Assessment of different methods for estimation of missing data in precipitation studies

Mohammad-Taghi Sattari, Ali Rezazadeh-Joudi and Andrew Kusiak

ABSTRACT

The outcome of data analysis depends on the quality and completeness of data. This paper considers various techniques for filling in missing precipitation data. To assess suitability of the different methods for filling in missing data, monthly precipitation data collected at six different stations was considered. The complete sets (with no missing values) are used to predict monthly precipitation. The arithmetic averaging method, the multiple linear regression method, and the non-linear iterative partial least squares algorithm perform best. The multiple regression method provided a successful estimation of the missing precipitation data, which is supported by the results published in the literature. The multiple imputation method produced the most accurate results for precipitation data from five dependent stations. The decision-tree algorithm is explicit, and therefore it is used when insights into the decision making are needed. Comprehensive error analysis is presented.

Key words | arid areas, arithmetic averaging, decision tree, missing precipitation data, multiple regression, partial least squares

Mohammad-Taghi Sattari
Department of Water Engineering, Agriculture Faculty,
University of Tabriz, Tabriz,
Iran

Ali Rezazadeh-Joudi (corresponding author)
Young Researchers and Elite Club,
Maragheh Branch, Islamic Azad University,
Maragheh,
Iran
E-mail: *alijoudi66@gmail.com*

Andrew Kusiak
Department of Mechanical and Industrial Engineering,
University of Iowa,
Iowa City,
IA,
USA

INTRODUCTION

Rainfall is an important part of the hydrological cycle. One of the first steps in any hydrological and meteorological study is accessing reliable quality data. However, precipitation data is frequently incomplete. The incompleteness of precipitation data may be due to damaged measuring instruments, measurement errors and geographical paucity of data (data gaps) or changes to instrumentation over time, a change in the measurement site, a change in data collectors, the irregularity of measurement, or severe topical changes in the climate.

The accurate planning and management of water resources depends on the presence of consistent and exact precipitation data in meteorology stations. In cases where it has not been possible to accurately and consistently record precipitation data in a particular time section, it is necessary to estimate the missing precipitation data before applying it in hydrological models. The data in climate studies also face this issue, as well as measurements in the ocean share similar data problems in regard to missing precipitation data (Lyman & Johnson 2008; Abraham *et al.* 2013; Cheng *et al.* 2015a, 2015b).

The estimation of missing data in hydrological studies is necessary for timely implementation of projects such as dam or canal construction. This information is extremely valuable in areas that deal with heavy precipitation events and floods. The accurate estimation of the missing data makes a great contribution to accurate assessment of the capacity of flood control structures in rivers and also dam spillovers. It reduces the risk of floods in the downstream of these structures. Abraham *et al.* (2015) observed precipitation changes in the United States and stated that observations and projections of precipitation changes can be useful in designing and constructing infrastructure to be more resistant to both heavy precipitation and flooding.

doi: 10.2166/nh.2016.364

Homogeneity and trend tests of data used in hydrological modeling or water resource analysis are essential. Numerous methods have been introduced for estimating and reconstructing missing data. They can be categorized as empirical methods, statistical methods, and function fitting approaches (Xia *et al.* 1999). Most of these methods derive the missing values using observations from neighboring stations. Selecting appropriate methods for estimating missing precipitation data may improve the accuracy of hydrological models. The literature points to rather arbitrary selection methods for estimating and reconstructing missing data (Hasanpur Kashani & Dinpashoh 2012). Some of the most significant studies involving estimation and reconstruction of missing rainfall data are discussed next.

Xia *et al.* (1999) estimated the missing data of daily maximum temperature, minimum temperature, mean air temperature, water vapor pressure, wind speed, and precipitation with six methods. They determined that the multiple regression analysis method was most effective in estimating missing data in the study area of Bavaria, Germany. Teegavarapu & Chandramouli (2005) applied a neural network, the Kriging method and the inverse distance weighting method (IDWM) for estimation of missing precipitation data. They demonstrated that a better definition of weighting parameters and a surrogate measure for distances could improve the accuracy of the IDWM. De Silva *et al.* (2007) used the aerial precipitation ratio method, the arithmetic mean method, the normal ratio (NR) method, and the inverse distance method to estimate missing rainfall data. The NR method was found to be most accurate. The arithmetic mean method and the aerial precipitation ratio method were most appropriate for the wet zone. You *et al.* (2008) compared methods for spatial estimation of temperatures. The spatial regression approach was found to be superior over the IDWM, especially in coastal and mountainous regions. Dastorani *et al.* (2009) predicted the missing data using the NR method, the correlation method, an artificial neural network (ANN), and an adaptive neuro-fuzzy inference system (ANFIS). The ANFIS approach performed best for the missing flow data. ANN was found to be more efficient in predicting missing data than traditional approaches. Teegavarapu (2009) estimated missing precipitation records by combining a surface interpolation technique and spatial and temporal association rules. The results suggested that this integrated approach improved the precipitation estimates. Teegavarapu *et al.* (2009) applied a genetic algorithm and a distance weighting method for estimating missing precipitation data. The genetic algorithm provided more accurate estimates over the distance weighting method. Kim & Pachepsky (2010) reconstructed missing daily precipitation data with a regression tree and an ANN. Better accuracy was accomplished with the combined regression tree and ANN rather than using them independently. Hosseini Baghanam & Nourani (2011) developed an ANN model to estimate missing rain-gauge data. The resulting feed-forward network was found to be accurate. Nkuna & Odiyo (2011) confirmed accuracy of the ANN in estimating the missing rainfall data. Hasanpur Kashani & Dinpashoh (2012) assessed accuracy of different methods of estimating missing climatological data. They concluded that although the ANN approach is more complex and time consuming, it outperformed the classical methods. Also, the multiple regression analysis method was found to be most suitable among the classical methods. Choge & Regulwar (2013) applied ANN to estimate the missing precipitation data. Che Ghani *et al.* (2014) estimated the missing rainfall data with the gene expression programming (GEP) method. The GEP approach was used to determine the most suitable replacement station for the principal rainfall station. Teegavarapu (2014) attempted to achieve statistical corrections for spatially interpolated missing precipitation data estimations.

The literature review indicates that there are no significant studies that evaluate various methods for estimating missing precipitation data in arid regions, such as southern parts of Iran and most of them have been performed in countries with almost mild or wet climates such as the studies of Xia *et al.* (1999), Teegavarapu & Chandramouli (2005), De Silva *et al.* (2007), You *et al.* (2008), Teegavarapu (2009), Teegavarapu *et al.* (2009), Kim & Pachepsky (2010), Che Ghani *et al.* (2014), and Teegavarapu (2014). Also most of the previous research is about the application of ANN and GEP methods in comparing classic methods, but there is not any remarkable study that evaluates the efficiency of the M5 model tree, which is one of the new and modern data mining methods.

