



Missing rainfall data estimation—an approach to investigate different methods: case study of Baghdad

Mukhalad Abdullah^{1,2} · Nadhir Al-Ansari²

Received: 21 July 2021 / Accepted: 12 November 2022 / Published online: 30 November 2022
© The Author(s) 2022

Abstract

The missing of the meteorological data in Iraq is common due to malfunction of measuring devices, security status, and human effects. The study tested 17 missing precipitation data estimation methods in Baghdad city as a case study, where, all the surrounding stations around Baghdad experienced the missing of data for various reasons, and some of the missing data are for a full year record. The methods examined in this study are based on different approaches, some of the methods are based upon the distances to the targeted station, others are upon regression factors, and there are also methods that combine several factors. There are also other types of missing data filling methods which depend on imputation and artificial intelligence. The investigation of the most accurate method to find the missing data will assist researchers and decision makers to fill the gap in their analysis in one of the most vulnerable countries in terms of drought and climate changes impacts. Results showed that Expectation Maximization (EM) method utilization has the best results with the least errors, and Multiple Linear Regression (MLR) method was ranked the second best method. In general, all of the applied methods had resulted acceptable interpolations, and it was clear that the combined methods have low significance on the results in comparison with others. All of these findings are limited to the study area meteorological and spatial conditions.

Keywords Missing precipitation · Expectation maximization · Multiple linear regression · Baghdad precipitation · Arid zones precipitation

Introduction

Baghdad station was selected as a case study in this research. Baghdad is the capital city and also the largest city in Iraq; it is also the main location of human activities in the Mesopotamian plain. The major irrigation projects are within Baghdad area (Abdullah and Al-Ansari 2021), where the accuracy of climate records is far important to plan and enhance irrigation practices; also, this area is the hub of transportation infrastructures in the country.

Iraqi Meteorological Organization and Seismology (IMOaS) is the official authority that manages the meteorological and seismology data. It is common to find missing data for many reasons, for instance, in the

years 2003 and 2004, most of the data are missed even in Baghdad station. Another example was in the period from 2014 to 2017; the meteorological stations in the western and northern parts of the country had stopped recording the data (IMOaS 2021).

The finding of the most suitable algorithm to perform the filling of missing rainfall data is essential. It is worth mentioning also that the climate and geospatial characteristics of Baghdad are greatly approaching other parts of Mesopotamia; thus, it can reliably extrapolate the results in other areas.

Many methods were successfully tested and adopted in other parts of the world, where the missing data filling models are based on several concepts, which are mainly the correlation with the surrounding stations, the spatial analysis with the surrounding stations, and the artificial intelligence.

Several efforts were made by researchers to predict the missing rainfall data, as it is one of the common problems. Tropical Rainfall Measuring Mission (TRMM) estimates were analyzed; the method showed some limitation near the water bodies and when the precipitation

Responsible Editor: Broder J. Merkel

✉ Nadhir Al-Ansari
nadhir.alansari@ltu.se; mukhaladabdullah@gmail.com

¹ Baghdad, Iraq

² Lulea University of Technology, Lulea, Sweden

is lower than 200 mm annually, but rather that, this provides results with a correlation coefficient that might reach 0.91 (Abdulrazzaq 2020). This method also provides some acceptable prediction in a specific circumstance, but with overestimation during dry months at different locations (Abdulrida and Al-Jumaily 2016).

Artificial neural network (ANN) method to test the estimation of four different stations, which are Basra, Baghdad, Mosul, and Rutba, had shown good results (Al-Salihi et al. 2013); testing of this method also applied in Sulaymaniyah city, northeast Iraq, and for the period of 2013–2018, the results of ANN provide a good estimation up to 91.5% of accuracy (Murad and Jaff 2020).

Isohyet method is adopted to estimate the missing rainfall data in Nineveh governorate, which includes Mosul City. The study was made for 8 weather stations, during 20 years from 2000 to 2019. The results are promising and generally good (Alozeer 2020).

Methodology and data

Study area

Baghdad station was chosen as a case study where it is located in Iraq. Iraq is located in the Middle East covering an area of 438,320 km². The climate of Iraq is mainly continental, subtropical semi-arid type, with the north and north-eastern mountainous regions having a Mediterranean climate. In most of the country territories, the rainfall is very seasonal, which starts in the winter from December to February. Regarding the north and northeast of the country, the rainfall starts from November to April. Average annual rainfall is about 216 mm; it varies from 1200 mm in the northeast to less than 100 mm over 60% of the country in the south. Regarding the temperatures, winters are cool to cold; the temperature in the day is 16 °C and dropping at night to 2 °C with a possibility of frost. Summers are dry and hot to extremely hot, with a shade temperature of over 43 °C during July and August, yet dropping at night to 26 °C (Frenken 2009).

As shown in Fig. 1, several stations exist and are managed by IMOaS; unfortunately, many of the data were missed in these stations, and some data are scattered and unreliable. Data for the years 2003 and 2004 are mostly not available, and the data for the years 2014 to 2017 were missed in Ramadi station.

The resort to another alternative was to compile the raw from the online database of Texas A&M University (globalweather.tamu.edu/). The data are generated from climate forecast system reanalysis (CFSR); this tool is a global, high-resolution, coupled atmosphere–ocean–land

surface–sea ice system designed to provide the best estimate of the state of these coupled domains over the targeted period. The system has several strengths and limitations; one of these limitations is the few relative evaluations that had been conducted.

The analysis was made for the period of 35 years, from 1980 to 2014. Figure 1 and Table 1 show the statistical and spatial characteristics of the study area stations. Twelve stations were gathered; these stations are Razazza, Jurf Al-Sakhar, Latifiyah, Suwairah, Falluja, Abu Ghraib, Baghdad, Nahrawan, Tharthar, Taji, Rashidiyah, and Khan Saad. Baghdad station is considered in this study the target station, it lies in the center of the area, and it is the most important one; there is no difference in elevation for most of the stations, except Tharthar stations which have an elevation of 82 m above sea level (a.s.l.). Also, the maximum diagonal distance between stations is 108 km, while the closest distance is around 29 km between Baghdad and Abu Ghraib stations. Figure 1 shows the geographical location of the study area stations.

In order to examine the hypothesized missing data in the analysis, Baghdad station was considered the target station where the missing occurred. Also, figures and conclusions were shown for the time period from 1988 to 1995, and the years 1982 and 2000, where these years are the wettest and driest, respectively, in terms of the time duration of the study.

There are 17 methods performed to analyze the proposed missing data findings, among which are simple and complicated methods, and some were made with help of computer software.

Missing data estimation methods

Arithmetic average (AA)

This method is one of the easiest and most widely used in hydrologic application. It is simply the mean of the surrounding stations to the targeted area in the study zone. According to Linsley et al. (1975), AA method will yield a good estimate in flat country; the surrounding stations are uniformly distributed, and the individual stations do not vary so far from the mean. Use of this method is limited when the topography is more complicated; Eq. (1) as following is the formula of the AA method:

$$Y = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

where: Y is the missing value at the target station, X_i is the measured value of the i th surrounding station, and n is the number of these stations.

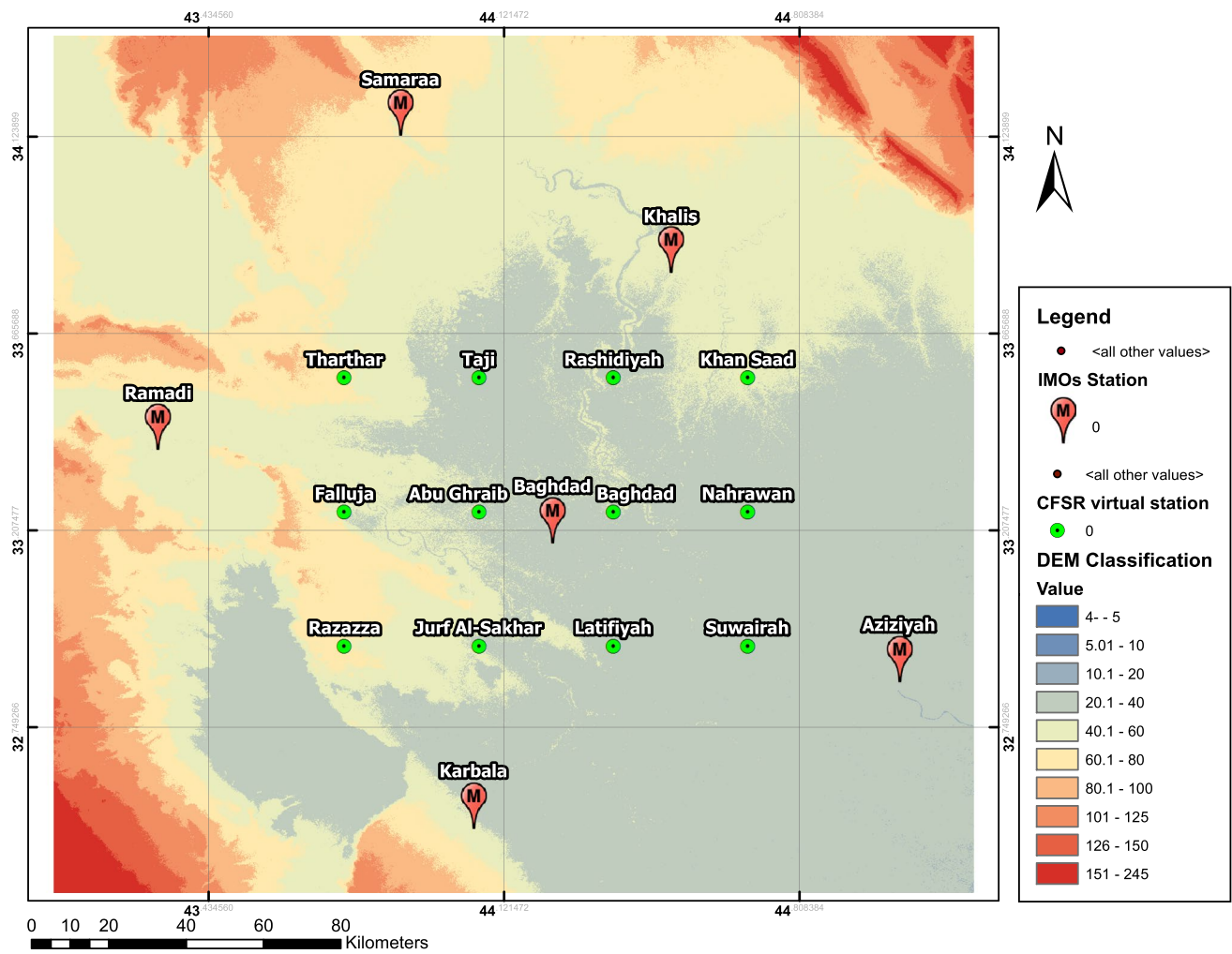


Fig. 1 Geographical location of the stations within the study area

Table 1 Statistical and spatial characteristics of the stations within the study area

