMSE suggests the minimum percentage of variation between observed values and estimated values of missing data for daily rainfall occurrences. So, the method with the minimum CVRMSE is considered as the best.

Standard deviation of error (ESD) The standard deviation of error (difference between the observed and estimated value) indicates the fluctuations of the deviations. The minimum ESD is used as the criterion to identify the best technique for estimating the missing value (Silva et al. 2007). Then, it is defined as

$$ESD = \sqrt{(n-1)^{-1}\sum_{i=1}^{n}\left(\varepsilon_i - \overline{\varepsilon}\right)^2}, \ where \ \overline{\varepsilon} = n^{-1}\sum_{i=1}^{n}\varepsilon_i \ and \ \varepsilon_i = Y_i^{obs.} - Y_i^{est.} \tag{18}$$

Similarity index (*S* index) S index is the criterion of agreement for assessing model performance which implies the percentage of agreement between the observed and estimated values. The values of S index lie between 0.0 and 1.0, where 0.0 indicates complete disagreement and 1.0 indicates perfect agreement (Wilmott 1981). The S index is used to find out

the best missing value estimation technique for rainfall data (Suhaila et al. 2008). The S index is stated below:

$$S\ \text{index} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{obs.} - Y_i^{est.}\right)^2}{\sum_{i=1}^{n}\left(\left|Y_i^{obs.} - \overline{Y}\right| + \left|Y_i^{est.} - \overline{Y}\right|\right)^2} \quad (19)$$

where $\overline{Y}$ is the mean of observed daily rainfall and n is the number of estimated or observed observations.

3 Results and discussions

To estimate the missing value of daily rainfall observations, different methods and their comparative techniques are already discussed in the previous section for identifying the suitable method. The performance of data for the study is also discussed in Sect. 2. In that section, the classification procedures of 27 selected stations into five climatic sub-zones and the selection of target and reference stations from each sub-zone are elaborately discussed. The nature of missing data distribution of these stations follows MAR, is also explained in Sect. 2. The results of daily rainfall missing data estimation of five target stations for different methods and the results of comparative techniques for identifying station-wise suitable method are discussed in this section, followed by a comparison of the present study to similar studies conducted in other parts of the world.

The results of the comparison criteria of missing value estimation techniques for target station Sylhet of north-east region, Chittagong of south-east region, Dhaka of mid region, Khulna of south-west region, and Rajshahi of north-west region are revealed in Tables 3, 4, 5, 6, and 7, respectively. However, the correlation coefficient between daily rainfall amount of target station and its nearest reference station is higher than that of all other reference stations. For example, the distance between target station Chittagong and reference station Ambagan is smallest (15 km), and their correlation coefficient is found to be 0.91559 and it is statistically significant (Table 2).

In Fig. 2, box plots for all the stations in each of the five regions are shown taking $n = 14$, 75, and 146 observations, respectively, which were randomly selected and set as missing observations considering 1, 5, and 10% missing data. Each row of the figure shows the box plots for each region for three different sample sizes (e.g., row 1 in the figure shows box plots for the stations in south-east region for 14, 75, and 146 observations, respectively), and each column shows the box plots for different regions of same sample size (e.g., column 2 shows the box plots of stations of each region considering 75 observations). So, it is obvious that the box plots in column 1 will have less number of outliers than those of columns 2 and 3 because of the least sample size considered. If we wish to look at the pattern in each region for all sample sizes, similar behavior can be noticed. For instance, the number of outliers for stations in each region are increasing with the increase in sample size (e.g., number of outliers for $n =$ 14, 75,146 in Dhaka station of mid-region are 3, 10, and 30, respectively and in Sylhet station of north-east region are 2, 13, and 26, respectively). However, if we want to compare the pattern of stations of different regions, that can be done looking at the same column for a specific sample size. Let us consider column 2 ($n = 75$), for south-east region, we can observe that there are a considerable number of outliers for each station and the rainfall observations are right skewed for all the stations (median is zero for all the stations). Similar patterns can be observed for north-east, mid, and south-west regions. There is one extreme station in south-west region, named Satkhira for which the third quartile is also very small (Q_3 for Satkhira = 1), which might be the result of the random choice of observations; different sample of observations would result in different box plots, but the pattern of right-skewed data remains same for all combinations of

Table 3 Results of comparison measures for the missing value estimation techniques applied to estimate 1, 5, and 10% missing values in north-east region

Methods	Percentage of missing data	Kolmogorov-Smirnov test statistic (D)	P value for D statistic	Bias or ME	RMSE	CVRMSE	MAE	ESD	S index
EM-MCMC	1%	0.2143	0.9048	13.371	54.761	1.7871	34.05	55.11	0.302
	5%	0.3836	0.0000	4.3476	46.664	2.2832	26.84	46.78	0.306
	10%	0.3082	0.0000	4.5869	41.780	2.3062	22.65	41.67	0.330
Single best estimator	1%	0.1429	0.9988	4.5256	20.590	2.3592	10.07	20.27	0.994
	5%	0.0685	0.9955	4.2432	22.650	2.2834	9.93	22.29	0.890
	10%	0.0753	0.8017	4.1878	22.936	2.2936	9.93	22.57	0.789
Linear regression	1%	0.7857	0.0004	0.2049	19.180	2.3317	12.63	18.86	0.964
	5%	0.7808	0.0000	−0.078	20.899	2.0904	12.49	20.87	0.789
	10%	0.7534	0.0000	−0.118	21.137	2.1059	12.46	21.13	0.688

