Digest 2016, December 2016, 1955-1972

Frequency Based Prediction of Büyük Menderes Flows†

Fatih DİKBAŞ[1]

ABSTRACT

In this study, a new method for the data driven prediction of interrelated and chaotic time series data showing seasonal fluctuations is proposed. The method produces predictions based on the temporal and quantitative relationships among the available data related with the frequencies of the value ranges of observed data. The method, which is called frequency based prediction, has a general approach and requires no testing/validation/adjustment/ weight determination steps. The developed method is used for predicting 9050 monthly total flow observations of 34 stations on Büyük Menderes River and for infilling 1210 missing data. High correlations obtained between the observations and predictions for all stations show that the proposed method is successful in the prediction of streamflow data.

Keywords: *Frequency based prediction, data-driven modeling, monthly total streamflow data, Büyük Menderes Basin, estimation of missing data.*

1. INTRODUCTION

Accurate, reliable and complete observations are required for the modeling and estimation of the components of the hydrologic cycle. Determination of the spatial and temporal quantitative variations of these data plays an important role for hydrological analysis and design of water resources systems. River flows constitute an important process of the hydrologic cycle and a vast amount of methods exist for the scientific evaluation of river flows. Monthly total river flows are frequently used in hydrologic studies and there are many random factors influencing the amount of flow rates. Though the river flows generally show seasonality, the high variability of the numerous influencing factors causes a chaotic and a relatively random behavior. This behavior makes the modelling and prediction of flow rates challenging.

In recent decades, with the developments in software technologies, traditional hydraulic and hydrologic models have been supported/complement by data-driven methods [1] (Solomatine 2008). A data-driven model involves the analysis of time series data but it should not be regarded as a computational exercise ignoring physical processes. Determination of the spatial and temporal interrelationships among time series-data like river flow rates is mathematically equivalent to the determination of the relationships between the functions generating flow rates. In fact, an observed flow rate is a result of a function that combines all of the parameters generating the flow rate. In this view, working

1 Pamukkale University, Denizli, Turkey - f_dikbas@pau.edu.tr

† Published in Teknik Dergi Vol. 27, No. 1 January 2016, pp: 7325-7343

directly on river flow time-series data becomes a study that does not ignore any of the contributing parameters of the river flow rate function (even though the relations and the variations of the parameters are not evaluated).

The power of basic data driven modelling techniques has already been proven and the researchers are working for making data-driven models more robust, understandable and really useful for managers [1]. Samples from the huge amount of data-driven modeling studies in literature on hydrological processes may be listed as follows:

Artificial Neural Networks (ANN) is a widely used method and was implemented in research subjects like modelling rainfall–runoff processes [2-4], river forecasting [5-6], estimation of suspended sediment concentration [7], modelling of evapotranspiration [8] and developing rainfall intensity-duration-frequency curves [9].

Fuzzy Rule Based Systems were used in areas like drought assessment [10], prediction of precipitation events [11], modelling of hydrological extremes [12], modelling rainfall-discharge dynamics [13] and flood forecasting [14].

Support Vector Machines has gained popularity among researchers in recent years and rainfall-runoff modelling [15], precipitation forecasting [16] and stream flow forecasting [17-19] are some of the application areas.

Instance-based learning [20]; runoff estimation by machine learning methods [21] and flood forecasting using ANN, Neuro-Fuzzy, and Neuro-GA Models [22] are among other remarkable data driven modelling studies. An experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology was presented by Elshorbagy et al. [23].

In most of the existing studies, the time series data is regarded as a one dimensional vector. Generally hydrological time series have an annual cycle of seasonality and a two-dimensional matrix containing a full cycle in each row represents the temporal behavior of the variable in a more comprehensible way than a one-dimensional vector. For example, river flows generally show fluctuations through a year but they are reluctant to be out of the observed range in the same month of different years. A conditionally formatted two-dimensional matrix perfectly illustrates this two-directional behavior. In this study, the river flow observations are used so that the months are in columns and the years are in rows of the matrices.

The proposed frequency based prediction method has a methodology developed for estimating and forecasting interrelated and chaotic time series data showing seasonal fluctuations. The estimations are deduced from the temporal and quantitative relationships among the available data by determining the frequencies of the value ranges of observed data. The method produces estimations on which observation range is possible at what probability for any missing value. This approach enables making multiple estimations for a single missing value by making use of the determined highest frequencies of observed ranges. The method has a general approach and requires no learning / testing / validation / adjustment / weight coefficient determination / smoothing steps contrary to many existing data driven methods. The estimations for the missing values are determined by using the available data in one step. The only value to be determined in advance is the maximum number of clusters and this value should be determined according to the quantitative structure of the available data.

The proposed method is used for the estimation of 9050 monthly total flow observations and 1210 missing values of 34 flow rate observation stations located on Büyük Menderes River. The stations are chosen so that the different regions of the river are well represented, this enabled testing the success of the method in the estimation of data from stations showing variations in data length and values. Figure 1 shows the locations of the chosen stations. No evaluation could be made on the branch flowing from Uşak region towards the station 065 located at the downstream exit of Adıgüzel Dam Lake as there is no observation station on the branch. Weak side branches and creeks are not shown on the figure.

The observation lengths of the evaluated stations vary between 8 years (station 07-114) and 41 years (stations 07-003, 004 and 010). The data of seven stations are complete and the number of missing values in the remaining stations vary between 12 and 122. The mean number of missing values is 36. In most of the stations, the minimum flow rate observation is zero and in some stations there is no zero observation in the investigated period. The mean monthly total flow observations for all stations is 18.9 hm^3. The station with the lowest mean in the whole observation period is the station 07-097 with an average value of 0.3 hm^3, and station with the highest mean is the station 07-062 with an average of 219.3 hm^3. The maximum monthly total flow rates observed in the monitored periods vary between 2.8 hm^3 (07-097) and 1121 hm^3 (07-062). The elevations of the stations are within the range 17 m (07-062) and 1145 m (07-111). The characteristics of the stations are presented together with the statistical comparisons of estimations and observations in Table 5.