Station	Latitude (N)	Longitude (E)	Elevation (m a.s.l.)	*Average rainfall (mm)	*Maximum rainfall (mm)	*Minimum rainfall (mm)	*Standard deviation
Razazza	32°56'24.37"	43°45'0.00"	54	7.7	109.0	0.0	13.7
Jurf Al-Sakhar	32°56'24.37"	44° 3'45.00"	30	7.5	80.5	0.0	12.8
Latifiyah	32°56'24.37"	44°22'30.00"	26	7.6	86.8	0.0	12.4
Suwaitrah	32°56'24.37"	44°41'15.00"	29	9.1	92.9	0.0	14.2
Falluja	33°15'8.64"	43°45'0.00"	38	8.6	121.4	0.0	14.7
Abu Ghraib	33°15'8.64"	44° 3'45.00"	33	8.5	111.2	0.0	14.4
Baghdad	33°15'8.64"	44°22'30.00"	33	9.2	82.3	0.0	14.5
Nahrawan	33°15'8.64"	44°41'15.00"	36	11.7	101.3	0.0	17.2
Tharthar	33°33'52.57"	43°45'0.00"	82	9.5	88.9	0.0	14.7
Taji	33°33'52.57"	44° 3'45.00"	42	9.9	95.4	0.0	15.3
Rashidiyah	33°33'52.57"	44°22'30.00"	35	11.6	98.2	0.0	17.2
Khan Saad	33°33'52.57"	44°41'15.00"	38	14.7	118.2	0.0	21.1

*Based on monthly data

Normal ratio (NR)

The NR method is applied when the annual mean of any surrounding station is no more than 10% of that for the target station; this method was adopted by the US National Weather Services (Anderson 1972); this method was firstly proposed by Paulhus and Kohler (1952), where the ratios between the targeted station and surrounding stations are the weighting factor as in equation below:

$$Y = \frac{1}{n} \sum_{i=1}^n \frac{N_s}{N_i} X_i \quad (2)$$

where N_s is the mean of available rainfall data at the target station, N_i is the mean of the available rainfall data at the i th surrounding stations, and n is the number of surrounding stations. Although, some stations in the study have a mean difference by more than 10% of Baghdad station, but these stations were considered in the calculation to examine the limitation of this criterion within the Baghdad area.

Geographical coordinates (GC)

Regarding the geographical coordinates method, it is weighting of the vertical and horizontal coordinates with reference to the total of all surrounding stations around target station (Yozgatligil et al. 2012). The inputs as in the equation below are the latitude and longitude of the stations; the GC method formula is:

$$Y = \sum_{i=1}^n \left[\frac{\frac{1}{x_i^2 + y_i^2}}{\sum_{i=1}^n \frac{1}{x_i^2 + y_i^2}} \right] X_i \quad (3)$$

where: x_i and y_i are the longitude and latitude of the i th surrounding station.

Normal ratio with geographical coordinates (NRGC)

This method is adopted to combine the weighting factors of mean ratios and geographical coordinates; some researchers find a slightly better accuracy when employing this method (Armanuos et al. 2020). The formula of NRGC method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{\left(\frac{1}{x_i^2 + y_i^2} \right) \left(\frac{N_s}{N_i} \right)}{\sum_{i=1}^n \left(\frac{1}{x_i^2 + y_i^2} \right) \left(\frac{N_s}{N_i} \right)} \right] X_i \quad (4)$$

Inverse distance weighting (IDW)

This method has been widely used, since it was first introduced by the United States Department of Agriculture

(USDA) to estimate the missing rainfall data by considering the reciprocal of the inverse of distances between the target station and the surrounding stations (Barbalho et al. 2014). The formula of IDW method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{\left(\frac{1}{d_i^k} \right)}{\sum_{i=1}^n \left(\frac{1}{d_i^k} \right)} \right] X_i \quad (5)$$

where: d_i is the distance from the target station to the i th surrounding station, and k is the distance of friction varying from 1 to 6; in this study, k was assumed to equal 1.

Correlation coefficient weighted (CCW)

According to Teegavarapu and Chandramouli (2005), this method yields a better result as long as the correlation between the target and surrounding stations is higher. This method gives the weight of the ratio of the correlation coefficient. Since the correlation factors between Baghdad station and the other stations are almost above 0.9, then, a promising result is expected. The formula of CCW method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{(r_i)}{\sum_{i=1}^n (r_i)} \right] X_i \quad (6)$$

where: r_i is the Pearson correlation coefficient (rPearson) between the target station and each surrounding station.

Linear regression (LR)

Simply, this method is used to establish a linear relation between the targeted station and the most correlated nearby station in terms of statistics (Armanuos et al. 2020). Once the linear equation is derived, the estimated values can be calculated using this formula. In this study, the linear equation was established between Abu Ghraib station and Baghdad Station, as the correlation between both is the highest. The formula of this method is as follows:

$$Y = a + bX_i \quad (7)$$

where: Y is the estimated rainfall data of the targeted station, and X_i is the observed rainfall value of the neighboring station; a is the intercept, and b is the regression coefficient.

Multiple linear regression (MLR)

MLR is based on the same concept as (LR) method, but the modification with this method is that the regression

is linked with all other stations in the study area (Teegaravapu, 2009). The factors were calculated using Excel-Microsoft Office software. The formula for MLR is as follows:

$$Y = b_0 + \sum_{i=1}^n b_i X_i \tag{8}$$

where: Y is the estimated rainfall data at the target station, X_i is the observed rainfall value of the i th surrounding station, b_i are the regression coefficients of the i th surrounding stations, and n is the number of the surrounding stations.

Multiple imputation (MI)

This method was first introduced by Rubin (1988) in 1988. It is based on the distribution of imputation that reflects uncertainty of the missing data, in order to overcome the underestimation of single imputation (Sattari and Rezazadeh Joudi 2016). There are different software applications to perform this method. In this study, SPSS Statistics software was adopted to conduct the missing data calculations.

Nonlinear iterative partial least squares (NIPALS) algorithm for missing data (NIPALS)

The NIPALS method was first introduced by Wold (1968). The algorithm of this method is to calculate the slope of the least squares line that crosses the origin of the points of the observed data. where eigenvalues are determined by the variance of the NIPALS components. In this study, SPSS Statistics software was adopted to conduct the missing data calculations using NIPALS method.

UK method (UK)

This method is adopted by the UK Meteorological Office to calculate missing data of meteorological components where the comparison was held with one of a single nearby station (Armanuos et al. 2020)(Kashani and Dinpashoh 2011). Since Abu Ghraib station has the highest correlation with Baghdad station, so this station was adopted for the application of the UK method. The estimated values were calculated by multiplying the values in Abu Ghraib station by the ratio of mean rainfall of Abu Ghraib station to that of Baghdad station.

Expectation maximization (EM)

This method was first proposed by Dempster et al. (1977); EM method is a multilayer perceptron type neural network and multiple imputation strategy using Monte Carlo Markov Chain based on expectation–maximization

(Yozgatligil et al. 2012). It is an iterative method both for the estimation of mean values and covariance matrices from incomplete data (Schneider 2001). In this study, SPSS Statistics software was adopted to perform EM method.

Closest station method (CSM)

This is the simplest and easiest method to predict the missing data of meteorological factors. After analyzing the long records of data, the missing values are replaced with the data from a nearby station that has the highest correlation coefficient (Bárdossy and Pegram 2014; Kanda et al. 2017). In the case under this study, the station with the best correlation is Abu Ghraib station. Also, it is worth to mention that this method is named with different jargons, but has the same algorithm.