Target station: Sylhet; reference station: Srimangal

Table 4 Results of comparison measures for the missing value estimation techniques applied to estimate 1, 5, and 10% missing values in south-east region

Methods	Percentage of missing data	Kolmogorov-Smirnov test statistic (*D*)	*P* value for *D* statistic	Bias or ME	RMSE	CVRMSE	MAE	ESD	*S* index
Arithmetic Average	1%	0.0714	0.9999	0.204	2.971	4.573	5.99	12.98	0.955
	5%	0.1096	0.7731	0.041	7.705	1.684	5.84	15.62	0.859
	10%	0.1370	0.1291	0.006	11.215	1.620	5.84	16.37	0.886
Normal ratio	1%	0.0714	0.9898	0.303	12.979	0.977	5.93	12.89	0.958
	5%	0.1096	0.7731	0.136	15.590	1.645	5.78	15.58	0.861
	10%	0.1370	0.1291	0.103	16.348	1.567	5.78	16.35	0.891
Normal ratio with weighted correlation	1%	0.0714	0.9957	0.160	9.702	0.715	4.39	9.65	0.977
	5%	0.1096	0.7731	0.124	11.654	1.189	4.33	11.65	0.929
	10%	0.1370	0.1291	0.093	12.094	1.331	4.31	12.09	0.926
Inverse distance	1%	0.2857	0.6172	9.185	22.841	2.457	9.18	21.60	0.335
	5%	0.3288	0.0007	8.893	27.467	2.480	8.89	26.11	0.325
	10%	0.3356	0.0000	8.877	28.530	2.735	8.87	27.17	0.283
EM-MCMC	1%	0.2857	0.6172	− 1.16	9.709	0.438	6.33	10.00	0.986
	5%	0.5714	0.0207	− 1.63	9.886	0.446	7.23	10.12	0.984
	10%	0.3014	0.0026	− 2.75	14.520	0.446	9.36	14.36	0.938
Single best estimator	1%	0.1429	0.9988	0.542	8.999	0.754	3.85	8.99	0.974
	5%	0.0685	0.9955	0.620	10.843	0.987	3.85	10.83	0.952
	10%	0.0274	0.9999	0.594	11.159	1.302	3.84	11.15	0.929
Linear regression	1%	0.7143	0.0016	− 0.03	9.016	0.796	4.05	9.00	0.972
	5%	0.6712	0.0000	0.058	10.828	1.013	4.06	10.83	0.951
	10%	0.6644	0.0000	0.031	11.138	1.324	4.04	11.14	0.929
Multiple regression	1%	0.6429	0.0061	6.539	24.006	0.681	11.2	23.69	0.979
	5%	0.6575	0.0000	3.990	28.945	1.073	12.5	28.79	0.947
	10%	0.6301	0.0000	1.390	33.656	1.267	14.7	33.68	0.940

Target station: Chittagong; reference stations: Ambagan, Cox's Bazar, Feni, Hatiya, Kutubdia, M. Court, Rangamati, Sandwip, Sitakunda

observations. Same explanations apply to the stations of north-west regions with very lower values of third quartiles (Q_3 for Rajshahi = 0, Q_3 for Ishwardi = 1, and Q_3 for Chuadanga = 2). The presence of outliers in stations in columns 1 and 3 can be explained similarly. This is to keep in mind that these box plots are representing the actual rainfall occurrences for the days those are considered missing in the present study; they are not representative for the whole data set. So, we cannot generalize the findings of the box plots to assess the geographic variation among the stations. These are presented only to help in assessing the performance of the missing value estimation techniques applied to estimate these observations.

3.1 North-east region

Only one reference station (Srimangal) is identified corresponding the target station Sylhet, which have very high elevation (Table 2). For single reference station, the methods EM-MCMC, SBE, and LR are applicable among the methods to estimate the missing values of daily rainfall data. In these methods, SBE for 1, 5, and 10% missing data and EM-MCMC for 1% missing data provides good fit following the KS test. The efficiency measurement technique CVRMSE provides a similar result (around 2.29) for SBE and EM-MCMC methods. SBE method provides the highest value of *S* index compared to other methods for 1, 5, and 10% missing values (Table 3). The correlation coefficient between target and reference stations for daily rainfall data is very low (0.3094) due to long distance (68 km) between target station and reference station (Table 2). For such relationship, the EM-MCMC and LR methods did not perform well. Therefore, the SBE method is the most suitable method for estimating the missing value of daily rainfall data for Sylhet station.

