Evaluation of Statistical Methods for Estimating Missing Daily Streamflow Data

Mustafa Utku YILMAZ[1]
Bihrat ÖNÖZ[2]

ABSTRACT

In this study, it was aimed to investigate the applicability of various statistical estimation methods for the Porsuk River basin, which has sparse streamflow observations. Estimations were performed using regression analysis (REG), the single donor station based drainage area ratio (DAR), the multiple donor stations based drainage area ratio (MDAR), standardization with mean (SM), standardization with mean and standard deviation (SMS), inverse distance weighted (IDW) methods. Two separate studies were conducted for both partially missing data and completely missing data. In order to estimate streamflow statistics for use in SM and SMS methods, logarithmic regression equations were suggested. The promising results obtained from ensemble approaches will provide a significant hydrological contribution to streamflow estimations.

Keywords: Missing data, Porsuk River basin, regression, streamflow estimation.

1. INTRODUCTION

Nowadays, estimation of missing streamflow data for gauged basins (partially missing data) or estimation of streamflow for ungauged basins (completely missing data) is a popular subject that has been researched extensively in hydrological studies because researchers need reliable data to make accurate analyzes [1,2]. In many basin-based studies, there are several reasons for missing data in historical records of meteorological data (precipitation, temperature, evaporation, snow etc.) and hydrological data (streamflow) provided by different institutions. The missing data due to various reasons such as station failures, transportation, and climatic difficulties, human-induced effects etc., creates significant problems in terms of effective water resources planning and management [3]. Especially in developing countries or regions such as Turkey, recorded data of natural events are limited. With these limited values, it is too difficult to reliably determine the hydrological behavior

Note:

- This paper has been received on May 04, 2018 and accepted for publication by the Editorial Board on December 24, 2018.
- Discussions on this paper will be accepted by January 31, 2020.

- https://dx.doi.org/10.18400/tekderg.421091

[1] Kirklareli University, Department of Civil Engineering, Kirklareli, Turkey - utkuyilmaz@klu.edu.tr
https://orcid.org/0000-0002-5662-9479

[2] Istanbul Technical University, Department of Civil Engineering, Istanbul, Turkey - onoz@itu.edu.tr
https://orcid.org/0000-0002-4531-2476

of a particular basin. There are significant problems of missing data in most river basins of Turkey. This case is one of the biggest challenges faced by scientists working on streamflow estimation in Turkey [4].

A station for which streamflow data is missing is called the target station and a station used to estimate the missing data is called the donor station. Selection of donor stations that are most likely to be hydrologically similar is the most critical step for a good streamflow estimation because their data will be used to estimate the missing values. The donor station is usually chosen as the nearest station for estimating missing values at the target stations. In other words, the distance is the main selection criterion for donor streamflow station selection. Choosing the nearest streamflow station as the donor streamflow station is preferred in widespread practice, but acceptance of the distance as the primary donor streamflow station selection criterion may not always be correct. Although the distance is currently used as the selection criterion for selecting a donor station, it may be possible to obtain better results by using the most correlated station as the donor station for estimation of target stations [4]. On the other hand, the use of multiple donor stations can provide improved streamflow estimation instead of using a single donor station [1,5].

Many statistical modeling methods require complete and uninterrupted data. In case of missing data, statistical analyses do not provide confidence. Various methods are used to estimate the missing data (partially or completely) using data from the donor station in hydrology sciences. The most common methods for estimating missing data are single and multiple regression analysis [6], interpolation such as kriging and inverse distance weighted (IDW) [7], time series analysis [8, 9], artificial neural networks [10,11]. Flow duration curves, which show the percentage of time that a specific streamflow is equaled or exceeded during a given period, can also be used for estimating missing data [5,12]. Moreover, algorithms such as expectation maximization and nearest neighbors in statistical software packages can be used to estimate missing data [13]. In addition, if there is missing data at the gauging station (partially missing data), the missing data can be replaced easily using the mean imputation method (the mean of the recorded data of the target station). Obviously, this method may not provide good representation [14]. Without depending on only one method, estimating missing data with as many different methods as possible and using the best method compared with others are of great importance.

Drainage area ratio (DAR) method is the most commonly preferred method to estimate streamflow for target stations where streamflow data are not available using data from a single donor station [15,16,17,18]. The multiple donor stations based DAR (MDAR) method was developed by Shu and Ouarda [1]. MDAR produces the streamflow estimations for a target station as the weighted average of the estimations from more donor stations. Their estimation performance was improved by a weighted combination of donor stations [1]. IDW interpolation method, which is a variant of the DAR method, is a widely used method for the estimation of missing data in hydrology [7,19]. In the DAR method, streamflow values are transferred from a single donor station to the target station. However, the IDW method is used for direct streamflow transfer to the target station from multiple donor stations. The estimation performance was improved significantly by expanding the DAR method to consider multiple donor stations [7]. Standardization by the mean streamflow (SM) [20] and maintenance of variance extension method introduced by Hirsch [15], which Farmer and Vogel [20] termed standardization with mean and standard deviation (SMS) are common in hydrology. These methods are used for estimation of streamflow time series for ungauged basins or gauged basins which have missing data. Regional regression models were developed by Farmer and Vogel [20] to estimate streamflow statistics for use in SM and SMS methods as a function of drainage area and some meteorological parameters. Hirsch [15] showed the suggested method, called the SMS method, has better performance than the DAR, but Farmer and Vogel [20] revealed that both the SM and the SMS methods may not always be superior to the DAR method. Although different methods are used to estimate the missing data, the correlation between the stations should be considered [21]. Thus, a comparison is made between more stations instead of a single station to act as a target station.

In this study, various statistical estimation methods were evaluated for estimating missing daily streamflow data at the selected stations located in the Porsuk River basin. Missing data estimation was applied in two ways as partially missing data and completely missing data (ungauged). First, the missing data at each station was completed with one of the linear and nonlinear regression analysis, DAR, SM and SMS methods (adapted to daily streamflows) according to the performance approach described in the application section. Secondly, DAR, SM, SMS, MDAR and IDW methods were applied to estimate completely missing data of stations. These methods used were preferred to methods such as the artificial intelligence methods because: (1) they do not require much data and (2) they are relatively simple to apply and (3) they are robust and effective methods. There were significant statistical difficulties in both applying and evaluating the methods. One of the biggest challenges of the study was low water potential and a large amount of missing daily data. Another difficulty was that all missing data at the target station could not be completed with the best donor station and another donor station was required.

In order to reduce the uncertainties in individual methods, two of the methods used were weighted according to their relative performance as measured by Nash-Sutcliffe efficiency (NSE) and then combined to obtain the recommended ensemble estimates for each station. The all possible pairwise combinations of the methods are used to employ ensemble approaches for each target station in the study. The performance of these ensemble approaches was compared with individual methods. Statistical parameters (mean and standard deviation) used in the SM and SMS methods were calculated by suggested regression equations based on logarithmic relationships between statistical parameters and drainage area. These regression relationships were obtained for each target station using data from all other stations. A method for estimating missing data should conserve streamflow characteristics such as mean, standard deviation, skewness coefficient. For this purpose, after the missing data for each station are replaced with estimated values, statistical characteristics between long-term completed data and long-term original data were compared. Thus, it was examined whether the statistical structure of the dataset was preserved after the completion process. In addition, it was determined whether there was a difference in the statistical characteristics by comparing the significant percentiles of the flow duration curves of the observed and the estimated dataset.

The objectives of this research are (1) to investigate the use of applied methods for estimating missing daily streamflows for such a difficult basin and (2) to improve the estimation performance of the methods by identifying the most appropriate donor stations and (3) to obtain ensemble approaches that provide more efficient results compared to individual methods.