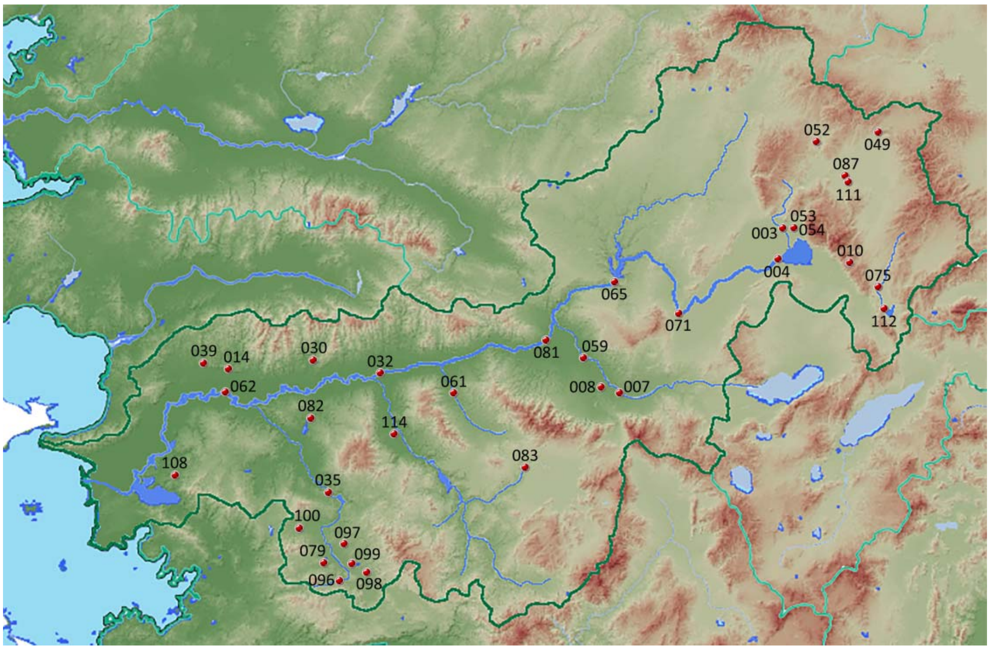


Figure 1. The flow rate observation stations chosen in Büyük Menderes Basin

2. FREQUENCY BASED PREDICTION OF FLOW RATE DATA

A pair of data in a set of data formatted as a matrix represents the temporal and quantitative behavior of the observed variable at the smallest scale. With statistical reasoning, valuable information can be extracted from the relationships within the data set and estimations on the missing values can be made. The main idea behind the proposed method is that all the adjacent pairs in the observed data set contain information about the temporal and quantitative variation of flow rates and possible value ranges of the neighboring observations might be estimated by using the extracted information.

The basic concept of the method is that any value in a data series showing periodic behavior has strong relationships with closer observations and weak quantitative associations with distant observations. The proposed frequency based prediction method produces estimations based on the recurrences of the quantitative relationships among neighboring cells covering a 7 x 7 sized area (Figure 2) around any data cell in the matrix. The neighborhood region does not have to be 7 x 7 in size but was sufficient for obtaining succesfull estimations for all of the 34 data sets investigated in this study. Use of a wider neighborhood region might unnecessarily increase the required computation time.

For example, the monthly total river flow series investigated in this study show seasonal fluctuations and lower values are experienced in summer months while higher values are observed in winter months. For this reason, when a data in January is being estimated, the frequencies of observed data pairs in a 7 month (October-April) and 7 year range in which the required month is at the center are used instead of the summer observations. Similarly, when a missing value in August is being estimated, the 7 month period between May-November is used. In this way, for the whole data series, estimations based on the frequencies of the data pairs with highest quantitative and temporal associations can be obtained while the adverse impacts that might be caused by the weakly or even inversely associated data pairs are prevented.

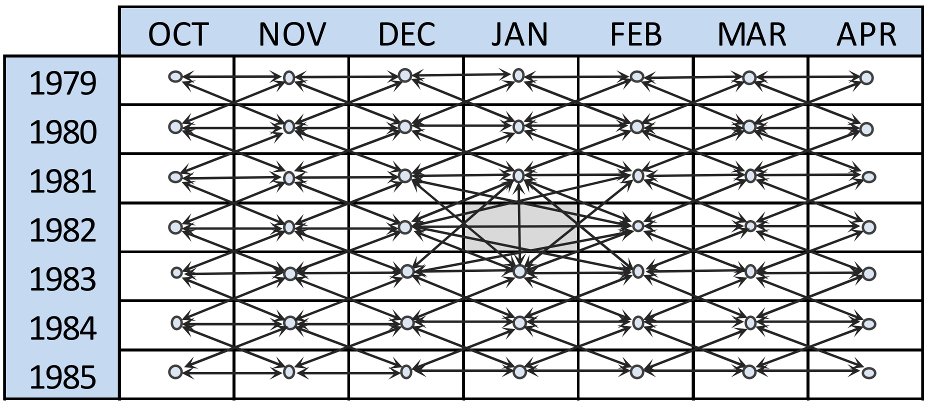
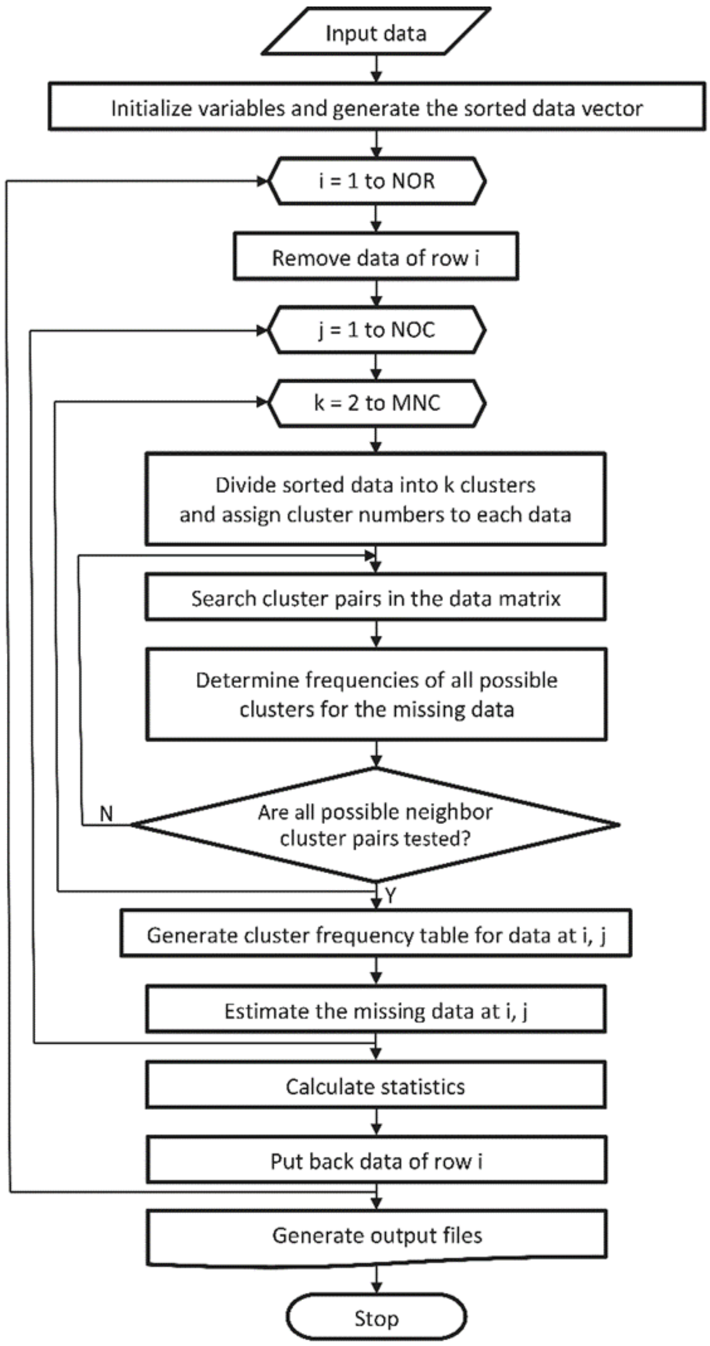


Figure 2. The data pairs to be searched in the data matrix for the purpose of determining the cluster frequencies.



