Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Türkiye'nin nehirlerinde eksik akım verilerinin tamamlanması için çeşitli veri odaklı tekniklerin performans değerlendirmesi

Performance evaluation of various data driven techniques for infilling missing streamflow data across Turkey's rivers

Muhammet Yılmaz [1*], Fatih Tosunoğlu [2]

[1] Erzurum Teknik Üniversitesi, Mühendislik ve Mimarlık Fakültesi, İnşaat Mühendisliği Bölümü, Erzurum, Türkiye
[2] Erzurum Teknik Üniversitesi, Mühendislik ve Mimarlık Fakültesi, İnşaat Mühendisliği Bölümü, Erzurum, Türkiye

Sorumlu Yazar / Corresponding Author *: muhammet.yilmaz@erzurum.edu.tr

Geliş Tarihi / Received: 06.01.2022 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 23.09.2022 *DOI:10.21205/deufmd.2023257405*
Atıf şekli/ How to cite: YILMAZ, M., TOSUNOĞLU, F. (2023). Performance evaluation of various data driven techniques for infilling missing streamflow data across Turkey's rivers. DEUFMD, 25(74),317-328.

Öz

Eksik veriler, su kaynaklarının etkin bir şekilde planlanması ve yönetilmesinin önünde her zaman bir engel teşkil etmektedir. Su kaynaklarının optimal tasarımı için eksiksiz ve güvenilir hidrolojik zaman serileri gereklidir. Türkiye genelinde 54 gözlem istasyonunun eksik akış verilerinin doldurulması için bir çalışma yapılmıştır. Doğrusal regresyon (LR), yapay sinir ağı (ANN), uyarlanabilir nöro-bulanık çıkarım sistemi (ANFIS), Destek vektör makinesi (SVM), Çok değişkenli uyarlanabilir regresyon eğrileri (MARS) ve K-en yakın komşu (KNN) kullanılarak tahminler gerçekleştirilmiştir. Yöntemlerinin performansları dört performans kriterine göre değerlendirilmiştir; bunlar, ortalama kare hata (RMSE), belirleme katsayısı (R^2), ortalama mutlak hata (MAE) ve Kling-Gupta verimliliği (KGE) dir. Bir istasyonda eksik akış verilerinin doldurulması için, çevredeki istasyonlardan alınan güvenilir ve uzun akış verileri girdi olarak seçilmiştir. Sonuçlar, tek bir yöntemin çalışma alanı için en uygun yöntem olarak belirlenemeyeceğini ortaya koymuştur. Test aşamasında, R^2 0,54 ile 0,99 arasında ve KGE aralığı 0,62 ile 0,98 arasındadır. Bu çalışma, özellikle SVM ve MARS yöntemlerinin Türkiye'deki nehirlerdeki eksik akış verilerinin tahmin edilmesi için uygun olduğunu göstermiştir. Bu bulgular, hidrolojik modelleme ve su kaynakları planlaması ve yönetiminde kullanılabilecek güvenilir akış verileri sağlayacaktır.

Anahtar Kelimeler: *Eksik değerler, akım, destek vektör makineleri, Çok değişkenli uyarlanabilir regresyon eğrileri, Turkiye*

Abstract

Missing data with gaps is always an obstacle to effective planning and management of water resources. Complete and reliable hydrological time series are necessary for the optimal design of water resources. A study was conducted to fill in missing streamflow data of 54 observation stations across Turkey. This process was done with the aid of various statistical estimation methods. Estimations were performed by using Linear regression (LR), Artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), Support vector machine (SVM), Multivariate Adaptive regression splines (MARS), and K-nearest neighbor (KNN) methods. Performances of infilling methods were evaluated based on four performance criteria; namely, root mean squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), and the Kling–Gupta efficiency (KGE) during training and test periods. Reliable and long streamflow data from surrounding

stations were selected as input to fill in missing streamflow data for an output station. The results revealed that a single method cannot be specified as the best-fit method for the study area. During the test phase, the R2 ranged from 0.54 to 0.99, and the KGE range was between 0.62 and 0.98. This study showed that especially SVM and MARS methods are suitable for estimating missing streamflow data in Turkey's rivers. These findings will provide reliable streamflow data that can be used in hydrological modeling and water resources planning and management.
Keywords: *Missing values; streamflow; Support vector machine; Multivariate adaptive regression splines; Turkey*