The purpose of this study is to investigate the capability of 10 different traditional and data-driven methods to estimate missing precipitation data in arid areas of southern

Iran and to identify the most appropriate method. The 10 examined methods include arithmetic averaging (AA), inverse distance interpolation, linear regression (LR), multiple imputations (MI), multiple linear regression analysis (MLR), non-linear iterative partial least squares (NIPALS) algorithm, NR, single best estimator (SIB), UK traditional (UK) and M5 model tree.

MATERIALS AND METHODS

Study area and data analysis

The studied region encompasses a spacious part of southern Iran and includes an area more than seventy thousand square kilometers. The studied region includes hot and dry areas and is impacted by arid and semi-arid climates. The weather of the coastal zone is extremely hot and humid in the summer, as the temperature occasionally exceeds 52 °C. The average annual temperature in this region is approximately 27 °C. The amounts of monthly precipitation at the six rain-gauge stations located in southern Iran, namely Bandar Abbas, Bandar Lengeh, Jask, Minab, Kish Island and Abomoosa Island between 1986 and 2014 are used in this investigation. There is no significant difference between the elevations of the studied areas (5 to 30 meters above the sea level). The climate at each station was determined using the De Martonne (1923) aridity index shown in Equation (1).

$$I = \frac{P}{T + 10} \tag{1}$$

where P and T are the average annual precipitation (mm) and temperature (°C), respectively. Figure 1 shows the geographical area of the studied region. Table 1 includes geographic coordinates of the examined weather stations, their elevations, and characteristics of the monthly precipitation data.

Normally, there are no particular issues regarding recording data at meteorology stations. However, the inconsistency of the data record may happen in certain time sections *per se*. Hence, in this study we have hypothesized that 10% of data might not be measured. It may need to be estimated.

In this study, the Bandar Lengeh and Bandar Abbas stations were considered the target stations. The Bandar Abbas station is likely to have a precipitation regime

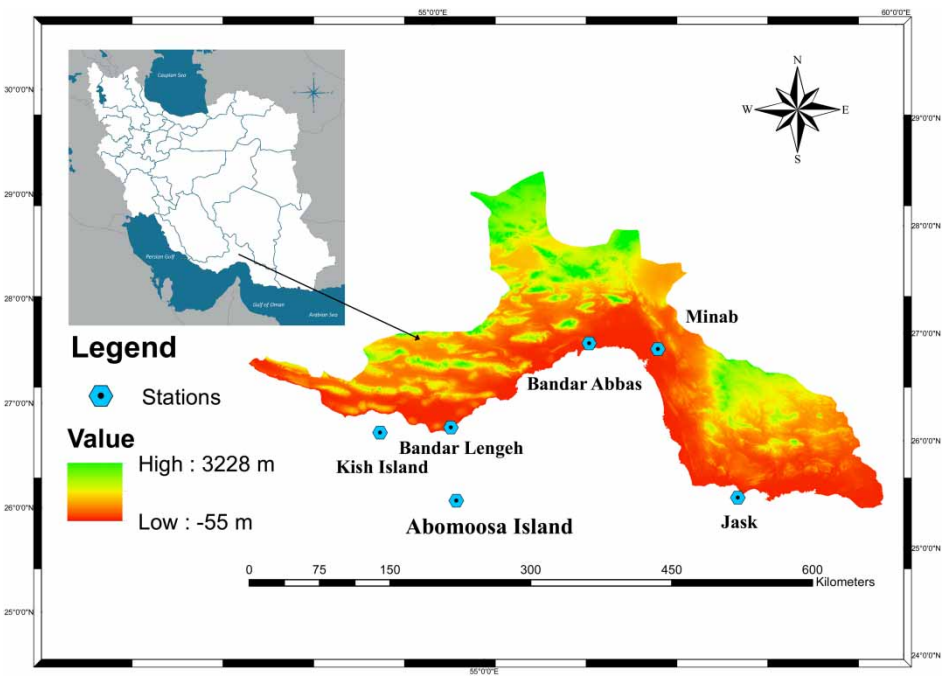

Figure 1 | Study area and location of stations.

Table 1 | Statistics of precipitation data and geographic position of selected rain gauge stations

		Abomoosa Island	Bandar Abbas	Jask	Bandar Lengeh	Kish Island	Minab
Geographic position	Latitude (N)	25 °50′	27 °13′	25 °38′	26 °32′	26 °30′	27 °6′
	Longitude (E)	54 °50′	56 °22′	57 °46′	54 °50′	53 °59′	57 °5′
	Elevation (m)	6.6	9.8	5.2	22.7	30.0	29.6
Statistics of precipitation data	Index of aridity	0.280	0.383	0.272	0.276	0.364	0.436
	Climate type	Dry	Dry	Dry	Dry	Dry	Dry
	Min rainfall (mm)	0	0	0	0	0	0
	Max rainfall (mm)	205	194.7	312	184.4	209.6	195.3
	Mean rainfall (mm)	10.653	14.128	10.181	10.435	12.805	16.804
	Standard deviation	28.037	33.402	29.876	27.126	30.78	35.979

different from other stations because it is affected by the elevation of Hormozgan Province. Thus, this station was not taken to be a target one. On the other hand, Bandar Lengeh is located almost in the middle of the zone regarding its latitude and longitude.

After statistical analysis and quality control of the available data, including homogeneity and trend tests, an attempt has been made to evaluate the efficiency of different classic statistical methods and a decision-tree model to estimate missing data.

Simple AA

This is the simplest method commonly used to fill in missing meteorological data in meteorology and climatology. Missing data is obtained by computing the arithmetic average of the data corresponding to the nearest weather stations, as shown in (2),

$$V_0 = \frac{\sum_{i=1}^{n} V_i}{N} \tag{2}$$

where V_0 is the estimated value of the missing data, V_i is the value of same parameter at ith nearest weather station, and N is the number of the nearest stations. The AA method is satisfactory if the gauges are uniformly distributed over the area and the individual gauge measurements do not vary greatly about the mean (Te Chow *et al.* 1988).

IDWM

The inverse distance (reciprocal-distance) weighting method (IDWM) (Wei & McGuinness 1973) is the method most commonly used for estimating missing data. This weighting distance method for estimating the missing value of an observation, which uses the observed values at other stations, is determined by

$$V_0 = \frac{\sum_{i=1}^{n} (V_i/D_i)}{\sum_{i=1}^{n} (1/D_i)} \tag{3}$$

where D_i is the distance between the station with missing data and the ith nearest weather station. The remaining parameters are defined in Equation (2).