NOTE: NOR: Number of rows, NOC: Number of Columns, MNC: Maximum number of clusters

Figure 3. The flowchart of the frequency based prediction method.

The flowchart in Figure 3 shows the general application procedure of the method. First, the observed values are read from the input file and a three-dimensional vector containing the sorted data and their coordinates on the data matrix is generated by sorting the data in ascending order. The coordinate information is crucial because the observation time of any value is very important in the investigation of the temporal and quantitative investigation of a time series data. Sorting and making statistical investigations on a variable without considering the observation times of each individual variable means ignoring information on the temporal relationships between observations. In the presented method, sorting is made to determine the range clusters of all observations.

A range cluster is obtained by dividing the observations sorted from the lowest to the highest into sets with as equal a number of elements as possible. The observed range in each station is divided into 2 to 12 range clusters and the method is applied for each cluster setup. The maximum number of clusters was set to 12 for the flow rate data used in this study and the obtained results seem to be sufficient, but a different number of clusters might be required in the investigation of other variables. The number of clusters should be chosen according to the behavior of the time-series and the amount of observations. As the method generates estimations by removing observed values from the series, the optimum number of clusters may easily be determined by trying different numbers of clusters.

Two different approaches may be used in the generation of the clusters and the determination of cluster indexes showing the cluster to which the observations are assigned. In the first approach, each cluster has as equal a number of elements as possible and the clusters have varying ranges. Equation 1 is used to assign observed values to clusters.

In the second approach, range values are equalized and the clusters have a varying number of elements. The bounds of the cluster ranges are the lowest and highest observations belonging to that range. Equation 2 is used to assign observed values to clusters.

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

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

In the above equations:

n_d: The total number of observations in the sorted data vector.

i: The index number of the observation (changes between 1 and n_d)

Cl_i: The cluster index to be assigned to the i^{th} observation (This value changes between 1 and 12 for the observations used in this study).

int(): The function converting a decimal number into an integer

n_{Cl}: The number of clusters used to divide the sorted data vector (This value is 12 for the observations used in this study).

X_i: The i^{th} observation in the sorted data series

X_{min}; X_{max}: The minimum and maximum observations in the whole data series.

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. But, in this situation, it must not be forgotten that the value ranges of the clusters with extreme values will increase and the higher values might be underestimated.

2.1. The Application of the Proposed Method

The primary aim of the method proposed in this study is the estimation of missing values in time-series data. As each hydrologic time series has a different set of values, the estimation success of a method varies from one station to another. The estimation success even varies for varios portions of a time series. This situation requires that the method proven to be giving good estimates for a station should be tested with existing observations prior to producing estimates for the missing values in another station. For testing the estimation success of the proposed method, each row in each data matrix is removed and estimated one by one by using the relationships among remaining data. This process is automatically implemented by the developed software and the estimation success of the method is statistically evaluated by comparing the obtained estimations and observations. In this process, the missing values in the dataset are also estimated together with the deliberately removed observations.

For a comprehensive explanation of the method, the estimation steps of the application of the method on the monthly total flow rate observations of station 07-010 Dinar-Irgıllı (Turkey) for the 1982 water year will be presented. The data set covers 466 observations between the years 1960-2000. The monthly mean flow in the observed period is 7.7 m^3/s and the average of monthly total flows is 5.92 hm^3. The highest value was observed in May 1970 (18.2 hm^3) and the lowest value was observed in June, July and August 1995 (0 hm^3). All observations in 1974 and 1975 and October and November observations of 1960 are missing. When the long year averages of monthly total flow rates are calculated, the highest value was obtained for March (8.97 hm^3) and the lowest value was obtained for August (1.73 hm^3).

When the data series is divided into clusters, a cluster index is assigned to each data. When the data series is divided into 12 clusters, the cluster index values assigned to the neighbors of October 1982 (the cell to be estimated, shown with blue border in Tables 1.a and 1.b) are shown in Table 1.b. No cluster index is assigned to the missing values and to the data removed for testing the estimation performance of the method.

It is possible to generate questions as follows within the neighborhood of the missing data and it is also possible to find the associations related with these questions in other regions of the matrix:

What might the total flow in October 1982 be when the total flow in September 1980 is 3.33 hm^3 and the total flow in October 1981 is 8.88 hm^3 (the red rectangle in Table 1.a)?

What might the total flow in October 1982 be when the total flow in August 1982 is 0.75 hm^3 and the total flow in September 1982 is 1.75 hm^3 (the yellow rectangle in Table 1.a)?

These two questions might be expressed as follows by using the cluster indexes:

What might the cluster value of the missing cell in October 1982 be when the cluster value in October 1980 is 6 and the cluster value in October 1981 is 10?

What might the cluster value of the missing cell in October 1982 be when the cluster value in August 1982 is 4 and the cluster value in September 1982 is also 4?

Table 1.a) The neighbors of the missing cell, b) The cluster indexes of the neighbors, c) The cluster indexes of the first matching region

	JUL	AUG	SEP	OCT	NOV	DEC	JAN
1979	0.20	0.20	0.96	3.29	4.13	7.43	10.20
1980	0.24	0.24	0.23	3.33	7.62	9.50	9.13
1981	1.74	1.95	4.94	8.88	10.70	10.90	13.30
1982	0.60	0.75	1.75				
1983				7.05	8.21	8.34	8.74
1984	0.25	0.83	0.73	2.94	7.96	10.00	9.17
1985	2.49	2.89	5.28	6.87	10.00	10.30	10.90

a

	JUL	AUG	SEP	OCT	NOV	DEC	JAN
1979	3	3	4	6	6	8	10
1980	3	3	3	6	9	10	10
1981	4	4	7	10	11	11	12
1982	4	4	4				
1983				8	9	9	10
1984	3	4	4	5	9	10	10
1985	5	5	7	8	10	11	11

b

	APR	MAY	JUN	JUL	AUG	SEP	OCT
1961	9	7	7	6	5	5	6
1962	9	8	7	6	6	6	7
1963	10	12	12	10	9	10	11
1964	10	9	8	4	3	5	6
1965	11	11	11	9	9	10	12
1966	12	12	11	10	9	10	12
1967	12	12	9	7	6	10	11

c

The two questions above investigate the value of the concerned missing data by assessing the quantitative relationships between the observed values within the neighborhood region of the data to be estimated. 158 similar unique questions may be asked about the searched value by using the horizontal, vertical and diagonal data pairs in the neighborhood of the missing data shown with a blue border. The adjacent data pairs in the neighborhood of October 1982 are shown in Table 1.a. As each cell has a different location on the data matrix, the 7 x 7 sized neighborhood region for each cell is also special to each cell.