2. STUDY AREA

Porsuk River basin located in the northwestern part of the Central Anatolia Region of Turkey was selected as the study area (Figure 1). This basin lies between north latitudes of 38°44' to 39°99' and east longitudes of 29°38' to 31°59'. Porsuk River basin is a sub-basin of the Sakarya basin and has a drainage area of about 11000 km². It is one of the most economically important basins. Almost half of the land in the Porsuk River basin is devoted to agricultural activities, and forests and meadows constitute the other half. Residential and industrial areas account for only about 5% of the total basin area. The extensive and intensive agricultural activities are due to the high fertility of the soil. Most of the basin has a typical continental climate, with hot and dry summers and cold and semi-humid winters. Porsuk River with a length of 460 km is one of the longest tributaries of Sakarya River. The river originates at the Murat Mountain and flows in an easterly direction until its confluence with the Sakarya River. Sakarya River with a length of 824 km is the third longest river in Turkey and flows into the Black Sea. The long-term average annual precipitation for the whole basin is approximately 450 mm. Thus, the water potential of the basin is slightly lower than Turkey's long-term (1970-2010) average annual precipitation of 643 mm [22].

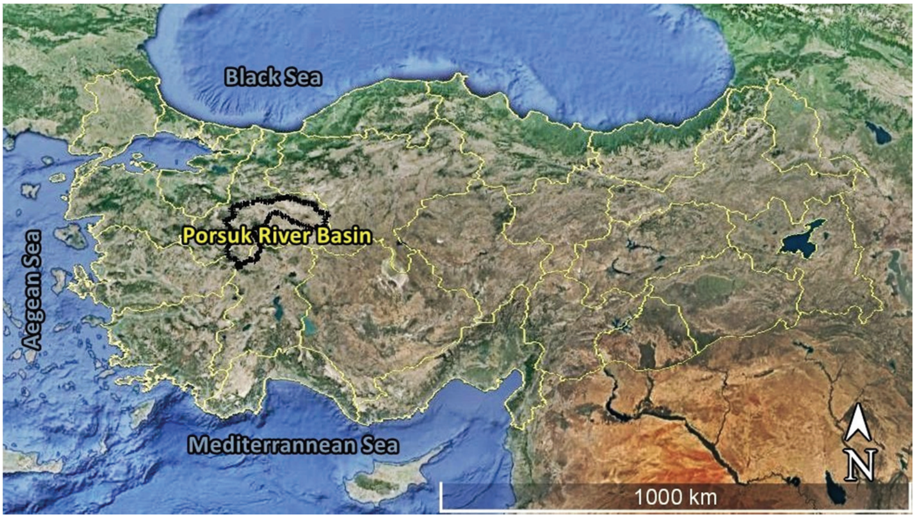

Figure 1 - Location map of Porsuk River basin in Turkey

3. DATA

Daily streamflow data used in this study were provided from the State Hydraulic Works (DSI) of Turkey. Some details of the stations and the statistical characteristics of the streamflow data considered in this study are presented in Table 1. Performance of missing data estimation methods was tested for selected streamflow gauging stations in the Porsuk River basin. Observed daily streamflow data were available for 6 stations on the same stream for the 21 years investigated period of 1990-2010. Figure 2 shows the locations of the 6 selected stations. The selected stations are not located downstream of a dam or reservoir. Data of the stations were missing for at least 2 years and at most 5 years. The years of measurement of the streamflow data for each station are given in Figure 3. Selected 6 stations have highly logarithmic linear relationships between statistical parameters (long-term mean and standard deviation) and their drainage area is as shown in Figure 4. The coefficients of determination (R^2) of each relationship are 0.9667 for mean and 0.9644 for standard deviation, respectively. The nearest and the most correlated stations for each station were identified and provided in Table 2.

Table 1 - Characteristics of the stations and statistics calculated from daily mean streamflow data

Station Number	D12A033	D12A055	D12A063	D12A093	D12A181	E12A003
Drainage Area (km^2)	2432	297	290,7	153,1	3810,5	3938,4
Elevation (m)	951	1062	826	1130	912	855
Observations (days)	7670	7670	7670	7670	7670	7670
Obs. with missing data (%)	9,5	19,0	9,5	23,8	14,3	0,0
Obs. without missing data (%)	90,5	81,0	90,5	76,2	85,7	100,0
Mean (m^3/s)	2,432	0,862	0,594	0,512	5,172	4,666
Std. deviation	4,815	1,734	1,348	1,022	6,323	4,535
Zero (%)	0,00	18,04	0,17	7,97	0,00	0,00

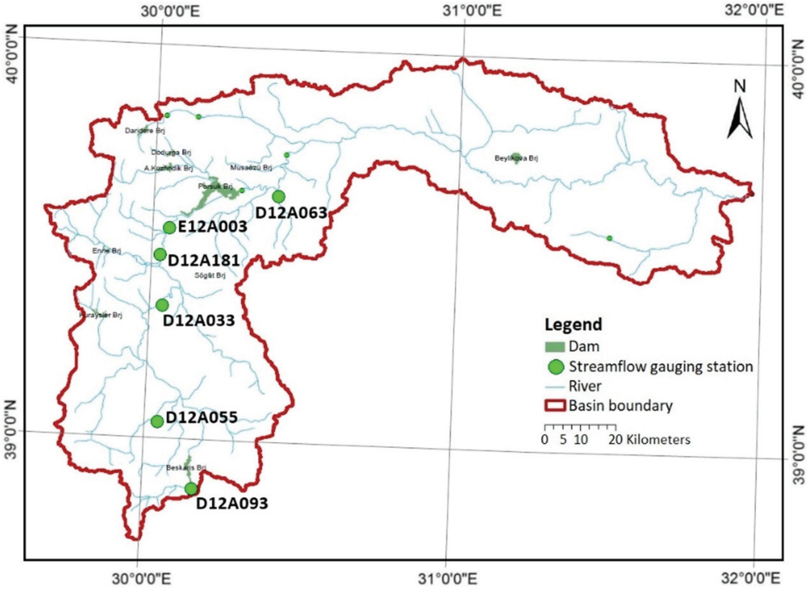

Figure 2 - Locations of the 6 selected streamflow gauging stations used for this study

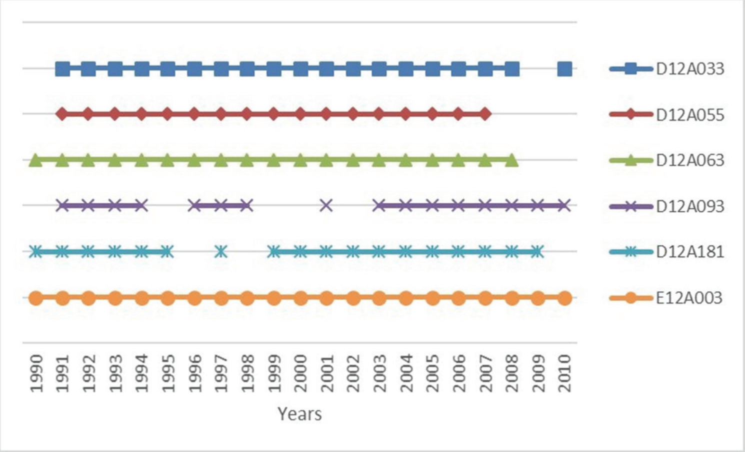

Figure 3 - The years of measurement of the streamflow data for each station

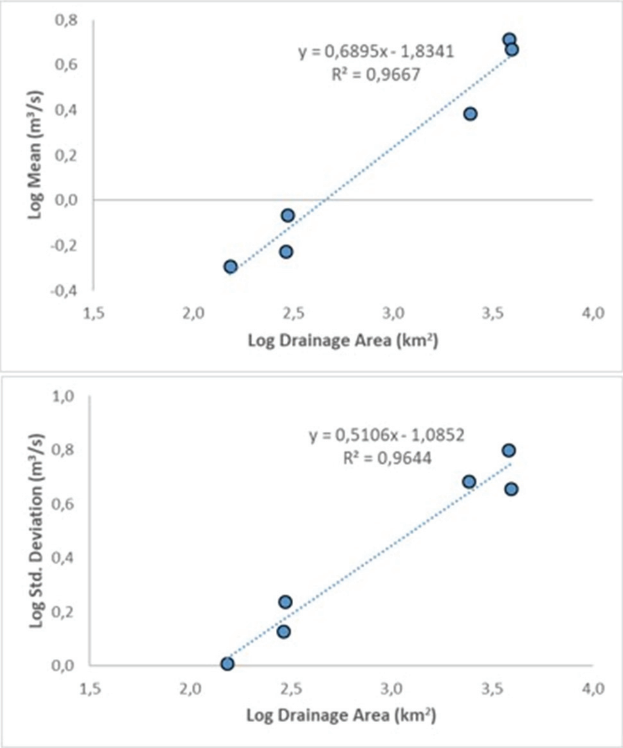

Figure 4 - Logarithmic linear relationship between mean and std. deviation of long-term daily streamflow and drainage area