3.2 South-east region

For this region, the nine rainfall stations are identified as reference stations surrounding to the target station, Chittagong (Table 2). Kolmogorov-Smirnov goodness-of-fit test provides satisfactory results for all missing value estimation methods of daily rainfall observations except regression methods, ID and

Table 5 Results of comparison measures for the missing value estimation techniques applied to estimate 1, 5, and 10% missing values in mid-region

Methods	Percentage of missing data	Kolmogorov-Smirnov test statistic (D)	P value for D statistic	Bias or ME	RMSE	CVRMSE	MAE	ESD	S index
Arithmetic average	1%	0.1429	0.9988	− 0.43	7.642	3.220	3.74	7.60	0.997
	5%	0.1507	0.3786	− 0.52	8.616	2.178	3.72	8.60	0.937
	10%	0.1644	0.0387	− 0.55	8.689	2.161	3.70	8.67	0.911
Normal ratio	1%	0.1429	0.9988	0.068	7.342	2.923	3.54	7.32	0.994
	5%	0.1507	0.3786	− 0.01	8.284	2.079	3.51	8.28	0.932
	10%	0.1644	0.0387	− 0.04	8.349	2.069	3.49	8.35	0.903
Normal ratio with weighted correlation	1%	0.0714	0. 9987	1.163	7.490	2.610	3.45	7.45	0.997
	5%	0.1233	0.6359	1.066	8.603	2.121	3.40	8.56	0.927
	10%	0.1438	0.0975	1.028	8.668	2.131	3.38	8.62	0.917
Inverse distance	1%	0.2143	0.9048	4.259	10.113	2.478	4.26	9.47	0.476
	5%	0.4247	0.0000	4.120	11.260	2.762	4.12	10.53	0.411
	10%	0.2945	0.0000	4.128	11.565	2.814	4.13	10.83	0.368
EM-MCMC	1%	0.2857	0.6172	− 0.12	12.856	2.106	6.80	13.25	0.904
	5%	0.2603	0.0142	− 0.12	13.453	2.106	6.73	13.55	0.708
	10%	0.0411	0.9987	− 0.12	10.112	2.106	3.92	10.17	0.906
Single best estimator	1%	0.1429	0.9988	0.079	8.860	3.548	4.04	8.85	0.998
	5%	0.0548	0.9999	0.066	9.783	2.478	3.95	9.78	0.958
	10%	0.0342	1.0000	0.060	9.878	2.464	3.92	9.88	0.921
Linear regression	1%	0.7857	0.0004	− 0.23	8.449	3.309	4.85	8.33	0.977
	5%	0.7671	0.0000	− 0.31	9.537	2.374	4.79	9.53	0.944
	10%	0.7466	0.0000	− 0.35	9.610	2.373	4.76	9.60	0.905
Multiple regression	1%	0.7857	0.0004	− 0.09	12.816	6.168	7.05	12.96	0.975
	5%	0.7671	0.0000	0.160	12.092	3.021	6.49	12.11	0.712
	10%	0.7466	0.0000	− 0.82	13.807	3.442	7.27	13.80	0.810

Target station: Dhaka; reference stations: Madaripur, faridpur, Chandpur, Tangail, Comilla

EM-MCMC methods for 5 and 10% data. The bias of the estimated missing values is found the minimum for all the fitted methods other than ID and MR methods. However, S index provides good performance for all the methods except ID method (Table 4).

The box plots of 1, 5, and 10% daily rainfall data for target and reference stations in this region indicate some outliers in reference stations (Fig. 2). In these stations, daily rainfall observations show high variation due to high discrimination of elevations (Table 2).The box plots also indicate the possibility of the existence of a pair-wise moderate correlation between daily rainfall observations of the reference stations (Fig. 2), so the regression models may not provide a good fit for estimation of missing values. The ID method does not provide significant result in this region due to considerable variation of the distance between the target and each of the reference stations (Table 2). Therefore, to estimate the missing value of daily rainfall data in Chittagong station, four methods (AA, NR, NRWC, and SBE) provided satisfactory performance.

3.3 Mid region

In this region, five reference stations are identified neighboring target station Dhaka. According to distance, Faridpur is the nearest reference station to the target station (distance 57 km), and the elevation of the reference stations and target station are almost similar except Chandpur station (Table 2). The KS test provides a good fit for all methods except LR and MR methods, and AA (for 10% missing data), NR (for 10% missing), ID (for 5 and 10% missing), and EM-MCMC (for 5% missing) methods. The EM-MCMC method for estimating these missing data of daily rainfall provide the higher RMSE, MAE, and ESD than that of other methods. However, the bias of the estimates is the lowest for SBE method and S indices are close to 1 for AA, NR, NRWC, and SBE methods (Table 5).