The answer for these two sample questions and the answers to the questions generated by using the remaining data pairs in the neighborhood of the missing data are searched by finding matches within the data matrix. For example, to find the answer to the first question, vertically adjacent values of 6 and 10 are searched through the whole matrix. The first match is found for the clusters of July 1962 and 1963 located in the data region shown in Table 1.c. It is also seen in the same table that the cluster indexes of September 1962 and 1963 are 6 and 10 respectively. The cluster values just below these two mathed pairs are 4 and 5 respectively. The search is resumed through the whole dataset and the frequencies of the clusters just below the horizontally adjacent 6 and 10 are determined. The search is repeated for all 158 diagonally and horizontally adjacent cluster pairs located within the neighborhood region of the questioned missing value and the frequencies of the clusters in the relative location of the October 1982 data are increased by one in each match. When the search for all cluster pairs is completed, the total frequencies of the 12 clusters will be found and the frequency values shown in the 12th column of the frequency table given for the month October in Table 2 are obtained. The cluster with the highest frequency is regarded as the cluster with the highest probability and the cluster with the lowest frequency is regarded as the cluster with the lowest probability.

The process for 12 clusters is repeated for all the procedures of generating 2 to 11 clusters and the cluster frequency table for October 1982 is obtained. For this purpose, the sorted data series is divided into two clusters so that one cluster includes the lower values and the other cluster includes the higher values. Then, each data point is assigned a cluster index: 1 for the data in the first cluster (the lower values) and 2 for the data in the second cluster (the higher values). After the assignment of the cluster indexes, the cluster frequencies are determined and estimations are made. When the process for two clusters is completed, the sorted data series is divided into three clusters so that each cluster has as equal number of observations as possible and again each data point is assigned a cluster index ranging from 1 to 3. Then new frequencies are calculated and estimations are made. These processes are repeated by increasing the cluster number by one at each step and the process ends after the estimations are made for the highest number of clusters. At the end of the clustering and frequency determination processes, the cluster frequency tables constituting the base of the missing data estimations will have been generated (Table 2).

The dark green regions in the frequency tables show the data ranges with higher probability and the regions with red color show the data ranges with lower probability. The columns with the titles "Min" and "Max" on the right of Table 2 show the value ranges for each cluster when the number of clusters is 12. For example, the value range for the 4th cluster, which has the highest frequency among the 12 clusters, is 2.0-3.3 hm^3.

Table 2 shows the cluster frequency tables generated for the months of 1982 water year for the station 07-010. The title numbers of the columns show the cluster numbers to which the data set is divided and the title numbers of the rows show the cluster indexes. For example, in the table generated for the month October, when the data series is divided into two clusters the frequency of the first cluster was 7491 and the frequency of the second cluster was 12345. These values indicate that the monthly total flow rate value of October 1982 might most probably be within the value range of the second cluster. When the data set is divided into 12 clusters, the highest frequency (112) was obtained for the 4th cluster. While the frequency tables obtained for the months showing a high rate of variability like October and November are blurry, the red and green paths follow a more apparent path in the frequency tables obtained for the months like March, April and August for which the variability is lower.

Table 2. The cluster frequency tables generated for the 1982 data of the station 07-010

OCT	2	3	4	5	6	7	8	9	10	11	12	NOV	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	7491	1601	525	234	161	78	27	22	0	0	0	1	6078	1085	283	148	78	48	14	15	0	0	0	0.0	0.0
2	12345	4106	1673	741	459	207	151	124	78	63	38	2	16304	4236	1567	541	260	131	96	72	38	35	11	0.0	0.1
3		4552	2030	1074	679	464	268	173	69	61	46	3		7219	2332	943	621	267	162	93	32	43	21	0.1	0.5
4			1963	1004	691	453	317	231	198	188	112	4			3296	1323	711	381	349	210	161	141	69	0.5	2.0
5				1042	521	395	314	167	167	127	87	5				1622	842	470	340	170	138	130	117	2.0	3.3
6					541	339	228	307	164	81	101	6					927	503	292	305	147	95	77	3.3	4.9
7						237	244	105	135	134	72	7						468	353	188	186	167	70	4.9	6.3
8							187	180	95	83	106	8							314	270	157	98	110	6.3	7.4
9								145	112	89	56	9								258	168	127	116	7.4	8.6
10									68	118	76	10									145	142	98	8.6	10.2
11										66	82	11										140	117	10.2	12.3
12											35	12											107	12.3	18.2

DEC	2	3	4	5	6	7	8	9	10	11	12	JAN	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	5756	773	244	87	54	23	13	2	0	0	0	1	5660	749	201	57	26	6	1	0	0	0	0	0.0	0.0
2	20701	4217	1354	450	138	77	42	41	15	18	4	2	23838	4555	1262	472	139	76	63	32	24	1	4	0.0	0.1
3		9706	2660	928	579	237	126	70	25	23	15	3		11845	2909	1076	531	209	130	56	27	24	16	0.1	0.5
4			5105	1564	799	355	286	141	139	127	39	4			7002	1809	836	324	257	153	134	110	36	0.5	2.0
5				2577	1131	463	335	191	125	95	91	5				3274	1276	439	338	253	139	103	118	2.0	3.3
6					1533	694	372	252	161	122	100	6					2202	782	381	293	203	178	99	3.3	4.9
7						810	457	300	219	143	73	7						1253	624	342	212	102	102	4.9	6.3
8							544	366	185	129	136	8							883	431	230	156	103	6.3	7.4
9								376	282	135	133	9								643	374	213	153	7.4	8.6
10									246	223	162	10									312	355	187	8.6	10.1
11										192	167	11										290	279	10.2	12.3
12											173	12											252	12.3	18.2

FEB	2	3	4	5	6	7	8	9	10	11	12	MAR	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	5579	802	226	59	13	7	6	0	0	0	0	1	5547	957	264	90	31	5	7	0	0	0	0	0.0	0.0
2	24891	4517	1111	444	146	65	51	11	9	4	2	2	22971	3905	957	518	233	75	44	25	11	4	4	0.0	0.1
3		12965	2842	987	401	194	115	52	36	26	17	3		11721	2541	858	369	113	98	96	43	18	26	0.1	0.5
4			7090	1874	780	322	193	137	94	74	36	4			7064	1723	729	325	151	94	56	48	42	0.5	2.0
5				3877	1393	362	335	234	133	91	67	5				3978	1366	319	260	170	114	59	53	2.1	3.3
6					2765	863	367	263	196	224	102	6					2881	915	342	211	168	180	75	3.3	4.9
7						1660	587	398	225	85	135	7						1664	605	302	169	103	115	4.9	6.3
8							1114	531	234	169	87	8							1178	565	258	134	98	6.3	7.4
9								800	501	274	171	9								716	512	257	116	7.4	8.6
10									377	453	225	10									419	450	233	8.6	10.2
11										338	344	11										334	373	10.2	12.3
12											264	12											251	12.3	18.2