NR method

The NR method, first proposed by Paulhus & Kohler (1952) and later modified by Young (1992), is a common method for estimation of rainfall missing data. This method is used if any surrounding gauges have normal annual precipitation exceeding 10% of the considered gauge. This weighs the effect of each surrounding station (Singh 1994). The estimated data is considered as a combination of parameters with different weights, as shown in Equation (4).

$$V_0 = \frac{\sum_{i=1}^{n} W_i V_i}{\sum_{i=1}^{n} W_i} \tag{4}$$

where W_i is the weight of ith nearest weather station expressed in Equation (6).

$$W_i = \left[R_i^2 \left(\frac{N_i - 2}{1 - R_i^2} \right) \right] \tag{5}$$

where R_i is the correlation coefficient between the target station and the ith surrounding station, and N_i is the number of points used to derive correlation coefficient.

SIB

In the SIB method, the closest neighbor station is used as an estimate for a target station. The target station rainfall is estimated using the same data from the neighbor station that has the highest positive correlation with the target station (Hasanpur Kashani & Dinpashoh 2012).

LR

LR is a method used for estimating climatological data at stations with similar conditions. In statistics, LR is an approach for modeling the relationship between scalar dependent variable y and one independent parameter denoted X. LR was the first type of regression analysis to be studied rigorously and to be used extensively in practical applications (Xin 2009). This is because models that depend linearly on their unknown parameters are easier to fit than models that are non-linearly related to their parameters because the statistical properties of the resulting estimators are easier to determine. In this study, the Kish Island station data was used to calculate the missing data of the target station (Bandar Lengeh) using the LR method.

Multiple linear regression

Multiple linear regression (MLR) is a statistical method for estimating the relationship between a dependent variable and two or more independent, or predictor, variables. MLR identifies the best-weighted combination of independent variables to predict the dependent, or criterion, variable. Eischeid *et al.* (1995) highlighted many advantages of this method in data interpolation and estimation of missing data. The missing data (V_0) is estimated from Equation (6).

$$V_o = a_0 + \sum_{i=1}^{n} (a_i V_i) \tag{6}$$

where a_i, $a_1, \ldots, a_n$ are the regression coefficients.

MI

A single imputation ignores the estimation of variability, which leads to an underestimation of standard errors and confidence intervals. To overcome the underestimation problem, multiple imputation methods are used, where each missing value is estimated with a distribution of imputation reflecting uncertainty about the missing data. MI lead to the best estimation of missing values. Since the rainfall data is skewed to the right, the data needs to be transformed by taking the natural logarithm of the observed data before the method is applied. In some cases, the data may not have a normal distribution with a logarithmic transformation. In these cases, other transformation methods such as the Box-Cox power transformations method (Box & Cox 1964) or the Johnson transformation method (Luh & Guo 2000) could be applied. Then, the average of imputed data is calculated to provide the missing data at the target station (Radi *et al.* 2015). In many studies, five imputed data sets are considered sufficient. For example, Schafer & Olsen (1998) suggested that in many applications, three to five imputations are sufficient. In this study, the statistical XLSTAT software was used to generate multiple imputations.

NIPALS algorithm for missing data

The NIPALS algorithm was first presented by Wold (1966) under the name NILES. It iteratively applies the principal component analysis to the data set with missing values. The main idea is to calculate the slope of the least squares line that crosses the origin of the points of the observed data. Here eigenvalues are determined by the variance of the NIPALS components. The same algorithm can estimate the missing data. The rate of convergence of the algorithm depends on the percentage of the missing data (Tenenhaus 1998). In this study, the statistical XLSTAT software is used to generate the NIPALS algorithm.

UK traditional method

This method traditionally used by the UK Meteorological Office to estimate missing temperature and sunshine data was based on comparison with a single neighboring station (Hasanpur Kashani & Dinpashoh 2012). In this study, the

ratio between the average rainfall at the target station (Bandar Lengeh) and the average rainfall at the station with the highest correlation (Kish Island) was calculated. Then, that ratio was multiplied by the rainfall at the station with the highest correlation to the target station.

Decision tree model

The M5 decision-tree model is a modified version of the Quinlan (1992) model, where linear functions rather than discrete class labels (Ajmera & Goyal 2012; Sattari *et al.* 2013) are used at the leaves. The M5 model is based on a divide-and-conquer approach, working from the top to the bottom of the tree (Witten & Frank 2005). This splitting criterion is based on the standard deviation reduction (*SDR*) expressed in Equation (7),

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \tag{7}$$

where T is the set of examples that reaches the node, T_i represents the subset of examples that have the ith outcome of the potential set, and *sd* represents the standard deviation. Applying this procedure results in reduction of standard deviation in child nodes. As a result, M5 chooses the final split as the one that maximizes the expected error reduction (Quinlan 1992). The M5 decision tree may become too large due to overfitting with test data. Quinlan (1992) suggested pruning the overgrown tree.

Performance metrics

In order to compare accuracy of the discussed methods for reconstructing missing monthly rainfall data, the following four metrics, Equations (8)–(11), are used.

$$E = 1 - \frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{8}$$

$$r_{\text{pearson}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \tag{9}$$

$$MAE = \frac{1}{n} \sum_{i=0}^{n} |X_i - Y_i| \tag{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n} (Y_i - X_i)^2}{N}} \tag{11}$$

where X is the observed value and Y denotes the computed value.

Computational results

Considering the importance of data accuracy in climate studies, the standard normal homogeneity test (SNHT) and the Mann-Kendall (MK) trend test were applied to the data sets using XLSTAT software (Table 2). The SNHT test was developed by Alexanderson (1986) to detect a change in a series of rainfall data. The purpose of the MK test (Mann 1945; Kendall 1975; Gilbert 1987) is to statistically assess if there is a monotonic upward or downward trend of the variable of interest over time.