1. Introduction

Observed streamflow records, which are the integrated results of all meteorological and hydrological processes in a basin, provide useful information in planning and designing hydraulic construction projects [1]. However, streamflow data are not fully available in catchments in Turkey and many regions around the world, owing to the malfunction of measuring equipment, human-induced factors, and extreme weather conditions. Short and intermittent data negatively affect scientific and administrative studies in the fields of agriculture, hydrology and water resources and can lead to wrong decision making [2].This situation is also one of the biggest obstacles faced by hydrologists working with flow data for developing countries such as Turkey [3]. Therefore, long and continuous data sets are required for the hydrological history of a basin, reconstruction of historical climate, and planning and operation of water resources systems [4].

Infilling missing flow data is typically done by reconstructing the missing values by using observations from the neighboring station [5]. Several methods reported in literature have been developed for reconstructing missing data. These methods can be categorized in three ways, namely empirical approaches, statistical approaches, and function fitting techniques [6]. Conventional statistical techniques range from simple (for example, listwise deletions or binary deletions) to advanced methods (for example, moving average and regression) [7]. A disadvantage of these methods is the assumption of linearity between the estimators and streamflow, which causes a failure to represent the nonlinear dynamics found in hydrological studies [8]. Because of their ability to determine complex nonlinear relationships between input and target data without a physical understanding of the modeled system, machine learning (ML) techniques have been used for better estimation in reconstructing of missing data [9]. Therefore, there has been a growing number of publications on works involving the reconstruction of missing streamflow data across the world. For instance, [10] used the correlation technique, artificial neural network (ANN), and an adaptive neuro-fuzzy inference system (ANFIS) to estimate missing streamflow data. They emphasized that the the ANFIS method provided the best results for missing data. [9] evaluated the accuracy of various methods of estimating missing streamflow data in the three sub-basins of the Euphrates Basin. They found that ANFIS and ANN methods provided more accurate estimates for streamflow estimation in the Upper and Lower Euphrates Basins, while genetic programming and ANFIS models were more effective in estimating missing data in the Middle Euphrates Basin. [11] confirmed the accuracy of the ANN for estimating the missing streamflow data in the Taehwa River watershed, Korea. [12] used an ANN model to estimate missing streamflow-data. The resulting multilayer perceptron type network was found to be correct.

The literature review showed that there are no significant studies evaluating various methods for infilling missing flow data comprehensively across Turkey. There is a huge gap in the solving problems related to missing flow data across the country. Therefore, this study aims to fill this gap in the literature by completing missing flow data across Turkey. In addition, most of the previous studies are related to the application of ANN and ANFIS methods in reconstructing of missing data, but there is no significant study evaluating the effectiveness of the new and modern data mining methods such as Support vector machine (SVM), Multivariate adaptive regression splines (MARS), and K-nearest neighbor (KNN).

The aim of this study is to evaluate the performance of various estimation methods under different model selection criteria to infill the missing data in the streamflow records across Turkey. For this purpose, six methods (LR, ANN, ANFIS, SVM, MARS, and KNN) were used to fill the gaps of streamflow time series from observational data obtained from the neighboring station. Besides, the performance of

those methods are compared based on root mean squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), and the Kling–Gupta efficiency (KGE) tests.

2. Material and Method

2.1. Linear regression

Linear regression (LR) is a statistical technique used to find a suitable relationship between a dependent variable and an independent variable [13]. The regression equation of LR can be written as:

$$Y = \beta_o + \beta_1 X \quad (1)$$

where, Y is the dependent variable, X represents the independent variable, and β_o and β_1 are the regression coefficients.

2.2. Artificial neural networks

Artificial neural networks (ANN) can be defined as a data-driven statistical approach that can quickly solve non-linear relationships between input and output data. ANN is a robust computing tool inspired by features of the human brain and nervous system. There are many types of artificial neural networks such as multilayer perceptron (MLP), radial basis function (RBF) networks, and recurrent neural networks (RNN) used by researchers in hydrological studies [14]. The type of network used in this study is MLP, which consists of three layers that are input layer, hidden layer, and output layer. Each layer can have many nodes, which are connected together by weights. Each node receives the weighted input which is the output of each node in the previous layer and transmits it to the nodes of the next layer by means of links for proper output after processing it with an activation function. Many researchers in hydrological issues studied different ANN paradigms. Among the applications, the feed forward back propagation (FFBP) is one of the most popular networks [15].