The box plots for 1, 5, and 10% data of daily rainfall provide the presence of outliers for every station (Fig. 2). The correlation coefficient of daily rainfall amount between the target station Dhaka and for each of the reference stations expect Faridpur station is found around 0.45. For such weaker

Table 6 Results of comparison measures for the missing value estimation techniques applied to estimate 1, 5, and 10% missing values in south-west region

Methods	Percentage of missing data	Kolmogorov-Smirnov test statistic (D)	P value for D statistic	Bias or ME	RMSE	CVRMSE	MAE	ESD	S index
Arithmetic average	1%	0.2857	0.6172	0.209	7.804	2.352	3.69	7.79	0.743
	5%	0.1233	0.6359	0.169	9.340	1.945	3.91	9.34	0.782
	10%	0.1164	0.2756	0.150	9.519	1.955	3.90	9.52	0.788
Normal ratio	1%	0.1761	0.8230	− 0.88	51.138	19.273	8.51	15.52	0.528
	5%	0.1438	0.2664	− 0.73	20.218	4.337	6.68	14.27	0.813
	10%	0.1233	0.2171	0.051	9.498	1.951	3.90	9.50	0.923
Normal ratio with weighted correlation	1%	0.1830	0.7858	− 0.95	51.328	19.400	8.55	15.57	0.538
	5%	0.1370	0.3206	− 0.76	20.217	4.339	6.66	14.26	0.817
	10%	0.1164	0.2756	0.125	9.264	1.902	3.73	9.27	0.943
Inverse distance	1%	0.2857	0.6172	4.928	11.043	2.448	4.93	10.33	0.343
	5%	0.3562	0.0002	4.927	13.633	2.791	4.93	12.78	0.285
	10%	0.3562	0.0000	5.466	13.918	2.848	5.47	13.06	0.274
EM-MCMC	1%	0.2857	0.6172	− 2.91	6.897	1.341	5.18	6.49	0.867
	5%	0.2603	0.1423	− 1.44	10.330	1.539	6.53	10.30	0.791
	10%	0.1301	0.1686	− 0.72	7.304	1.201	3.26	7.29	0.895
Single best estimator	1%	0.1429	0.9988	0.39	8.951	2.483	3.90	8.94	0.683
	5%	0.0685	0.9955	− 0.08	11.490	2.432	4.37	11.49	0.742
	10%	0.0274	0.9982	− 0.13	11.716	2.424	4.37	11.72	0.754
Linear regression	1%	0.9953	0.0000	0.249	8.212	2.281	4.69	8.15	0.622
	5%	0.6986	0.0000	0.075	10.096	2.085	4.96	10.09	0.695
	10%	0.6918	0.0000	0.039	10.309	2.113	4.94	10.31	0.703
Multiple regression	1%	0.9981	0.0000	3.724	10.821	2.559	5.10	10.46	0.309
	5%	0.6438	0.0000	− 0.14	15.980	3.405	8.27	16.02	0.250
	10%	0.6438	0.0000	− 0.72	16.815	3.500	8.67	16.82	0.248

Target station: Khulna; reference stations: Jessore, Barisal, Patualkhali, Mongla, Satkhira

relationship, LR and MR methods may not be provided good fit. Again, for Dhaka and Faridpur station, this correlation is found 0.603. Due to this relationship, SBE method can be considered as the best estimator to estimate the missing value of rainfall data for Dhaka station on the basis of lowest bias and the higher value of S index compared to all other methods.

3.4 South-west region

For this region, five stations are identified as reference stations surrounding to the target station Khulna. For these stations, elevation is almost similar (around 2.1 m). In respect of distance, the nearest station is Mongla (35 km) to the target station (Table 2). The methods AA, NR, NRWC, EM-MCMC, and SBE demonstrate good fit to estimate the missing value of daily rainfall data following KS test. The bias and MAE of the estimates are found lower for AA method, and CVRMSE is observed lower for EM-MCMC compared to other methods. The value of S index for EM-MCMC method indicates the highest (S index close to 1) than that of other methods (Table 6).

The box plots for daily rainfall observations of the south-west region indicate a large number of outliers for all stations (Fig. 2). For this reason, the regression methods do not work well to estimate missing data of daily rainfall data. Further, the ID method also does not provide good fit due to the long distance between the target and reference stations. Therefore, the EM-MCMC method is found to be the best estimator for Khulna station to estimate the missing value of daily rainfall data.

3.5 North-west region

For this region, two rainfall stations are identified as reference station against target station Rajshahi. Ishwardi is the nearest reference station to the target station according to distance. The correlation coefficient between the target and its nearest reference station for daily rainfall data is 0.508 (Table 2).The methods AA, NR, NRWC, and SBE provide good fit to estimate the missing value of daily rainfall data following KS test. The bias of the estimates is found lowest for AA and SBE methods, and CVRMSE is found lowest for AA method than

Table 7 Results of comparison measures for the missing value estimation techniques applied to estimate 1, 5, and 10% missing values in mid-region and north-west region