Modified correlation coefficient with inverse distance weighting (MCCIDW)

The IDW and CCW methods are combined in a single formula [18]. The MCCIDW method gives a power for the correlation coefficient and the distance which is symbolled p , ranging from 1 to 6 (Armanuos et al. 2020), and for the purpose of calculation, p is considered to be 1. The formula of MCCIDW method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{\left(\frac{r_i^p}{d_i^p} \right)}{\sum_{i=1}^n \left(\frac{r_i^p}{d_i^p} \right)} \right] X_i \tag{9}$$

Modified old normal ratio with inverse distance (ONRID)

As in the previous method, this method adopted another approach by combining the effect of distance and mean ratios between stations (Azman, Zakaria, & Ahmad Radi, 2015; Syed Jamaludin et al. 2008); the formula of this method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{\left(\frac{N_s}{N_i} \cdot d_i^{-2} \right)}{\sum_{i=1}^n \left(\frac{N_s}{N_i} \cdot d_i^{-2} \right)} \right] X_i \tag{10}$$

Normal ratio inverse distance weighting with correlation (NRIDC)

In this method, a new combination is proposed by Azman et al. (2015), by considering the superimposition of NR,

Table 2 rPearson coefficient matrix of the studied stations

	Razazza	Jurf Al-Sakhar	Latifiyah	Suwairah	Falluja	Abu Ghraib	Baghdad	Nahrawan	Tharthar	Taji	Rashidiyah	Khan Saad
Razazza	1.00											
Jurf Al-Sakhar	0.97	1.00										
Latifiyah	0.87	0.96	1.00									
Suwairah	0.78	0.87	0.96	1.00								
Falluja	0.97	0.94	0.85	0.78	1.00							
Abu Ghraib	0.97	0.97	0.91	0.83	0.98	1.00						
Baghdad	0.91	0.95	0.96	0.92	0.92	0.97	1.00					
Nahrawan	0.81	0.87	0.93	0.96	0.82	0.88	0.96	1.00				
Tharthar	0.89	0.88	0.83	0.78	0.95	0.93	0.89	0.83	1.00			
Taji	0.91	0.91	0.87	0.81	0.96	0.97	0.95	0.88	0.98	1.00		
Rashidiyah	0.86	0.88	0.87	0.85	0.90	0.93	0.96	0.94	0.91	0.97	1.00	
Khan Saad	0.78	0.82	0.85	0.86	0.81	0.86	0.93	0.96	0.85	0.90	0.97	1.00

Table 3 Results of Pettit's test to examine data homogeneity

Station	Pettit's test		Degree of confidence
	p value	Risk of rejecting H ₀ (%)	
Razazza	0.759	75.92	95%
Jurf Al-Sakhar	0.879	87.86	95%
Latifiyah	0.849	84.88	95%
Suwairah	0.748	74.84	95%
Falluja	0.671	67.1	95%
Abu Ghraib	0.878	87.76	95%
Baghdad	0.662	66.22	95%
Nahrawan	0.536	53.6	95%
Tharthar	0.441	44.08	95%
Taji	0.402	40.16	95%
Rashidiyah	0.268	26.76	95%
Khan Saad	0.191	19.14	95%

IDW, and CCW in the same formula. The formula of NRIDC method is as follows:

$$Y = \sum_{i=1}^n \left[\frac{\left(r_i^p \cdot \frac{N_s}{N_i} \cdot d_i^{-2} \right)}{\sum_{i=1}^n \left(r_i^p \cdot \frac{N_s}{N_i} \cdot d_i^{-2} \right)} \right] X_i \tag{11}$$

where the power of the correlation coefficient P should be more than 4.

Modified normal ratio based on square root distance (MNR-T)

This method was first proposed by Tang et al. (1996); it also combines the weighting of mean ratios and the distance to target station as in the ONRID method, but with another formulation. The MNR-T formula is as follows:

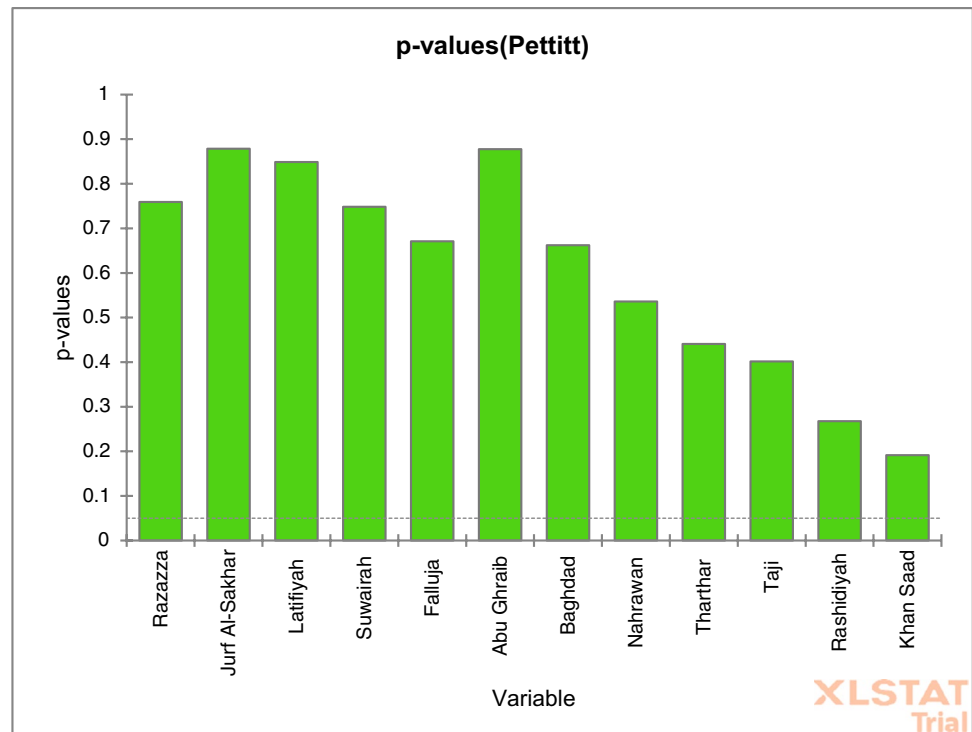
$$Y = \sum_{i=1}^n \frac{N_s}{N_i} \cdot \left[\frac{\left(d_i^{\frac{1}{p}} \right)}{\sum_{i=1}^n \left(d_i^{\frac{1}{p}} \right)} \right] X_i \tag{12}$$

where: the power of the distance p ranges from 1.5 to 2, where for the purpose of calculations, p is considered to be equal to 1.75.

Metrics of performance

In order to evaluate the performance of each of 17 proposed methods in this study, several error measurements were conducted to find the error between the predicted

Fig. 2 *p* values of the tested stations



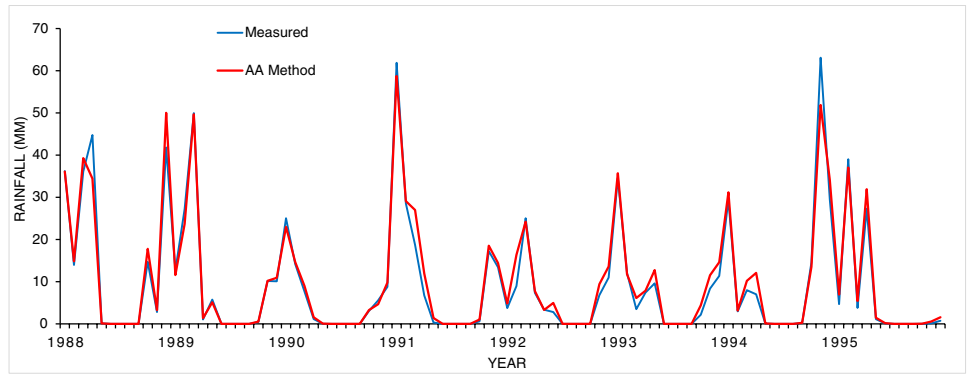
and observed values. This study uses six methods: mean absolute error (MAE) which is one of the measures of error, the output varies from 0 to ∞ , less values mean better results (Azman et al. 2015)(C. Willmott et al. 2009); root mean square error (RMSE): this measure

is very common in meteorological application, and it is very similar to (MAE); coefficient of efficiency (CE), output values of CE range from -1 to $+1$, the value of 1.0 shows a perfect estimation, while on the contrary, as approaching -1 , means not a good estimation (Kashani

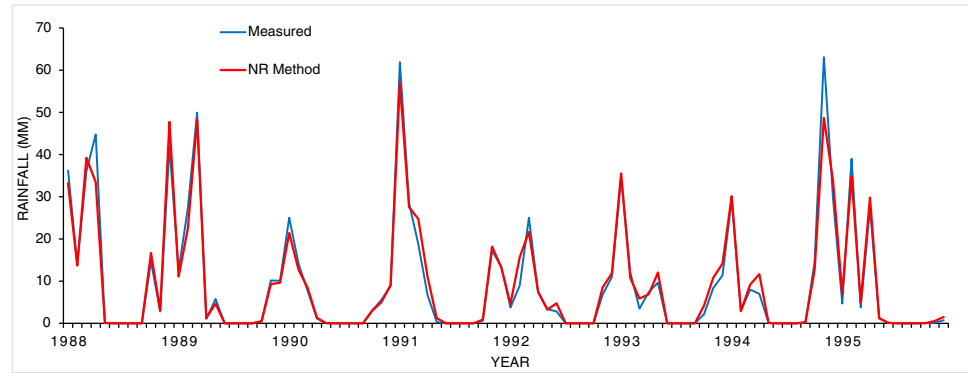
Table 4 Results of performance metrics of the applied methods

Method	MAE	RMSE	CE	S-index	SS	$R_{pearson}$
AA	1.189	2.338	0.974	0.993	0.974	0.987
NR	1.170	2.408	0.972	0.993	0.972	0.986
GC	1.188	2.343	0.974	0.993	0.974	0.987
NRGC	1.176	2.419	0.972	0.993	0.972	0.986
IDW	1.095	2.079	0.979	0.995	0.979	0.991
CSM	1.760	3.736	0.933	0.983	0.933	0.968
CCW	1.178	2.313	0.974	0.994	0.974	0.988
LR	2.061	3.646	0.936	0.983	0.936	0.968
MLR	0.245	0.455	0.999	1.000	0.999	1.000
UK	1.742	3.735	0.933	0.983	0.933	0.968
MCCIDW	1.088	2.062	0.980	0.995	0.980	0.991
ONRID	1.353	2.429	0.972	0.993	0.972	0.990
NRIDC	1.347	2.417	0.972	0.993	0.972	0.991
MNR-T	1.287	2.665	0.966	0.991	0.966	0.983
MI	0.328	0.959	0.980	0.995	0.980	0.995
NIPALS	0.344	1.234	0.967	0.992	0.967	0.984
EM	0.061	0.209	0.999	1.000	0.999	1.000