Table 2 - The nearest and the most correlated stations to each station

	The Nearest		Distance (km)		The Most Correlated		Correlation Coefficent (r)	
Station Number	**1st station**	**2nd station**	**1st station**	**2nd station**	**1st station**	**2nd station**	**1st station**	**2nd station**
D12A033	D12A181	E12A003	13,20	20,58	D12A181	E12A003	0,85	0,83
D12A055	D12A093	D12A033	22,62	34,56	E12A003	D12A181	0,74	0,73
D12A063	E12A003	D12A181	32,27	37,24	D12A055	D12A093	0,62	0,56
D12A093	D12A055	D12A033	22,62	54,30	E12A003	D12A181	0,75	0,73
D12A181	E12A003	D12A033	7,60	13,20	E12A003	D12A033	0,95	0,85
E12A003	D12A181	D12A033	7,60	20,58	D12A181	D12A033	0,95	0,83

4. METHODS

The methods used to estimate the missing data discussed in this study are given below.

a) Regression analysis (REG): Regression analysis is a statistical method that is commonly used to explain relationships among variables. In this study, the linear, exponential and logarithmic regression analysis are applied. The mathematical relationship with the highest R^2 value between the two stations was taken into account to estimate daily streamflow at the target station. The daily streamflow, Q (m^3/s) between target and donor station is represented by

$$Q_{target} = \alpha Q_{donor} + \beta \tag{1}$$

$$Q_{target} = \alpha Q_{donor}^{\beta} \tag{2}$$

$$Q_{target} = \alpha ln Q_{donor} + \beta \tag{3}$$

where α and β are the coefficients of the regression equations. They are calculated using concurrent daily streamflow data at the donor and the target stations.

b) Drainage area ratio (DAR): The applicability of the DAR method is closely related to the hydrological similarity (similar drainage area, climate, and geographical conditions) between two stations. This method assumes that the streamflow per unit area of hydrologically similar basins is equal. That is, for any given days,

$$Q_{target} = \frac{A_{target}}{A_{donor}} Q_{donor} \tag{4}$$

where the ratio of drainage areas for the target and donor station, A_{target}/A_{donor}, is used to transfer streamflow at the donor station, Q_{donor}, to that at the target station, Q_{target}.

DAR method is commonly used when the stations are located on the same stream and the ratio between the drainage areas of the donor station and the target station (A_{target}/A_{donor}) is between 0.5 and 1.5 [23].

c) Standardization with mean (SM): This method takes into account the ratio of streamflow to the mean streamflow. Mathematically,

$$Q_{target} = \frac{\mu_{Q_{target}}}{\mu_{Q_{donor}}} Q_{donor} \tag{5}$$

where Q is the daily streamflow at the subscripted station and μ is the mean of the flows at the subscripted station.

In this study, the SM method applied with annual (SM1) and monthly (SM12) variations. SM1 standardizes the daily time series with the annual mean streamflow, while SM12 standardizes the daily time series with monthly mean streamflow.

d) Standardization with mean and standard deviation (SMS): This method is based on assumption that the standardized streamflows at both a target and a donor station are approximately equal, and can be expressed as:

$$Q_{target} = \mu_{Q_{target}} + \sigma_{Q_{target}} \left(\frac{Q_{donor} - \mu_{Q_{donor}}}{\sigma_{Q_{donor}}} \right) \tag{6}$$

where μ and σ are the mean and standard deviation of the streamflows at the subscripted station.

In this study, the SMS method applied with annual (SMS1) and monthly (SMS12) variations. The distinction between these variations of the SMS method is similar to the distinction between SM1 and SM12.

Two standardization methods (SM and SMS) described above require streamflow statistics (mean and standard deviation) for the target station. Linear regression equations to estimate streamflow statistics (mean and standard deviation) at the target station for use in SM and SMS methods are suggested in this study. Mean and standard deviation of streamflow data should be highly related to drainage area on account of its statistic stationarity. Considering this assumption, suggested equations were derived from logarithmic relationships between statistical parameters and drainage area. After some algebraic manipulations, these base 10 logarithmic equations can be shown as:

$$\theta = 10^{\beta} A^{\alpha} \tag{7}$$

where θ is mean or standard deviation of the streamflows and A is the drainage area, α and β are the slope of the regression line (regression coefficient) and intercept value of regression line (constant), respectively. A single regression equation is required for the annual methods (SM1 and SMS1) whereas a regression equation for each month is required for the monthly methods (SM12 and SMS12).

e) Multiple donor stations based DAR (MDAR): This method produces the streamflow estimations at a target station as the weighted average of the estimations from n donor stations which are calculated as:

$$Q_{target} = \frac{\sum_{i=1}^{n} w_i Q_{d_i}}{\sum_{i=1}^{n} w_i} \tag{8}$$

$$w_i = \frac{\frac{1}{d_i}}{\sum_{i=1}^{n}\frac{1}{d_i}} \tag{9}$$

$$d_i = \sqrt{\left(A_{target} - A_{donor}\right)^2} \tag{10}$$

where Q_{d_i} is the estimation from the donor station i, and w_i is the weight assigned to the donor station i, and d_i is the similarity distance measure between the target station and donor station i, and A is the drainage area at the subscripted station.

f) Inverse distance weighted (IDW): This method is mathematically expressed as:

$$q_{target} = \sum_{i=1}^{n} w_i q_{donor_i} \tag{11}$$

$$w_i = \frac{\frac{1}{{d_i}^p}}{\sum_{i=1}^{n}\frac{1}{{d_i}^p}} \tag{12}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{13}$$

where q is the area normalized streamflow value ($m^3/s/km^2$) at the subscripted station and n is the total number of donor stations considered for the interpolation. The distance between the target and donor station d is calculated individually for each of the n donor stations. d is the distance between two stations, p is power parameter equal to 2 in this study and w is the interpolation weights. The sum of the weights assigned to each donor station is equal to 1. Distance, d, between two stations was calculated using a variation of the Haversine Formula [24]:

$$d = arccos\left(sin(lat_1)\ x\ sin(lat_2) + cos(lat_1)\ x\ cos(lat_2)\ x\ cos(lon_2 - lon_1)\right)\ x\ r \tag{14}$$

where lat_1 and lon_1 represent latitude and longitude in radians of the station 1, respectively and lat_2 and lon_2 represent the latitude and longitude in radians of the station 2, respectively. The radius of the earth, r, is approximately 6378.1 km.

4.1. Proposed Method (Ensemble Streamflow Estimation)

In the study, the ensemble approaches combining two methods were proposed in order to get more consistent estimation and reduce uncertainties in the individual methods used. These approaches were applied by considering the weighted average of any two individual methods. All possible double combinations of individual method 1 and 2 (i.e. DAR-SM, DAR-SMS, DAR-IDW, MDAR-IDW, IDW-SM, IDW-SMS etc.) were considered. The mathematical expression of estimated streamflow, $\hat{Q}$, using the weighted ensemble approach (adapted from Farmer and Vogel) [20] is given in Equation (15).

$$\hat{Q}_{ensemble} = w\hat{Q}_{method_1} + (1 - w)\hat{Q}_{method_2} \tag{15}$$

where $\hat{Q}$ is the daily streamflow estimated from subscripted method and w is a weight (bounded by 0 and 1) which is based on the relative efficiency of any two methods. The weight was estimated by using the Langmuir equation as the ratio of NSE of any two methods as given in Equation 16 and Equation 17.

$$w = \frac{\varphi}{1+\varphi} \tag{16}$$

$$\varphi = \frac{\left(NSE_{method_2}-1\right)^2}{\left(NSE_{method_1}-1\right)^2} \tag{17}$$

4.2. Construction of Flow Duration Curves

A flow duration curve (FDC) was constructed from the daily streamflow for each station. The FDC was separated into three segments which represent different hydrological conditions: high-flow ($Q_{0.1}$, $Q_{0.5}$, Q_2, Q_5, Q_{10}), middle-flow (Q_{20}, Q_{30}, Q_{40}, Q_{50}, Q_{60}), low-flow (Q_{70}, Q_{80}, Q_{90}, Q_{95}, Q_{99}). Each percentile (streamflow exceedence probabilities) represents a different segment of the FDC.

