Stoch Environ Res Risk Assess (2017) 31:2415–2434
DOI 10.1007/s00477-016-1356-x

ORIGINAL PAPER

Frequency based imputation of precipitation

Fatih Dikbas[1]

Published online: 19 November 2016

Abstract Changing climate and precipitation patterns make the estimation of precipitation, which exhibits two-dimensional and sometimes chaotic behavior, more challenging. In recent decades, numerous data-driven methods have been developed and applied to estimate precipitation; however, these methods suffer from the use of one-dimensional approaches, lack generality, require the use of neighboring stations and have low sensitivity. This paper aims to implement the first generally applicable, highly sensitive two-dimensional data-driven model of precipitation. This model, named frequency based imputation (FBI), relies on non-continuous monthly precipitation time series data. It requires no determination of input parameters and no data preprocessing, and it provides multiple estimations (from the most to the least probable) of each missing data unit utilizing the series itself. A total of 34,330 monthly total precipitation observations from 70 stations in 21 basins within Turkey were used to assess the success of the method by removing and estimating observation series in annual increments. Comparisons with the expectation maximization and multiple linear regression models illustrate that the FBI method is superior in its estimation of monthly precipitation. This paper also provides a link to the software code for the FBI method.

Keywords Frequency based imputation · Data-driven modelling · Precipitation · Estimation of missing data

✉ Fatih Dikbas
f_dikbas@pau.edu.tr

[1] Civil Engineering Department, Pamukkale University, Denizli, Turkey

1 Introduction

The importance of accurate and reliable modeling, estimation and forecasting of precipitation is becoming increasingly apparent as the rapid worldwide increase in population and water demand puts pressure on limited water resources and dwindling water supplies (Leconte et al. 2013; Popp et al. 2016). Accurate and reliable observations of precipitation are essential to the performance of valid hydrologic studies; yet, many precipitation records are incomplete. Complete records improve the ability of these studies to determine spatial, temporal and quantitative variations in precipitation data, which is crucial to the design of water supply systems. Changes in the water cycle and precipitation patterns, coupled with a warming climate (Hou et al. 2014; Reager and Famiglietti 2009), increase the need for stronger precipitation models (Zhang et al. 2010).

Developments in software technologies in recent decades have allowed traditional hydraulic and data-driven models to support/complement hydrologic models (Solomatine et al. 2008). Data-driven models analyze time series data, but they should not be regarded as computational methods that ignore physical processes. Determining the spatial and temporal interrelationships between precipitation time series data is mathematically equivalent to determining the relationships between the drivers of precipitation. In other words, precipitation is a function of its contributing variables. Thus, the analysis of precipitation time series data comprises the consideration of all variables that contribute to precipitation (though the relationships and variations of the variables are not evaluated); and the success of making accurate estimations of missing data is directly related to the level of understanding of the temporal and quantitative relationships between observed data.

Though precipitation is generally seasonal, the high variability in numerous influencing factors sometimes indicates the existence of a chaotic (Jayawardena and Lai 1994; Sivakumar 2000; Sivakumar et al. 1999) and relatively random behavior. This nonstationary and sometimes erratic behavior results in distinct variations in precipitation across space and time and makes the observation, quantification, estimation and forecasting of precipitation challenging (Wang and Lin 2015). Consequently, although there are a vast number of data-driven modeling studies that estimate hydrologic processes such as streamflow (which generally occur continuously), a very limited number of studies address the data-driven estimation of missing precipitation records. Some prominent studies that have utilized data-driven methods to estimate precipitation have applied artificial neural networks (ANNs), fuzzy rule based systems (FRBSs), genetic algorithms (GAs), support vector machines (SVMs), particle swarm optimization (PSO) and expectation maximization (EM) in the computation of results.

Lack of generality and overfitting are two of the most important problems associated with existing data-driven methods, as discussed in detail by Remesan and Mathew (2015). Both issues result in model failure when the training and testing period ranges change. Unfortunately, most data-driven hydrologic modeling studies do not even mention (or test) these issues. Another problem associated with existing methods is that time series data is generally regarded as a one-dimensional vector. This results in a failure to acknowledge the variation of behavior seen through time series data. For example, hydrological time series generally indicate an annual cycle of seasonality, with values observed in the winter months varying greatly from those observed during the summer months. Instead of using a one-dimensional time series to represent this data, a two-dimensional matrix containing a full cycle in each row would better express this temporal hydrological variability in a more comprehensible way and would enable the investigation of the two-dimensional behavior of time series data (Dikbas 2016b). Detailed information about the concepts, approaches, experiences and problems associated with the data-driven modeling of hydrologic variables exist in literature (Elshorbagy et al. 2010a, b; Maier and Dandy 2000; Maier et al. 2010; Remesan and Mathew 2015; Sikorska et al. 2015; Solomatine et al. 2008; Solomatine 2006; Yozgatligil et al. 2013).

This paper discusses the implementation of the Frequency Based Imputation (FBI) method to analyze observation data from 70 precipitation stations in Turkey. The method was first used to analyze all streamflow observations from 34 stations on the Buyuk Menderes River (Turkey) (Dikbas 2016a). This approach is based on the assumption that an individual observation in a time series is more closely and quantitatively linked to data observed within a short period of time and with data from the same subsection of other periods if the time series is periodic (i.e., same season in different years). The method searches neighboring data cluster pairs of missing data within an observed series, and then estimates the probable range and value of the missing data by utilizing temporal relationships. It is direct and uses all existing raw data to obtain estimates of missing values; and it requires no training/testing periods or input parameters to execute the applied procedure.

2 Materials and methods

2.1 Description of the frequency based imputation method

When precipitation observations are placed on a matrix with months in columns and years in rows, we expect annual fluctuations in the horizontal direction and values similar to each other in the vertical direction. In this setup, the smallest scale representing the temporal and quantitative behavior of precipitation is an adjacent pair of data on the two-dimensional matrix. This micro-statistical reasoning allows the FBI method to extract valuable information based on relationships within the dataset and provides information on the possible range of missing observations.

Figure 1 illustrates the logic behind the FBI method. The blue cell at the center of Fig. 1e (January 1985) is the missing value to be estimated. The method considers that the neighbors within the 7×7 matrix surrounding the missing value contain the strongest clues about the expected range of the missing cell. A wider field would add cells with a poorer relationship to the data point in question (like trying to determine the influence of values in September or May on a value in January which are less likely to be as influential as the considered temporally closer values from October to April); and a narrower field would remove cells with potential relationship (like ignoring the influences of October and April on the January value). Similarly, expanding the field vertically would result in the consideration of observations four or more years preceding or following the missing value, even though these values are less likely to relate to the value in question when compared to the values in closer years. The numbers in each cell in Fig. 1 are cluster values calculated by using Eq. 1 or 2 after the observed series was sorted and divided into range clusters (Appendix 1).

After the cluster index values for each cell are determined, the process of generating a cluster frequency table for each missing value begins. To this end, all adjacent cluster pairs within the neighborhood of a missing cell

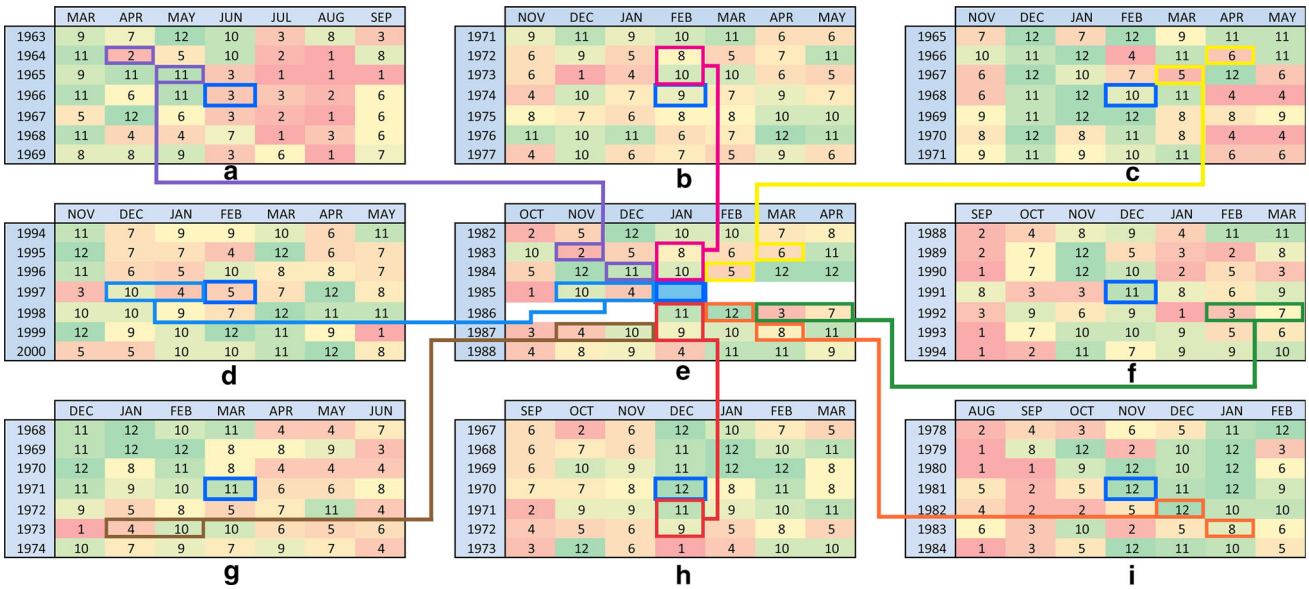

Fig. 1 The missing observation to be estimated (the *blue cell*) and eight example cluster pairs to be searched in the data matrix (each pair is shown in a *different color*) (**e**), matching cluster pairs found in different sections of the data matrix (take careful note of the relative location of the missing value) (**a–d, f–i**) and the probable values of the missing data (cells with blue borders at the center of **a–d, f–i**)

are searched using a data matrix. Figure 1e shows eight of the many cluster pairs in the neighborhood of the missing value for January 1985. The remaining subfigures show the locations of the matching cluster pairs. The aim of the search for matching cluster pairs is to deduce the highest probable cluster value for the missing cell. This task is accomplished by looking at the cluster values of the blue-bounded cells at the relative location of the missing January 1985 cell. These clusters show the probable values for the missing cell in January 1985 by answering the questions constructed using the searched and matched cluster pairs. One of the eight questions illustrated in Fig. 1 is:

"What might the cluster value of the missing cell in January 1985 be when the cluster value in January 1983 is 8 and the cluster value in January 1984 is 10?"

The goal here is to find the third cluster value of three vertically aligned cells when the first value is 8 and the second value is 10. One of the answers to this question is shown in Fig. 1b and is written as follows:

"The cluster value for February 1974 is 9 when the cluster value for February 1972 is 8 and the cluster value for February 1973 is 10". In other words, the cluster value for January 1985 might be 9 based on previously observed series values.

For all eight cluster pairs in Fig. 1e, the probable cluster values at the relative January 1985 location in the remaining figures are found to be: 12 (2 times), 11 (2 times), 10, 9, 5 and 3. When the search for all pairs in the neighborhood of the missing value is completed, the cluster with the highest frequency is considered to have the highest probability of being the missing value. The estimated precipitation value is calculated by taking the average of the observations that generated the greatest cluster frequency. Details of how the cluster frequencies were determined and generated are provided in Appendix 2.