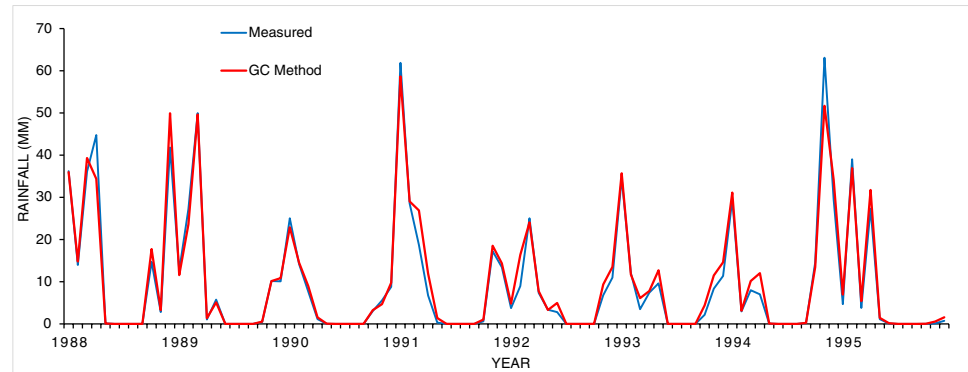
Fig. 3 Time series comparison of monthly precipitation prediction with measured values for the years 1988 to 1995, methods abbreviation as following: **a.** Arithmetic average (AA); **b.** normal ration (NR); **c.** geographical coordinates (GC); **d.** normal ration with geographical coordinates (NRGC); **e.** inverse distance weighted (IDW); **f.** correlation coefficient weighted (CCW); **g.** linear regression (LR); **h.** multiple linear regression (MLR); **i.** multiple imputation (MI); **j.** nonlinear iterative partial least square (NIPALS); **k.** UK method (UK); **l.** expectation maximization (EM); **m.** closet station method (CSM); **n.** modified correlation coefficient with inverse distance weighting (MCCIDW); **o.** modified old normal ration with inverse distance (ONRID); **p.** normal ration inverse distance weighting with correlation (NRIDC); **q.** modified normal ration based on square root distance (MNR-T)



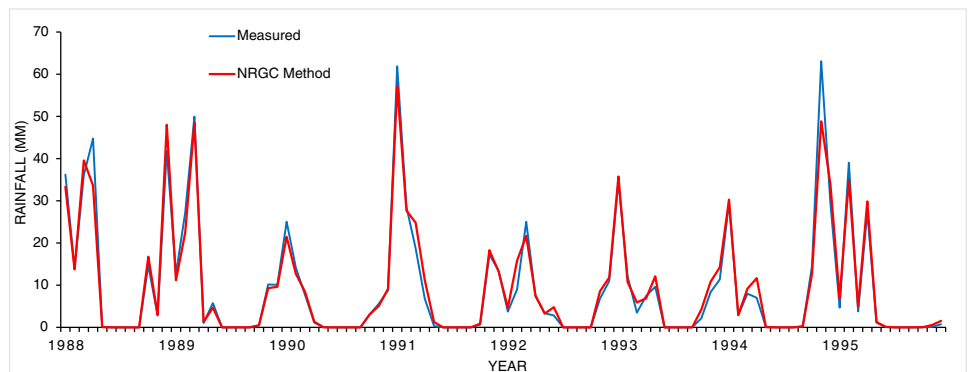
(a)



(b)



(c)



(d)

Fig. 3 (continued)

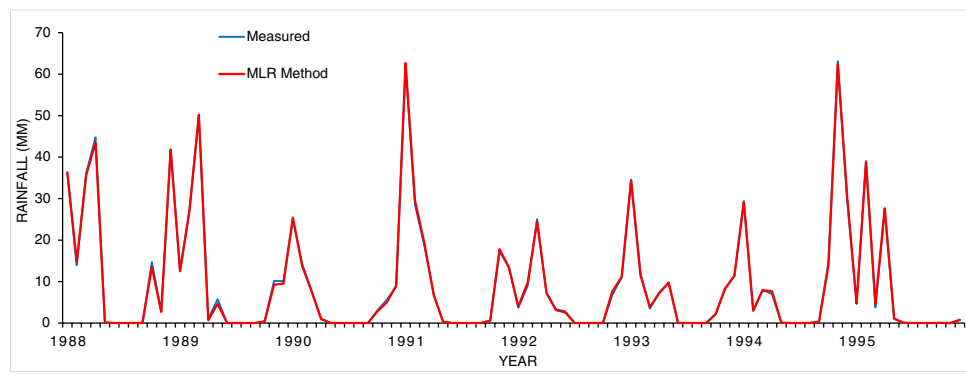
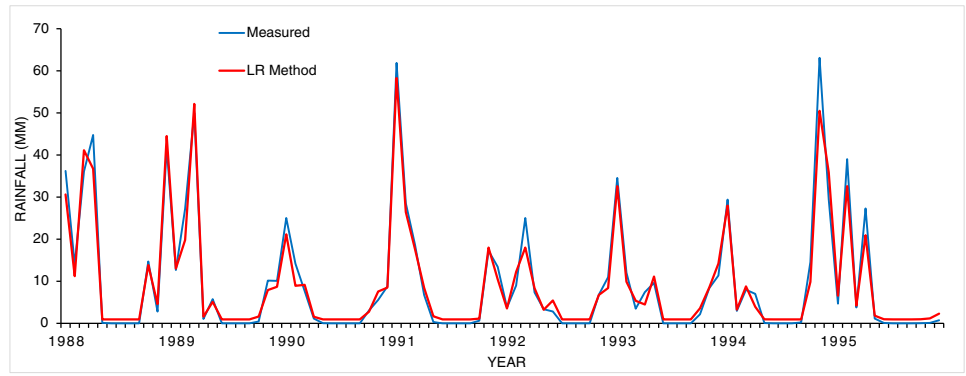
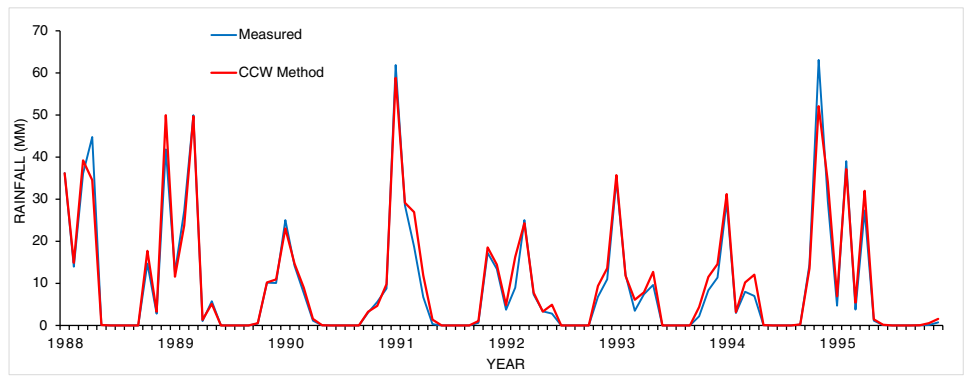
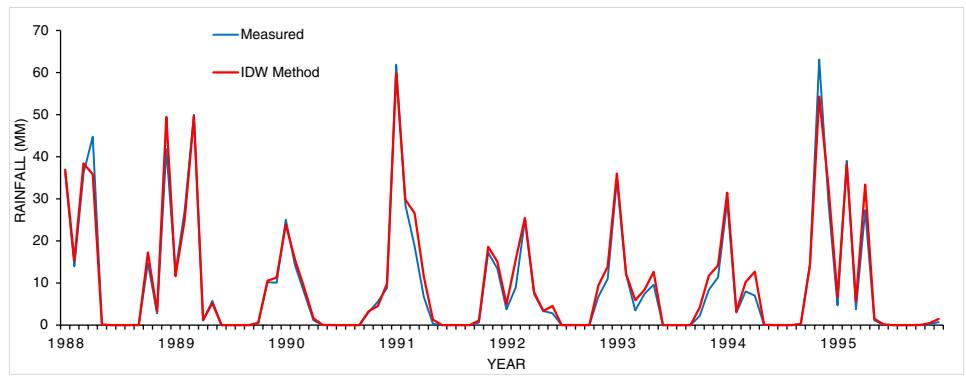


Fig. 3 (continued)