4.3. Evaluation Criteria

Nash-Sutcliffe efficiency (NSE) [25] and root mean square error (RMSE) were used to evaluate the performance of each method. Their formulations are given as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{est}\right)^2}{\sum_{i=1}^{n}\left(X_i^{obs}-\overline{X^{obs}}\right)^2} \tag{18}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{est}\right)^2}{n}} \tag{19}$$

where X_i^{obs} is the i-th observed value, X_i^{est} is the i-th estimated value and $\overline{X^{obs}}$ is the mean of all the observed data for a time series of length n. The ideal values NSE and RMSE are 1 and 0, respectively.

In this study, the methods were tested using a jack-knife approach on the original data. In jack-knife (leave-one-out) approach, each station is in turn considered in common period as ungauged (completely missing) for obtaining a streamflow estimation in that station.

5. APPLICATION OF THE METHODS

5.1. Completion of Missing Data (Partially Missing Data)

For this application, data from the most appropriate donor station was needed to complete the missing data of each target station. Each station whose data are available was selected as a donor station of the target station for REG, DAR, SM and SMS methods. Because the estimated missing data were not compared with the actual data, the estimation performance of methods was evaluated on values before and after the missing data, where observed values are available. Although this evaluation approach does not provide a real estimation performance, it provides a strong emphasis in favour of the estimated values. The most appropriate donor stations of the target station were identified, giving the best NSE and RMSE results according to each method. Then, the missing data at each station was completed with the estimated values obtained from the most appropriate donor stations. In this way, the complete 21-year daily streamflow data were obtained for each station. The linear (Equation 1), exponential (Equation 2) and logarithmic (Equation 3) regression analysis (REG) were applied to express mathematically the relationships between stations. Coefficients of determination, R^2, were calculated to assess the relationship between stations. The mathematical relationship with the highest R^2 value between the two stations was used in the study. As an example, the relationships obtained for one station was given in Figure 5. Prior to logarithmic regression analysis, streamflow data that were zero were replaced with one percent of the minimum observed daily flow from the dataset of each station in order to avoid problems related to log-transforming values of zero. If streamflow data from the most appropriate donor station were also missing, streamflow data from the next most appropriate donor station were used to estimate missing streamflow data at the target station. When the intercept value (constant) was negative in the calculated regression equations, these equations generated some negative streamflow values. These negative values were replaced with zero streamflow.

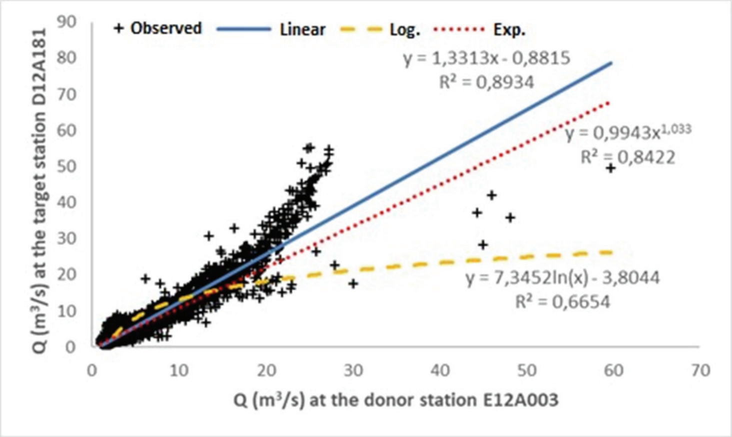

Figure 5 - Investigation of relationships between stations and selection of mathematical equations

5.2. Estimation of Streamflow for Ungauged (Completely Missing Data)

Two separate study periods for ungauged application were conducted, one for simultaneous common years of all stations and the other for 21 years (completed data) of all stations. The daily streamflow data were estimated for assumed completely missing data (ungauged). First, simultaneous common years of all stations were determined in a 11 year period (1991, 1992, 1993, 1994, 1997, 2001, 2003, 2004, 2005, 2006, 2007). Each of the stations was considered ungauged in turn. In other words, the data record of one of the stations in this study area was kept out from the database and considered as an ungauged station. Each of the other stations in turn was selected as the donor station of the ungauged station for the DAR, SM and SMS methods. All of the other stations (all potential donors) were selected as the donor stations of the ungauged station for the MDAR and IDW methods. Then, the daily streamflow data for the ungauged station were estimated using the data from the donor stations. This process was repeated for all target stations considered. The actual daily streamflow data were used to evaluate the estimation performance of methods. Secondly, the same procedure was repeated using 21 years of data for each station. Ensemble estimations were generated using the best possible results of individual methods for 21 year period. In order to find the combination of two methods that gave the best streamflow estimates of the target stations, the all possible ensemble approaches combining two methods were tried.

In this study, it was necessary to estimate the mean and standard deviation of each target station for use in SM and SMS methods. Therefore, each station in turn was considered as a target station, and logarithmic linear relationships between statistical parameters (mean and standard deviation) and drainage areas were obtained for each target station using data from all other stations. For example, the logarithmic linear regression equations obtained for station D12A063 have 0.972 R^2 value for mean, and 0.9622 R^2 value for standard deviation, respectively (Figure 6a). For station D12A181, R^2 values of each relationship are 0.9654 for mean and 0.9618 for standard deviation, respectively (Figure 6b). The obtained logarithmic equations were converted into algebraic equations (Equation 7) as explained in the SM and SMS methods. The drainage area of the target station was replaced by the A parameter in these equations, and then the long-term mean and standard deviation of the target station were calculated. The equations of the SMS method generated negative values when the difference between streamflow value and mean was larger than the estimated mean value. When the equations generated negative values, the negative values were replaced with zero streamflow.

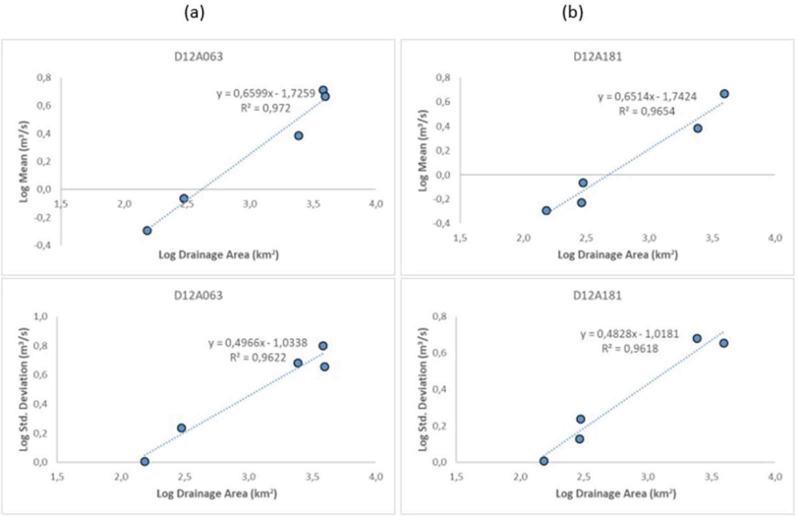

Figure 6 - Logarithmic linear relationships obtained (a) for station D12A063 and (b) for station D12A181

5.3. Estimation of the Flow Duration Curves

The flow duration curves (FDC) of the estimated values giving the best NSE results for each individual method were estimated. The selected 15 percentiles for different hydrological conditions of 11 year and 21 year FDCs of observed and estimated daily streamflow were compared and will be presented in section 6.

6. RESULTS AND DISCUSSION

The results of partially missing (completion of missing data) and completely missing (assumed ungauged) applications were described below, respectively.