In the SNHT, the null hypothesis (H_0) was homogeneity of the data and the alternative hypothesis (H_1) was heterogeneity of the data. In the MK trend test, the null

Table 2 | Results of homogeneity and trend test of selected stations

	SNHT		MK trend test			
Station	***p*-value**	**Risk of rejecting H_0 (%)**	***p*-value**	**Kendal's tau**	**Risk of rejecting H_0 (%)**	***a***
Abomoosa Island	0.444	44.39	0.448	−0.03	44.76	0.05
Bandar Abbas	0.214	21.40	0.085	−0.067	8.46	0.05
Jask	0.201	20.09	0.310	−0.041	30.95	0.05
Bandar Lengeh	0.168	16.81	0.446	−0.03	44.57	0.05
Kish Island	0.159	15.9	0.206	−0.05	20.63	0.05
Minab	0.640	64.03	0.510	−0.026	50.95	0.05

hypothesis was randomness and absence of any trends in data, and the alternative hypothesis was non-randomness and presence of trends in the data. If the p-value is more than significance level (α), the null hypothesis is confirmed; otherwise, the alternative hypothesis is acceptable. The results in Table 2 show that the data related to monthly precipitation is homogeneous and random at all stations and can be used with confidence. The correlation of monthly precipitation at different stations is important and applicable in modeling. Hence, the correlation between the monthly precipitation at different stations was investigated (Table 3). The synoptic station of Bandar Lengeh was used as the target station.

Table 3 | Correlation matrix of investigated stations

	Bandar Abbas	Minab	Jask	Abomoosa Island	Kish Island	Bandar Lengeh
Bandar Abbas	1	0.837	0.569	0.708	0.721	0.794
Minab	0.837	1	0.529	0.697	0.672	0.743
Jask	0.569	0.529	1	0.623	0.660	0.740
Abomoosa Island	0.708	0.697	0.623	1	0.751	0.793
Kish Island	0.721	0.672	0.660	0.751	1	0.852
Bandar Lengeh	0.794	0.743	0.740	0.793	0.852	1

Table 4 | The rules produced by the M5 model tree for monthly precipitation estimation

Rule no.	If	Then	Equation no.
1	$B.\,Abbas(P) \leq 3.55$ and $Kish(P) \leq 0.15$	$B.\,Lenge(P) = (0.0127{*}\text{B. Abbas(P)}) + (0.0087{*}\text{B. Jask(P)}) + (0.0225{*}\text{Abomoosa(P)}) + (0.0277{*}\text{Kish(P)}) + 0.03$	(12)
2	$Kish(P) \leq 24.7$	$B.\,Lenge(P) = (0.1063{*}\text{B. Abbas(P)}) + (0.0271{*}\text{Bandarjask(P)}) + (0.2112{*}\text{Abomoosa(P)}) + (0.2875{*}\text{Kish(P)}) + 0.036$	(13)
3	*otherwise*	$B.\,Lenge(P) = (0.3675{*}\text{B.Abbas(P)}) + (0.3516{*}\text{B. jask(P)}) + (0.2328{*}\text{Kish(P)}) + (0.096)$	(14)

Note: *B represents Bandar in all equations.

Table 5 | Performance criteria values for different methods of estimating missing monthly rainfall data

Method		R	N-S	RMSE (mm)	MAE (mm)	Mean (mm)	Variance (mm)
	Test phase	–	–	–	–	5.682	14.838
Classical statistical methods	AA	0.95	0.86	5.65	2.78	7.136	17.135
	MLR	0.93	0.87	5.49	2.61	5.897	13.026
	NIPALS	0.94	0.86	5.61	3.25	7.149	15.602
	NR	0.90	0.73	7.992	3.82	8.32	17.091
	IDWM	0.90	0.75	7.70	3.85	8.065	16.792
	MI	0.83	0.53	10.56	8.41	12.508	13.3
	LR	0.65	0.46	11.22	50.00	4.985	8.142
	UK	0.65	0.47	11.16	4.97	4.525	8.846
	SIB	0.65	0.47	11.19	4.60	5.553	10.855
Data mining method	M5	0.95	0.89	5.01	2.48	4.621	12.293

As shown in Table 3, the precipitation at the Bandar Lengeh station is most correlated with the Kish Island, Bandar Abbas, and Abomoosa Island stations, respectively. Latitude is a key factor behind varying precipitation levels across different regions. Precipitation correlation values in different stations are, therefore, positively correlated with their respective latitude. As Table 3 indicates, precipitation correlation values are greater between Bandar Lengeh and