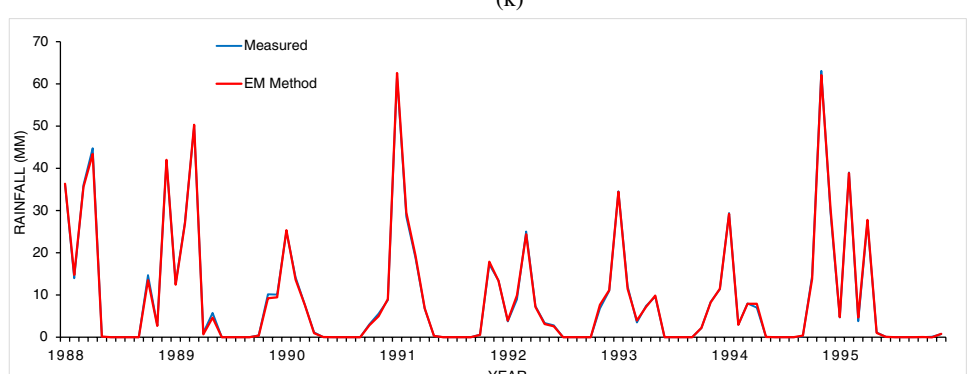
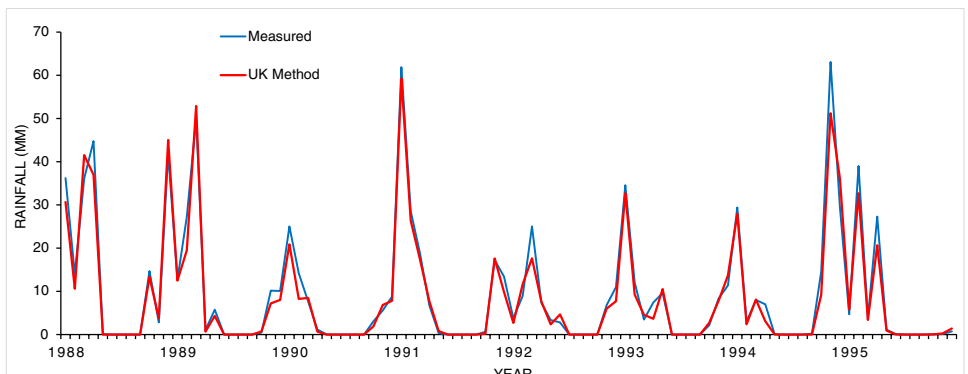
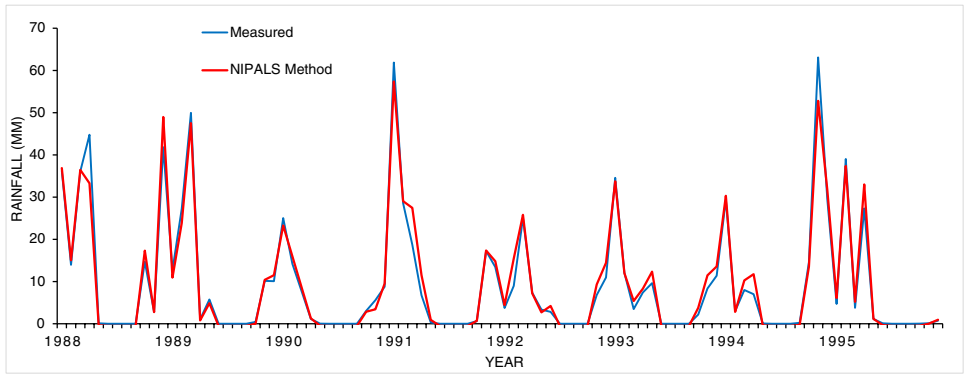
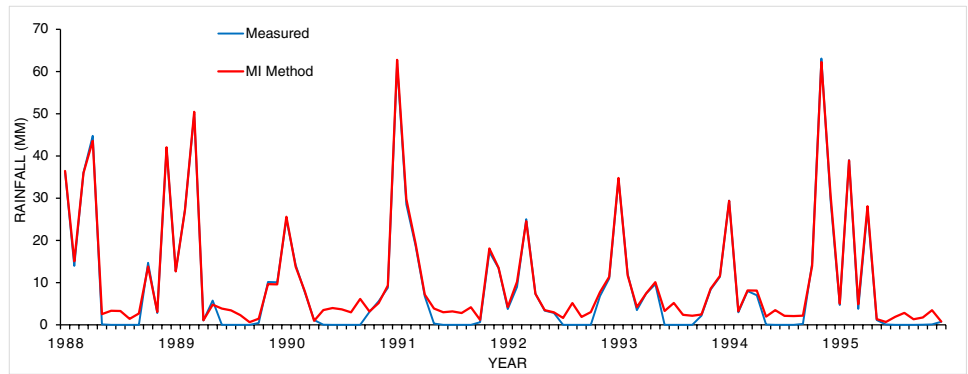
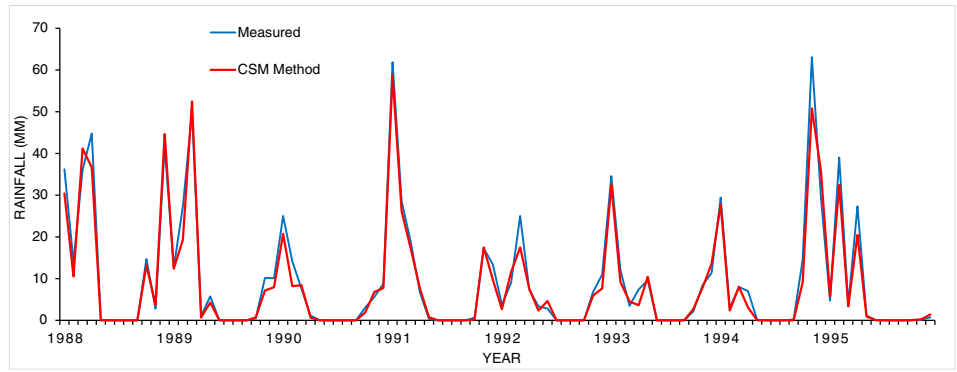
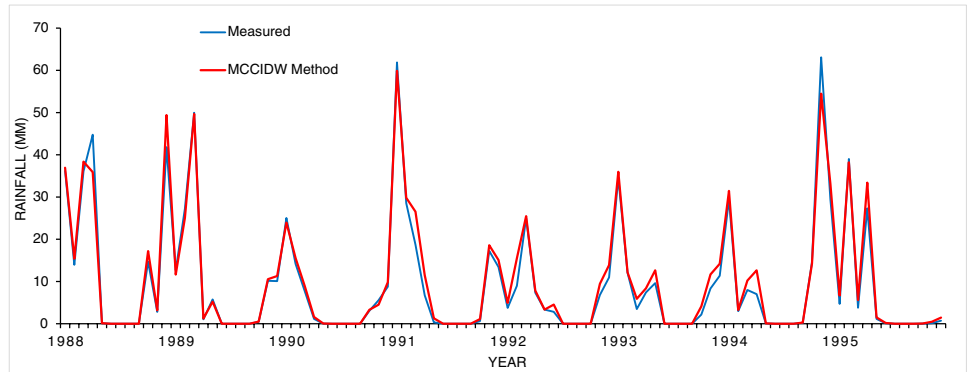


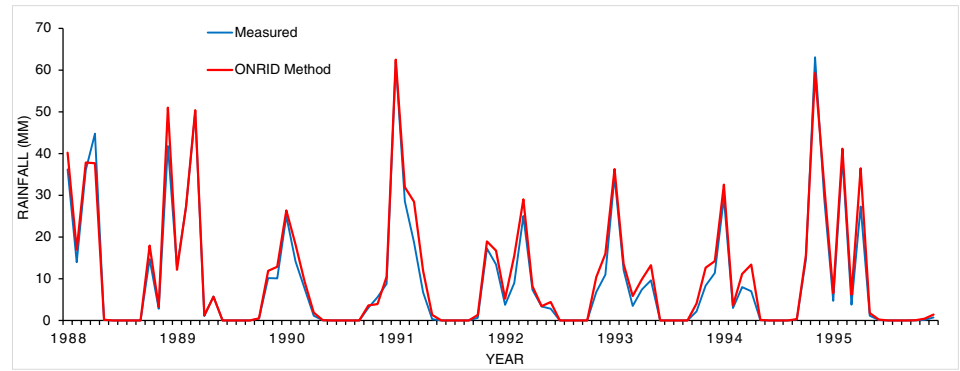
Fig. 3 (continued)



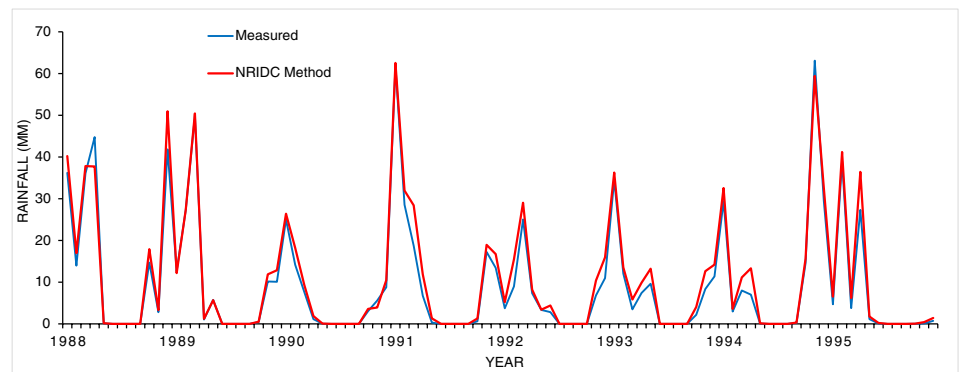
(m)



(n)

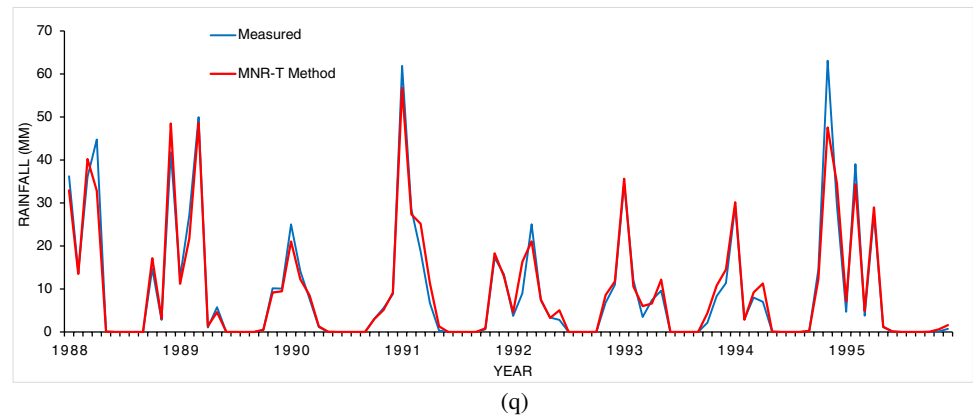


(o)



(p)

Fig. 3 (continued)



and Dinpashoh, 2011); similarity index (S-index), the values of similarity index range from 0 to 1, where the value of 1 means perfect results (Willmott 1981); skill score (SS), which is another index of efficiency, where the output ranges from 0 to 1, the value of 1 is perfect results, while as approaching 0, there is a drop in the efficiency of matching (Carvalho et al. 2016); rPearson coefficient which is very common in statistical application.