In order to determine the missing data completion method, Figure 7 shows the range of the best NSE and RMSE values of each station for each individual methods. Results showed that the REG method for each station gave better NSE and RMSE values than the other methods, according to the evaluation approach of the performance of the methods to complete the missing data. The mean of the NSE values was 0.66 and the mean of the RMSE values was 1.67.

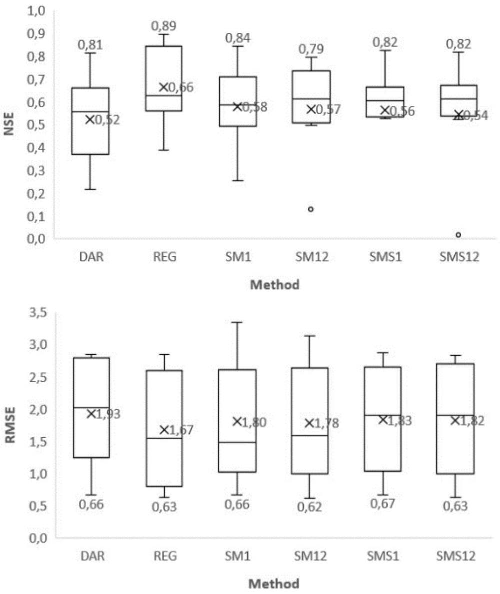

Figure 7 - Performance evaluation of the methods for missing data completion

Also, the REG method gave satisfactory results in terms of estimating missing data in preserving streamflow characteristics (mean, standard deviation, skewness) (Table 3). Unfortunately, there was no station in the study area with the highest correlation to the data of stations D12A063. Thus, missing data at this station were completed using the next most correlated station whose data was available. The results after completion of data indicated that the obtained estimation was statistically significant and the statistical structure of the data for each station was mostly preserved after the completion process (Table 3).

Table 3 - Long-term statistical parameters for completed data of each station

Target station	Method	Donor station	R^2	Mean	Std
D12A033	REG-Exp	D12A181	0,81	2,413	4,716
D12A055	REG-Linear	E12A003	0,55	0,851	1,646
D12A063	REG-Exp	D12A093	0,33	0,608	1,302
D12A093	REG-Linear	E12A003	0,56	0,568	1,033
D12A181	REG-Linear	E12A003	0,89	5,330	6,333

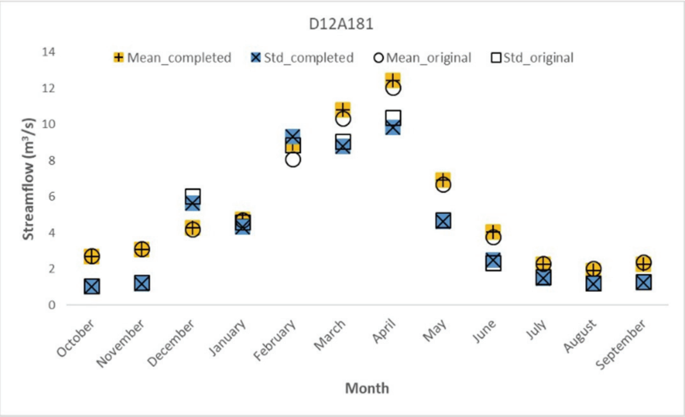

Figure 8 - Comparison of long-term statistical parameters between the completed data and original data of station D12A181

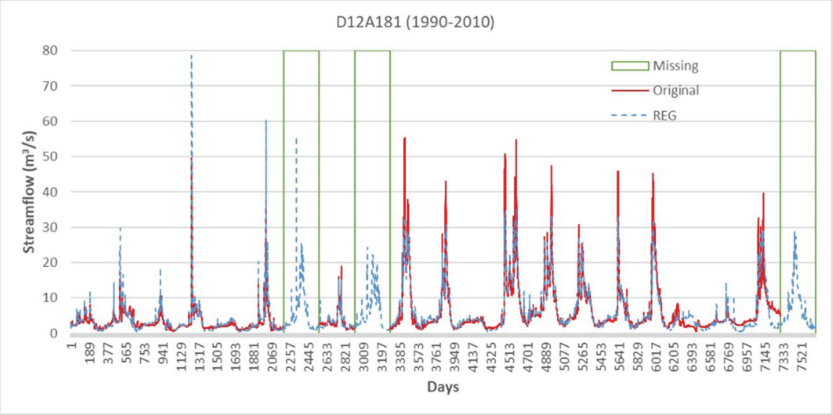

Figure 9 - The observed and estimated daily streamflow data of station D12A181

For example, as shown in Figure 8, when long-term monthly mean and standard deviation between completed data and original data of station D12A181 were compared, it was determined that there were no significant differences. The missing data of the station D12A181 were completed by a mathematical relation depending on correlation and the complete 21-year data of the station were obtained (Figure 9).

Figure 10 contains all estimates obtained for each station using each individual method for 11 year period and for 21 year period. For the 11 year period, the graphical overview for the estimation performance of each model was presented in the box plots (Figure 10a and Figure 10b). The ranges of NSE values resulted using a single donor from DAR, SM1, SM12, SMS1, and SMS12 methods for all stations were; −24.43 (outlier) to 0.82, −1.37 (outlier) to 0.85, −1.44 (outlier) to 0.79, −1.23 (outlier) to 0.81 and −0.83 (outlier) to 0.79, respectively. The ranges of NSE values resulted using all potential donors from MDAR and IDW methods for all stations were; −0.48 to 0.80, and 0.50 to 0.86 (outlier), respectively. The mean NSE values of these seven methods were −2.06, 0.17, 0.24, 0.23, 0.28, 0.36, and 0.59, respectively. For the 21 year period, the graphical overview for the estimation performance of each model was presented in the box plots (Figure 10c and Figure 10d). The ranges of NSE values resulted using a single donor from DAR, SM1, SM12, SMS1, and SMS12 methods for all stations were; −29.51 (outlier) to 0.82, −2.41 (outlier) to 0.87, −0.87 to 0.85, −0.62 to 0.84 and −0.33 to 0.83, respectively. The ranges of NSE values resulted using all potential donors from MDAR and IDW methods for all stations were; −0.14 to 0.82, and 0.27 (outlier) to 0.89 (outlier), respectively. The mean NSE values of these seven methods were −3.16, 0.09, 0.24, 0.28, 0.29, 0.40, and 0.55, respectively.

For both application periods, the DAR method demonstrated the greatest variability, while the IDW method exhibited the least variability. Where the ratio of the drainage areas of the target and donor station were greater than 1.5, almost all NSE values resulted from DAR method were negative. The best NSE results of each station were given in Table 4 for both 11 year period and for 21 year period. According to the results, the methods using multiple donors were superior than the methods using a single donor. For all target stations in the study area, the D12A181 or E12A003 mainstream stations were found to be good donor station options. Estimation performance of the individual methods for the 21 year period at 4 out of 6 stations was better than those for the 11 year period.

Table 4 - The best NSE results of each station

	11 year period			21 year period		
Target Station	**Method**	**Donor Station**	**NSE**	**Method**	**Donor Station**	**NSE**
D12A033	**MDAR**	All potential donors	0,56	**SM12**	D12A181	0,70
D12A055	**SM12**	D12A181	0,64	**SMS12**	E12A003	0,62
D12A063	**IDW**	All potential donors	0,53	**IDW**	All potential donors	0,27
D12A093	**SM12**	D12A181	0,58	**SMS1**	E12A003	0,63
D12A181	**IDW**	All potential donors	0,86	**IDW**	All potential donors	0,89
E12A003	**SM1**	D12A181	0,85	**SM1**	D12A181	0,87

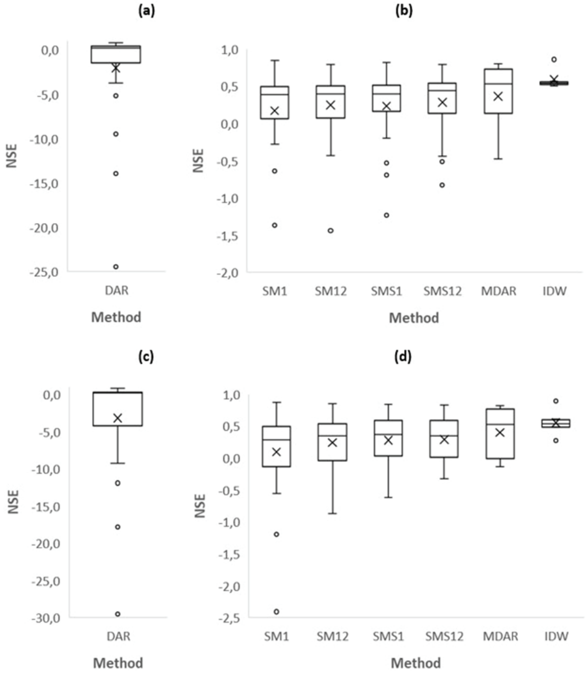

Figure 10 - Statistical summary of NSE performance of the individual methods (a) DAR method for 11 year period, (b) other methods for 11 year period, (c) DAR method for 21 year period, (d) other methods for 21 year period

The performance of the ensemble approaches, which combine two individual methods, showed a slight improvement compared to the best individual methods. NSE values of the best ensemble approaches for each station were given in Table 5. As shown Figure 11 the best ensemble approach for each station was better or almost of equal performance compared to the best individual method for each station.