2.3. Adaptive Neuro-Fuzzy Inference System

Adaptive Neuro-Fuzzy Inference System (ANFIS) was proposed by combining a fuzzy inference system (FIS) and ANN [16]. Typical ANFIS is a multilayer network consisting of five components, namely input nodes, output nodes, fuzzy system generator, fuzzy inference system, and adaptive neural network [17]. One of the main objectives of the ANFIS is to optimize the parameters of FIS by using input-output data sets via a learning algorithm. It captures the learning ability to optimize parameters of membership functions and adjust rules directly from data. The performance of the fuzzy inference system depends on the predicted parameters. The full explanation of ANFIS can be found here [18].

2.4. Support vector machine

Support vector machine (SVM) as a non-parametric technique, which was proposed firstly by [19], can be used for both classification and regression (SVR) problems. The basic idea behind SVR is to realize the principle of structural risk minimization to recognize the model between predictive and predicted values [20]. The SVR nonlinearly models primary data points from the input space in a higher dimensional feature space by using an appropriate kernel function. There are four types of commonly used kernels; namely, polynomial, linear, sigmoid, and radial basis function (RBF). Several studies have shown that the RBF performs better than other kernel functions [21, 22]. Hence, we used the RBF kernel in the present study.

2.5. Multivariate Adaptive Regression Spline

Multivariate adaptive regression splines (MARS), first introduced by [23], are classified in non-parametric regression methods. In this technique, the time series data is separated into a different number of subsets, and then the suitable basis functions are fitted to the available data. Spline is a function that is specifically defined at a certain interval, and its two-headed points are called knots. The basic function is implemented to show the data for each spline, which is specified at each knot. The MARS model technique is applied in a two-step procedure. In the first, numerous basic functions (BFs) are added to the model until the sum of squared errors is significantly reduced. The first model tends to be overfitting, so the backward direction eliminates the unnecessary variables and prevents the model from overfitting. Finally, a generalized cross-validation criterion is implemented in order to choose the most suitable BFs [24]. The general form of the MARS model can be described by the given equation below:

$$f(x) = \delta_o + \sum_{n=1}^{N} \delta_n h_n(X) \quad (2)$$

where, $h_n(X)$ indicates spline functions, δ is coefficient that is calculated by minimizing the residual errors. N represents the number of

functions. Please refer to [25] and [26] for further details about the MARS model.

2.6. K-Nearest Neighbor

K-nearest neighbor (K-NN) is a nonparametric method that can be utilized for both classification and regression problems. It is based on the idea that the outcome of a case is the same as the outcome of its nearest neighbor cases. K represents the amount of the closest neighbors of the queried point. The value of the queried point is equal to the average of its closest neighbors. The KNN regression method used in this study is developed as follows: 1) Calculate the Euclidian distance between the predictor examples and the queried example. 2) Sort the observation data in ascending order based on distance. 3) Calculate an inverse distance weighted average of the K-nearest neighbors. 4) Determine the most appropriate K number of the nearest neighbors [27].

In this algorithm, the number of neighbors (K) affects the prediction results; therefore, their amounts must be accurately calculated in order to obtain optimum results. For the first time in the KNN literature, a robust global method i.e. Differential Evolution optimization (DE) algorithm was applied, in order to determine the optimal K. The DE algorithm was first developed by [28] to avoid complex mathematical procedures and to provide optimum solutions to engineering and finance problems [29]. The DE tool is available in R as DEoptim package [30]. There are very few articles on hydrological topics that use the DE algorithm [31, 32, 33]. In order to determine the most appropriate K, the minimization of the NSE function is chosen as the objective function. However, as it is known, the convergence of NSE to 1 means that the success of the model is high. Thus, for the minimum of the objective function, the NSE function was revised as in Eq. 3.

$$\mathrm{NSEabs} = \left| \frac{\sum_{i=1}^{n}(Q_o - Q_s)^2}{\sum_{i=1}^{n}(Q_o - \overline{Q_o})^2} \right| \quad (3)$$

in which, Q_o andQ_s represent the observed and simulated values, respectively. The DE algorithm is run 1000 times on each model to identify the best value of the objective function.