The above mentioned metrics formulas are shown below in Eqs. (13) to (17):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i| \quad (13)$$

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - x_i)^2}{n}} \quad (14)$$

$$CE = 1 - \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (15)$$

$$S \text{ index} = 1 - \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (|y_i - \bar{y}| + |x_i - \bar{y}|)^2} \quad (16)$$

$$SS = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\sum_{i=1}^n (\bar{y} - y_i)^2} \quad (17)$$

Results and discussion

First, raw data were tested to examine the homogeneity and correlation between the stations within the command area. The main goal of this paper is to examine different methods of missing precipitation data estimation;

therefore, the comparisons were performed between the predicted values and the measured values.

Examining of raw data

To determine how far the studied stations are statistically correlated, Table 2 below shows the rPearson coefficient between stations. The targeted station in this study is Baghdad station, mostly the correlation is above 0.9, and 6 stations correlation are above 0.95. Abu Ghraib station has the largest value; further it is the closest station to Baghdad station in terms of distance, while Tharthar station has the lowest value of correlation, which is 0.89. Generally, rPearson coefficients between stations are above 0.8, except 3 cases, which are Razazza and Suwairah, Razazza and Khan Saad, and Tharthar and Suwairah. All of the last mentioned cases have the same value of 0.78. The largest correlation value in the table is 0.98 between Abu Ghraib and Falluja stations. It can be concluded that the distance has the largest effect on the value of the correlation factor, where the later varies inversely with distance, keeping in mind that station elevations within Baghdad have variance.

In order to examine the homogeneity of the data (monthly precipitation data), Pettit's test was performed utilizing (XLSTAT) software: in this test, the null hypothesis H_0 : Data are homogeneous, and alternative hypothesis H_a : There is a date at which there is a change in the data. When the computed p value is greater than the significance level $\alpha = 0.05$, one cannot reject the null hypothesis H_0 . Table 3 shows the results of Pettit's test; the highest p value was observed at Abu Ghraib station with a value of 0.878, while the lowest p value was observed at Khan Saad station with value of 0.191. For all stations, and since p values are greater than 0.05, one cannot reject the null hypothesis; therefore, and according to Pettit's test, data for the examined stations are homogenous. Figure 2 also shows the diagram of p values for the tested stations.

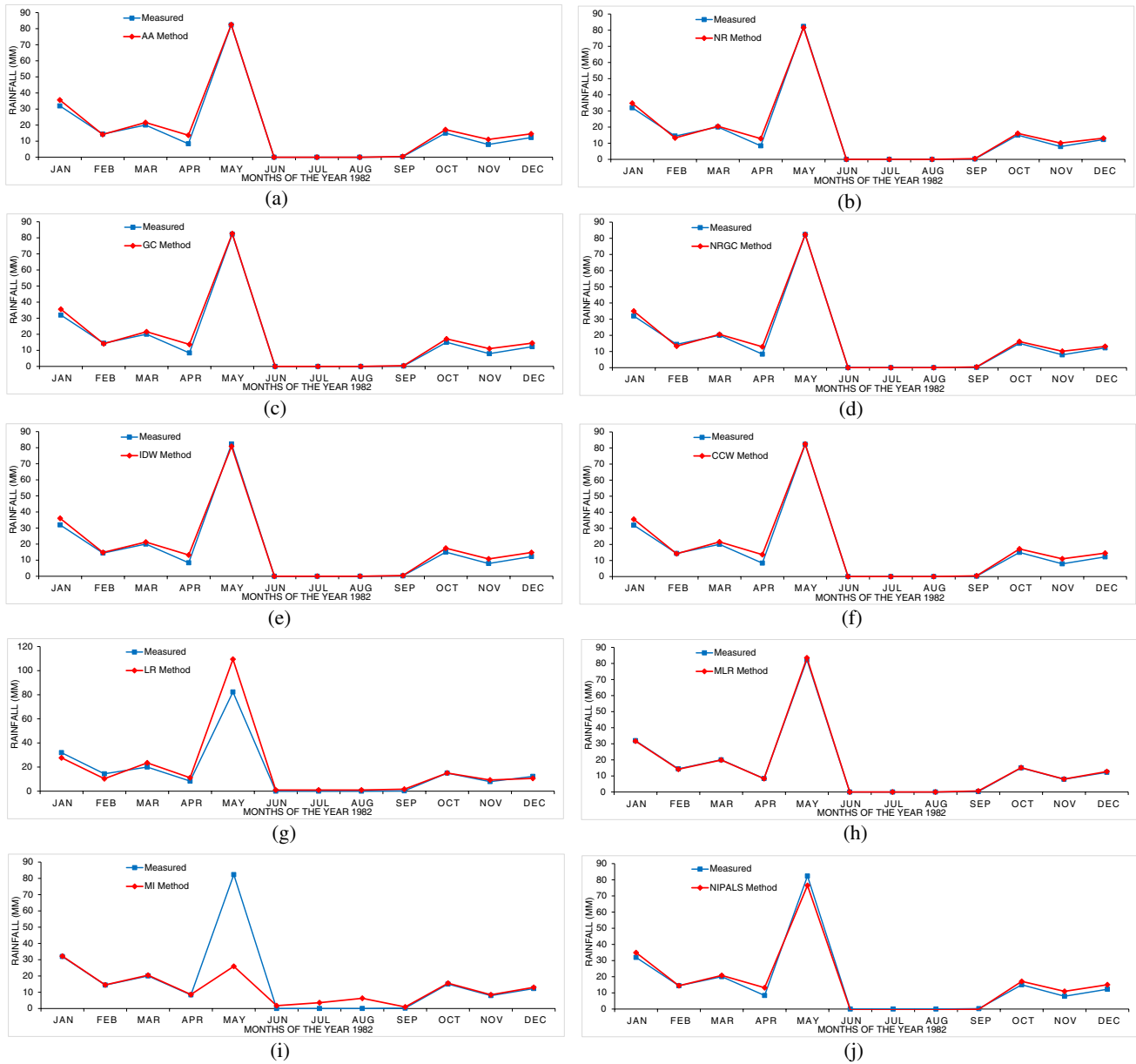


Fig. 4 Comparison of monthly precipitation prediction with measured values for the year 1982, methods abbreviation as following: **a.** Arithmetic average (AA); **b.** normal ration (NR); **c.** geographical coordinates (GC); **d.** normal ration with geographical coordinates (NRGC); **e.** inverse distance weighted (IDW); **f.** correlation coefficient weighted (CCW); **g.** linear regression (LR); **h.** multiple linear regression (MLR); **i.** multiple imputation (MI); **j.** nonlinear iterative

partial least square (NIPALS); **k.** UK method (UK); **l.** expectation maximization (EM); **m.** closet station method (CSM); **n.** modified correlation coefficient with inverse distance weighting (MCCIDW); **o.** modified old normal ration with inverse distance (ONRID); **p.** normal ration inverse distance weighting with correlation (NRIDC); **q.** modified normal ration based on square root distance (MNR-T)

Comparisons between 17 methods

As stated in this paper, 17 missing data methods were applied; the target station was Baghdad station, which is about in the center of the study area. Some methods are simple like CSM and AA; others are depending on spatial characteristics, averages, and regression with

other stations. In addition, some methods are combining the weighting factors of 2 or 3 characteristics in one method, such as ONRID method. Also, some methods, like EM and MI, employ artificial intelligence, where it was computed using advanced software. In this study, it was assumed that all the data of Baghdad station were missed, i.e., the monthly data from 1980 to 2014; the