APR	2	3	4	5	6	7	8	9	10	11	12	MAY	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	6530	1275	442	179	92	27	7	1	0	0	0	1	8551	1980	770	336	188	78	38	18	0	0	0	0.0	0.0
2	20588	3512	1064	623	360	126	105	81	27	19	22	2	15795	3416	1389	907	653	300	259	172	82	77	66	0.0	0.1
3		10195	2250	780	364	201	141	120	72	59	48	3		7205	1922	819	465	358	254	135	92	149	89	0.1	0.5
4			6377	1490	653	353	136	79	74	97	39	4			4400	1098	610	390	148	153	125	123	58	0.5	2.1
5				3612	1191	282	285	147	60	43	55	5				2617	903	215	321	196	69	67	103	2.1	3.3
6					2729	832	289	180	136	178	66	6					2018	647	236	151	168	172	68	3.3	4.9
7						1438	620	261	155	83	137	7						1156	436	211	110	84	130	5.1	6.3
8							1149	570	248	134	69	8							904	423	197	76	43	6.4	7.4
9								646	458	303	99	9								507	380	228	80	7.5	8.6
10									397	417	232	10									256	330	164	8.7	10.3
11										305	332	11										197	267	10.3	12.2
12											223	12											183	12.3	18.2

JUN	2	3	4	5	6	7	8	9	10	11	12	JUL	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	11197	2978	1288	571	291	133	81	34	0	0	0	1	12988	4131	1653	822	464	209	133	54	0	0	0	0.0	0.0
2	11025	3099	1786	1339	1013	629	514	290	155	174	103	2	7405	2158	1845	1369	1441	1009	738	448	310	319	172	0.0	0.1
3		4578	1506	780	574	456	352	255	189	255	199	3		2180	1090	704	520	406	389	417	222	305	274	0.1	0.5
4			2842	773	525	374	190	200	166	161	82	4			1118	446	375	312	189	148	129	186	84	0.5	2.1
5				1756	585	220	235	246	79	78	124	5				687	349	176	169	157	65	63	109	2.1	3.3
6					1289	435	221	149	85	105	52	6					475	221	163	80	86	74	43	3.4	5.1
7						788	331	158	102	63	91	7						352	127	107	42	34	71	5.1	6.4
8							659	293	145	56	39	8							269	111	66	52	15	6.4	7.5
9								303	283	150	73	9								129	101	59	66	7.5	8.7
10									216	214	101	10									78	90	26	8.7	10.3
11										142	200	11										59	73	10.3	12.2
12											122	12											54	12.3	18.2

AUG	2	3	4	5	6	7	8	9	10	11	12	SEP	2	3	4	5	6	7	8	9	10	11	12	Min	Max
1	13633	4712	1660	905	471	225	151	51	0	0	0	1	14059	4061	1359	729	405	183	107	19	0	0	0	0.0	0.0
2	5891	2333	2311	1552	1671	1184	679	519	359	289	152	2	9620	3465	2816	1451	1430	997	460	403	280	210	132	0.0	0.1
3		1502	920	812	615	429	432	438	248	289	256	3		2826	1581	1124	861	570	493	457	258	289	166	0.1	0.5
4			610	408	357	315	341	140	151	147	170	4			1182	674	545	449	511	230	189	147	231	0.5	2.1
5				253	230	157	126	151	104	113	84	5				544	352	370	258	192	217	188	120	2.1	3.3
6					121	131	91	70	76	112	75	6					259	229	172	171	92	155	102	3.4	5.1
7						97	99	73	52	56	45	7						183	137	116	103	146	92	5.1	6.4
8							90	40	37	31	31	8							179	63	88	69	87	6.4	7.5
9								39	32	26	31	9								79	68	48	68	7.5	8.7
10									19	25	24	10									42	47	38	8.7	10.3
11										16	26	11										29	48	10.3	12.2
12											7	12											18	12.3	18.2

2.2. Calculation of the Missing Values According to the Frequency Tables

Various approaches may be used for estimating the missing values in the dataset by making use of the cluster frequency tables. For example, instead of making a definite estimation, the value ranges of the clusters with the highest frequency might be evaluated as estimation ranges. Furthermore, the location where the direction of the regions with distinct green tones in the frequency tables cut the value range table might be used as the estimation value of the missing data. In this study, the method followed in the estimation of the missing values and the values deliberately removed from the data set is as follows: The sums of the observations generating the frequency of each cluster in the frequency table generation process are calculated. When the process of frequency table generation ends, the sum of the observations generating the frequency values will be determined. The highest 5 frequencies among the frequencies obtained for the 12 clusters are determined and sorted in descending order. As seen in Table 2, for the data of October 1982, the 5 clusters with highest frequencies are 4th, 8th, 6th, 5th and 11th clusters respectively. The most probable 5 estimations for October 1982 are calculated by dividing the total observation values obtained for each cluster to the cluster frequency values. The real observed value in October 1982 is 4.42 hm^3 and this value is within the value range of the 6th cluster and the estimated value for this cluster is 3.77 hm^3.

The estimation process is repeated for the remaining months of the year and the first five most probable estimation values are obtained for each month. Table 3.a shows the first 5 most probable values obtained for each month of 1982 together with the real observations and the values closest to the observed real values among the 5 estimations. The correlations between the estimations and observations are provided on the right of the table. The nearest estimations to the observations are indicated with bold font. Among the 12 estimated values, 6 of the nearest estimations are obtained in the first estimation series, 2 of them are obtained in the second estimations and the remaining 4 nearest values are obtained in the third estimations. The correlation between the first estimations and the observations is 0.859 while the correlations between the best estimations among the first three estimations and the observed values is 0.980. As it is clearly seen, the first three estimations were sufficient for obtaining the best estimations for the data of the evaluated year and the 4th and 5th estimations did not contribute to the improvement of correlation between the estimations and observations. It must again be noted that the observed values of the year 1982 were removed from the data set prior to the implementation of the method and these values are not known in any step of the estimation process. The proposed method generates multiple estimations for both the missing values and the values deliberately removed from the data set.

For testing the estimation success of the proposed method, 5 different estimation series for the 1982 observations of the station 07-010 are generated by using multiple linear regression. In the calculations, the observed values in 1982 are removed from the data set, as was done in the proposed method, and estimations are obtained by using the remaining observations. The observed values and the correlations between the observed and estimated values are presented in Table 3.b together with the best estimation values. It is observed that the estimations of the proposed method are generally closer to the observations when compared to the estimations obtained by multiple linear regression.