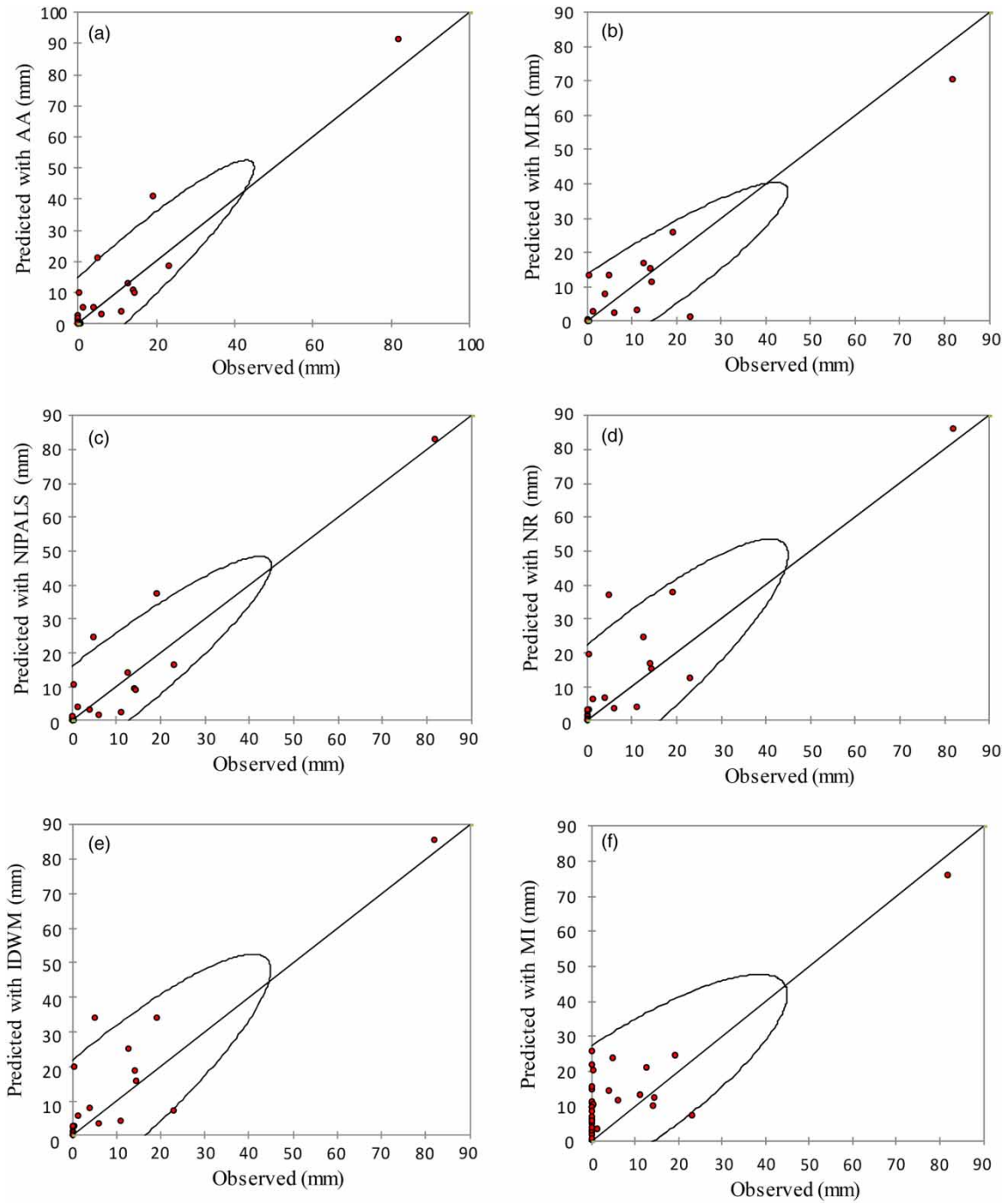

Figure 2 | Scatter diagram of predicted and observed precipitation values generated by (a) AA, (b) MLR, (c) NIPALS, (d) NR, (e) IDWM, (f) MI, (g) LR, (h) UK, (i) SIB, (j) M5. *(Continued.)*

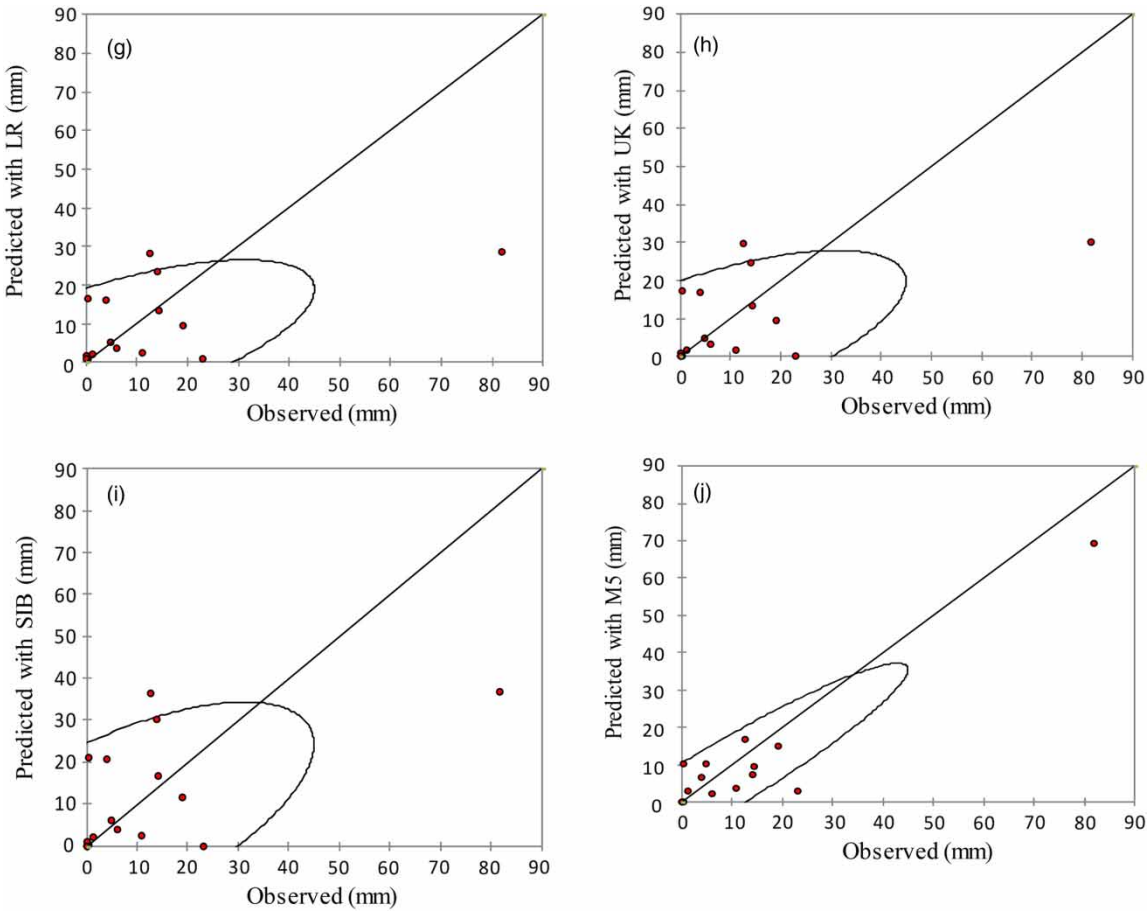

Figure 2 | Continued.

Jask stations than between the Jask and Bandar Abbas stations. This could be attributed to latitudinal proximity of Bandar Abbas to Jask as well as to the evident comparability of the two cities in terms of condition, which also applies to other stations. Out of the total precipitation data at each station, 10% was randomly assumed to be missing. The missing data was used as a test section and the residual one for training. The number of neighboring stations employed in different methods was dependent on the method. For example, in the methods of LR, UK and SIB, only one station's data highly correlated with target station's data was employed, but in the AA, IDWM, MLR, NR and NIPALS methods, all of neighboring stations were used. In the M5 and MI methods, different combinations of input parameters varying from one station to five stations were used to see which one had better performance.

In the multiple imputation method, the best results were obtained for data at five stations. In estimating the missing values of precipitation at Bandar Lengeh, the M5 decision-tree model was selected. The best results were obtained when the data related to monthly precipitation at the stations of Bandar Abbas, Jask, Abomoosa and Kish Islands was used. The M5 model in the form of three decision rules (involving linear Equations (12)–(14)) estimates the monthly precipitation at the Bandar Lengeh station with relatively acceptable accuracy. These rules are provided in Table 4.

Decision rule (1) above states that if the amount of monthly precipitation at Bandar Abbas is equal or less than 3.55 mm and the monthly precipitation at Kish Island is equal or less than 0.15 mm, then monthly precipitation in Bandar Lengeh is calculated from Equation (12). Rule 2 states that if the monthly precipitation at Kish Island is equal to or less than 24.7, then the monthly precipitation

at Bandar Lengeh is calculated using Equation (13). According to rule 3, in other situations, the amount of monthly precipitation at Bandar Lengeh is computed using Equation (14). The results obtained from various classic statistical methods and the M5 decision tree model are presented in Table 5.