2.7. Models performance criterion

The performance of the models has been examined utilizing different evaluation criteria found in the literature. The criteria used in the present study include the root mean squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), and the Kling–Gupta efficiency (KGE); their equations have been shown as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_i^{pre}\right)^2}{n}} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_i^{pre}\right)^2}{\sum_{i=1}^{n}\left(Q_i^{obs} - \bar{Q}\right)^2} \quad (5)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Q_i^{obs} - Q_i^{pre}\right| \quad (6)$$

where, n denotes the number of data points used; Q_i^{obs}, $\bar{Q}$ and Q_i^{pre} represent observed values, the average of the observed values, and predicted data, respectively.

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{pre}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{pre}}{\mu_{obs}} - 1\right)^2} \quad (7)$$

in which r denotes the linear correlation between observations and predictions; σ_{obs} and σ_{pre} are the standard deviations in observations and predictions, respectively; μ_{obs} is the observation mean, and μ_{pre} is the simulation mean.

The values close to 0 for MAE and $RMSE$ and the values close to 1 for the R^2 and KGE are indicative of a more desirable performance for the model.

2.8. Study area and data

Turkey, located between latitudes 36–42°N and longitudes26–45°E, is selected as the study site, in order to apply the aforementioned methods. The country covers a catchment area of about 780,576 km² with a mean elevation of 1141 m. Turkey is characterized by a heterogeneous landscape with four main climate types because of its temperate and subtropical zones. In general, the Mediterranean climate is seen in the Mediterranean and Aegean regions, with mild, wet winters and warm to hot, dry summers. The climate in the Black Sea region is characterized by high annual precipitation in all seasons, and its soil is mainly characterized by brown forest soils. Central, eastern, southeastern, and west-central of Turkey have a typical land climate, with hot and dry summers and a quite cold snowy winter. Finally, the Marmara region, which connects the Aegean Sea and the Black Sea, experiences a transitional climate between

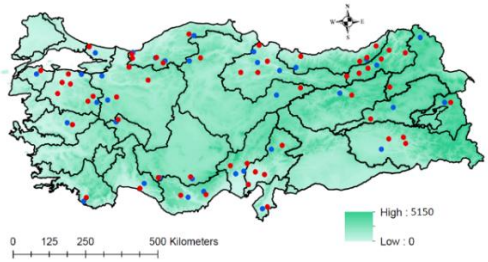
Tablo 1. Türkiye nehir havzalarının temel özellikleri

Table 1. Basic characteristics of the river basins, Turkey

Basin No	Basin Name	Area of Basin (x1000 km²)	Basin average height (m)	Average Precipitation (mm/year)	Total streamflow (km³/year)
1	Meriç	14.56	56.63	604	1.33
2	Marmara	24.1	42.25	728.7	8.33
3	Susurluk	22.399	201.56	711.6	5.43
4	Aegean	10.003	63.75	624.2	2.09
5	Gediz	18	220.06	603	1.95
6	Little Menderes	6.907	4	727.4	1.19
7	Big Menderes	24.976	413.83	664.3	3.03
8	West Mediterranean	20.953	383.47	875.8	8.93
9	Central Mediterranean	19.577	248.85	100.4	11.06
10	Burdur Lake	6.374	910	446.3	0.5
11	Afyon	7.605	1016.67	451.8	0.49
12	Sakarya	58.16	508.62	524.7	6.4
13	West Black Sea	29.598	325.67	811	9.93
14	Yeşilırmak	36.114	695.63	496.5	5.8
15	Kızılırmak	78.18	748.48	446.1	6.48
16	Middle Anatolia	53.85	1139.37	416.8	4.52
17	East Mediterranean	22.048	269.05	745	11.07
18	Seyhan	20.45	749.68	624	8.01
19	Hatay	7.796	159.17	815.6	1.17
20	Ceyhan	21.982	684.81	731.6	7.18
21	Euphrates	127.304	1009.87	540.1	31.61
22	East Black Sea	24.077	443.24	1198.2	14.9
23	Çoruh	19.872	757.39	629.4	6.3
24	Aras	27.548	1652.65	432.4	4.63
25	Van Lake	19.405	1829.29	474.3	2.39
26	Tigris	57.614	844.79	807.2	21.33
		Total	**Average**	**Average**	**Average**
		779.452	591.49	658.86	186.05

the Black Sea and Mediterranean types with uniformly rainy, but hot and slightly rainy in summer. For a detailed description of the climate of Turkey, refer to [34] and [35].