2.2 Study area and data

To test the applicability of the developed method and provided software on various climate zones, a total of 34,330 monthly total precipitation observations from 70 stations across 21 different basins in Turkey were estimated (Fig. 2). Turkey has a moderately dry climate. Average precipitation tends to be high in the coastal regions of Turkey and decreases towards the inland regions. The area around Rize on the coast of the Black Sea receives an average annual precipitation of 2200 mm, while Salt Lake region receives 250–300 mm. The Aegean and Mediterranean coasts are wet in the winter but dry during the summer. The Black Sea coastline is the only region in Turkey that receives precipitation throughout the year. Figure 2 illustrates the average annual precipitation in Turkey between 1981 and 2010. The selected stations represent the majority of the climate and elevation zones, and cover nearly all hydrological basins in Turkey.

The General Directorate of State Hydraulic Works of Turkey observes precipitation throughout the country using pluviographs capable of measuring liquid (rainfall) and solid (snow, hail, freezing rain, grain, etc.) precipitation. Therefore, the observations used in this study include liquid precipitation and water equivalents of solid precipitation.

Table 1 outlines the descriptive statistics for all stations, including percentiles and best-fitting distributions. The

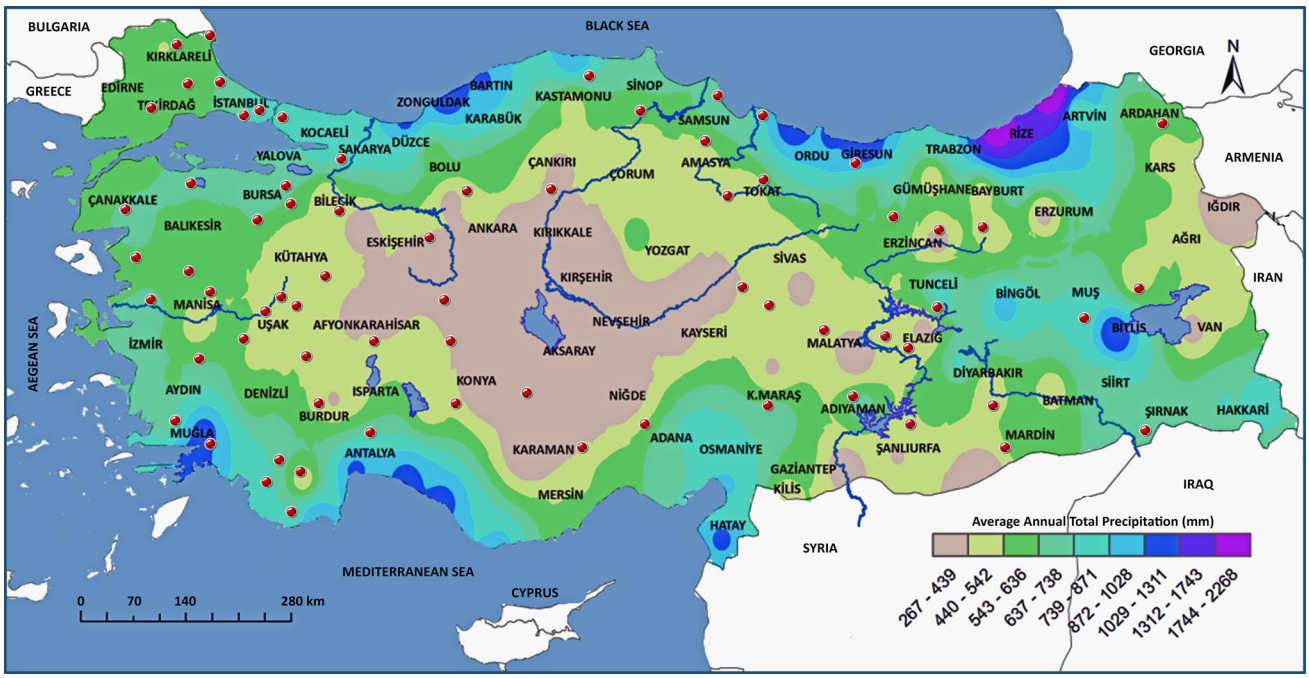

Fig. 2 Map of 1981–2010 average annual precipitation in Turkey, including the locations of the 70 stations used in this study

highest and lowest values (excluding 0.0) are shown in bold in all tables throughout the article. The majority of precipitation series from all stations (48/70) were found to fit the Wakeby distribution. The skewness and excess kurtosis measures indicate that the probability distributions for all stations are positively skewed and leptokurtic (except 21-007). A majority of the stations (67/70) have a minimum monthly precipitation of 0. A total of 59% (41/70) of stations registered zero monthly precipitation for at least 5% of the year, while 36% (25/70) of stations measured zero monthly precipitation more than 10% of the year and 4% (3/70) of stations measured zero monthly precipitation data during more than 25% of the year.

A comprehensive explanation of the applied steps for the estimation of the monthly total precipitation is presented below for the observations of station 07-016 in Çivril-Denizli (Turkey). The first seven values from 1962 are missing, and the total number of existing observations at station 07-016 is 521. When 12 observations (a year of data) are removed from the set to test the model's ability to make estimations, this number decreases to 509, resulting in a missing data rate of 3.6% (Fig. 3).

The details of the estimation process are presented using the observed values from 1985. The entire estimation process was repeated for each missing data point. First, the software removed and estimated data for each year between 1962 and 1984. Then, the 1985 values were removed from the set and estimated. The January value was estimated first. Figure 4 shows the observed values for those months and years surrounding January 1985. The October–December columns represent values from the previous calendar year (current water year).

To assess the quantitative relationships between the observations, the observed series are sorted and divided into 2–12 clusters, as explained in Appendix 1. The greatest number of clusters (12) was chosen based on the length and variability of the time series. The results show that this number was sufficient to generate successful results. Figure 5 shows the cluster values for the field surrounding January 1985 at each clustering step. Lower values are shown in shades of red and higher values are shown in shades of green. When the observed data series is divided into two clusters, the first cluster contains the lower precipitation values (0–25.8 mm) from the sorted observations, and the second cluster contains the higher values (26.0–204.8 mm). Each data point is assigned a cluster index: 1 for the data in the first cluster and 2 for the data in the second cluster, as shown in the first table of Fig. 5.

In the remaining cluster divisions (3–12), the January 1985 (84.9 mm) value is always located within the highest range of observations and thus the last cluster (bounded in blue in Fig. 5). The temporal and quantitative relationships between the horizontally, vertically and diagonally adjacent cluster pairs in the neighborhood of the missing data are determined as explained in Sect. 2. Then, the relationships are used to estimate the probable cluster value of the deliberately removed data in the center of the neighborhood.

When the sorted observations of the station are divided into 12 clusters, 496 cluster pairs matching with the adjacent cluster pairs in the neighborhood of January 1985

Table 1 Descriptive statistics, percentiles and best-fitting distributions for all stations

Station Elevation (m)	01-004 90	01-005 395	01-008 395	02-004 35	02-009 40	02-011 **10**	02-012 180	02-018 30	03-009 20	03-013 770	03-027 240	04-003 320
Statistic												
Sample Size	490	485	472	446	482	463	442	452	469	528	500	485
Missing	2	7	8	10	10	5	14	40	11	0	16	7
Mean	42.3	54.0	46.4	69.3	77.8	47.2	82.5	61.0	46.3	55.6	52.2	69.5
Variance	1139	1627	1331	3015	4301	1541	5073	2537	1755	2680	1420	4076
Std. Error	1.52	1.83	1.68	2.60	2.99	1.82	3.39	2.37	1.93	2.25	1.69	2.90
Skewness	1.11	1.04	1.45	1.18	1.55	1.27	1.66	1.19	1.50	1.44	0.78	1.35
Excess Kurtosis	1.06	1.01	3.56	2.07	2.76	2.21	4.02	1.42	4.14	2.91	0.28	2.06
Percentiles												
Min	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5%	1.2	3.0	4.6	4.2	7.0	1.7	4.6	2.0	0.0	0.3	1.2	0.4
10%	4.6	8.0	7.8	7.8	13.4	4.4	10.2	8.0	2.3	2.3	7.7	3.8
25% (Q1)	17.0	24.2	18.5	26.8	30.1	17.1	31.8	22.9	13.2	15.2	22.6	17.6
50% (Median)	33.8	46.1	38.7	56.5	61.8	38.0	64.5	47.9	36.2	42.9	46.0	53.8
75% (Q3)	60.0	77.9	65.0	101.1	103.3	70.5	114.5	88.9	66.8	83.0	74.7	101.8
90%	89.6	107.6	93.3	142.0	165.6	98.0	171.1	130.8	107.8	129.6	105.0	150.4
95%	111.1	134.3	117.6	174.5	211.2	117.8	225.2	165.0	123.6	154.3	122.4	196.1
Max	170.1	205.0	264.3	365.5	351.0	242.5	476.5	272.0	315.9	320.5	184.1	342.3
Best-Fit Distribution*	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	BETA	WAK	BETA

Station Elevation (m)	04-008 500	05-001 930	05-004 1020	05-007 100	05-008 340	05-012 715	05-016 670	06-005 380	07-013 885	07-016 825	07-022 1095
Statistic											
Sample Size	431	511	461	530	524	521	528	516	516	521	523
Missing	1	17	19	10	4	7	0	12	12	7	5
Mean	78.4	39.3	38.6	45.8	46.0	37.3	49.3	70.3	94.9	36.5	53.0
Variance	8145	1297	1337	2946	2878	1105	2398	6615	11,595	959	2455
Std. Error	4.35	1.59	1.70	2.36	2.34	1.46	2.13	3.58	4.74	1.36	2.17
Skewness	1.92	1.25	1.30	1.72	1.82	1.15	1.25	1.69	1.62	1.13	1.44
Excess Kurtosis	5.13	1.76	1.96	3.44	4.25	1.11	1.33	2.68	2.82	1.85	2.62
Percentiles											
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10%	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.2
25% (Q1)	8.8	10.2	10.9	2.9	5.3	10.2	9.0	10.5	11.4	11.8	15.5
50% (Median)	47.5	32.0	29.9	27.3	29.0	30.2	37.0	41.3	57.9	29.5	39.3
75% (Q3)	117.5	55.8	59.9	70.5	65.6	53.4	76.0	99.2	137.2	54.4	78.5
90%	196.7	90.6	91.5	122.7	121.8	84.6	119.7	184.3	258.1	78.8	117.7
95%	254.7	110.3	110.7	158.3	155.4	105.9	142.2	256.6	326.7	93.3	147.0
Max	630.4	208.2	214.5	333.1	362.9	159.9	254.6	405.2	602.3	204.8	291.5
Best-Fit Distribution*	GEV	WAK	G.PAR	GEV	GEV	WAK	G.PAR	WAK	GEV	WAK	G.PAR

Station Elevation (m)	08-006 730	08-008 240	08-010 1300	08-013 230	08-014 1410	09-014 310	10-007 850	11-002 1085	12-003 744	12-011 250	12-012 1100	12-014 40
Statistic												
Sample size	507	528	471	504	534	518	510	484	508	501	498	494
Missing	9	0	9	12	6	10	6	8	8	3	6	10
Mean	**130.6**	74.0	58.6	84.9	37.3	113.7	32.1	46.9	28.9	31.7	36.3	78.1