Methods	Percentage of missing data	Kolmogorov-Smirnov test statistic (D)	P value for D statistic	Bias or ME	RMSE	CVRMSE	MAE	ESD	S index
Arithmetic average	1%	0.0714	0.9988	0.312	7.554	2.521	3.51	7.53	0.652
	5%	0.0685	0.9955	− 0.06	9.164	2.817	3.49	9.17	0.690
	10%	0.0616	0.9443	− 0.09	9.379	2.804	3.51	9.38	0.706
Normal ratio	1%	0.0714	0.9999	3.068	9.889	2.914	4.36	9.66	0.343
	5%	0.0685	0.9955	0.035	13.901	4.356	6.17	13.94	0.224
	10%	0.0616	0.9443	− 0.29	14.109	4.247	6.38	14.13	0.233
Normal ratio with weighted correlation	1%	0.2143	0.9048	0.385	12.455	8.219	6.10	12.62	0.992
	5%	0.0685	0.9955	− 0.03	13.971	4.381	6.22	14.01	0.224
	10%	0.0616	0.9443	− 0.41	14.418	4.345	6.49	14.44	0.228
Inverse distance	1%	0.0714	0.9980	3.947	9.855	2.635	3.95	9.33	0.320
	5%	0.2192	0.0600	3.476	10.666	3.131	3.48	10.14	0.254
	10%	0.2466	0.0003	3.453	10.859	3.184	3.45	10.32	0.245
EM-MCMC	1%	0.3571	0.3338	− 0.31	11.029	1.576	5.75	11.44	0.856
	5%	0.3425	0.0004	0.248	14.190	2.063	7.20	14.29	0.719
	10%	0.1712	0.0277	0.124	10.034	2.037	3.60	10.07	0.809
Single best estimator	1%	0.0714	0.9985	− 0.03	8.308	2.851	3.57	8.26	0.646
	5%	0.0411	0.9924	− 0.17	10.661	3.282	3.66	10.67	0.655
	10%	0.0274	0.9989	− 0.21	11.066	3.326	3.70	11.07	0.666
Linear regression	1%	0.9965	0.0000	0.351	7.544	2.246	4.35	7.49	0.564
	5%	0.8219	0.0000	0.017	8.655	2.580	4.24	8.65	0.572
	10%	0.8082	0.0000	− 0.02	8.865	2.616	4.24	8.86	0.584
Multiple regression	1%	0.9990	0.0000	2.099	9.419	2.667	4.86	9.39	0.281
	5%	0.8219	0.0000	0.035	11.739	3.588	5.81	11.76	0.204
	10%	0.8082	0.0000	− 0.19	11.965	3.565	5.94	11.97	0.21

Target station: Rajshahi; reference stations: Chuadanga, Ishwardi

that of other methods. However, the value of S index is found almost same (around 0.65) for AA and SBE methods (Table 7).

The box plots indicate high variation among the stations' rainfall data in this region (Fig. 2); due to this, the methods LR, MR, ID, and EM-MCMC do not provide satisfactory results in terms of comparison criteria. Besides, for long distance from the target to reference stations (Table 2), the ID method does not perform adequately. Therefore, the AA and SBE methods provide well fit in respect of lowest bias and high S index value to estimate the missing value of daily rainfall data in Rajshahi station.

3.6 Comparison with other similar studies

The present study has been conducted to suggest a suitable method to estimate the missing values in daily rainfall data in Bangladesh. The study employed eight different methods found in different literature and compared the performances of the methods using seven techniques. To the best of our knowledge, this is the first study making an attempt to find the appropriate missing value estimation technique for Bangladesh till date. However, this study was inspired by similar studies conducted in other parts of the world. For instance, there have been studies to find out the best method to estimate missing values in Turkish meteorological data (Yozgatligil et al. 2013), daily precipitation data from Brazil (Ferrari and Ozaki 2014), rainfall data from Malaysia (Suhalia et al. 2008), Italy (Lo Presti et al. 2010), Andes region in Venezuela (Garcia et al. 2006), etc.

Garcia et al. (2006) performed a cluster analysis to find two closest stations corresponding to a rainfall station and fill in the missing value of the target station from those closest station. They applied their method to daily, weekly, bi-weekly, and monthly data of 106 rainfall stations in Andes region in Venezuela and assessed the performance of the proposed method using mean error (ME), MAE, RMSE, coefficient of correlation (r), and Willmott agreement index (d). The author did not compare the proposed method with any other methods.

Yozgatligil et al. (2013) suggested EM-MCMC algorithm as best technique in case of Turkish meteorological data after