The results in Table 5 indicate that among the classical statistical methods, simple AA, MLR, and the NIPALS algorithm are most accurate. The accuracy of the AA method could be due to the fact that the stations under study were located at similar elevation conditions (about 5 to 30 meters above sea level) and followed a rather similar precipitation pattern. The AA and MLR methods may be used in arid areas with similar elevation conditions. The decision tree model provides quite accurate predictions with the correlation coefficient of 0.95,

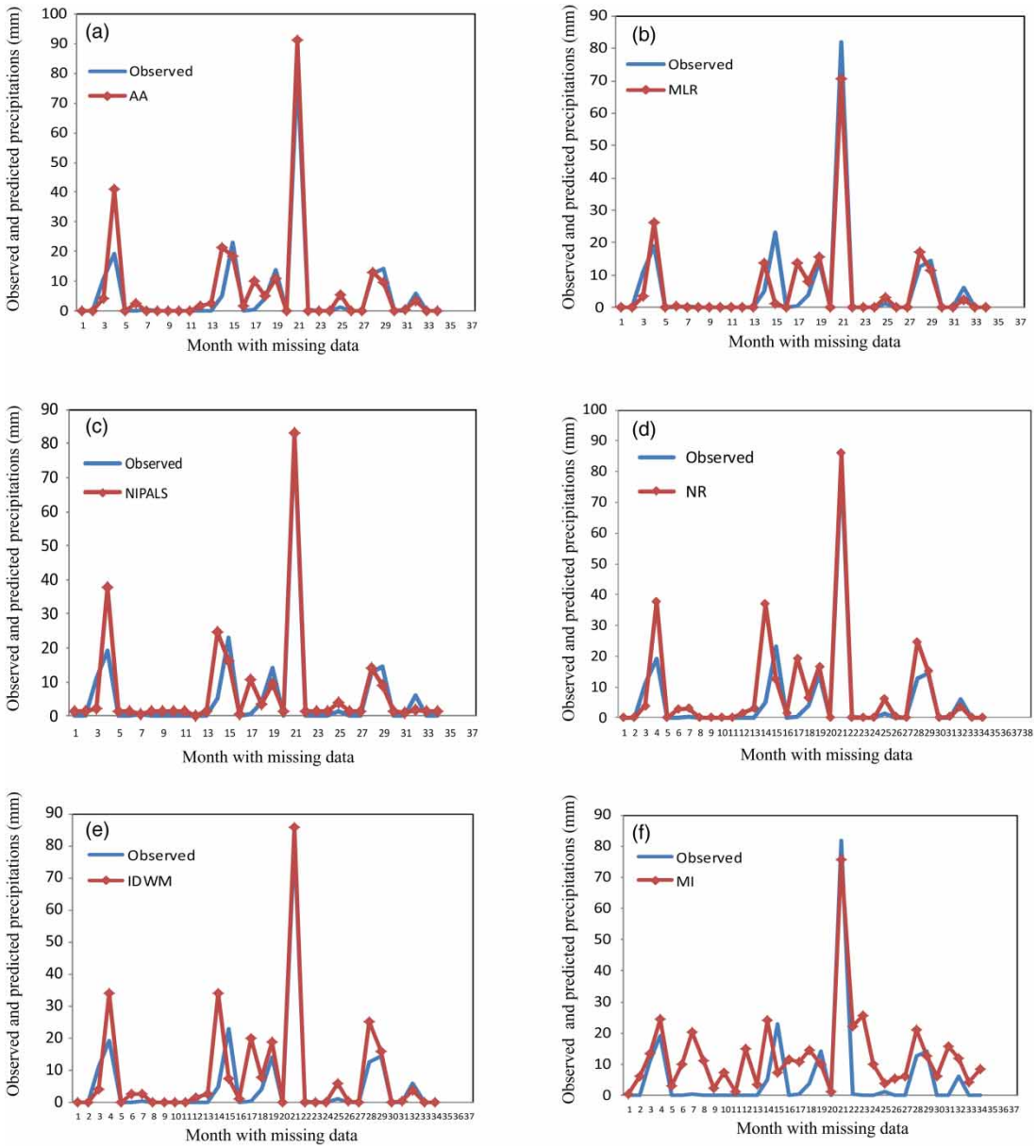

Figure 3 | Time series of predicted and observed values of precipitation generated by (a) AA, (b) MLR, (c) NIPALS, (d) NR, (e) IDWM, (f) MI, (g) LR, (h) UK, (i) SIB, (j) M5. *(Continued.)*