Table 1 continued

Station Elevation (m)	08-006 730	08-008 240	08-010 1300	08-013 230	08-014 1410	09-014 310	10-007 850	11-002 1085	12-003 744	12-011 250	12-012 1100	12-014 40
Variance	**25,647**	7419	4021	12,568	1078	17,627	808	1630	520	595	729	2803
Std. error	**7.11**	3.75	2.92	4.99	1.42	5.83	1.26	1.84	1.01	1.09	1.21	2.38
Skewness	1.75	1.81	2.22	2.03	1.32	1.85	1.42	1.51	1.02	1.19	0.86	1.14
Excess kurtosis	3.34	3.62	7.79	5.22	1.97	4.21	3.19	3.30	1.09	2.51	0.58	2.73
Percentiles												
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5%	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.4	0.5	0.5	9.9
10%	0.0	0.0	1.3	0.0	2.5	0.5	1.5	4.2	2.0	3.7	4.2	16.9
25% (Q1)	9.3	9.6	12.0	2.1	12.3	15.0	9.7	17.3	11.4	12.1	15.7	37.9
50% (Median)	69.8	47.2	39.4	39.3	29.0	66.5	25.8	37.5	25.2	27.4	31.8	71.5
75% (Q3)	**195.9**	104.8	83.8	130.5	53.5	167.3	47.3	67.1	40.7	46.7	51.3	107.6
90%	**373.6**	187.7	132.6	233.9	83.7	293.8	71.0	97.8	60.7	62.9	73.6	148.2
95%	**472.6**	256.7	175.4	309.3	102.2	395.8	85.7	120.1	70.4	75.5	87.8	178.9
Max	**893.5**	468.9	476.6	725.2	187.5	879.5	194.7	247.8	124.5	150.9	137.6	399.9
Best-fit distribution*	GEV	J.SB	WAK	GEV	G.PAR	BETA	WAK	WAK	WAK	WAK	WAK	WAK

Station Elevation (m)	12-042 981	12-047 950	12-049 900	14-005 1600	14-007 830	14-017 870	14-018 **10**	14-019 635	15-008 1330	15-010 800	15-019 460
Statistic											
Sample size	462	465	509	451	531	**566**	519	517	387	500	479
Missing	18	15	7	17	9	10	9	11	33	4	13
Mean	31.5	94.1	29.1	41.0	44.1	54.2	75.0	35.5	43.1	28.2	41.5
Variance	835	5559	652	983	1253	1362	2607	798	1258	527	831
Std. error	1.34	3.46	1.13	1.48	1.54	1.55	2.24	1.24	1.80	1.03	1.32
Skewness	1.39	1.48	1.43	1.29	1.95	1.09	1.28	0.96	1.78	1.05	0.85
Excess kurtosis	2.71	3.38	3.39	2.88	**10.58**	1.70	2.13	0.63	4.31	0.96	0.27
Percentiles											
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5%	0.0	7.5	0.0	1.9	1.2	6.1	12.0	0.0	4.3	0.5	4.0
10%	1.3	15.3	1.5	6.3	4.6	12.7	20.0	2.4	8.5	3.5	9.2
25% (Q1)	8.9	**41.2**	8.8	17.8	17.9	27.0	36.7	12.8	19.4	9.9	19.5
50% (Median)	24.6	**76.1**	22.3	35.1	39.1	48.4	65.5	31.7	33.9	23.1	36.5
75% (Q3)	48.1	130.4	43.4	57.3	61.6	73.9	98.5	51.6	56.8	41.6	58.9
90%	68.8	197.3	62.4	79.7	89.1	102.1	145.9	75.1	84.5	60.5	83.8
95%	87.2	227.2	79.8	100.1	109.7	125.2	167.5	95.5	117.0	75.1	97.1
Max	184.3	474.9	177.8	224.1	351.0	238.1	305.6	139.7	223.1	**123.0**	140.1
Best-fit distribution*	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK

Station Elevation (m)	15-020 25	16-013 1155	16-016 1450	16-019 1150	16-030 1350	18-003 1740	18-013 1225	18-016 1670	20-009 895	21-003 1475	21-004 1180	21-006 **2040**
Statistic												
Sample size	515	492	483	462	536	553	472	437	514	401	497	500
Missing	1	0	11	18	16	11	8	7	2	33	19	16
Mean	64.1	26.5	24.7	38.5	32.0	25.1	52.9	35.6	79.4	**18.7**	26.1	30.6
Variance	2251	756	627	968	791	535	3571	950	6870	**499**	697	933
Std. error	2.09	1.24	1.14	1.45	1.22	**0.98**	2.75	1.47	3.66	1.12	1.18	1.37
Skewness	1.46	2.15	1.44	1.10	1.35	1.57	**2.50**	1.52	1.30	1.64	1.41	1.80
Excess kurtosis	4.14	8.37	2.16	1.36	3.61	3.40	10.17	3.29	1.53	2.84	1.98	3.92

Table 1 continued

Station Elevation (m)	15-020 25	16-013 1155	16-016 1450	16-019 1150	16-030 1350	18-003 1740	18-013 1225	18-016 1670	20-009 895	21-003 1475	21-004 1180	21-006 **2040**
Percentiles												
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5%	8.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10%	13.4	0.0	0.0	3.5	0.0	1.8	0.4	3.0	0.0	0.0	0.0	1.0
25% (Q1)	29.7	6.2	4.0	15.0	9.2	7.5	10.6	13.1	10.4	1.4	5.1	9.7
50% (Median)	52.7	19.7	18.5	32.5	26.5	19.7	38.8	29.4	52.8	**10.5**	19.3	21.5
75% (Q3)	88.6	39.7	36.7	57.6	49.2	35.1	73.3	48.6	124.4	**26.7**	38.4	41.6
90%	127.6	62.4	57.3	81.3	67.9	58.1	120.4	76.6	198.1	**52.0**	62.2	71.2
95%	153.4	75.9	76.1	97.6	88.8	71.4	173.6	94.1	246.8	**65.4**	80.9	98.1
Max	377.6	229.6	133.0	183.8	216.0	154.5	492.2	189.9	436.1	130.2	135.4	186.9
Best-fit distribution*	WAK	WAK	WAK	WAK	WAK	WAK	WAK	WAK	BETA	GEV	WAK	WAK
Station Elevation (m)	21-007 1350	21-017 1200	21-025 1120	21-027 432	21-029 800	21-031 1880	21-034 1258	21-046 680	22-001 1700	24-013 2000	26-005 815	26-019 530
Statistic												
Sample size	511	501	508	480	507	501	432	451	546	**362**	451	481
Missing	5	3	8	24	9	27	0	5	18	10	17	**47**
Mean	53.4	29.2	39.0	35.5	69.1	41.7	58.6	36.3	72.6	58.2	36.4	53.5
Variance	2508	806	1419	1827	6419	1163	2711	1813	2313	1497	1549	4254
Std. error	2.22	1.27	1.67	1.95	3.56	1.52	2.51	2.00	2.06	2.03	1.85	2.97
Skewness	**0.77**	1.21	1.24	1.47	1.36	1.27	0.81	1.51	1.71	1.18	1.22	1.47
Excess kurtosis	**-0.31**	1.66	1.73	2.17	1.59	2.09	0.09	3.18	6.30	2.05	1.42	2.30
Percentiles												
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**5.2**	2.0	0.0	0.0
5%	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	**15.8**	8.8	0.0	0.0
10%	0.0	0.0	0.0	0.0	0.0	4.8	0.1	0.0	**21.7**	15.4	0.0	0.0
25% (Q1)	6.4	4.4	7.0	0.0	1.2	15.1	10.3	0.0	38.0	30.1	0.4	0.0
50% (Median)	40.8	23.7	31.0	21.5	44.4	35.8	48.3	23.0	62.1	52.3	28.0	25.3
75% (Q3)	90.3	44.8	59.0	57.5	109.6	58.6	91.0	61.9	98.2	78.4	59.7	88.1
90%	128.5	69.3	93.6	95.1	189.2	88.5	137.1	98.9	133.6	110.7	89.7	149.3
95%	148.5	84.7	108.1	126.0	238.4	109.6	162.1	121.4	157.2	132.9	112.7	177.4
Max	217.4	171.0	220.8	222.7	399.9	188.4	260.0	290.1	421.4	251.4	216.6	380.4
Best-fit distribution*	GEV	GEV	GEV	WAK	GEV	WAK	GEV	GEV	WAK	WAK	WAK	WAK

* *WAK* Wakeby, *BETA* beta, *GEV* generalized extreme value, *G.PAR* generalized pareto, *J.SB* Johnson SB

were found in the data matrix. Eight examples of the searched pairs in the neighborhood of the missing data, and matched pairs from various regions of the data matrix are shown in Fig. 1. The process described above for 12 clusters is repeated for 2–11 clusters, and a cluster frequency table is obtained for each month of 1985 (Fig. 6).

From left to right, each column in each table shows the frequencies obtained after dividing the observed value range for station 07-016 into 2–12 clusters. Each column heading indicates the number of clusters into which the observed data range is divided. Each row heading indicates the cluster indices. The Min and Max columns on the right show the cluster ranges when the number of clusters is 12. For example, Cluster 1 includes 0 values, Cluster 2 includes values from 0.1 to 4.5 mm, Cluster 11 includes values from 65.1 to 80.3 and Cluster 12 includes the highest values (80.8–204.8 mm).

The frequency table for each month provides information on the possible value of the missing data point in that month. For example, the first column of the frequency table for January 1985 shows the frequency values obtained for the first (the lower values) and the second (the higher values) clusters when the data series is divided into two clusters. The frequency value of the second cluster (10,075) is higher than the frequency value of the first cluster (6623). This shows that it is more probable that the

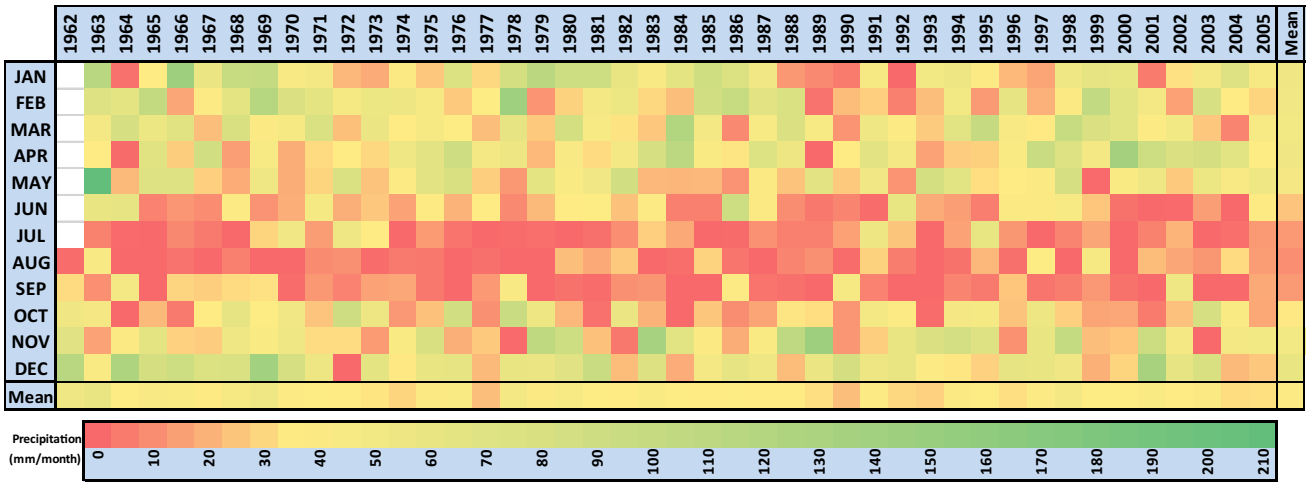

Fig. 3 Heat map of monthly precipitation observations for station 07-016

	OCT	NOV	DEC	JAN	FEB	MAR	APR
1982	2.1	22.8	92.4	58.1	54.9	31.2	46.2
1983	55.8	3.9	21.5	41.7	28.6	24.4	78.5
1984	18.7	131.0	71.5	63.6	22.1	118.1	107.0
1985	0.0	65.6	17.3	84.9	84.4	46.2	39.6
1986	24.0	37.7	46.6	71.9	95.4	8.5	31.9
1987	9.2	17.5	61.1	49.3	64.4	42.2	71.3
1988	15.2	38.8	51.0	12.2	74.1	73.6	52.5

Fig. 4 Observed values for months and years around January 1985

January 1985 data point range was identified by the second cluster (within the 29.0–204.8 mm range).