Table 3.a. The observations of the station 07-010 in 1982; the first 5 most probable estimations determined by the proposed frequency based prediction method; correlations between the estimations and the observations.

Month	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	
Observation	**4.42**	**7.98**	**10.30**	**10.60**	**9.55**	**11.70**	**11.30**	**10.40**	**5.65**	**0.56**	**0.93**	**1.73**	**Corr.**
Estimation 1	1.15	2.70	14.47	**11.13**	11.13	**11.16**	**11.21**	**11.22**	11.31	**0.26**	0.25	**0.86**	0.859
Estimation 2	6.66	11.36	**11.12**	14.36	14.32	14.34	9.32	14.37	0.28	0.04	**0.97**	0.25	0.898
Estimation 3	**3.77**	**7.95**	9.31	9.27	**9.22**	9.39	14.42	9.40	**2.72**	2.68	0.04	0.05	0.926
Estimation 4	2.63	6.80	6.77	7.98	8.02	8.08	5.56	5.66	14.40	1.27	2.76	2.84	0.522
Estimation 5	11.19	14.53	7.90	2.61	5.59	5.66	8.15	2.62	0.03	11.31	4.17	4.03	-0.106
Nearest Est.	**3.77**	**7.95**	**11.12**	**11.13**	**9.22**	**11.16**	**11.21**	**11.22**	**2.72**	**0.26**	**0.97**	**0.86**	**0.980**

Table 3.b. The observations of the station 07-010 in 1982; 5 estimations determined by using multiple linear regression; correlations between the estimations and the observations.

Month	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	
Observation	**4.42**	**7.98**	**10.30**	**10.60**	**9.55**	**11.70**	**11.30**	**10.40**	**5.65**	**0.56**	**0.93**	**1.73**	**Corr.**
Estimation 1	1.07	4.13	17.50	6.13	2.32	**8.74**	**9.90**	14.70	**2.68**	0.21	0.31	5.28	0.681
Estimation 2	0.50	16.30	**8.40**	**7.11**	13.60	15.10	6.91	**11.90**	0.24	10.20	0.31	0.79	0.581
Estimation 3	2.97	**10.70**	7.49	5.20	**7.23**	7.61	12.90	6.19	0.28	0.06	0.00	**2.53**	0.798
Estimation 4	0.04	11.60	3.10	15.00	2.44	8.01	17.70	6.19	14.20	**0.42**	**0.57**	0.04	0.603
Estimation 5	**3.21**	2.48	17.50	3.27	6.29	15.10	6.91	6.25	0.24	9.82	0.18	10.20	0.294
Nearest Est.	**3.21**	**10.70**	**8.40**	**7.11**	**7.23**	**8.74**	**9.90**	**11.90**	**2.68**	**0.42**	**0.57**	**2.53**	**0.892**

The number of estimations to be made might be decreased or increased according to the variability of the evaluated data set. As it is well known that the river flow series show a relatively chaotic behavior and the most probable flow rate value might not become the observed flow rate. For this reason, having multiple estimations at hand for a missing value will be very helpful for the researchers, practitioners and the administrators. Generation of 5 estimations for the flow rate series of the stations evaluated in this study was sufficient for obtaining successful estimations.

The increase of correalations between the observations and the estimations are evaluated according to the increasing estimation number for testing the advantage of calculating more than one estimation for a missing data. Table 4 shows the correlations between the observed series of the station 07-010 and the nearest estimations within the first 2, 3, 4 and 5 estimations for each year and for the whole series.

Annual correlations over 0.7 occurred between the observed values and the nearest estimates in the first two estimations in 77% of cases (30/39). This rate increased to 90% (35/39) for the first three and four estimates and to 97% (38/39) for the first five estimates. The rate of correlations over 0.9 for the first 2, 3, 4 and 5 estimations are 38% (15/39), 56% (22/39), 77% (30/39) and 85% (33/39) respectively.

The last column of Table 4 shows the correlations between the whole series consisting of 466 observations and the estimations nearest to the observations in the first 2, 3, 4 and 5 estimations. The correlation of 0.84 obtained for the first two estimations might be regarded to be sufficient in practice but the correlations became higher than 0.9 when the number of estimations were increased and the correlation for the whole series became 0.97 when 5 estimations were calculated. This correlation value might be considered as a high and reliable value for the estimation of monthly total flow series.

These results show that increasing number of estimations provide a higher reliability and precision but it must not be forgotten that in some cases even when the estimations come closer to the observations, the correlation value might decrease. For example, in Table 4, it is seen that the correlation values for the year 1966 decrease with the increase of the number of estimations. This situation is caused by the function used in the calculation of correlation but the situation in which the correlation value decreases when the estimations become closer to the observations is rarely experienced and generally better estimations produce better correlations. As the purpose in the modelling of hydrologic variables is usually obtaining estimations close to the observations, the correlation coefficient alone is not sufficient for making statistical evaluations. For this reason, looking at more than one statistical parameters enable making better assessments.

Table 4. The correlations between the observations of the station 07-010 and the best estimations within the first five estimation series

Estimations	**1960**	**1961**	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**	**1968**	**1969**	**1970**	**1971**
1-2	0.46	0.70	0.22	0.64	0.63	0.22	0.93	0.58	0.74	-0.35	0.83	0.95
1-3	0.36	0.72	0.41	0.94	0.85	0.72	0.89	0.83	0.69	0.04	0.86	0.85
1-4	0.34	0.61	0.55	0.97	0.90	0.87	0.88	0.82	0.93	0.68	0.86	0.88
1-5	0.89	0.89	0.70	0.98	0.97	0.97	0.88	0.94	0.92	0.68	0.88	0.93

Estimations	**1972**	**1973**	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**
1-2	0.95	0.90			0.88	0.62	0.91	0.83	0.94	0.81	0.91	0.84
1-3	0.96	0.92			0.96	0.89	0.89	0.96	0.97	0.93	0.98	0.87
1-4	0.97	0.94			0.98	0.91	0.97	0.97	0.97	0.95	0.98	0.96
1-5	0.97	0.91			0.99	0.99	0.99	0.97	0.97	0.98	0.98	0.99

Estimations	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**
1-2	0.94	0.76	0.79	0.78	0.73	0.79	0.59	0.97	0.72	0.95	0.95	0.99
1-3	0.96	0.95	0.93	0.92	0.84	0.91	0.85	0.98	0.78	0.96	0.96	0.99
1-4	0.98	0.97	0.95	0.91	0.94	0.94	0.92	1.00	0.93	0.97	0.98	0.99
1-5	0.99	0.98	0.97	0.92	0.98	0.95	0.93	1.00	0.93	0.97	0.98	0.99

Estimations	**1996**	**1997**	**1998**	**1999**	**2000**	**Whole**
1-2	0.87	0.87	0.90	0.90	0.92	**0.84**
1-3	0.94	0.93	0.95	0.90	0.93	**0.91**
1-4	0.96	0.98	0.97	0.94	0.97	**0.94**
1-5	1.00	0.97	0.97	0.98	0.97	**0.97**

The graphs in Figure 4 compare the estimations of the frequency based prediction method with the observations of the station 07-010. Even though the flow rates have shown significant variations within the observation period, a good fit between the observations and estimations was obtained. The estimation performance for the extreme values observed in 1966, 1969 and 1970 were relatively low. This situation is caused by the approach implemented by the method. As the method tries to estimate the most probable value for a missing data, the low probabilities of the extreme values which are observed only a few times through the observation period cause the estimations to remain low. The capability of the method in estimation of the extreme values might be improved by considering the temporal and spatial variations of hydrologic variables like precipitation which are directly associated with flow rates.