Table 5 - NSE values of the best ensemble approach for each station

Target station	The best ensemble approach	NSE
D12A033	MDAR-SM12	0,72
D12A055	MDAR-SMS1	0,65
D12A063	DAR-SMS1	0,35
D12A093	IDW-SMS1	0,67
D12A181	IDW-SMS1	0,89
E12A003	IDW-SM1	0,87

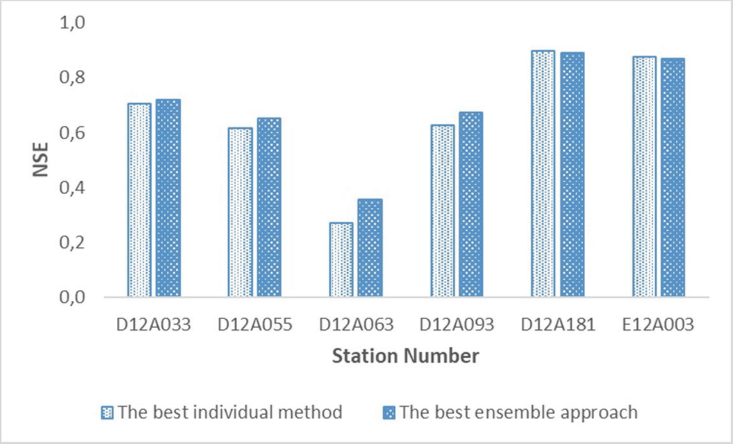

Figure 11 - Comparison of NSE performance between the best individual methods and the best ensemble approaches for each station.

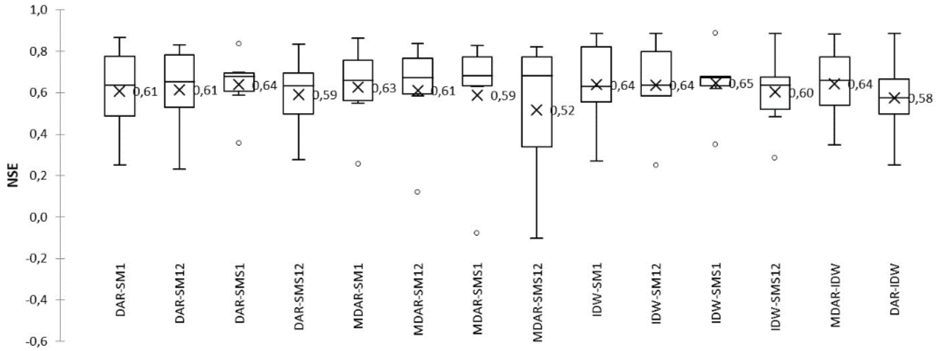

Figure 12 - Comparison of NSE values of all possible ensemble approaches

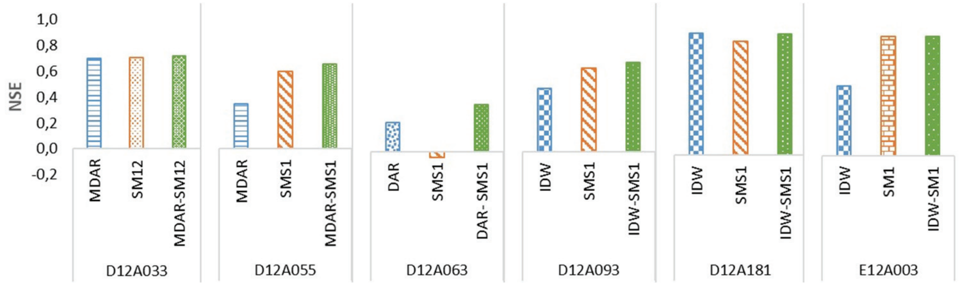

Figure 13 - Comparative performance evaluation of the individual methods and the best ensemble approach which is a combination of these individual methods

The box plots of NSE values for all stations using all possible ensemble approaches were presented in Figure 12. The mean NSE value was slight higher for IDW-SMS1 than the other ensemble approaches. Also, IDW-SMS1 had the least variability. As shown in Figure 13, the ensemble approaches combining two individual methods produced generally better results than individual methods.

The scatter plots between observed and estimated streamflow using DAR, SMS1, IDW methods and the best ensemble approach (IDW-SMS1) for station D12A181 were shown in Figure 14. The estimated streamflow plotted against the observed streamflow show a good match except for observed peak values of streamflow (Figure 14). The estimated values matched the observed values reasonably well, with R^2 of 0.93 for the best ensemble approach (IDW-SMS1) for station D12A181. The Nash-Sutcliffe efficiency (NSE) indicates how well the plot of observed against estimated streamflow data fits the 1:1 line. DAR, SMS1, IDW, and IDW-SMS1 have good method performance with NSE values of 0.82, 0.84, 0.89 and 0.89, respectively.

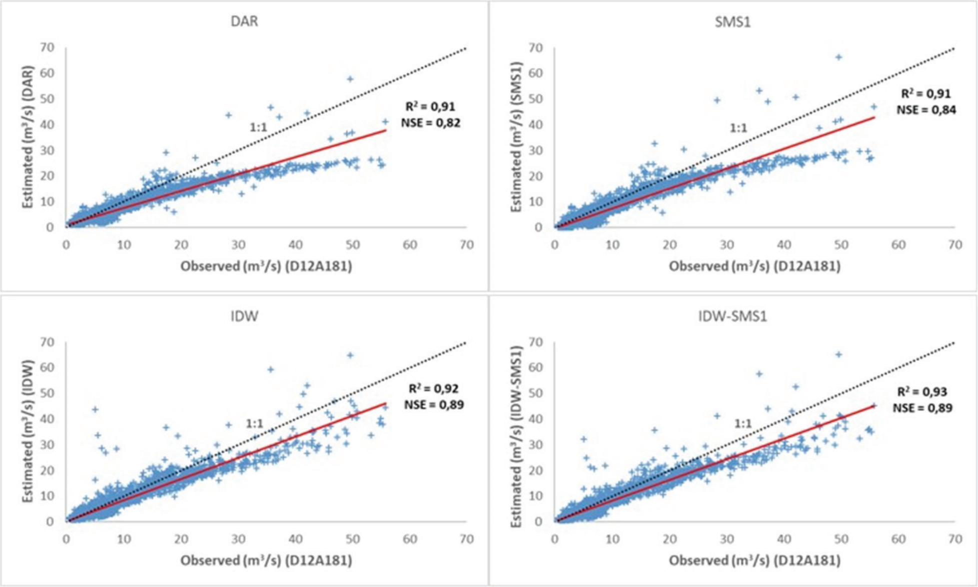

Figure 14 - Comparison between observed and estimated streamflow using DAR, SMS1, IDW method and the best ensemble approach (IDW-SMS1) for station D12A181

Estimation performance evaluation was also conducted for flow duration curves (FDC). Figure 15 illustrated the worst (D12A063) and the best (D12A181) station results for estimated FDC. The NSE values at each hydrological condition (high flow, middle flow, low flow) were calculated. The results indicated that estimated 21-year flow duration curves have better agreement with observed (original) flows than estimated 11-year flow duration curves.