The division into three clusters yields frequencies of 1388, 2965 and 3397, respectively. The high value of the third cluster indicates that the desired value is most probably within the 45.9–204.8 mm range. Similarly, for the remaining clusters, the higher frequencies trend toward the bottom of the January 1985 cluster frequency table, indicating that the missing data point is most probably in the higher observation range.

The larger the number of clusters, the smaller the data range covered by each cluster. The increase in the number of clusters results in a green path that highlights the highest frequencies generally observed. This green path shows the clusters with the highest probability of representing the missing value range; in contrast, the red cells indicate those clusters with a lower probability of representing the missing value. Months with highly variable observations (like

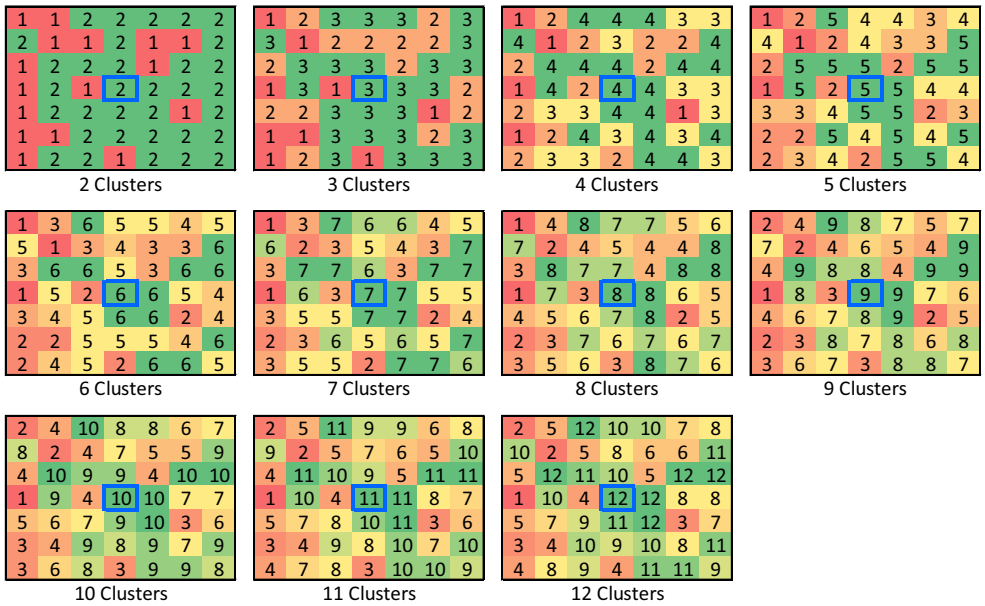

Fig. 5 Cluster numbers for data points near January 1985 at each clustering step

JAN	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6623	1388	519	210	112	73	35	40	13	1	0	0.0	0.0
2	10075	2965	1211	555	214	139	96	55	48	41	28	0.1	4.5
3		3397	1242	668	414	204	124	63	31	39	13	4.5	11.0
4			1609	632	354	240	168	120	65	33	30	11.1	16.9
5				801	378	254	167	106	85	59	39	16.9	23.2
6					413	281	133	75	62	42	63	23.2	29.0
7						275	206	109	80	54	32	29.0	37.5
8							181	168	99	45	51	37.6	45.3
9								134	84	52	43	45.9	53.6
10									96	74	70	53.6	65.0
11										98	63	65.1	80.3
12											64	80.8	204.8

FEB	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6904	1479	502	227	112	79	19	16	14	9	0	0.0	0.0
2	10626	3485	1361	547	260	138	102	85	42	37	37	0.1	4.5
3		3776	1301	756	432	187	126	72	58	41	21	4.5	11.0
4			1732	665	414	299	208	105	72	49	47	11.1	16.9
5				941	407	230	175	128	98	46	36	16.9	23.2
6					525	288	149	116	93	72	78	23.2	29.0
7						350	185	100	71	79	66	29.0	37.5
8							266	167	56	48	40	37.6	45.3
9								139	118	65	45	45.9	53.6
10									105	89	61	53.6	65.0
11										62	61	65.1	80.3
12											47	80.8	204.8

MAR	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	7184	1558	604	276	113	57	28	18	10	0	0	0.0	0.0
2	10679	3362	1234	556	311	206	155	96	65	44	24	0.1	4.5
3		3711	1444	698	405	190	125	74	64	55	64	4.5	11.0
4			1559	781	443	220	175	112	69	36	45	11.1	16.9
5				946	414	301	152	101	59	69	45	16.9	23.2
6					486	278	209	136	77	68	40	23.2	29.0
7						319	190	134	100	55	44	29.0	37.5
8							201	97	84	57	74	37.6	45.3
9								148	110	74	49	45.9	53.6
10									110	59	49	53.6	65.0
11										63	44	65.1	80.3
12											49	80.8	204.8

APR	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	7731	1631	650	290	144	73	53	20	13	5	0	0.0	0.0
2	10961	3524	1366	663	313	226	148	82	77	55	35	0.1	4.5
3		3772	1347	815	468	257	164	107	80	55	56	4.5	11.0
4			1479	729	415	268	198	116	67	51	43	11.1	16.9
5				867	435	295	172	141	77	65	57	16.9	23.2
6					541	246	163	111	81	62	60	23.2	29.0
7						294	187	141	95	62	30	29.0	37.5
8							207	146	57	73	56	37.6	45.3
9								120	142	56	53	45.9	53.6
10									93	97	36	53.6	65.0
11										57	73	65.1	80.3
12											56	80.8	204.8

MAY	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6299	1563	526	273	145	94	53	28	9	3	0	0.0	0.0
2	8220	2634	946	537	329	231	140	78	62	50	32	0.1	4.5
3		2756	1064	658	378	182	127	67	66	56	71	4.5	11.0
4			1051	607	369	248	170	162	107	42	22	11.1	16.9
5				572	422	208	189	111	74	54	52	16.9	23.2
6					306	224	110	99	95	56	53	23.2	29.0
7						197	143	92	57	90	73	29.0	37.5
8							131	113	68	51	41	37.6	45.3
9								113	52	44	26	45.9	53.6
10									55	57	57	53.6	65.0
11										63	29	65.1	80.3
12											36	80.8	204.8

JUN	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6161	1613	590	260	168	83	60	22	5	0	0	0.0	0.0
2	6842	2218	853	525	348	214	178	85	77	89	61	0.1	4.5
3		2083	793	455	284	187	150	106	89	64	46	4.5	11.0
4			780	477	298	167	146	127	53	51	41	11.1	16.9
5				543	360	128	90	81	65	51	55	16.9	23.2
6					266	228	135	84	40	37	49	23.2	29.0
7						192	114	95	65	35	17	29.0	37.5
8							143	118	85	33	35	37.6	45.3
9								72	78	58	50	45.9	53.6
10									69	59	48	53.6	65.0
11										63	43	65.1	80.3
12											33	80.8	204.8

JUL	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	5180	1572	612	283	163	92	69	19	20	2	0	0.0	0.0
2	5163	1684	531	416	266	172	171	106	86	59	55	0.1	4.5
3		1239	713	429	250	117	85	76	71	47	31	4.5	11.0
4			408	352	275	142	66	60	80	44	26	11.1	16.9
5				263	205	123	139	110	72	31	25	16.9	23.2
6					127	150	115	65	90	58	22	23.2	29.0
7						111	71	63	26	48	49	29.0	37.5
8							85	85	46	22	26	37.6	45.3
9								46	25	68	28	45.9	53.6
10									24	38	43	53.6	65.0
11										21	22	65.1	80.3
12											17	80.8	204.8

AUG	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	4818	1579	548	257	113	75	40	31	3	2	0	0.0	0.0
2	4807	1284	560	410	287	204	139	104	94	42	45	0.1	4.5
3		1139	474	291	179	116	95	109	93	71	38	4.5	11.0
4			452	303	197	102	89	58	36	44	44	11.1	16.9
5				218	197	148	94	60	42	41	19	16.9	23.2
6					131	127	72	75	44	46	19	23.2	29.0
7						90	79	49	24	49	45	29.0	37.5
8							99	35	69	22	23	37.6	45.3
9								44	25	49	28	45.9	53.6
10									23	22	24	53.6	65.0
11										23	17	65.1	80.3
12											15	80.8	204.8

SEP	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	4936	1580	526	288	124	54	47	24	6	1	0	0.0	0.0
2	5402	1485	574	347	264	221	178	170	92	77	38	0.1	4.5
3		1421	603	388	217	108	100	108	79	59	63	4.5	11.0
4			527	337	261	152	72	57	54	52	25	11.1	16.9
5				266	222	158	144	63	53	30	25	16.9	23.2
6					139	121	79	72	84	37	26	23.2	29.0
7						98	71	86	77	76	50	29.0	37.5
8							82	38	43	41	34	37.6	45.3
9								53	27	39	36	45.9	53.6
10									45	15	25	53.6	65.0
11										33	15	65.1	80.3
12											29	80.8	204.8

OCT	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	4862	1384	471	252	146	31	17	8	1	0	0	0.0	0.0
2	6400	1794	662	425	306	198	175	90	74	59	30	0.1	4.5
3		1869	733	378	218	163	97	94	93	77	69	4.5	11.0
4			843	410	238	148	101	81	67	58	33	11.1	16.9
5				424	267	172	128	114	42	53	53	16.9	23.2
6					218	180	118	112	60	33	25	23.2	29.0
7						132	115	85	65	47	21	29.0	37.5
8							107	60	50	43	37	37.6	45.3
9								77	48	46	38	45.9	53.6
10									59	35	31	53.6	65.0
11										37	24	65.1	80.3
12											38	80.8	204.8

NOV	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	5539	1262	434	260	107	74	59	15	13	2	0	0.0	0.0
2	7276	2295	905	388	188	176	122	72	68	54	34	0.1	4.5
3		2344	891	459	308	165	138	54	56	24	23	4.5	11.0
4			998	562	258	208	144	136	62	32	26	11.1	16.9
5				535	294	181	133	118	69	61	46	16.9	23.2
6					300	164	130	111	64	73	48	23.2	29.0
7						178	116	78	67	34	53	29.0	37.5
8							119	102	63	62	25	37.6	45.3
9								92	63	47	32	45.9	53.6
10									92	60	48	53.6	65.0
11										55	33	65.1	80.3
12											44	80.8	204.8

DEC	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6101	1410	559	281	124	57	42	19	7	4	0	0.0	0.0
2	8137	2467	884	464	251	249	138	102	45	66	27	0.1	4.5
3		2924	989	570	343	158	89	88	68	39	34	4.5	11.0
4			1220	536	318	241	192	116	64	38	33	11.1	16.9
5				591	364	190	126	94	104	58	47	16.9	23.2
6					312	227	140	86	47	52	51	23.2	29.0
7						200	161	138	67	45	45	29.0	37.5
8							150	122	86	54	39	37.6	45.3
9								87	106	77	50	45.9	53.6
10									65	62	59	53.6	65.0
11										38	38	65.1	80.3
12											37	80.8	204.8

Fig. 6 Cluster frequency tables of the months of 1985

June) result in fuzzy frequency tables, while months with low variability (like August) produce more distinguishable red and green patterns. For 1985, the green trends are more apparent in the January–April and July–September frequency tables.