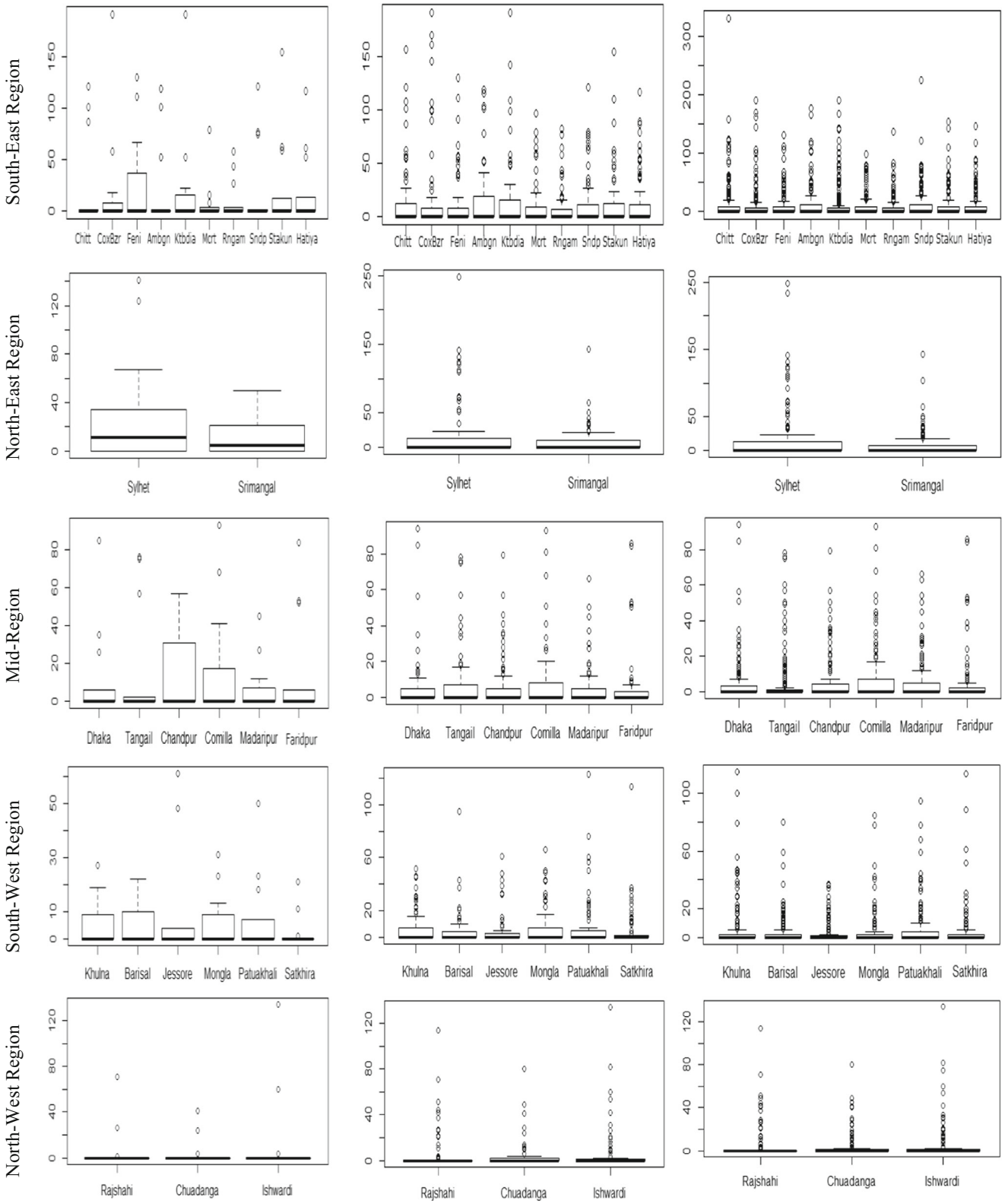

Fig. 2 Box plots for daily rainfall data of five regions of Bangladesh (vertical axes indicate the amount (mm) of daily rainfall occurrences, and horizontal axes indicate names of stations)

comparing simple and weighted arithmetic average methods, multilayer perceptron neural network, and MCMC-based multiple imputation methods. The comparison criteria used in the study were RMSE, coefficient of variation of RMSE (CVRMSE), and correlation dimension technique (branch of nonlinear dynamic time series analysis).

Ferrari and Ozaki (2014) compared nearest neighbor method, inverse distance-weighted ratio method, and linear regression method for imputation of missing values in precipitation data from the state of Parana in southern of Brazil according to the value of RMSE. The author stated the inverse distance weighted ratio method to be most appropriate for imputing missing precipitation data from 484 stations in the area of interest.

Silva et al. (2007) compared arithmetic mean method, normal ratio method, and inverse distance method to estimate the missing rainfall data in Sri Lanka according to the measurements of descriptive statistics of error, RMSE, mean absolute percentage of error, and correlation coefficient. The authors also proposed a new method named aerial precipitation ratio which selected stations representing each of seven major ecological zones in Sri Lanka, and monthly rainfall data was estimated using abovementioned methods taking the surrounding stations in each zone into account. The authors suggested different methods to be suitable for different zones in Sri Lanka with no indication of single best method.

Lo Presti et al. (2010) proposed methods to fill in the missing observations in daily rainfall data in the Candelaro River Basin (Italy) in two stages. In the first stage, the authors assessed the missingness mechanism present in the data and then applied four different regression methods (simple substitution, parametric regression, ranked regression, and Theil method) to estimate the missing daily rainfall data. By studying the absolute error distribution, the authors indicated the Theil method to be the most suitable one, though a very complex method. Simple substitution method was also marked as acceptable method.

The present study employed eight methods to estimate the missing rainfall data from a target station from each of the five climatic sub-zones of Bangladesh, so the methods are applicable to all other stations. This kind of climatic or ecological division was only made by Silva et al. (2007). Before estimating the missing daily rainfall amount in the target stations, the missingness mechanism of the missing rainfall data was tested following the suggestions of Lo Presti et al. (2010). To the best of our knowledge, none of the other studies have tested the missingness mechanism. The present study chose the eight methods, which is highest among all studies reported here, from all the mentioned literature on the basis of relevance, simplicity, and relative performance in other regions. The comparison of such a high number of methods allowed flexibility in making choice of the best method to estimate the missing data in daily rainfall observations. Also, seven comparison criteria used in the present study were combined from the previous studies. The K-S test to determine the goodness-of-fit test, apart from the present study, was only applied once (Simolo et al. 2010) to assess the performance of missing value estimation techniques in rainfall data. The result of K-S test has significant contribution in choosing the most suitable method in the present article. One of the unique element of this study was the inclusion of box plots for the selected missing observations in target and reference stations which helped to understand the actual scenario in different stations across Bangladesh on the days chosen to be missing, which has effect on the performance of particular missing value estimation technique. Though this study did not propose any new method, it integrated a wide range of methods and comparison criteria along with some descriptive measures to be able to estimate the missing data in daily rainfall which will give rise to further scientific studies involving continuous rainfall data in future.