Regional some descriptive statistics as the minimum, mean, maximum, and standard deviation of the estimates given in Table 6 and Table 7 showed a close match between the observed and estimated values. Both for 11 year period and for 21 year period (1990-2010), except for $Q_{0.1}$ and Q_{99}, all of the percentiles of the estimates obtained by the method with the best performance for each station, showed high coefficient of determination (linear relationship between mean observed and mean estimated values).

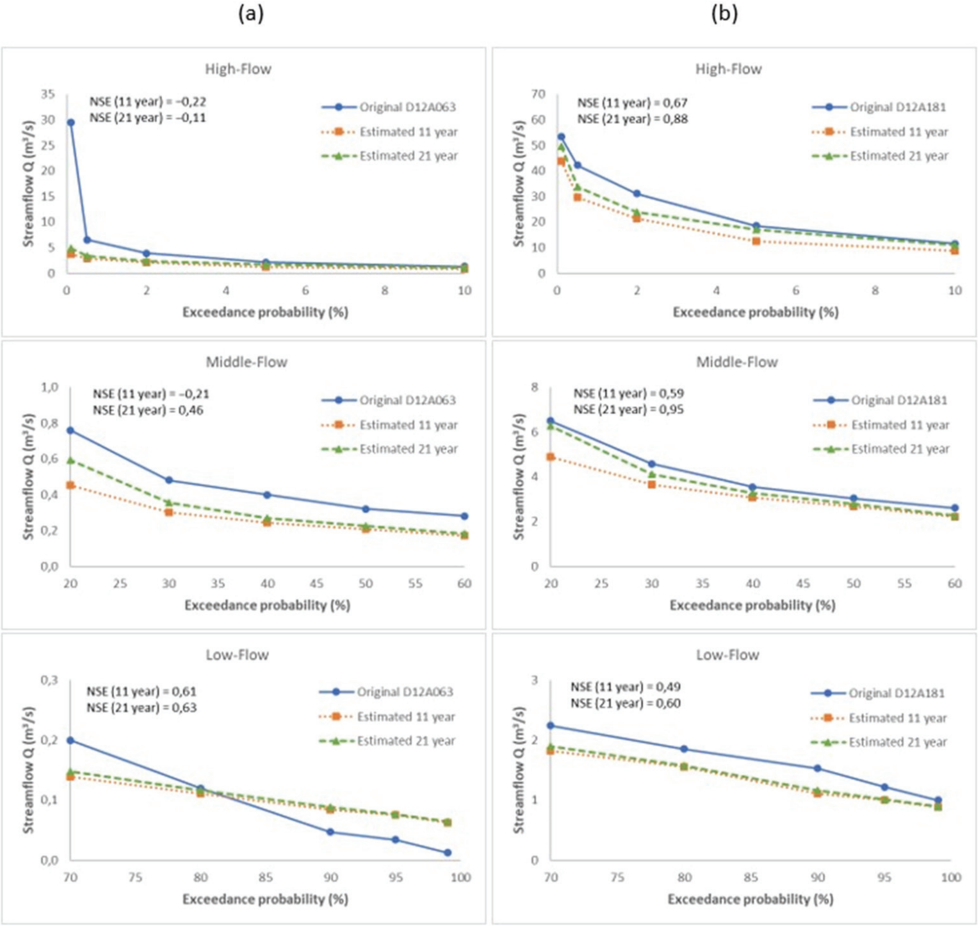

Figure 15 - Comparison of observed and estimated 15 percentiles of different hydrological conditions using the best individual method (a) for station D12A063 (the worst) and (b) for station D12A181 (the best)

Table 6 - Regional results of the estimation of the 15 percentile flows for 11 year period

			Observed				Estimated			
	Percentile Flow	**R^2**	**Min**	**Mean**	**Max**	**Std**	**Min**	**Mean**	**Max**	**Std**
High-flow	$Q_{0.1}$	0,57	7,5	30,7	62,7	23,5	3,7	19,9	43,8	17,4
	$Q_{0.5}$	0,81	5,0	17,7	36,7	13,7	2,9	15,2	30,3	12,6
	Q_2	0,89	3,3	12,3	25,4	9,6	2,1	10,7	21,5	9,0
	Q_5	0,99	1,8	6,7	13,9	5,2	1,2	6,1	12,6	5,1
	Q_{10}	0,98	1,0	4,5	9,9	3,9	0,8	4,2	8,7	3,7
Middle-flow	Q_{20}	0,95	0,5	2,5	5,9	2,4	0,5	2,4	4,9	2,1
	Q_{30}	0,93	0,3	1,8	4,2	1,8	0,3	1,8	3,6	1,6
	Q_{40}	0,90	0,2	1,5	3,5	1,6	0,2	1,4	3,1	1,3
	Q_{50}	0,87	0,1	1,2	3,1	1,4	0,2	1,2	2,7	1,2
	Q_{60}	0,84	0,1	1,0	2,6	1,2	0,1	1,0	2,2	1,0
Low-flow	Q_{70}	0,83	0,0	0,9	2,3	1,1	0,1	0,9	1,8	0,9
	Q_{80}	0,81	0,0	0,7	1,9	0,9	0,0	0,7	1,6	0,7
	Q_{90}	0,81	0,0	0,5	1,5	0,7	0,0	0,6	1,3	0,6
	Q_{95}	0,81	0,0	0,5	1,3	0,6	0,0	0,5	1,1	0,5
	Q_{99}	0,77	0,0	0,4	1,1	0,5	0,0	0,4	0,9	0,4

Table 7 - Regional results of the estimation of the 15 percentile flows for 21 year period (1990-2010)

			Observed				Estimated			
	Percentile Flow	**R^2**	**Min**	**Mean**	**Max**	**Std**	**Min**	**Mean**	**Max**	**Std**
High-flow	$Q_{0.1}$	0,72	8,2	34,9	56,3	18,5	4,8	25,5	49,7	20,0
	$Q_{0.5}$	0,79	6,3	21,0	42,1	13,9	3,4	18,5	33,7	14,8
	Q_2	0,91	3,8	14,7	29,0	10,0	2,4	13,0	23,8	10,0
	Q_5	0,96	2,0	9,3	19,0	6,9	1,7	8,9	17,2	6,8
	Q_{10}	0,94	1,2	6,0	12,2	4,6	1,1	5,6	11,0	4,4
Middle-flow	Q_{20}	0,87	0,7	3,4	7,1	2,7	0,6	3,1	6,3	2,6
	Q_{30}	0,85	0,5	2,2	4,8	1,9	0,4	2,0	4,1	1,8
	Q_{40}	0,82	0,3	1,7	3,7	1,6	0,2	1,6	3,3	1,5
	Q_{50}	0,82	0,2	1,4	3,2	1,4	0,1	1,3	2,8	1,3
	Q_{60}	0,80	0,1	1,1	2,7	1,2	0,0	1,1	2,3	1,1
Low-flow	Q_{70}	0,78	0,0	0,9	2,3	1,1	0,0	0,9	1,9	0,9
	Q_{80}	0,78	0,0	0,7	2,0	0,9	0,0	0,7	1,6	0,7
	Q_{90}	0,77	0,0	0,5	1,5	0,7	0,0	0,5	1,2	0,6
	Q_{95}	0,80	0,0	0,5	1,3	0,6	0,0	0,5	1,0	0,5
	Q_{99}	0,66	0,0	0,4	1,2	0,5	0,0	0,4	0,9	0,4

7. CONCLUSION AND RECOMMENDATION

In this study, two separate application trials were conducted in case of both partially missing data and completely missing data (ungauged). In partially missing data application, REG, DAR, SM and SMS methods were used to complete the missing data at the target station. With the evaluation approach described in the application section, the most appropriate completing method was selected for each station. Missing data at each target station were completed with the selected methods. Thus, the complete time series of 21 years was obtained for each station. In completely missing data application (each of the target stations was in turn considered as ungauged), DAR, SM, SMS, MDAR, and IDW methods were used to estimate daily streamflow at the ungauged station. Individual methods were assessed with observed 11 year period and completed data (21 year period). The results showed that the use of multiple donor stations significantly improved the estimates at the target station. The estimates obtained from the IDW method were found generally to be superior to other individual methods. The DAR method had an unacceptable performance with negative NSE values for most estimations in this study area. Ensemble approaches combining two individual methods were proposed in order to obtain more significant estimates in the study. Proposed ensemble approaches are carried out for only 21 year period. All possible ensemble approaches combining individual methods used in the study were tested. Thus, promising estimation results were obtained for each station. In the study, it is not claimed that any method can give always much better results than other methods. However, ensembles approaches combining individual methods can work well and achieve good estimation performance, although they are not always superior to individual methods.