2.3 Estimation of missing values based on cluster frequencies

The 12th column in the frequency table for January 1985 (Fig. 6) is used to estimate the missing data for that date. The clusters that occur most often provide the most likely ranges of value for the missing data. In the January 1985 example, the highest frequency (70) occurs in cluster 10, which represents the precipitation range between 53.6 and 65.0 mm. The average of the 70 observations (60.04 mm) used to generate this frequency is the most likely estimation of the missing January 1985 value. The obtained estimate will always be within the range of the averaged cluster. In the present example, the actual observed value for the January 1985 data point was 84.9 mm (within the range of the 12th cluster).

The second highest frequency (64) obtained by the example model occurred in cluster 12 (80.8–204.8 mm range). The average of the 64 observations used to generate this frequency is 110.6 mm and is the second probable estimate for the January 1985 value. The third highest frequency (63) occurred in clusters 6 and 11, which represent the third and fourth most likely estimates (71.2 and 25.2 mm) of value. The green path in the January 1985 frequency table indicates that the most likely value will be within the range of clusters 10–12; and, of the first five estimations, the third estimate obtained (cluster 11) is the nearest to the real observed value. This approach is repeated for the five highest total frequency values for each month analyzed, and the five most likely estimates for each month are written in a correlation tables output file by the software. As previously stated, precipitation is relatively chaotic, and the most likely precipitation might not be the experienced precipitation. Therefore, generating multiple precipitation values with a high likelihood of occurrence is very useful to scientists and practitioners who work with precipitation data.

The three lowest frequencies obtained for the 12 clusters occurred in clusters 1–3, indicating that the range 0.00–11.0 mm is the least likely to represent the total precipitation that occurred in January 1985. The actual 1985 data points to be tested were removed prior to the application of the method and were not known by the software at any stage of the estimation process.

The ability of the FBI method to estimate precipitation values can be compared to estimates generated using the EM and MLR methods, which are also direct methods. EM is an iterative method used to identify the maximum likelihood estimates of parameters in statistical models (Dempster et al. 1977). It also enables parameter estimation in probabilistic models with incomplete data. A good introduction to the mathematical foundations and applications of the EM method is provided by Do and Batzoglou (2008). As with the FBI method, the EM and MLR methods have the ability to generate estimates for a series by using existing observations in the series itself; they do not require preprocessing of data, and unlike methods such as ANN, they do not require the adjustment of any input parameters to improve the results. To compare these two models with the FBI method, all existing station 07-016 observations were estimated using the EM and regression modules in the missing value analysis toolbox of the IBM S.P.S.S. software. The same approach used to estimate values in the FBI method was applied. The data from each year was removed and estimated using both methods. Table 2 shows the estimates obtained for the test year using the FBI, EM and regression methods, together with the long-term monthly averages. The correlations obtained using the EM (0.713), regression (0.778) and long-term average (0.733) methods are significantly lower than the correlation found using the FBI method (0.976).

To test the advantages of generating multiple estimates for a missing value, the increase in correlation with the increase of the number of estimations is assessed for all observations of the station 07-016 annually. Table 3 shows the correlations between the observed values and the best estimates generated within the first 2, 3, 4 and 5 estimations for each year. Annual correlations over 0.7 occurred between the observed values and the nearest estimates in the first two estimations in 58% of cases (25/43). This rate

Table 2 Correlations between the observed values and the best estimates from the FBI, EM, regression and long-term average methods for 1985

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
Observation	**84.9**	84.4	46.2	39.6	20.0	6.0	**0.0**	27.2	0.0	24.0	37.7	46.6	**Corr.**
Best FBI Est.	71.2	**73.1**	49.4	27.0	19.9	**1.8**	**1.8**	34.1	2.4	20.4	31.8	48.8	**0.976**
EM	49.8	48.8	48.0	49.6	45.2	23.8	12.6	**9.2**	12.8	32.7	48.6	**57.6**	0.713
Regression	**82.0**	53.2	24.3	25.6	**−12.2**	−2.5	12.7	40.9	20.6	37.0	40.5	56.0	0.778
Long-term Av.	**51.8**	49.4	48.1	49.7	44.9	24.6	16.1	**13.2**	16.7	34.0	48.4	58.7	0.733

Table 3 Correlations between observed values from station 07-016 and the estimates generated using the clusters with the five highest frequencies

Estimations	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
1–2	0.807	0.679	0.797	0.806	0.583	0.763	0.881	0.898	0.793	0.265	0.698
1–3	0.925	0.896	0.876	0.884	0.809	0.884	0.963	0.883	0.938	0.590	0.816
1–4	0.929	0.933	0.950	0.904	0.950	0.898	0.964	0.916	**0.985**	0.842	0.972
1–5	0.929	0.952	0.956	0.945	0.980	0.970	0.964	0.949	0.984	0.924	0.972
Estimations	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
1–2	0.627	0.823	0.561	0.694	0.773	0.834	0.841	0.735	0.552	0.743	0.461
1–3	0.867	0.851	0.773	0.878	0.936	0.878	0.861	0.908	**0.582**	0.903	0.643
1–4	0.939	0.860	0.840	0.915	0.962	0.931	0.937	0.922	**0.766**	0.916	0.920
1–5	0.944	0.960	0.979	0.942	0.957	0.975	0.967	0.929	0.922	0.966	0.980
Estimations	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
1–2	0.717	0.737	0.539	0.880	0.770	0.652	0.430	**0.139**	**0.944**	0.873	0.792
1–3	0.955	0.813	0.856	**0.972**	0.832	0.707	0.720	0.832	0.964	0.922	0.840
1–4	0.973	0.865	0.901	0.981	0.962	0.873	0.872	0.932	0.982	0.929	0.947
1–5	0.976	0.865	**0.904**	**0.992**	0.962	0.931	0.930	0.935	0.982	0.928	0.964
Estimations	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Whole
1–2	0.706	0.553	0.574	0.747	0.814	0.901	0.672	0.777	0.468	0.479	**0.696**
1–3	0.812	0.929	0.795	0.815	0.826	0.937	0.916	0.882	0.908	0.683	**0.843**
1–4	0.877	0.940	0.847	0.910	0.884	0.959	0.912	0.944	0.928	0.903	**0.912**
1–5	0.924	0.969	0.933	0.954	**0.992**	0.961	0.917	0.951	0.950	0.983	**0.944**

increased to 91% (39/43) when three estimates were generated and to 100% when four or five estimates were produced. Similarly, the rate of annual correlations over 0.8 was 28, 81, 98 and 100% for the first 2, 3, 4 and 5 estimates, respectively; and the rate of annual correlations over 0.9 was 5, 33, 74 and 98%, respectively. These results indicate that increasing the number of estimates generated increases the model's reliability and accuracy.

The last column in the table (titled "Whole") shows the correlations between the entire observed series and the series of best estimates derived from the first 2, 3, 4 and 5 estimations. A correlation value of 0.843 obtained for the first three estimations might be regarded as sufficient to estimate precipitation. Increasing the number of estimates to 4 produces a correlation of 0.912, while increasing the number to 5 yields a correlation of 0.944 for the entire series. These correlations indicate the production of extremely reliable precipitation estimates.

Table 4 presents the correlations between the observed values from station 07-016 and the estimates derived using the FBI, EM and regression methods, as well as the long-term averages for each year. For all years, the correlations between the FBI method estimates and the observed values exceed the correlations between the EM, regression and long-term average values and the observed data. While 98% (42/43) of the annual correlations between the FBI method and the observed values are over 0.9, all annual correlations with the compared methods are under 0.9.

The highest and lowest correlations produced by each method are shown in bold. The obtained results reveal that the estimates produced using the EM method tend to be more similar to the long-term averages than the observed values. This resulted similar correlation values for both the EM method and the long-term averages across the years. The correlations of the compared methods follow a similar pattern. Generally, the correlations increase or decrease together over the years. For example, the lowest annual correlations with the EM method (0.015) and the long-term averages (0.030) occurred in 1972. This year represented the sixth lowest annual correlation for the FBI method (0.904) and the third lowest for the regression (0.006) method.

To compare the general performance of the methods used to estimate precipitation values at station 07-016, five statistical measures (correlation (r), Nash–Sutcliffe efficiency coefficient (E), root mean squared error (RMSE), mean absolute error (MAE) and mean bias error (MBE)) were calculated and presented in Table 5. The FBI method performed best using all statistical measures except the MBE. The negative E value obtained using the regression method indicates that the observed mean is a better indicator of value than the regression method. The other

Table 4 Correlations between the observed values for station 07-016 and the estimates from the FBI, EM, regression methods, and the long-term averages

Year	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
FBI	0.929	0.952	0.956	0.945	0.980	0.970	0.964	0.949	0.984	0.924	0.972
EM	0.282	0.352	**0.881**	0.564	0.707	0.633	0.703	0.577	0.696	**0.015**	0.461
Regression	0.598	0.126	0.538	0.691	0.548	0.277	0.382	0.240	0.456	0.006	−0.007
Long-term Av.	0.334	0.403	**0.884**	0.623	0.744	0.667	0.752	0.608	0.701	**0.030**	0.484
Year	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
FBI	0.944	0.960	0.979	0.942	0.957	0.975	0.967	0.929	0.922	0.966	0.980
EM	0.868	0.794	0.633	0.675	0.433	0.589	0.744	0.705	0.398	0.533	0.546
Regression	0.764	0.591	0.756	0.356	0.379	0.029	0.323	0.433	0.227	0.507	0.257
Long-term Av.	0.865	0.791	0.662	0.666	0.486	0.626	0.764	0.738	0.412	0.566	0.575
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
FBI	0.976	**0.865**	0.904	**0.992**	0.962	0.931	0.930	0.935	0.982	0.928	0.964
EM	0.713	0.277	0.856	0.572	0.382	0.288	0.458	0.287	0.625	0.718	0.238
Regression	0.778	0.248	0.356	0.271	0.785	**−0.012**	0.299	0.416	**0.839**	0.624	0.092
Long-term Av.	0.733	0.316	0.854	0.579	0.422	0.342	0.492	0.303	0.638	0.720	0.268

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
FBI	0.924	0.969	0.933	0.954	0.992	0.961	0.917	0.951	0.950	0.983
EM	0.601	0.458	0.765	0.325	0.652	0.653	0.435	0.613	0.546	0.675
Regression	0.674	0.387	0.267	0.361	0.266	0.191	0.262	0.328	0.197	0.644
Long-term Av.	0.614	0.476	0.765	0.348	0.673	0.694	0.483	0.643	0.569	0.666

statistical measures also reveal that utilization of long-term averages is preferred to use of the regression method. As expected, the MBE for long-term averages was zero, while the MAE was lowest for the FBI method, suggesting that the FBI method estimates are closer to the observed values. The MAE and MBE statistics should be considered together because equal averages for estimates and observed values does not generally mean that the estimations are sufficiently close to the observations. The averages may be similar even though there are significant positive and negative differences between the estimates and observed values. These differences can be detected by calculating the MAE, which has advantages over the RMSE and MBE in assessing average model performance (Willmott and Matsuura 2005).

The graphs shown in Fig. 7 compare the observed values from station 07-016 with the estimates produced using the FBI, EM and regression methods. A very good fit is seen between the FBI method estimates and the observed values across the time series, indicating that the method is sensitive to the variations in precipitation. On the other hand, the estimates produced using the EM and regression methods lack generality and sensitivity. Figure 7 also shows that the FBI method provides lower estimates for rarely observed high precipitation values even though it produces better estimates compared to the EM and regression methods. Low estimations of extreme values occur as a result of the estimation logic behind the FBI method, which considers the frequency of observed values; it is well known that the frequency of extreme precipitation is generally low. The graphs also show that the estimates obtained for extreme values are always higher than the remaining estimates. This may be considered a disadvantage of the method; however, its ability to estimate extreme values might be improved by considering observations from nearby stations.