4 Conclusion

A suitable method of estimating missing rainfall value is of great interest to the researchers worldwide. The reason behind such interest is to make use of rainfall data from long series where occasional missing values pose formidable difficulty in using such data. In the present paper, a comparison of different methods has been conducted in order to suggest the best possible choice under certain specific criteria. Although the focus of the current paper is Bangladesh, the statistical criteria that are being used in this study can be generalized on the basis of underlying statistical reasoning that are highlighted here. To estimate the missing value of daily rainfall observations for five climatic regions' target stations of the country, the eight methods and seven comparison techniques are employed to identify the best suitable method for each of the stations. For performing these methods, three sets of daily rainfall missing data sample (1, 5, and 10%) with 1000 times repetitions are considered (Sect. 2). The performance of the estimation methods according to the comparison techniques are shown in Tables 3, 4, 5, 6, and 7. On the basis of these results, the discussions are made in Sect. 3. From the results and discussions of the study, the following conclusions can be drawn. We have made an attempt to find a single method that can be suggested for all the stations in Bangladesh. To examine whether the findings of this study hold for other countries, studies can be repeated for other countries as well. This may provide a consensual technique under varied conditions prevail in the nature and extent of missing values in the time series data on rainfall.

Let us consider five measures of comparison (out of seven measures included in this study) for identifying the best estimation technique, namely, (i) K-S test statistic, (ii) bias, (iii) RMSE, (iv) MAE, and (v) S index. Two other measures of comparison CVRMSE and ESD are ignored due to inclusion of similar measures RMSE and MAE. The Kolmogorov-Smirnov test statistic shows that among all the estimation techniques, only SBE provides consistently acceptable estimation technique for all the regions. Other measures of

comparison such as bias, RMSE, MAE, and *S* index also confirm that SBE is consistently better as a technique of estimating missing values. In some cases, arithmetic average, EM-MCMC, provides good estimate along with the linear or multiple regression estimates but the results are not consistent for all the regions. Garcia et al. (2006) observed that closest station method as the best one to fill in the missing observations of rainfall data in different time scales in Andes region in Venezuela. Lo Presti et al. (2010) stated the simple substitution method, which is same as the SBE described in the present study, to be an acceptable technique of missing value estimation in daily rainfall data in the Candelaro River Basin (Italy) when the similarity value is particularly high and significant. In the present study, from Table 2, it can be observed that the target station had significant positive correlation with all its reference stations. For the SBE method, a single station is chosen for each target station, which has highest significant correlation with the daily rainfall observations in the target station, and it also happens to be the nearest station to the target station according to the distance (Table 2). Thus, the consistent performance of SBE method has both statistical reasoning and practical significance. Hence, we may conclude that the technique of single best estimator is singled out in this study as the possible choice of estimating missing values.

Acknowledgements This study is supported under the HEQEP sub-project, CP-3293, in the Department of Applied Statistics, East West University funded by World Bank and implemented by University Grants Commission of Bangladesh (UGC). The authors are also grateful to Bangladesh Meteorological Department (BMD) for providing the data. We acknowledge the critical comments from anonymous reviewers and editor.

References

Ahrens B (2006) Distance in spatial interpolation of daily rain gauge data. Hydrol Earth Syst Sci 10:197–208

Asati SR (2012) Analysis of rainfall data for drought investigation at Agra U. P. Int J Life Sci Biotechnol Pharm Res 1(4):81–86

Bangladesh Economic Review (2016) Economic adviser's wing, finance division, Ministry of Finance, Government of the People's Republic of Bangladesh

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250

Chen FW, Liu CW (2012) Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy Water Environ 10(3):209–222

Chowdhury MRK (2013) Country report: Bangladesh meteorological department (BMD), People's republic of Bangladesh

Collins LM, Schafer JL, Kam CM (2001) A comparison of inclusive and restrictive strategies in modern missing-data procedures. Psychol Methods 6:330–351

Cong RG, Brady M (2012) The interdependence between rainfall and temperature: copula analyses. Sci World J 2012:1–11

Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. J Hydrol 341:27–41

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Dumedah G, Coulibaly P (2011) Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. J Hydrol 400(1–2):95–102

Eischeid JK, Baker CB, Karl TR, Diaz HGF (1995) The quality control of long-term climatological data using objective data analysis. J Appl Meteorol 34:2787–2795