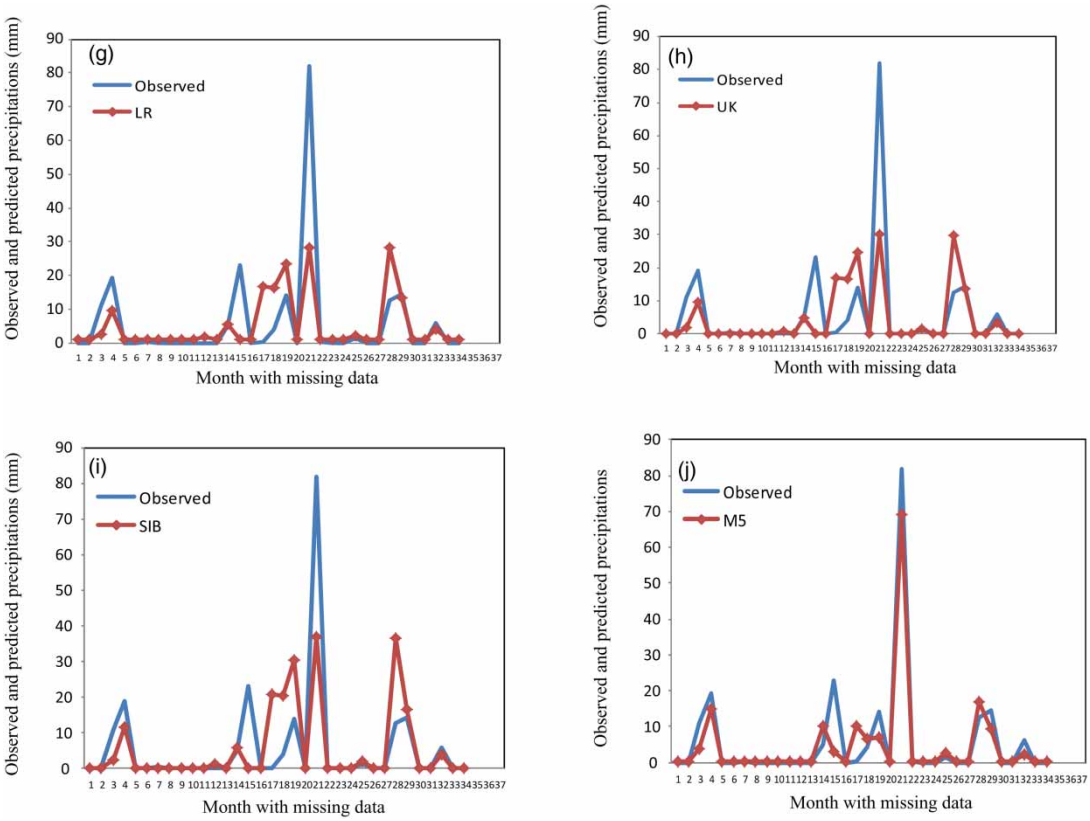

Figure 3 | Continued.

N-S coefficient of 0.891, the root mean square error of 5.066 mm, and the mean absolute error of 2.48 mm. Scatter diagrams and time-series charts produced by various methods are presented in Figures 2 and 3.

Figures 2 and 3 demonstrate that the decision tree algorithms developed with the data preprocessed with the AA method provided better results at the Bandar Lengeh station compared with other approaches studied in this research.

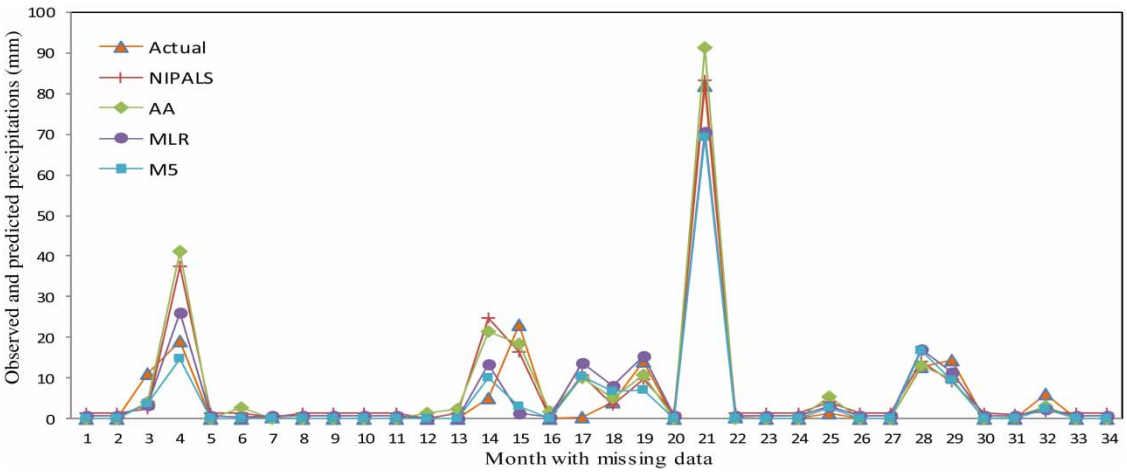

Figure 4 | Time series of values predicted with four models with missing precipitation data.

Figure 4 illustrates the prediction results generated by the (NIPALS) algorithm, AA, MLR, and the decision tree (M5) algorithm. The data used by the models in Figure 4 originated at the Bandar Lengeh station, and it contained missing values. The examination of the results shows that the SIB, LR, and UK methods have minimum accuracy among all methods under the study. This can be due to the nature of these methods, that is, only the precipitation data from one station having maximum correlation with the target station is used.

CONCLUSION

In the study reported in this paper, the monthly precipitation data at six stations located in arid areas was considered. The data collected was homogeneous, and no trends were found. However, numerous values were missing. Different methods were applied to fill in the missing data. The computational results demonstrated that among classical statistical methods, AA, MLR, and the NIPALS algorithm performed best. The high performance of AA might be related to the location of research stations at a similar elevation (between 5 to 30 meters above sea level). Therefore, using the AA method in arid areas with similar elevation is suggested. The results indicated that the MLR method was found to be suitable for estimating missing precipitation data. This result supports the findings of Eischeid *et al.* (1995), Xia *et al.* (1999), and Hasanpur Kashani & Dinpashoh (2012). Furthermore, Shih & Cheng (1989) stated that the regression technique and the regional average can be applied to generate missing monthly solar radiation data. They found the regression technique and AA satisfactory in interpolating missing values. The multiple imputation method performed best when precipitation data from five dependent stations was used. This finding was supported by the results reported in Radi *et al.* (2015). The research reported in this paper has demonstrated that the results if-then rules produced by the decision-tree algorithm provided high accuracy results with the correlation coefficient of 0.95, Nash-Sutcliffe coefficient of 0.89, root mean square error of 5.07 mm, and the mean absolute error of 2.48 mm. Due to its simplicity and high accuracy, the decision-tree model was suggested for estimating the missing values of precipitation in non-arid climates. Although the results reported in this paper were derived from regions in a single country, the results would be applicable to arid and semi-arid regions in other countries. This is due to the fact that all arid and semi-arid regions share the same or similar climate conditions.

REFERENCES

Abraham, J. P., Baringer, M., Bindoff, N. L., Boyer, T., Cheng, L. J., Church, J. A., Conroy, J. L., Domingues, C. M., Fasullo, J. T., Gilson, J., Goni, G., Good, S. A., Gorman, J. M., Gouretski, V., Ishii, M., Johnson, G. C., Kizu, S., Lyman, J. M., Macdonald, A. M., Minkowycz, W. J., Moffitt, S. E., Palmer, M. D., Piola, A. R., Reseghetti, F., Schuckmann, K., Trenberth, K. E., Velicogna, I. & Willis, J. K. 2013 A review of global ocean temperature observations: implications for ocean heat content estimates and climate change. *Reviews of Geophysics* **51**, 450–483.

Abraham, J. P., Stark, J. R. & Minkowycz, W. J. 2015 Extreme weather: observations of precipitation changes in the USA, forensic engineering. *Proceedings of the Institution of Civil Engineers* **168**, 68–70.

Ajmera, T. K. & Goyal, M. K. 2012 Development of stage–discharge rating curve using model tree and neural networks: an application to Peachtree creek in Atlanta. *Expert Systems with Applications* **39**, 5702–5710.

Alexanderson, H. 1986 A homogeneity test applied to precipitation data. *International Journal of Climatology* **6**, 661–675.