2.4 Application of the FBI method using the remaining 69 precipitation stations

The above discussion was generated based on estimates and observations for a single station (07-016). A method's ability to estimate values for a single station is not sufficient to claim that it will be successful in estimating values for other stations. To test the FBI method's application across multiple stations, we used the above method to estimate precipitation values for 70 stations across 21 different basins in Turkey. Stations were chosen based on location and the variation in observed values. The stations reflect various climates in Turkey, ranging from dry to wet (see the descriptive statistics of the observed series in Table 1). Table 6 presents the statistical measures (r, E,

Table 5 Statistical measures used to compare all observed values from station 07-016 with the estimates generated using the FBI, EM, and regression methods and the long-term averages

	r	E	RMSE	MAE	MBE
FBI	**0.944**	**0.889**	**10.322**	**6.764**	0.084
EM	0.513	0.263	26.566	19.987	−0.031
Regression	**0.351**	**−0.258**	**34.698**	**26.216**	**0.557**
Long-term Av.	0.544	0.296	25.954	19.532	**0**

normalized root mean squared error (NRMSE), mean absolute scaled error (MASE), MAE and MBE) generated for each station based on a comparison between produced FBI method estimates and the observed values from each station. The number of years data was available for each station is also presented in the table.

The correlations between the results of the FBI method and the observations exceeded 0.9 for 24% (17/70) of stations and exceeded 0.85 for 79% (55/70) of stations. The minimum correlation was 0.795, and the maximum correlation was 0.944. 11 of the 15 stations with the highest correlations are located in basins 4, 5, 6, 7, 8 and 9, which are all located within the Eastern Aegean and Eastern Mediterranean regions of Turkey. Similarly, 8 of the 15 stations with the lowest correlations are in basins 12, 13,

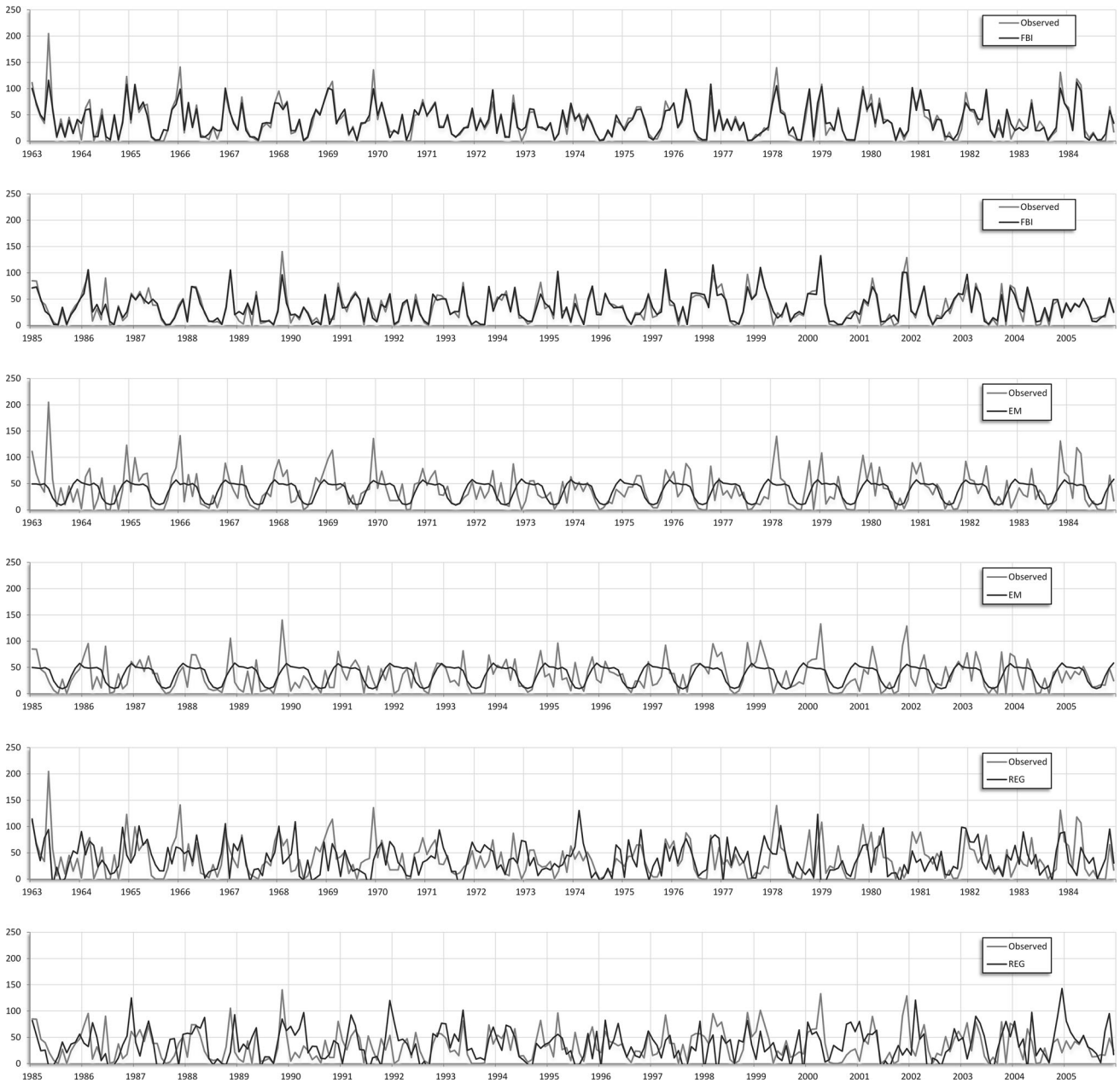

Fig. 7 Comparisons between estimates produced using the FBI, EM and regression methods with observed data from station 07-016

Table 6 Statistical measures of the comparisons between the estimated and observed values of the stations

Station	r	E	NRMSE	MASE	MAE	MBE	Obs. years
01-004	0.898	0.796	0.090	0.303	9.981	−2.112	41
01-005	0.851	0.690	0.109	0.336	13.418	−2.940	41
01-008	0.797	**0.591**	0.088	0.338	12.744	−4.764	40
02-004	0.900	0.778	0.071	0.289	14.335	−6.426	38
02-009	0.843	0.671	0.107	0.309	18.675	−8.929	41
02-011	0.880	0.754	0.080	0.305	10.954	−3.123	39
02-012	0.885	0.737	0.077	0.290	18.730	−8.344	38
02-018	0.873	0.730	0.096	0.297	14.287	−4.753	41
03-009	0.886	0.764	0.064	0.296	10.662	−2.868	40
03-013	0.897	0.780	0.076	0.290	12.462	−4.552	44
03-027	0.862	0.717	0.109	0.324	12.377	−3.041	43
04-003	0.904	0.788	0.086	0.283	15.604	−6.746	41
04-008	0.914	0.798	0.064	0.263	18.091	−8.746	36
05-001	0.880	0.737	0.089	0.332	10.439	−4.249	44
05-004	0.901	0.784	0.079	0.300	9.457	−3.527	40
05-007	0.904	0.765	0.079	0.303	12.001	−7.172	45
05-008	0.886	0.749	0.074	0.303	12.290	−6.260	44
05-012	0.899	0.785	0.096	0.299	9.008	−1.786	44
05-016	0.923	0.829	0.079	0.268	10.700	−3.583	44
06-005	0.913	0.784	0.093	0.287	17.830	−9.468	44
07-013	0.933	0.835	0.073	0.279	20.599	−10.19	44
07-016	**0.944**	**0.889**	**0.050**	**0.237**	6.764	**0.084**	44
07-022	0.894	0.778	0.080	0.297	12.531	−3.243	44
08-006	0.942	0.841	0.071	0.259	**27.571**	**−17.67**	43
08-008	0.927	0.817	0.079	0.274	17.121	−8.567	44
08-010	0.877	0.735	0.068	0.292	14.102	−5.691	40
08-013	0.915	0.794	0.070	0.284	19.726	−12.28	43
08-014	0.899	0.785	0.081	0.289	8.553	−1.984	45
09-014	0.920	0.799	0.068	0.290	27.246	−14.89	44
10-007	0.874	0.735	0.075	0.310	8.191	−1.160	43
11-002	0.865	0.714	0.087	0.313	11.050	−3.379	41
12-003	0.864	0.726	0.096	0.322	7.032	−1.514	43
12-011	0.842	0.689	0.090	0.304	7.201	−2.006	42
12-012	0.867	0.726	0.103	0.321	8.509	−1.797	42
12-014	0.834	0.668	0.076	0.336	17.169	−4.751	42
12-042	0.837	0.665	0.091	0.320	8.693	−3.378	41
12-047	0.885	0.758	0.077	0.264	17.407	−6.972	41
12-049	0.847	0.686	0.080	0.312	7.376	−2.284	40
14-005	0.897	0.781	0.065	0.284	8.214	−2.219	38
14-007	0.866	0.725	0.053	0.291	9.294	−2.161	41
14-017	0.820	0.644	0.092	0.366	12.702	−2.881	39
14-018	0.860	0.699	0.092	0.317	15.768	−4.890	38
14-019	0.888	0.759	0.099	0.294	7.850	−2.205	41
15-008	0.851	0.670	0.091	0.361	11.040	−4.767	40
15-010	0.859	0.702	0.102	0.330	7.399	−1.929	44
15-019	0.886	0.762	0.100	0.308	8.934	−2.400	43
15-020	0.820	0.636	0.076	0.356	15.755	−6.051	41
16-013	**0.795**	0.595	0.076	0.346	8.029	−3.333	36
16-016	0.824	0.630	**0.114**	0.364	7.880	−3.905	44
16-019	0.880	0.751	0.084	0.307	8.929	−2.286	40

Table 6 continued

Station	r	E	NRMSE	MASE	MAE	MBE	Obs. years
16-030	0.870	0.736	0.067	0.316	8.052	−1.986	45
18-003	0.873	0.713	0.080	0.316	**6.367**	−2.757	44
18-013	0.832	0.653	0.071	0.308	14.830	−6.633	44
18-016	0.867	0.720	0.086	0.334	8.828	−2.225	44
20-009	0.930	0.841	0.076	0.253	15.794	−6.863	44
21-003	0.860	0.662	0.100	0.347	6.447	−4.225	44
21-004	0.870	0.713	0.104	0.316	6.890	-3.395	44
21-006	0.887	0.724	0.086	0.308	7.778	−3.802	44
21-007	0.930	0.851	0.089	0.275	10.057	−4.089	43
21-017	0.891	0.764	0.081	0.283	6.712	−2.955	44
21-025	0.866	0.726	0.089	0.311	9.702	−3.665	40
21-027	0.835	0.647	0.114	0.387	11.140	−7.945	43
21-029	0.912	0.788	0.092	0.325	17.278	−11.24	45
21-031	0.898	0.784	0.084	0.272	8.363	−2.649	44
21-034	0.922	0.838	0.080	0.263	11.201	−2.927	43
21-046	0.845	0.672	0.084	0.380	11.289	−7.292	41
22-001	0.857	0.713	0.062	0.299	12.892	−3.891	43
24-013	0.882	0.752	0.077	0.310	11.213	−2.679	42
26-005	0.831	0.660	0.106	0.345	10.124	−6.399	42
26-019	0.812	0.620	0.106	**0.396**	17.423	−11.98	42
Minimum	0.795	0.591	0.050	0.237	6.367	−17.67	36
Average	0.877	0.737	0.084	0.308	12.072	−4.910	42
Maximum	0.944	0.889	0.114	0.396	27.571	0.084	45

14, 15, 16 and 18, which are located in the central and northern regions of Turkey. While the lowest 14 correlation values occurred for stations fitting to the Wakeby distribution, none of the 10 stations with the highest correlations fit this distribution. 8 of the 15 best correlated stations (including the first 3) instead fit the GEV distribution.