Eischeid JK, Pasteris PA, Diaz HF, Plantico MS, Lott NJ (2000) Creating a serially complete, national daily time series of temperature and precipitation for the western United States. J Appl Meteorol 39(9): 1580–1591

Ferrari GT, Ozaki V (2014) Missing data imputation of climate datasets: implications to modeling extreme drought events. Rev Bras Meteorol 29(1):21–28

Garcia B, Sentelhas P, Tapia L, Sparovek G (2006) Filling in missing rainfall data in the Andes region of Venezuela, based on a cluster analysis approach. Rev Bras Agrometeorol 14(2):225–233

Garcia M, Peters-Lidard CD, Goodrich DC (2008) Spatial interpolation in a dense gauge network for monsoon storm events in the southwestern United States. Water Resour Res 44: W05S13. https://doi.org/10.1029/2006WR005788

Goodison B, Louie PYT, Yang D (1998) WMO solid precipitation measurement inter comparison. Final report

Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL (1997) Analysis with missing data in prevention research. The science of prevention: methodological advances from alcohol and substance abuse research, 1, pp 325-366

Hubbard KG (1994) Spatial variability of daily weather variables in the high plains of the USA. Agric For Meteorol 68:29–41

Kemp WP, Burnell DG, Everson DO, Thomson AJ (1983) Estimating missing daily maximum and minimum temperatures. J Climate Appl 22:1587–1593

Kripalani RH, Inamdar S, Sontakke NA (1996) Rainfall variability over Bangladesh and Nepal: comparison and connections with features over India. Int J Climatol 16(6):689–703

Lam NSN (1983) Spatial interpolation methods : a review. Am Cartographer 10(2):129–149

Lennon JJ, Turner JRG (1995) Predicting the spatial distribution of climate: temperature in Great Britain. J Anim Ecol 64:370–392

Li X, Z Zhao (2001) Measures of performance for evaluation of estimators and filters. Proc. 2001 SPIE Conf. on Signal and Data Processing, (July–August), pp 1–12

Little JRA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Lo Presti R, Barca E, Passarella G (2010) A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ Monit Assess 160:1–22

Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. JASA 46(253):68–78

National Hurricane Center of USA n.d. http://www.nhc.noaa.gov/gccalc.shtml

Paulhus JLH, Kohler MA (1952) Interpolation of missing precipitation records. Mon Weather Rev 80(8):129–133

Rashid H-e (1991) Geography of Bangladesh (2nd edition). In: Dhaka University Press Limited, Dhaka

Rubel F, Hantel M (1999) Correction of daily gauge measurements in the Baltic Sea drainage basin. Nord Hydrol 30:191–208

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (1978) Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section, ASA, pp 20–34

Rubin DB (1987) Multiple imputation for non-response in surveys. Wiley, New York

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Scheffer J (2002) Dealing with missing data. Res Lett Inf Math Sci 3:53–160

Shepard D (1968) A two-dimensional interpolation functions for irregularly spaced data. Proceeding of the Twenty-Third National Conference of the ACM, Washington, DC, pp 517–524

Silva RP, Dayawansa NDK, Ratnasiri MD (2007) A comparison of methods used in estimating missing rainfall data. J Agric Sci 3(May):101–108

Simanton JR, Osborn HB (1980) Reciprocal-distance estimate of point rainfall. J Hydraul Eng 106:1242–1246

Simolo C, Brunetti M, Maugeri M, Nanni T (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. Int J Climatol 30:1564–1576

Suhalia J, Sayang MD, Jemain AA (2008) Revised spatial weighting methods for estimation of missing rainfall data. Asia-Pac J Atmos Sci 44(2):93–104

Tabios GQ, Salas JD (1985) A comparative analysis of techniques for spatial interpolation of precipitation. Water Resour Bull 21:365–380

Tabony RC (1983) The estimation of missing climatological data. J Climatol 3:297–314

Tang WY, Kassim AHM, Abubakar SH (1996) Comparative studies of various missing data treatment methods-Malaysian experience. Atmos Res 42:247–262

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. JASA 82(398):528–540

Teegavarapu RSV, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312:191–206

Tronci N, Molteni F, Bozzini M (1986) A comparison of local approximation methods for the analysis of meteorological data. Arch Meteorol Geophys Bioclimatol A 36:189–211

Walther BA, Moore JL (2005) The concept of bias, precison and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimators. Ecography 28:815–829

Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press, New York

Williams P (1998) Modelling seasonality and trends in daily rainfall data. Adv Neural Inf Proces Syst 10:985–991

Wallis JR, Letten Mayer DP, Wood EF (1991) A daily hydro climatological data set for the continental United States. Water Resour Res 27: 1657–1663

Wilmott CJ (1981) On the validation of models. Phys Geogr 2:194–194

Xia Y, Fabian P, Stohl A, Winterhalter M (1999) Forest climatology: estimation of missing values for Bavaria, Germany. Agric For Meteorol 96:131–144

Yim C (2015) Imputing missing data with SAS. SAS Global Forum 2015, April 26–29, 2015, Dallas, pp 1–21

Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 112(1–2):143–167

Young KC (1992) A three way model for interpolating monthly precipitation values. Mon Weather Rev 120:2561–2569