In this study, the aim is to fill the missing flow data with various statistical estimation methods in Turkey's rivers. Records of 54 flow stations operated by DSI (General Directorate of State Hydraulic Works) are used for applications. The gaps are filled with neighboring stations to the station with missing values.



Şekil 1. Çalışma alanı ve akım istasyonlarının yerleri. Kırmızı ve mavi daireler, sırasıyla eksik veriye sahip istasyonları ve tahmin edici istasyonları (komşu istasyon) gösterir.

Figure 1. Locations of the study area and streamflow stations. Red and blue circles indicate stations with missing data and predictive stations (neighboring station), respectively.

The record lengths range from 29 to 79 years, which can be considered statistically valid. The location map of the basins and the gauge stations under study are shown in Figure 1. In addition, the closest neighboring stations, which are used to infill missing flow data in the reference station, are shown in Figure 1, and additionally,

For missing data, nearest stations and training - test time information are given in the supplementary file.

Basic information and some important statistical characteristics of the basins are presented in Table 1.

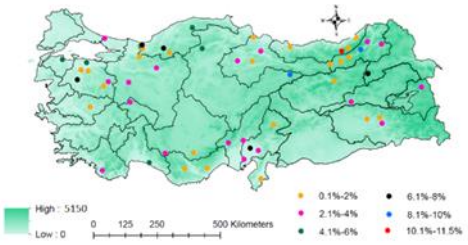
The annual average precipitation varies approximately between 100.4 (Central Mediterranean) and 1198.2 (East Black Sea) mm/year. The total streamflow varies from 0.5 to 31.61 km3/year and reached its highest value at Euphrates basin.

3. Results

Developed models are implemented for the estimation of observations of 54 stations across basins of Turkey. On average, the stations showed a gap of about 0.17% to 11.5% during the observation period (see Figure 2). One-input-one-output models were developed with one neighboring station for each station with missing data. In this study, 70% of the observations was used to train the models, while the remaining 30% was considered as a validation dataset.

In order to find the most successful ANN model, tangent sigmoid and linear transfer functions were used in the hidden layer and output layer, respectively. The Bayesian regularization backpropagation algorithm was employed in the network training process. Three neurons in the hidden layer were selected for this study by using the trial-and-error technique. ANFIS models were produced by using the same input data sets as the ANN models. In this study, Sugeno rule-based model with two sigmoid-shaped membership functions was adopted, and the fuzzy membership parameters were optimized via a back-propagation algorithm. After applying the procedures mentioned in section 2 for the other methods used in this study, the RMSE, R2, MAE, and KGE performance parameters of the most successful models were calculated and compared with each other. To evaluate the comparison results of six methods in detail, four model performance statistics were calculated during the testing period, and the results were listed in Tables 2, 3, 4 and 5. To avoid replication in results, only one basin example (for East Black Sea) of these results is showed. Bold values indicate the values of four model performance tests for the best results based on the six estimation techniques. Also, the RMSE and MAE results given in the tables are in m3/s.It can be seen from Table 2 that the MARS model has higher values of R2 (0.936) and KGE (0.937) and lower values of RMSE (3.749) and MAE (2.203) than those of the other models. Table 2 indicates that the MARS model has a better performance than the other models in terms of RMSE, KGE, MAE, and R2. Table 3 indicates that the MARS model provided more accurate estimation results than the other models for infilling missing flow data of Station 2232, as it is supported by more performance criteria. For Station 2247, the performance statistics of six models in the test period are presented in Table 4. It can be seen from the

table the ANN model, which has R2 of 0.797 and RMSE of 13.626, shows the highest performance among the other models in infilling missing flow data. However, when MAE and KGE are taken into account as performance indices, the LR and KNN model were found to be better than the other models in terms of MAE and KGE, respectively. As a result, ANN is more successful than the other models as it is supported by more criteria. Table 5 gives the performance statistics results of six models for Station D22a007. From Table 5, it is found that low RMSE and MAE values (1.388 and 0.790, respectively) and high R2 (0.903) value are obtained for the SVM model when compared to other models. It can be seen from Table 5 that the MARS estimation method gives the most successful estimation with maximum KGE (0.901). However, the SVM method is chosen to estimate the missing data of Station D22a007 because it is supported by more criteria.