All Nash–Sutcliffe efficiency coefficients exceeded 0.591; 73% (51/70) were over 0.70 and 11% (8/70) were over 0.80. The highest Nash–Sutcliffe efficiency coefficient was 0.889. The highest NRMSE value was 0.114; 83% (58/70) of the NRMSE values fell below 0.10, while the lowest NRMSE value was 0.050. All MASE values fell below 0.40; 90% (63/70) of these values were under 0.35 and 41% (29/70) were under 0.30. The lowest MASE value was 0.237. The MAE values ranged between 6.367 (obtained for 18-003) and 27.571 (obtained for 08-006), suggesting that the high correlation value (0.942) obtained for station 08-006 may be misleading because the MAE and MBE values for the station are higher than those for the remaining stations. MBE values ranged between 0.084 and −17.667, with data from 69 stations generating negative MBE values. This indicates that the precipitation estimates generated using the FBI method have a slight negative bias. Greater bias occurred at stations where extreme values and variations were much higher than at other stations, resulting in greater differences between the estimated and observed values. Future studies might investigate ways to obtain average estimates closer to average observations to eliminate bias errors without increasing MAE. A method to overcome this bias might be to multiply all estimates by the ratio between the averages of the observed values and the estimated values for each station. This intervention should only be made if the MAE between the observed and estimated values also decreases. Furthermore, though this intervention may improve the estimation of higher values, a much larger number of values in the lower ranges might be overestimated. A selected bias correction method will not produce the best results for all data series (Ajaaj et al. 2016); thus, the selection of a bias correction method should be left to the users of the FBI method where necessary.

3 Discussion and conclusions

This article assesses the ability of the FBI method to estimate non-continuous monthly precipitation data without the use of observation from neighboring stations. The goodness of fit measures calculated between the observed and estimated series show that the FBI method is capable of estimating monthly precipitation data obtained from various climatic zones. However, it is impossible to claim

that the method will always successfully estimate values for stations in other regions without first applying the method to observations from those stations. The practical experiences in the literature show that no data driven methodology is perfect enough to provide the best results for all stations or for all variables.

This method may also be used to estimate weekly or daily precipitation data; however, given that the randomness of precipitation generally increases with decreasing observation periods, it is anticipated that the success of the method will be lower for precipitation estimates at the weekly or daily scale. The inclusion of observations from highly correlated neighboring stations improve the generation of estimates with shorter sampling frequencies. Further studies may investigate the influence of neighboring stations on the estimation power of the presented method.

As noted above, the method analyzed in this study may not be suitable for the estimation of extreme observations that occur at a very low frequency. Values that occur with a very low frequency in a data series also have a low occurrence probability and will not occur frequently enough to be determined among the highest possible values. As is valid for most data-driven methods, the length of the data series used may influence the performance of the proposed method. The method may be less useful when applied to short data series, as the estimates produced by the presented method are based on the frequencies of the observed value ranges. The input dataset should have at least seven rows of input data (i.e., 7 years for monthly data) and more data will generally provide more information about the frequencies of the observations, consequently supporting the possibility of better estimations.

Another limitation of the method is that writing a software code for its implementation might not be easy for every user. With this in mind, a link to the source code written in Visual Basic is provided to the readers in Appendix 3. This will enable users to implement the FBI method on other datasets or in other research areas. Users of other programming languages or operating systems will need to convert the code.

The FBI method may be applied in many scientific disciplines, as it is a generally applicable, direct analysis method that requires no determination of input parameters, nor does it require any preprocessing of data. While most existing methodologies are one-dimensional, the FBI method is two-dimensional and has been shown to perform better when compared to the EM and MLR methods in the estimation of precipitation.

Acknowledgements I would like to thank The General Directorate of the State Hydraulic Works of Turkey for providing the data used in this study and the editors and reviewers for their valuable contributions and comments, which greatly improved the manuscript.

Appendix 1: Determination of range clusters

For various reasons, there are generally gaps in any time series dataset, and the reliable estimation of the missing data has great value. In the FBI method, the missing data value at the center of the matrix in Fig. 8 (cell i, j) has temporal and quantitative relationships with nearby cells.

To estimate the probable range of the missing value at node i, j, the value ranges of all existing observations in the dataset should be determined. First, the observed data is sorted in ascending order and a three-dimensional vector containing the sorted data and associated coordinates in the data matrix is generated. The coordinate of each data point used in this study is the observed month (column) and year (row) of the data and is unique for each observation. The coordinate information is crucial because the observation time of a given value affects the temporal and quantitative investigation of time series data. Sorting and investigating statistical relationships for a variable without considering the observation times of each individual variable mean ignoring information about the temporal relationship between observations.

After sorting the observations, the observed time series range is divided into 2 to n range clusters to evaluate and estimate the possible clusters into which the missing data point may fall. The value of n may increase with the amount of available data; this increase would provide more precise results, as the value range for each cluster would be narrower. The number of clusters should be chosen so that the distribution of the observed values is sufficiently represented. Currently, the maximum number of clusters is determined by running the software for various number of clusters. It must be noted that the selected cluster number may not be optimum for obtaining the best results, although the method may still produce successful results. A good approach to determine the maximum number of clusters might be to start with a high number of clusters (like 50). Then, the cluster number that produces sufficient frequency values and cluster ranges might be chosen by looking at the generated frequency tables. Future studies should propose a method for determining the optimum number of clusters based on the number and variability of

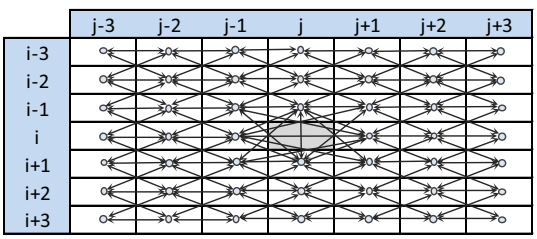

Fig. 8 Pairs to be searched in the data matrix

observations to further improve the successful estimation of missing values.

Clusters may be generated using two different approaches. In the first approach, each cluster has as equal a number of elements as possible (the clusters have varying ranges). Observed values are assigned to clusters using Eq. (1).

$$Cl_i = int\left(\frac{i * n_{cl}}{n_d}\right) + 1 \quad (1)$$

In the second approach, range values are equalized (the clusters have a varying number of elements). The bounds of the cluster ranges are the lowest and highest observations belonging to that range. Observed values are assigned to clusters using Eq. (2).

$$Cl_i = int\left(\frac{(X_i - X_{min}) * n_{cl}}{X_{max} - X_{min}}\right) + 1 \quad (2)$$

In the above equations: n_d is the total number of observations in the sorted data vector, i is the rank (index number) of the observation in the sorted data vector (changes between 1 and n_d), n_{cl} is the number of clusters used to divide the sorted data vector, Cl_i is the cluster index to be assigned to the i-th observation (changes between 1 and n_{cl}), $int()$ is the function converting a decimal number into an integer, X_i is the i-th observation in the sorted data series, $X_{min}; X_{max}$ the minimum and maximum observations.

Both approaches have advantages and disadvantages over each other. Selection of the appropriate clustering method completely depends on the diversity of the observed time series. For example, if the number of elements in certain clusters become too high compared to other clusters, then it would be better to generate clusters with an equal number of elements. For the precipitation data used in this paper, the first approach was used; each cluster included a similar number of elements. For example, for station 07-016, the first 11 clusters cover the range 0.0–80.3 mm while the 12th cluster covers the range 80.8–204.8 mm (1.54 times greater than the cumulative range of the first 11 clusters).

Appendix 2: Generation of the cluster frequency table

The clustering process explained in Appendix 1 assigns a cluster index to each observation. The cluster index value of each cell is the key to finding the cluster value of the missing cell. When the observed range is divided into two clusters, the first cluster includes the lower values and has a cluster index of 1, and the second cluster includes the higher values and has a cluster index value of 2. All adjacent cluster pairs in the data matrix near the missing cell are searched. Frequency values for the probable clusters are set to zero prior to the initiation of the search process. At the first clustering step, there are two possible clusters (1 or 2) into which the missing data may fall. When a match for a cluster pair is found in the matrix, the frequency of the cluster value at the relative location of the missing data point is increased by one. The maximum number of unique cluster pairs near the missing data point is 158. This number decreases if there is more than one missing data point in the neighborhood. The following rules provide three examples of the 158 unique rules used to find matching cluster pairs.

1. If $[Cl(X_{i,j-2}) = a \;\&\; Cl(X_{i,j-1}) = b]$ and if $[Cl(X_{p,q-2}) = a \;\&\; Cl(X_{p,q-1}) = b \;\&\; Cl(X_{p,q}) = c]$ then freq(c) = freq(c) +1.
2. If $[Cl(X_{i-2,j}) = a \;\&\; Cl(X_{i-1,j}) = b]$ and if $[Cl(X_{p-2,q}) = a \;\&\; Cl(X_{p-1,q}) = b \;\&\; Cl(X_{p,q}) = c]$ then freq(c) = freq(c) +1.
3. If $[Cl(X_{i-2,j-2}) = a \;\&\; Cl(X_{i-1,j-1}) = b]$ and if $[Cl(X_{p-2,q-2}) = a \;\&\; Cl(X_{p-1,q-1}) = b \;\&\; Cl(X_{p,q}) = c]$ then freq(c) = freq(c) +1.

In the above rules, Cl(X) is the cluster index of the observed value X; i and j are the row and column numbers of the missing node at the center of the 7 × 7 cell field; p and q are the row and column numbers of the cell at the relative location of the missing data at i, j and a, b and c are the cluster numbers of the related cells. When the entire dataset is divided into two clusters, a, b and c might have values of 1 or 2; for n clusters, they may have values ranging between 1 and n. The values of a, b and c may differ for each rule because they may represent different locations within the data matrix. The above three rules represent the horizontal cluster pair to the left of the missing node, the vertical cluster pair above the missing node and the diagonal cluster pair to the top left of the missing node, as shown in Fig. 9a in orange, yellow and green, respectively. Figure 9b shows the location of the first pair match for the first rule. With the first match, the frequency of the cluster number of the cell at the relative location of the missing data point is increased by one (the cell at p, q shown in pink). This is done because the cluster value at cell p, q is a probable value for the missing node at i, j, given that both cells have the same cluster pairs to the left. The search for the same pair then continues until all matching pairs are found and the frequencies of the clusters at the corresponding cells p, q are increased by one (for each match, the values of p and q might be different because the matching pairs will be at different locations within the data matrix).