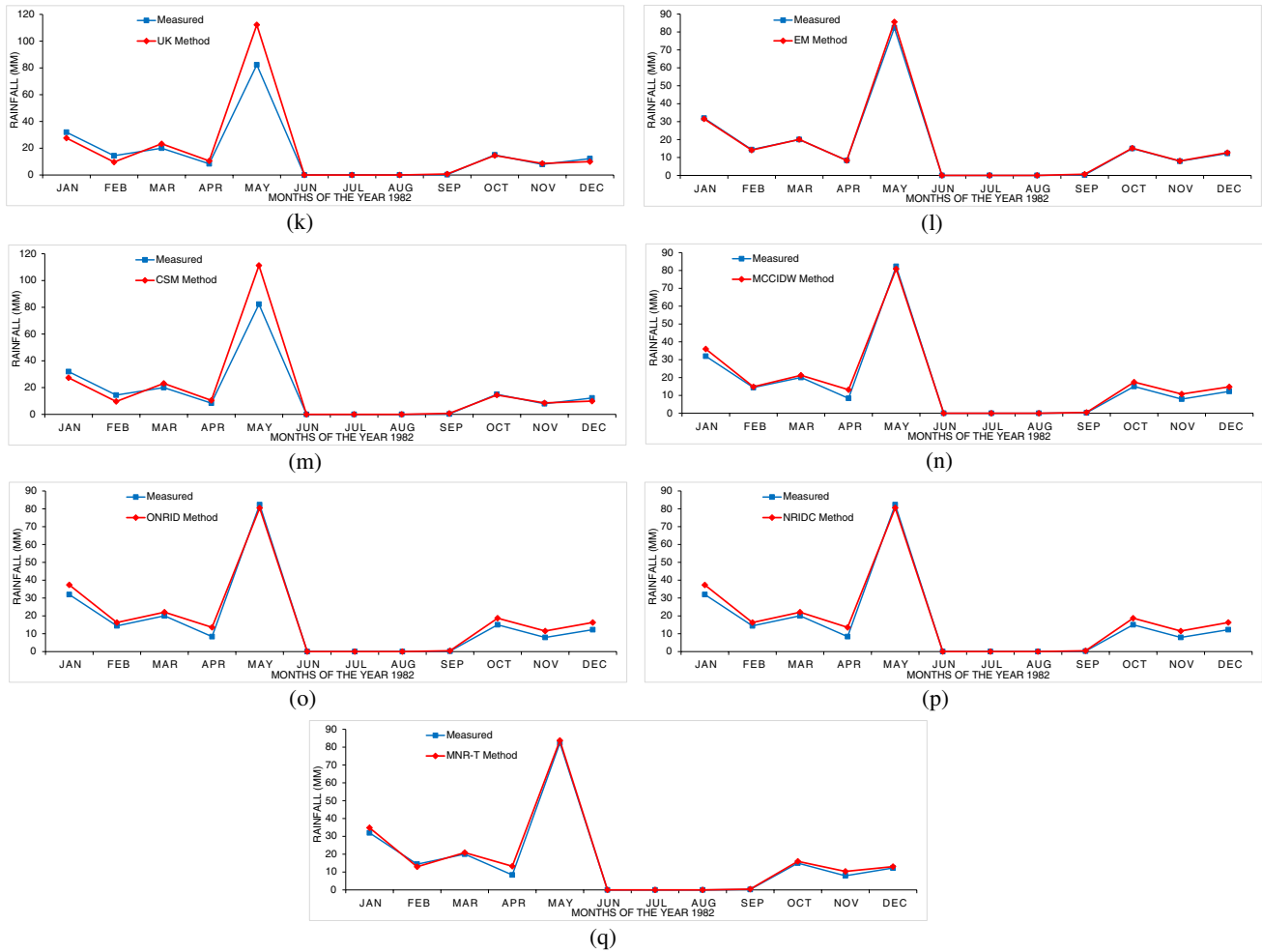


Fig. 4 (continued)

calculations of error indexes are based on these results. The exception was with MI, NIPALS, and EM methods, where the calculation was made for the years 1988 to 1995, as there is a need to have some existing values to perform these methods.

Table 4 shows the results of performance metrics of the applied methods. In this table, EM is the best method in terms of CE, S-index, SS, and rPearson metrics, and it has the lowest values of MAE and RMSE. MLR has the same performance, but it has a bit higher MAE and RMSE.

In general, all the methods have good estimation of missing data in Baghdad station; all can be adopted with the acceptable level of trust. The lowest values were founded with CSM, LR, and UK methods, where all has a regression factor of 0.968, and values of MAE and RMSE are more than 1.7 and 3.6, respectively. However, these indices seem good.

The methods of combined weighting factors, which are NRG, MCCIDW, ONRID, NRIDC, MNR-T made

no tangible difference in comparison with other methods that depend on a single factor, which are NR, GC, IDW. On another hand, the multiple regression method (MLR) has a better result than the single linear regression (LR).

In Figs. 3, 4 and 5, the results were visualized for a selected year to show how each method is performing. Figure 3 shows the time series of monthly precipitation for the years 1988 to 1995; it is easily noticed that all methods have a good matching with the measured data in Baghdad station, even with peak values.

Figures 4 and 5 show comparison between the measured and the predicted values in the 17 methods for the years 1982 and 200, respectively. The year 1982 was selected as it is the wettest year during the study period, while the year 2000 is the driest year during the study period. Again, good results were observed, except at some peaks with some methods in the dry year 2000, but for the year 1982, graphs were showing good estimations.

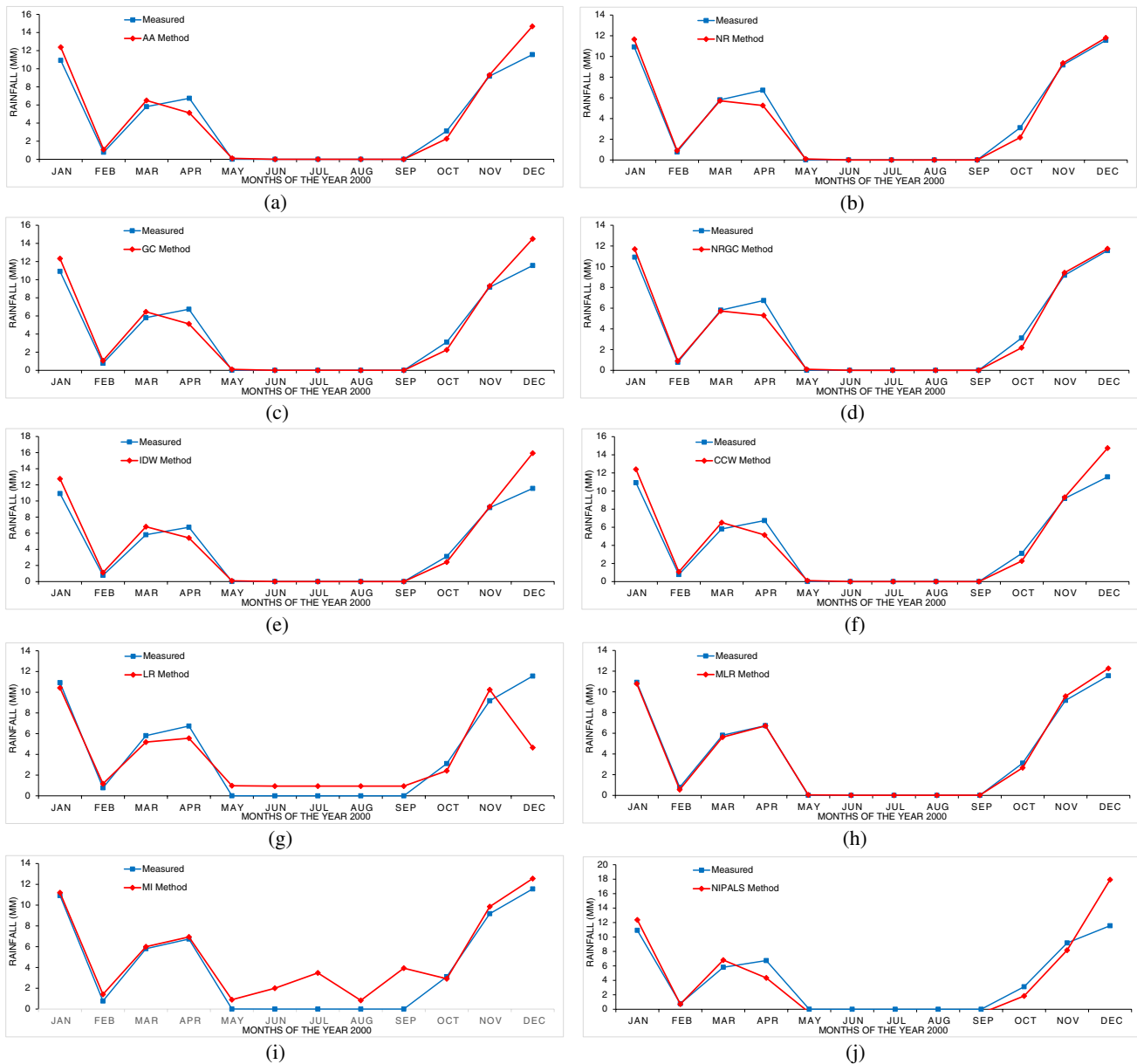


Fig. 5 Comparison of monthly precipitation prediction with measured values for the year 2000, methods abbreviation as following: **a.** Arithmetic average (AA); **b.** normal ration (NR); **c.** geographical coordinates (GC); **d.** normal ration with geographical coordinates (NRGC); **e.** inverse distance weighted (IDW); **f.** correlation coefficient weighted (CCW); **g.** linear regression (LR); **h.** multiple linear regression (MLR); **i.** multiple imputation (MI); **j.** nonlinear iterative

partial least square (NIPALS); **k.** UK method (UK); **l.** expectation maximization (EM); **m.** closet station method (CSM); **n.** modified correlation coefficient with inverse distance weighting (MCCIDW); **o.** modified old normal ration with inverse distance (ONRID); **p.** normal ration inverse distance weighting with correlation (NRIDC); **q.** modified normal ration based on square root distance (MNR-T)

Conclusions

Several conclusions can be derived from this study. The methods that yield the best result with the least error are EM, then MLR methods. Generally, all the 17 methods produce good predictions of the proposed missing data. Also, there are no tangible significant differences

between the methods that employ a single factor, such as location mean value, with that employing several combined factors, where this is limited with study area in the Baghdad zone. Errors of the predictions increase as the values of precipitation in the area decrease, where this was noticed in the results' comparison of the dry year 2000. In general, these good results might be attributed