Improving streamflow estimations is critical to a more effective and sustainable management of water resources in ungauged or limited gauged basins. Ensemble streamflow estimations can be used to improve the reliability of hydrological estimations. It is our suggestion that future studies should focus on the issue concerning which can increase the reliability of estimations in ungauged basins. In this study, both estimation of daily streamflows and flow duration curves, which have an important role in the development and operation of water resources, were implemented successfully in the Porsuk basin where the streamflow data are missing and the number of stations with streamflow data in the common long period (at least 10 years) is limited. Finally, successful results obtained in such a poorly gauged basin are expected to contribute to the streamflow estimation literature.

Symbols

A	: Drainage area
A_{donor}	: Drainage area for the donor station
A_{target}	: Drainage area for the target station
d	: The distance between two stations
d_i	: Similarity distance measured between the target station and donor station i
lat	: Latitude in radians of the station
lon	: Longitude in radians of the station
n	: The total number of donor stations
p	: The power parameter
q	: The area normalized streamflow value
Q	: Daily streamflow

Q_{donor}	: Daily streamflow for the donor station
Q_{target}	: Daily streamflow for the target station
$\widehat{Q}$	: Daily streamflow estimated from the subscripted method
r	: The radius of the earth
R^2	: The coefficient of determination
w	: Weight
w_i	: The weight assigned to the donor station i
X_i^{obs}	: The i-th observed value
X_i^{est}	: The i-th estimated value
$\overline{X^{obs}}$	: The mean of all the observed data for a time series of length n
α	: The coefficients of the regression equations
β	: The coefficients of the regression equations
θ	: The mean or standard deviation of the streamflows
μ	: The mean of the streamflow
σ	: The standard deviation of the streamflow
φ	: The ratio

Acknowledgements

This study was supported by the Research Fund of Istanbul Technical University under the project ‘Improvement of Streamflow Estimation in Ungauged Basins’ (Project number: 40717). The authors thank the State Hydraulic Works (DSI) of Turkey for providing data.

References

[1] Shu C. & Ouarda T. B. M. J., Improved methods for daily streamflow estimates at ungauged sites. Water Resour. Res., 48, p. W02523, 2012. https://doi.org/10.1029/2011WR011501

[2] Razavi T. & Coulibaly P., An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. Can. Water Resour. J., 42, pp. 2–20, 2017. https://doi.org/10.1080/07011784.2016.1184590

[3] Kalteh A. M. & Hjorth P., Imputation of missing values in precipitation-runoff process database. Hydrology Research, 40 (4), pp. 420-432, 2009. https://doi.org/10.2166/nh.2009.001

[4] Ergen K. & Kentel E., An integrated map correlation method and multiple-source sites drainage-area ratio method for estimating streamflows at ungauged catchments: A case study of the Western Black Sea Region, Turkey. Journal of Environmental Management, 166, 309–320, 2016. https://doi.org/10.1016/j.jenvman.2015.10.036

[5] Hughes D. A. & Smakhtin V., Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. Hydrological Sciences Journal, 41: 851–871, 1996. https://doi.org/10.1080/02626669609491555

[6] Tencaliec P., Favre A. C., Prieur C. & Mathevet T., Reconstruction of missing daily streamflow data using dynamic regression models. Water Resour. Res., 51 (2015), pp. 9447-9463, 2015. https://doi.org/10.1002/2015WR017399

[7] Patil S. & Stieglitz M., Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment. Hydrology and Earth System Sciences 16: 551–562, 2012. https://doi.org/10.5194/hess-16-551-2012

[8] Elshorbagy A. A., Panu U. S. & Simonovic S. P., Group-based estimation of missing hydrological data: I. Approach and general methodology. Hydrological Sciences Journal, 45, 849–866, 2000. https://doi.org/10.1080/02626660009492388

[9] Panu U. S., Khalil M. & Elshorbagy A., Streamflow data infilling techniques based on concepts of groups and neural networks. In: Govindraju, R.S., Rao, A.R. (Eds.). Artificial Neural Networks in Hydrology. Kluwer Academic Publishers, Dordrecht, 235-258, 2000.

[10] Elshorbagy A., Simonovic S. P. & Panu U. S., Estimation of missing streamflow data using principles of chaos theory. Journal of Hydrology, 255, pp. 123-133, 2002. https://doi.org/10.1016/S0022-1694(01)00513-3

[11] Dastorani M.T., Moghadamnia A., Piri J., & Rico-Ramirez M., Application of ANN and ANFIS models for reconstructing missing flow data. Environmental Monitoring and Assessment 166: 421–434, 2010. https://doi.org/10.1007/s10661-009-1012-8

[12] Mohamoud Y. M., Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrological Sciences, 53 (4), pp. 706-724, 2008. https://doi.org/10.1623/hysj.53.4.706

[13] Giustarini L., Parisot O., Ghoniem M., Hostache R., Trebs I., & Otjacques B., A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. Environmental Modelling & Software, 82, 308-320, 2016.

[14] Gill M. K., Asefa T., Kaheil Y. & McKee M., Effect of missing data on performance of learning algorithms for hydrologic prediction: Implication to an imputation technique. Water Resour. Res., 43, W07416, 2007.

[15] Hirsch R. M., An evaluation of some record reconstruction techniques. Water Resour. Res., 15, 1781–1790, 1979. https://doi.org/10.1029/WR015i006p01781

[16] Wiche G. J., Benson R. D. & Emerson D. G., Streamflow at selected gaging stations on the James River in North Dakota and South Dakota, 1953–1982, with a section on climatology, Water Resources Investigations Report 89-4039, US Geological Survey, 99 pp, 1989.

[17] Emerson D. G., Vecchia A. V. & Dahl A. L., Evaluation of drainage-area ratio method used to estimate streamflow for the Red River of the North Basin, North Dakota and Minnesota. US Geological Survey Scientific Investigative Report 2005–5017, 13 pp, 2005.

[18] Asquith W. H., Roussel M. C. & Vrabel J., Statewide analysis of the drainage-area ratio method for 34 streamflow percentile ranges in Texas. US Geological Survey Scientific Investigative Report 2006–5286, 34 pp, 2006.

[19] Chen T., Ren L., Yuan F., Yang X., Jiang S., Tang T., Liu Y., Zhao C. & Zhang L., Comparison of spatial interpolation schemes for rainfall data and application in hydrological modeling. Water, 9, 342, 2017. https://doi.org/10.3390/w9050342

[20] Farmer W. H & Vogel R. M., Performance-weighted methods for estimating monthly streamflow at ungauged sites. Journal of Hydrology 477: 240–250, 2013. https://doi.org/10.1016/j.jhydrol.2012.11.032

[21] Burgess T. M. & Webster R., Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of Soil Science, 31 pp. 315-331, 1980. https://doi.org/10.1111/j.1365-2389.1980.tb02084.x

[22] DSI, Management Plan for Porsuk Watershed, Final Report, State Water Works, Ankara, 2001.

[23] Hortness J. E., Estimating low flow frequency statistics for unregulated streams in Idaho. US Geol. Survey. Sci. Invest. Report 2006-5035, 2006.

[24] Sinnott R. W., Virtues of the Haversine. Sky and Telescope, vol. 68, no. 2, p. 159, 1984.

[25] Nash J. E. & Sutcliffe J. V., River flow forecasting through conceptual models. Part I-A discussion of principles. Journal of Hydrology, 10 (3), 282–290, 1970. https://doi.org/10.1016/0022-1694(70)90255-6