Box, G. E. P. & Cox, D. R. 1964 An analysis of transformations. *Journal of Royal Statistical Society, Series B (Methodological)* **26**, 211–252.

Che Ghani, N., Abuhasan, Z. & Tze Liang, L. 2014 Estimation of missing rainfall data using GEP: case study of raja river, Alor Setar, Kedah. *Advances in Artificial Intelligence.* http://dx.doi.org/10.1155/2014/716398, p. 5.

Cheng, L., Abraham, J., Goni, G., Boyer, T., Wijffels, S., Cowley, R., Gouretski, V., Reseghetti, F., Kizu, S., Dong, S., Bringas, F., Goes, F., Houpert, L., Sprintall, J. & Zhu, J. 2015a XBT science: assessment of XBT biases and errors. *Bulletin of the American Meteorological Society.* Doi: 10.1175/BAMS-D-15-00031.1.

Cheng, L., Zhu, J. & Abraham, J. P. 2015b Global upper ocean heat content estimation: recent progresses and the remaining challenges. *Atmospheric and Oceanic Science Letters* **8**, 333–338.

Choge, H. K. & Regulwar, D. G. 2013 Artificial neural network method for estimation of missing data. *International Journal of Advanced Technology in Civil Engineering* **2**, 1–4.

Dastorani, M. T., Moghadamnia, A., Piri, J. & Rico-Ramirez, M. 2009 Application of ANN and ANFIS models for reconstructing missing flow data. *Environment Monitoring Assessment.* doi:10.1007/s10661-009-1012-8.

De Martonne, E. 1923 Aridité et Indices D'Aridité. *Académie Des Sciences. Comptes Rendus* **182**, 1935–1938.

De Silva, R. P., Dayawansa, N. D. K. & Ratnasiri, M. D. 2007 A comparison of methods used in estimating missing rainfall data. *Journal of Agricultural Sciences* **3**, 101–108.

Eischeid, J. K., Baker, C. B., Karl, T. R. & Diaz, H. F. 1995 The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology and Climatology* **34**, 2787–2795.

Gilbert, R. O. 1987 *Statistical Methods for Environmental Pollution Monitoring.* Wiley, NY.

Hasanpur Kashani, M. & Dinpashoh, Y. 2012 Evaluation of efficiency of different estimation methods for missing climatological data. *Journal of Stochastic Environment Research Risk Assessment* **26**, 59–71.

Hosseini Baghanam, A. & Nourani, V. 2011 Investigating the ability of artificial neural network (ANN) models to estimate missing rain-gauge data. *Journal of Recent Research in Chemistry, Biology, Environment and Culture* **19**, 38–50.

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

Kim, J. & Pachepsky, A. Y. 2010 Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology* **394**, 305–314.

Luh, W. M. & Guo, J. H. 2000 Johnson's transformation two-sample trimmed t and its bootstrap method for heterogeneity and non-normality. *Journal of Applied Statistics* **27**, 965–973.

Lyman, J. & Johnson, G. 2008 Estimating annual global upper-ocean heat content anomalies despite irregular in situ ocean sampling. *J. Climate* **21**, 5629–5641.

Mann, H. B. 1945 Non-parametric tests against trend. *Econometrica* **13**, 163–171.

Nkuna, T. R. & Odiyo, J. O. 2011 Filling of missing rainfall data in Luvuvhu river catchment using artificial neural networks. *Journal of Physics and Chemistry of Earth* **36**, 830–835.

Paulhus, J. L. H. & Kohler, M. A. 1952 Interpolation of missing precipitation records. *Monthly Weather Review* **80**, 129–133.

Quinlan, J. R. 1992 Learning with Continuous Classes. In: *Proceedings AI'92* (Adams & Sterling, eds), World Scientific, Singapore, pp. 343–348.

Radi, N., Zakaria, R. & Azman, M. 2015 Estimation of missing rainfall data using spatial interpolation and imputation methods. *AIP Conference Proceedings* **1643**, 42–48.

Sattari, M. T., Pal, M., Apaydin, H. & Ozturk, F. 2013 M5 model tree application in daily river flow forecasting in Sohu stream, Turkey. *Water Resources* **40**, 233–242.

Schafer, J. L. & Olsen, M. K. 1998 Multiple imputations for multivariate missing-data problems: a data analysis perspective. *Multivariate Behavioral Research* **33**, 545–571.

Shih, S. F. & Cheng, K. S. 1989 Generation of synthetic and missing climatic data for Puerto Rico. *Water Resources Bulletin* **25**, 829–836.

Singh, V. P. 1994 *Elementary Hydrology.* Prentice Hall of India, New Delhi.

Te Chow, V., Maidment, D. R. & Mays, L. W. 1988 *Applied Hydrology.* McGraw-Hill, New York.

Teegavarapu, R. S. V. 2009 Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics* **11**, 133–146.

Teegavarapu, R. S. V. 2014 Statistical corrections of spatially interpolated missing precipitation data estimates. *Hydrological Process* **28**, 3789–3808.

Teegavarapu, R. S. V. & Chandramouli, V. 2005 Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology* **312**, 191–206.

Teegavarapu, R. S. V., Tufail, M. & Ormsbee, L. 2009 Optimal functional forms for estimation of missing precipitation data. *Journal of Hydrology* **374**, 106–115.

Tenenhaus, M. 1998 *La Régression PLS Théorie et Pratique.* Editions Technip, Paris.

Wei, T. C. & McGuinness, J. L. 1973 *Reciprocal Distance Squared Method: A Computer Technique for Estimating Area Precipitation*. Technical Report ARS-Nc-8. US Agricultural Research Service, North Central Region, OH, USA.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco.

Wold, H. 1966 Nonlinear Estimation by Iterative Least Square Procedures. In: *Research Papers in Statistics* (F. David, ed.). Wiley, New York, pp. 411–444.

Xia, Y., Fabian, P., Stohl, A. & Winterhalter, M. 1999 Forest climatology: estimation of missing values for Bavaria, Germany. *Agricultural and Forest Meteorology* **96**, 131–144.

Xin, Y. 2009 *Linear Regression Analysis: Theory and Computing.* World Scientific, London, Vol. 1–2.

You, J., Hubbard, K. G. & Goddard, S. 2008 Comparison of methods for spatially estimating station temperatures in a quality control system. *International Journal of Climatology* **28**, 777–787.

Young, K. C. 1992 A three-way model for interpolating monthly precipitation values. *Monthly Weather Review* **120**, 2561–2569.

First received 10 February 2016; accepted in revised form 3 August 2016. Available online 30 September 2016