Figure 4. The comparison of the observations of the station 07-010 with the estimations of frequency based prediction method.

3. APPLICATION OF THE FREQUENCY BASED PREDICTION METHOD ON THE REMAINING 33 STATIONS

All the above considerations were about the estimations of the observations of a single station (07-016). One might propose that the success of a method in the estimation of the values of a single station is not sufficient to claim that it will be successful in the estimations of other stations. To test this, the proposed method was used to estimate the observations of the remaining 33 stations located in the Büyük Menderes Basin. As can be seen in Figure 1, the stations are selected so that various flow properties in various regions of the river are well represented.

Table 5. Statistical comparison of the observed and estimated flow rate series and some characteristics of the stations

STATION	r	Na-Su.	NRMSE	MASE	Obs.Yr.	n	Miss.	Min.	Mean	Max	Elev.
07-003	0.93	0.83	0.06	0.35	41	480	12	0.0	7.1	95.5	837
07-004	0.93	0.86	0.06	0.36	41	456	36	0.0	33.3	162.0	814
07-007	0.99	0.98	0.05	0.20	16	177	15	0.0	8.0	23.6	260
07-008	0.97	0.93	0.05	0.40	22	264	0	0.0	5.5	12.9	300
07-010	0.97	0.93	0.06	0.48	41	466	26	0.0	5.9	18.2	841
07-014	0.90	0.79	0.06	0.35	39	419	49	0.0	5.7	65.3	70
07-030	0.91	0.82	0.08	0.37	39	408	60	0.0	2.7	20.6	177
07-032	0.97	0.94	0.05	0.37	38	396	60	4.6	112.1	447.0	68
07-035	0.94	0.87	0.06	0.32	36	402	30	0.1	24.1	216.0	112
07-039	0.85	0.68	0.05	0.35	36	381	51	0.0	1.8	36.9	73
07-049	0.93	0.86	0.07	0.38	29	240	108	0.0	2.4	17.7	1025
07-052	0.91	0.81	0.07	0.39	33	381	15	0.0	1.2	11.6	980
07-053	0.94	0.89	0.07	0.61	37	322	122	0.7	9.9	29.8	829
07-054	0.99	0.97	0.05	0.25	30	288	72	0.0	1.6	6.9	829
07-059	0.98	0.96	0.05	0.27	33	321	75	0.0	22.6	61.2	160
07-061	0.88	0.71	0.08	0.53	31	252	120	0.1	5.4	31.3	197
07-062	0.95	0.89	0.06	0.33	33	355	41	4.7	219.3	1121.0	17
07-065	0.93	0.86	0.08	0.39	31	324	48	0.0	64.5	206.0	307
07-071	0.96	0.90	0.08	0.33	31	372	0	1.0	27.5	86.4	758
07-075	0.94	0.88	0.07	0.36	24	239	49	0.0	0.7	4.9	1010
07-079	0.91	0.77	0.08	0.43	21	171	81	0.0	2.0	15.0	355
07-081	0.89	0.78	0.08	0.58	20	216	24	10.0	64.3	249.0	150
07-082	0.89	0.75	0.10	0.45	18	202	14	0.0	2.5	21.7	111
07-083	0.75	0.47	0.10	0.94	18	216	0	0.2	2.8	29.8	855
07-087	0.95	0.89	0.08	0.30	15	180	0	0.0	1.0	5.5	1067
07-096	0.89	0.77	0.08	0.35	13	143	13	0.0	0.7	6.2	450
07-097	0.95	0.85	0.08	0.29	11	120	12	0.0	0.3	2.8	425
07-098	0.78	0.58	0.09	0.38	13	156	0	0.0	0.4	5.7	500
07-099	0.88	0.75	0.09	0.40	13	139	17	0.0	0.9	7.2	395
07-100	0.91	0.79	0.07	0.32	13	156	0	0.0	0.8	8.8	325
07-108	0.90	0.77	0.11	0.39	11	120	12	0.0	0.8	5.2	160
07-111	0.85	0.70	0.07	0.34	10	120	0	0.0	1.0	11.4	1145
07-112	0.82	0.64	0.15	0.93	10	84	36	0.5	4.1	8.1	1005
07-114	0.84	0.66	0.09	0.38	8	84	12	0.0	1.5	16.0	140
Min:	0.75	0.47	0.05	0.20	8	84	0	0.0	0.3	2.8	17
Mean:	0.91	0.81	0.08	0.41	25.1	266.2	36	0.65	18.9	90.2	492.6
Max:	0.99	0.98	0.15	0.94	41	480	122	10.0	219.3	1121.0	1145

The method was used in the estimation of 9050 observed and 1210 missing monthly total flow values. The observation series are located in matrices as explained in Section 2 and estimations were obtained after each row was removed from the matrices in turn. Table 5 presents the statistical evaluations of the obtained results together with some characteristics of the stations. For testing the success of the method in the estimation of observed values, correlation coefficient, Nash-Sutcliffe efficiency coefficient, normalized root mean squared error (NRMSE) and mean absolute scaled error (MASE) statistics which are frequently used in the statistical evaluation of hydrologic variables are calculated for all stations. The observation periods of of the stations as years, the total number of existing monthly observations (n), the number of missing values, the minimum, mean and maximum monthly total flow values and the elevations of the stations are presented together with the statistical evaluations.

While the correlation value exceeds 0.75 for all stations, it exceeds 0.85 for 88% and 0.9 for 68% of the stations. The highest correlation value was 0.99 and was obtained for the stations 07-007 and 07-054. Likewise, the correlation value exceeded 0.9 for all stations with observation periods longer than 21 years. This situation shows that the increase in observation length has a positive impact on the estimation performance. The Nash-Sutcliffe efficiency coefficients vary between 0.47 (station 07-083) and 0.98 (station 07-007) and 86% exceeds 0.7, 65% exceeds 80% and 21% exceeds 0.90. The NMRSE values vary between 0.05 (stations 07-007, 008, 032, 039, 054 and 059) and 0.15 (station 07-112) and 88% of them are under 0.1. The MASE values range between 0.2 (station 07-007) and 0.94 (07-083) and 71% of them are under 0.4.