Şekil 2. Türkiye'nin farklı havzalarında bulunan akım ölçüm istasyonları için eksik veri oranı.

Figure 2. Proportion of missing data for streamflow gauging stations located in different basins of Turkey.

Tablo 2. Test süresi için 2215 istasyonunda altı modelin performans karşılaştırması. Girdi istasyonu 2233'dür.

Table 2. Performance comparison of six models at Station 2215 for test period. The input station is 2233.

	Model Names					
Performances	**LR**	**ANN**	**ANFIS**	**MARS**	**SVM**	**KNN**
RMSE	3.958	3.766	3.766	**3.749**	3.771	3.8
R^2	0.929	0.936	0.936	**0.936**	0.93	0.93
MAE	2.236	2.217	2.222	**2.203**	2.218	2.23
KGE	0.898	0.935	0.935	**0.937**	0.92	0.93

Tablo 3. Test süresi için 2232 istasyonunda altı modelin performans karşılaştırması. Girdi istasyonu 2233'dür.

Table 3. Performance comparison of six models at Station 2232 for test period. The input station is 2233.

	Model Names					
Performances	**LR**	**ANN**	**ANFIS**	**MARS**	**SVM**	**KNN**
RMSE	11.34	10.97	10.97	**10.97**	11.182	11.059
R^2	0.803	0.815	0.815	**0.816**	0.808	0.813
MAE	**7.298**	7.441	7.447	7.463	7.315	7.498
KGE	0.809	0.843	0.842	**0.844**	0.811	0.841

Tablo 4. Test süresi için 2247 istasyonunda altı modelin performans karşılaştırması. Girdi istasyonu 2238'dir.

Table 4. Performance comparison of six models at Station 2247 for test period. The input station is 2238.

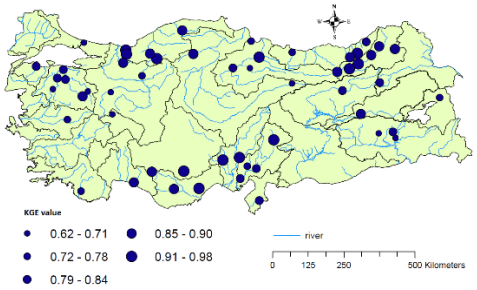
	Model Names					
Performances	**LR**	**ANN**	**ANFIS**	**MARS**	**SVM**	**KNN**
RMSE	15.04	**13.626**	13.82	13.691	14.486	13.786
R^2	0.754	**0.797**	0.791	0.795	0.771	0.792
MAE	**7.494**	7.843	8.021	7.915	7.746	7.931
KGE	0.696	0.783	0.778	0.776	0.704	**0.783**

Tablo 5. Test süresi için D22a007 istasyonunda altı modelin performans karşılaştırması. Girdi istasyonu 2233'dür.

Table 5. Performance comparison of six models at Station D22a007 for test period. The input station is 2233.

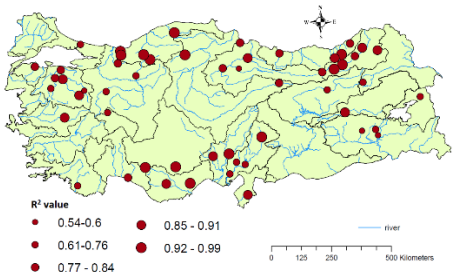
	Model Names					
Performances	**LR**	**ANN**	**ANFIS**	**MARS**	**SVM**	**KNN**
RMSE	1.545	1.413	1.40	1.412	**1.388**	1.425
R^2	0.88	0.899	0.900	0.8997	**0.903**	0.897
MAE	0.829	0.826	0.83	0.828	**0.79**	0.835
KGE	0.838	0.898	0.897	**0.901**	0.863	0.897

For the stations with missing data, the KGE values and R^2 values calculated according to the most successful model are given in Figure 3 and Figure 4, respectively. From Figure 3, it is found that the KGE values vary between approximately 0.62 and 0.98, and its highest value is observed at Stations 1340 and 2320. As can be seen from Figure 4, the magnitude of the R^2 values ranges from 0.54 to 0.99. The highest R^2 value is found at Stations 1340 and 2320 for streamflow estimation.