After the search for the first cluster pair is completed, the above process is repeated for the next pair until all pairs near the missing data point have been searched and the total frequencies for each probable cluster determined. The clusters with the highest frequencies will be the most likely

Fig. 9 **a** The cluster pairs (*orange, yellow* and *green*) for which rules 1, 2 and 3 are written, **b** a matching cluster pair for the first rule

a

	j-3	j-2	j-1	j	j+1	j+2	j+3
i-3	Cl(i-3,j-3)	Cl(i-3,j-3)	Cl(i-3,j-1)	Cl(i-3,j)	Cl(i-3,j+1)	Cl(i-3,j+2)	Cl(i-3,j+3)
i-2	Cl(i-2,j-3)	Cl(i-2,j-2)	Cl(i-2,j-1)	Cl(i-2,j)	Cl(i-2,j+1)	Cl(i-2,j+2)	Cl(i-2,j+3)
i-1	Cl(i-1,j-3)	Cl(i-1,j-2)	Cl(i-1,j-1)	Cl(i-1,j)	Cl(i-1,j+1)	Cl(i-1,j+2)	Cl(i-1,j+3)
i	Cl(i,j-3)	Cl(i,j-2)	Cl(i,j-1)		Cl(i,j+1)	Cl(i,j+2)	Cl(i,j+3)
i+1	Cl(i+1,j-3)	Cl(i+1,j-2)	Cl(i+1,j-1)	Cl(i+1,j)	Cl(i+1,j+1)	Cl(i+1,j+2)	Cl(i+1,j+3)
i+2	Cl(i+2,j-3)	Cl(i+2,j-2)	Cl(i+2,j-1)	Cl(i+2,j)	Cl(i+2,j+1)	Cl(i+2,j+2)	Cl(i+2,j+3)
i+3	Cl(i+3,j-3)	Cl(i+3,j-2)	Cl(i+3,j-1)	Cl(i+3,j)	Cl(i+3,j+1)	Cl(i+3,j+2)	Cl(i+3,j+3)

b

	q-3	q-2	q-1	q	q+1	q+2	q+3
p-3	Cl(p-3,q-3)	Cl(p-3,q-3)	Cl(p-3,q-1)	Cl(p-3,q)	Cl(p-3,q+1)	Cl(p-3,q+2)	Cl(p-3,q+3)
p-2	Cl(p-2,q-3)	Cl(p-2,q-2)	Cl(p-2,q-1)	Cl(p-2,q)	Cl(p-2,q+1)	Cl(p-2,q+2)	Cl(p-2,q+3)
p-1	Cl(p-1,q-3)	Cl(p-1,q-2)	Cl(p-1,q-1)	Cl(p-1,q)	Cl(p-1,q+1)	Cl(p-1,q+2)	Cl(p-1,q+3)
p	Cl(p,q-3)	Cl(p,q-2)	Cl(p,q-1)	Cl(p,q)	Cl(p,q+1)	Cl(p,q+2)	Cl(p,q+3)
p+1	Cl(p+1,q-3)	Cl(p+1,q-2)	Cl(p+1,q-1)	Cl(p+1,q)	Cl(p+1,q+1)	Cl(p+1,q+2)	Cl(p+1,q+3)
p+2	Cl(p+2,q-3)	Cl(p+2,q-2)	Cl(p+2,q-1)	Cl(p+2,q)	Cl(p+2,q+1)	Cl(p+2,q+2)	Cl(p+2,q+3)
p+3	Cl(p+3,q-3)	Cl(p+3,q-2)	Cl(p+3,q-1)	Cl(p+3,q)	Cl(p+3,q+1)	Cl(p+3,q+2)	Cl(p+3,q+3)

clusters into which the missing node will fall. Some cluster frequencies might remain at zero, indicating that it is unlikely that the missing data point will fall within that cluster.

In the first step, the observed data range was divided into two clusters. After the determination of the frequencies of both clusters, the observed range is divided into three clusters. This time, the cells in the data matrix will have cluster values ranging from 1 to 3. The process used to assign values to the two clusters above is repeated for the three clusters. For the missing value, the frequency of the three probable clusters will be zero to start. Then, all cluster pairs near the missing data point will be searched, and the frequencies of the clusters found at the relative location of the missing data point will be increased by one for each cluster pair match. The clustering, searching and cluster frequency determination process continues until the process has been applied for the greatest number of clusters. During this process, a cluster frequency table is generated to show the frequencies of the clusters determined at each clustering step. The highest frequency values in this table indicate the most likely clusters into which the missing data point will fall.

A dataset might have more than one missing value. The above method can be applied to each missing data point in the set and a frequency table generated for each missing cell. As the locations of the missing data points in the matrix will be different from one another, the neighbors of each missing cell will be unique; consequently, the frequency table for each missing data point will also be unique. To avoid repetition, cluster frequency table samples and details about how the estimates are calculated using the cluster frequencies are presented in the Application of the FBI Method section.

Appendix 3: The frequency based imputation software

The software developed to implement the method used in this study was written in Visual Basic in the Microsoft Visual Studio environment. The software is a console application that makes use of the interoperability feature, which enables synchronous operation of Microsoft Visual Basic and Microsoft Excel. The flowchart in Fig. 10 shows the general application procedure of the developed method and the software.

The first step in the application of the method is to read all observed values in the selected time series from the input file. The file is an Excel spreadsheet containing a two-dimensional matrix of the observed data. In this study, the columns in the data file represent months and the rows represent years. For each run, all observed data for a single station is evaluated. The method requires no preprocessing of data and uses all observed values from a station to generate the frequency tables for each observation; estimations are then made for the entire series. No observations are ignored and no smoothing occurs.

The software generates four output files containing the frequency tables, the estimations and their correlations with removed observations and statistical measures comparing the observed and estimated series to one another. Conditional formatting is used in the output files to visualize the

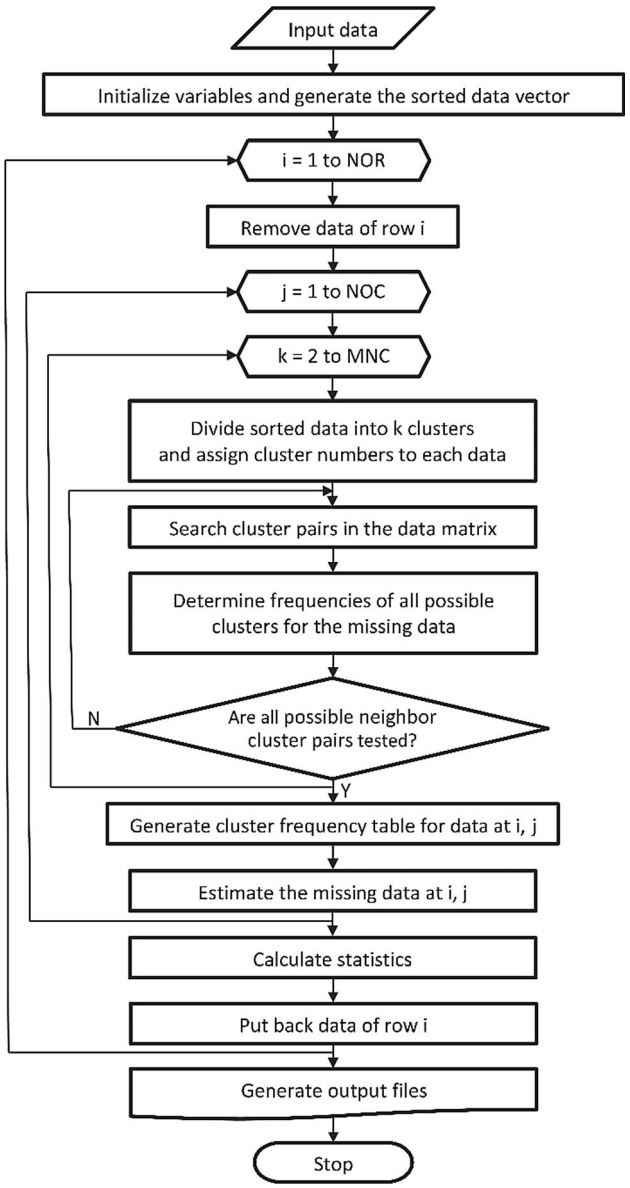

Fig. 10 Primary steps in the FBI method

differences between the values. The code is separated into distinct sections and explanations about the implementation of the method by the software are provided in the code itself.

The frequency based imputation software is distributed under the terms of the GNU General Public License version 3, and a copyright notice is provided at the beginning of the code. The software code may be downloaded using the following link: https://www.dropbox.com/s/l9eavvjiywipl19/FrequencyBasedImputation.vb?dl=0.

References

Ajaaj AA, Mishra AK, Khan AA (2016) Comparison of BIAS correction techniques for GPCC rainfall data in semi-arid climate. Stoch Environ Res Risk A 30:1659–1675. doi:10.1007/s00477-015-1155-9

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

Dikbas F (2016a) Frequency based prediction of Buyuk Menderes flows. Tek Dergi 27:7325–7343

Dikbas F (2016b) Three-dimensional imputation of missing monthly river flow data. Sci Iran 23:45–53

Do CB, Batzoglou S (2008) What is the expectation maximization algorithm?. Nat Biotech 26:897–899. http://www.nature.com/nbt/journal/v26/n8/suppinfo/nbt1406_S1.html

Elshorbagy A, Corzo G, Srinivasulu S, Solomatine DP (2010a) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 1: concepts and methodology. Hydrol Earth Syst Sci 14:1931–1941. doi:10.5194/hess-14-1931-2010

Elshorbagy A, Corzo G, Srinivasulu S, Solomatine DP (2010b) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 2: application. Hydrol Earth Syst Sci 14:1943–1961. doi:10.5194/hess-14-1943-2010

Hou AY et al (2014) The global precipitation measurement mission. Bull Am Meteorol Soc 95:701–722. doi:10.1175/BAMS-D-13-00164.1

Jayawardena AW, Lai F (1994) Analysis and prediction of chaos in rainfall and stream flow time series. J Hydrol 153:23–52. doi:10.1016/0022-1694(94)90185-6

Leconte J, Forget F, Charnay B, Wordsworth R, Pottier A (2013) Increased insolation threshold for runaway greenhouse processes on earth-like planets. Nature 504:268–271. doi:10.1038/nature12827

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124. doi:10.1016/S1364-8152(99)00007-9

Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ Model Softw 25:891–909. doi:10.1016/j.envsoft.2010.02.003

Popp M, Schmidt H, Marotzke J (2016) Transition to a Moist Greenhouse with CO_2 and solar forcing. Nat Commun. doi:10.1038/ncomms10627

Reager JT, Famiglietti JS (2009) Global terrestrial water storage capacity and flood potential using GRACE. Geophys Res Lett. doi:10.1029/2009GL040826

Remesan R, Mathew J (2015) Hydrological data driven modelling: a case study approach. Springer, Switzerland. doi:10.1007/978-3-319-09235-5

Sikorska AE, Montanari A, Koutsoyiannis D (2015) Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. J Hydrol Eng. doi:10.1061/(ASCE)HE.1943-5584.0000926

Sivakumar B (2000) Chaos theory in hydrology: important issues and interpretations. J Hydrol 227:1–20. doi:10.1016/S0022-1694(99)00186-9

Sivakumar B, Liong SY, Liaw CY, Phoon KK (1999) Singapore rainfall behavior: chaotic? J Hydrol Eng 4:38–48. doi:10.1061/(ASCE)1084-0699(1999)4:1(38)

Solomatine DP (2006) Data-driven modeling and computational intelligence methods in hydrology. Encyclopedia of hydrological sciences. Wiley, Hoboken. doi:10.1002/0470848944.hsa021

Solomatine D, See LM, Abrahart RJ (2008) Data-driven modelling: concepts, approaches and experiences. In: Abrahart R, See L, Solomatine D (eds) Practical hydroinformatics, vol 68, Water

science and technology library. Springer, Berlin, p 17. doi:10.1007/978-3-540-79881-1_2

Wang XL, Lin A (2015) An algorithm for integrating satellite precipitation estimates with in situ precipitation data on a pentad time scale. J Geophys Res Atmos 120:3728–3744. doi:10.1002/2014JD022788

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82. doi:10.3354/cr030079

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 112:143–167. doi:10.1007/s00704-012-0723-x

Zhang Q, Xu C-Y, Tao H, Jiang T, Chen YD (2010) Climate changes and their impacts on water resources in the arid regions: a case study of the Tarim River basin, China. Stoch Environ Res Risk A 24:349–358. doi:10.1007/s00477-009-0324-0