The statistical evaluations show that the estimation performance of the method is generally at a very good level. These results which were obtained by comparing the observed series and the estimation series provide sufficient proof for the success of the method in the estimation of monthly total flow series of the evaluated 34 stations. As was mentioned above, any method should be implemented on the available data prior to claiming that the method will be successful in the estimation of the considered series. Still, the observed results are so promising that the method might be successful in the estimation of other flow series. As the developed method has a general approach, it has a potential of being applied on other hydrologic variables or various time series data from other scientific disciplines like biostatistics, economics and social sciences.

4. DISCUSSION AND CONCLUSIONS

This study presents a data driven methodology named frequency passed prediction and the method was used for the estimation of 9050 monthly total flow rate observations and imputation of 1210 missing values from 34 stations on Büyük Menderes Basin. The observations are removed from the data sets annually in groups of 12 and estimated by using the remaining observations of the evaluated station. Estimation of missing data by using the observation series is the main aim of the developed method. The statistical comparisons of the estimations and observations show that the method successfully generates estimations for the deliberately removed observations of all of the 34 stations. Through the implementation of the method, the missing values in the data set were also estimated. The advantages of the proposed method may be summarized as follows:

- The method has a general approach and can be applied on any two dimensional data in one step without making any calibration, smoothing or weight determination.
- A pre-determined number of multiple estimations are determined for all missing values. The obtained estimations are the most probable estimations according to the proposed approach and the comparisons with the observed series show that the calculated estimations are successful for all series of from all stations. This feature is especially useful in evaluating variables with chaotic behavior like streamflow.
- The obtained results for the 34 flow rate observation stations show that the method can be used reliably in the estimation of monthly total flow rate records.

Acknowledgement

The author would like to thank The General Directorate of the State Hydraulic Works of Turkey (DSİ) for providing the data used in this study and the editorial board and the reviewers for their valuable contributions and comments, which greatly improved the manuscript.

References

[1] Solomatine, D. P., Abrahart, R. J., See, L. M., Data-Driven Modelling: Concepts, Approaches and Experiences. Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, R. J. Abrahart, L. M. See and D. P. Solomatine, (editörler), Springer, Berlin, 17-30, 2008.

[2] Dawson, C. W., Wilby, R., An Artificial Neural Network Approach to Rainfall-Runoff Modelling. Hydrol. Sci. J., 43(1), 47-66, 1998.

[3] Govindaraju, R. S., Ramachandra, R.A., Artificial Neural Networks in Hydrology, Kluwer, Dordrecht, 2001.

[4] Tayfur, G., Singh, V. P., ANN and Fuzzy Logic Models for Simulating Event-Based Rainfall-Runoff. J. Hydraul. Eng., 132(12), 1321-1330, 2006.

[5] Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., Wilby, R. L., Two Decades of Anarchy? Emerging Themes and Outstanding Challenges for Neural Network River Forecasting. Prog. Phys. Geogr., 36(4), 480-513, 2012.

[6] Huo, Z., Feng, S., Kang, S., Huang, G., Wang, F., Guo, P., Integrated Neural Networks for Monthly River Flow Estimation in Arid Inland Basin of Northwest China. J. Hydrol., 420-421, 159-170, 2012.

[7] Bhattacharya, B., Van Kessel, T., Solomatine, D. P., Spatio-Temporal Prediction of Suspended Sediment Concentration in the Coastal Zone Using Artificial Neural Network and a Numerical Model. J. Hydroinform., 14(3), 574-594, 2012.

[8] Adamala, S., Raghuwanshi, N. S., Mishra, A., Tiwari, M. K. Evapotranspiration Modeling Using Second-Order Neural Networks. J. Hydrol. Eng., 19(6), 1131-1140, 2014.

[9] Mirhosseini, G., Srivastava, P., Fang, X., Developing Rainfall Intensity-Duration-Frequency Curves for Alabama Under Future Climate Scenarios Using Artificial Neural Networks. J. Hydrol. Eng., 10.1061/(ASCE)HE.1943-5584.0000962 , 04014022, 2013.

[10] Pesti, G., Shrestha, B. P., Duckstein, P., Bogárdi, I. A., Fuzzy Rule-Based Approach to Drought Assessment. Water Resour. Res., 32(6), 1741-1747, 1996.

[11] Abebe, A. J., Solomatine, D. P., Venneker, R., Application of Adaptive Fuzzy Rule-Based Models for Reconstruction of Missing Precipitation Events. Hydrol. Sci. J., 45 (3), 425-436, 2000.

[12] Bogardi, I., Duckstein, L., Pongracz, R., Galambosi, A., Experience With Fuzzy-Rule-Based Modeling of Hydrological Extremes. Proc., Risk-Based Decisionmaking in Water Resources IX, ASCE, Reston, VA, 44-60, 2001.

[13] Vernieuwe, H., Georgieva, O., De Baets, B., Pauwels, V. R. N., Verhoest, N. E. C., De Troch, F. P., Comparison of Data-Driven Takagi–Sugeno Models of Rainfall–Discharge Dynamics. J. Hydrol., 302(1-4), 173-186, 2005.

[14] Nayak, P. C., Explaining Internal Behavior in a Fuzzy If-Then Rule-Based Flood-Forecasting Model. J. Hydrol. Eng., 15(1), 20-28, 2010.

[15] Bray, M., Han, D., Identification of Support Vector Machines for Runoff Modelling. J. Hydroinf., 6, 265-280, 2004.

[16] Chen, X., Zhu, S., Improved Hybrid Model Based on Support Vector Regression Machine for Monthly Precipitation Forecasting. J. Computers, 8(1), 232-239, 2013.

[17] Lin, J. Y., Cheng, C. T., Chau, K. W., Using Support Vector Machines for Long-Term Discharge Prediction. Hydrol. Sci. J., 51(4), 599-612, 2006.

[18] Kişi, Ö., Çimen, M., A Wavelet-Support Vector Machine Conjunction Model for Monthly Streamflow Forecasting. J. Hydrol., 399(1-2), 132-140, 2011.

[19] Karahan, H., İplikçi, S., Gürarslan, G., River Flow Estimation From Upstream Flow Records Using Support Vector Machines, J. Appl. Math., 2014

[20] Solomatine, D. P., Shrestha, D. L., Maskey, M., Instance Based Learning Compared to Other Data-Driven Methods in Hydrological Forecasting. Hydrol. Process., 22, 275-287, 2007.

[21] Yılmaz, A. G., Muttil, N., Runoff Estimation by Machine Learning Methods and Application to the Euphrates Basin in Turkey. J. Hydrol. Eng., 19(5), 1015-1025, 2014.

[22] Mukerji, A., Chatterjee,C., Raghuwanshi, N. S., Flood Forecasting Using ANN, Neuro-Fuzzy, and Neuro-GA Models. J. Hydrol. Eng., 14(6), 647-652, 2009.

[23] Elshorbagy, A., Corzo. G., Srinivasulu, S., Solomatine, D. P., 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, 2010.