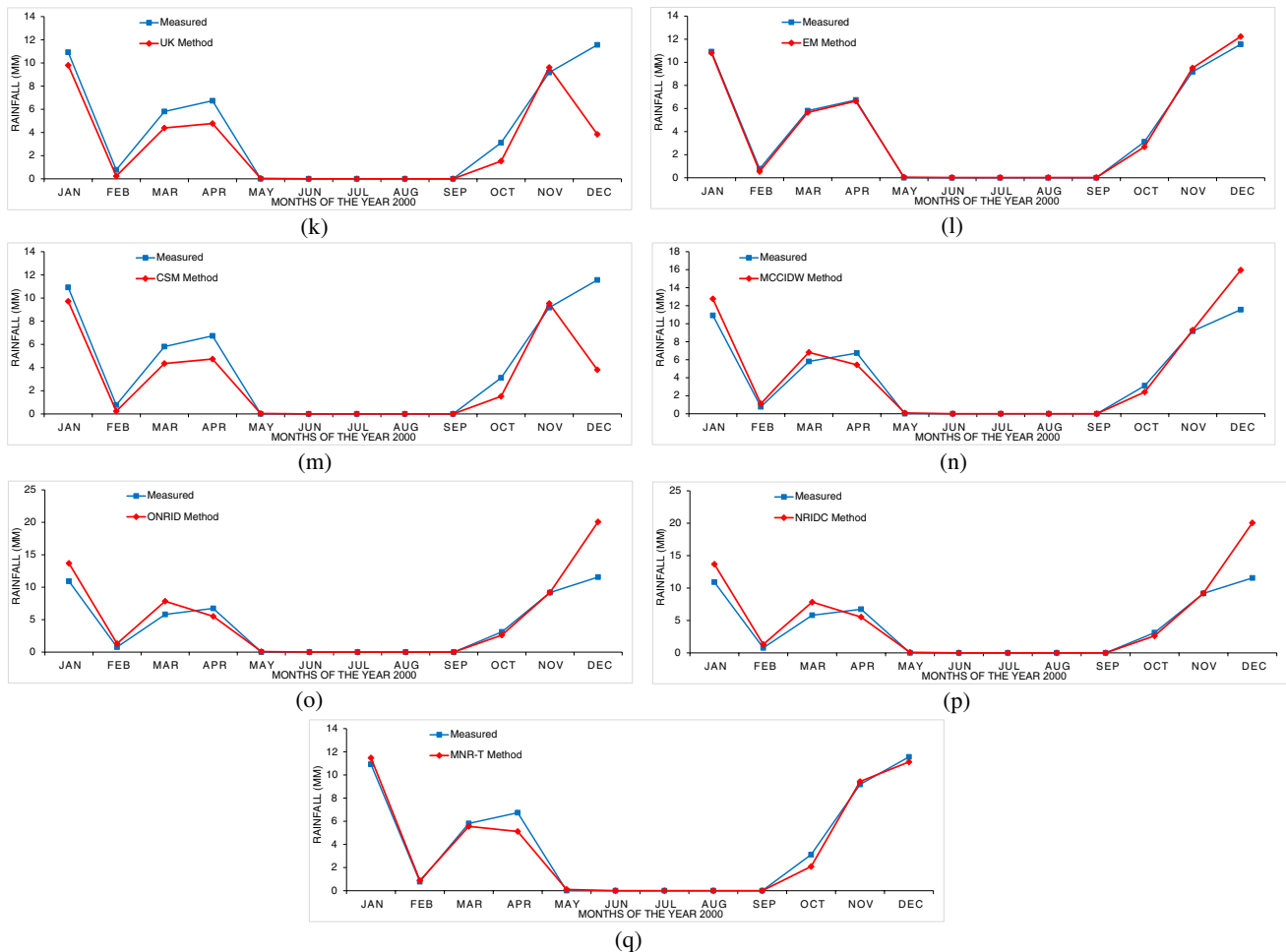


Fig. 5 (continued)

to the nature of the Baghdad area, where the topography is flat. Also, differences were observed between the results of the tested methods in other researchers, but with more complicated terrain.

These methods will be useful as Baghdad location considered within the drylands, where most of the previous tested methods showed comparatively less accuracy in the arid region in the middle and south of Iraq during the dry years, as well as the observed overestimates during the dry conditions. Also, it might be essential to consider to future data gathering, where it was expected that the climate change and rainfall trend variations might bring other facts.

Author contributions Conceptualization, MAH and NAA; methodology, MAH and NAA; software, MAH and NAA; validation, MAH and NAA; formal analysis, MAH and NAA; investigation, MAH and NAA; data curation, MAH and NAA; writing—original draft preparation,

MAH and NAA; writing—review and editing, MAH and NAA; visualization, MAH and NAA; all authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Lulea University of Technology.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- (IMOaS), Iraqi Meteorological Organization and Seismology (2021) Unpublished Data of Meteorological Stations in Iraq
- Abdullah M, Al-Ansari N (2021) Irrigation projects in Iraq. *J Earth Sci Geotech Eng* 11(2):35–160. <https://doi.org/10.47260/jesge/1123>
- Abdulrazzaq Z (2020) The feasibility of using TRMM satellite data for missing terrestrial stations in Iraq for mapping the rainfall contour lines. *Civ Eng Beyond Limits* 1:15–19. <https://doi.org/10.36937/cebel.2020.003.003>
- Abdulrida MA, Al-Jumaily K (2016) Comparisons of monthly rainfall data with satellite estimates of TRMM 3B42 over Iraq. *Int J Sci Res Publications* 6(1):494–501
- Al-Salihi AM, Al-Lami AM, Mohammed AJ (2013) Prediction of monthly rainfall for selected meteorological stations in Iraq using back propagation algorithms. *J Environ Sci Technol* 6(1):16–28. <https://doi.org/10.3923/jest.2013.16.28>
- Alzozeer A (2020) Estimation of mean areal rainfall and missing data by using GIS in Nineveh, Northern Iraq. *Iraqi Geological Journal* 53:93–103. <https://doi.org/10.46717/igj.53.1E.7Ry-2020-07.07>
- Anderson EA (1972) National weather service river forecast system forecast procedures. NOAA Tech Memo NWS HYDRO-14
- Armanuos A, Al-Ansari N, Yaseen Z (2020) Cross assessment of twenty-one different methods for missing precipitation data estimation. *Atmosphere* 11:1–35. <https://doi.org/10.3390/atmos11040389>
- Azman MA-Z, Zakaria R, Ahmad Radi NF (2015) Estimation of missing rainfall data in Pahang using modified spatial interpolation weighting methods. *AIP Conference Proceedings* 1643(2015):65–72. <https://doi.org/10.1063/1.4907426>
- Barbalho F, Silva G, Formiga K (2014) Average rainfall estimation: methods performance comparison in the Brazilian semi-arid. *J Water Resour Prot* 06:97–103. <https://doi.org/10.4236/jwarp.2014.62014>
- Bárdossy A, Pegram G (2014) Infilling missing precipitation records—a comparison of a new copula-based method with other techniques. *J Hydrol* 519:1162–1170. <https://doi.org/10.1016/j.jhydrol.2014.08.025>
- Carvalho J, Nakai A, Monteiro JE (2016) Spatio-temporal modeling of data imputation for daily rainfall series in homogeneous zones. *Revista Brasileira De Meteorologia* 31:196–201. <https://doi.org/10.1590/0102-778631220150025>
- Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc* 39(1):1–38
- Frenken K (2009) Irrigation in the Middle East region in figures AQUASTAT Survey-2008. Water Reports. Food and Agriculture Organization of the United Nations, Rome
- Kanda N, Negi H, Shekhar M, Rishi M (2017) Performance of various techniques in estimating missing climatological data over snow-bound mountainous areas of Karakoram Himalaya. *Meteorol Appl* 25. <https://doi.org/10.1002/met.1699>
- Kashani M, Dinpashoh Y (2011) Evaluation of efficiency of different estimation methods for missing climatological data. *Stoch Environ Res Risk Assess* 26. <https://doi.org/10.1007/s00477-011-0536-y>
- Linsley Jr RK, Kohler MA, Paulhus JL (1975) *Hydrology for engineers*. McGraw Hill
- Murad S, Jaff Y (2020) Comparable investigation for rainfall forecasting using different data mining approaches in Sulaymaniyah city in Iraq. *Int J Environ Sci Technol*. <https://doi.org/10.18488/journal.72.2020.41.11.18>
- Paulhus JL, Kohler MA (1952) Interpolation of missing precipitation records. *J Monthly Weather Review* 80(8):129–133
- Rubin DB (1988) An overview of multiple imputation. *Proceedings of the survey research methods section of the American statistical association*. Citeseer, pp 79–84
- Sattari M, Rezazadeh Joudi A (2016) Assessment of different methods for estimation of missing data in precipitation studies. *Hydrol Res* 48. <https://doi.org/10.2166/nh.2016.364>
- Schneider T (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J Clim* 14:853–871. [https://doi.org/10.1175/1520-0442\(2001\)014%3c0853:AOICDE%3e2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014%3c0853:AOICDE%3e2.0.CO;2)
- Syed Jamaludin SS, Deni S, Jemain A (2008) Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pac J Atmos Sci* 44:93–104
- Tang W, Kassim A, Abubakar S (1996) Comparative studies of various missing data treatment methods—Malaysian experience. *Atmos Res* 42:247–262. [https://doi.org/10.1016/0169-8095\(95\)00067-4](https://doi.org/10.1016/0169-8095(95)00067-4)
- Teegavarapu R (2009) Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *J Hydro* 11. <https://doi.org/10.2166/hydro.2009.009>
- Teegavarapu R, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J Hydrol* 312:191–206. <https://doi.org/10.1016/j.jhydrol.2005.02.015>
- Willmott C, Matsuura K, Robeson S (2009) Ambiguities inherent in sums-of-squares-based error statistics. *Atmos Environ* 43:749–752. <https://doi.org/10.1016/j.atmosenv.2008.10.005>
- Willmott CJ (1981) On the validation of models. *J Physical Geography* 2(2):184–194
- Wold HOA (1968) Nonlinear estimation by iterative least square procedures
- Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2012) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoret Appl Climatol* 112. <https://doi.org/10.1007/s00704-012-0723-x>