Şekil 3. En başarılı modele göre hesaplanan KGE değerinin büyüklük haritası

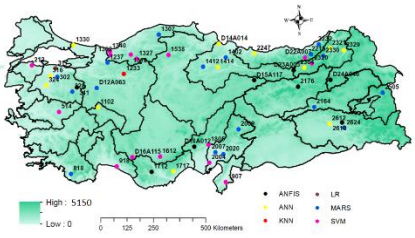
Figure 3. Map of magnitudes of the KGE value calculated according to the most successful model



Şekil 4. En başarılı modele göre hesaplanan R^2 değerinin büyüklük haritası.

Figure 4. Map of magnitudes of the R^2 value calculated according to the most successful model.

The most appropriate estimation methods chosen for stations with missing data are given in Figure 5. In most of the tests applied, the most suitable model is determined by considering the method that gives the best results.. Figure 5 showed that a single method has not emerged as the best method for all gauging stations. As seen in the Figure, LR, ANN, ANFIS, SVM, MARS and KNN were found to be the best models for the missing values of 1, 12, 8, 15, 17 and 1 stations, respectively. In addition, the tabulated version of Figure 5 is given in the supplementary file.



Şekil 5. Çalışmada kullanılan eksik verili istasyonlar için en iyi tahmin yönteminin gösterilmesi.

Figure 5. Demonstration of the best estimation method for stations with missing data used in the study.

4. Discussion and Conclusion

The results obtained from infilling missing series analysis by six popular models indicate that, based on four model selection criteria, no single method could be determined as the most appropriate method to complete missed data in Turkish river basins. However, the MARS and SVM methods most frequently provided consistent or robust reconstruction results, while LR and KNN were determined as the least chosen methods. As seen in Figures 3 and 4, the results are generally satisfactory when looking at the most successful model results in terms of KGE and R^2. Especially, in the north and south of Turkey, quite successful results were obtained in the estimation of the missing streamflow. This may be due to the topography in the mountainous regions and the high number of stations close to each other in the basin.

While ANN, ANFIS and LR methods were generally used as data driven techniques to fill in the missing streamflow data elsewhere in Turkey and the world, KNN, SVM and MARS methods were used less frequently. However, it was important for reference purposes to compare the accuracy of the reconstruction of the missing streamflow data in this study with the results obtained in other reconstruction studies. [9] applied ANFIS, ANN, genetic programming (GP) and LR methods to reconstruct the daily streamflow across the Euphrates Basin and they emphasized that ANFIS and ANN methods were the most appropriate methods to complete the missed data in the Upper and Lower Euphrates Basins,

whereas GP and ANFIS models were the best in the Middle Euphrates Basin. [36], who applied different types of ANN to fill monthly streamflow missing data, found correlation coefficient ranging from 0.56 to 0.73 in the testing phase. [12] investigated four different types of ANN to fill the missing data from monthly average streamflow and provided R^2 ranging from 0.94. This study showed that different estimation methods may be appropriate for estimating missing flow data in the same site. The successful performance of different methods for estimating missing flow data in a basin with the same climate, basin and hydrological characteristics highlights the necessity of using various estimation methods. For example, in Coruh basin, ANN for stations 2304, 2321 and 2329, SVM for station 2320, MARS for station 2330, ANFIS for station D23A003 were chosen as the best estimation method. As a result, it can be concluded that the use of different estimation techniques is a quite efficient and appropriate approach for estimating missing streamflow data. Complete records of streamflow data are essential and critical to effectively manage water resources. However, collecting such series may be very difficult, given the many reasons why gaps can occur in the observed data. Over the past decades, researchers have proposed methods to reconstruct these series using a variety of approaches, such as parametric and nonparametric techniques. The aim of this paper is to develop different models to infill the missing data in the flow records of the stations in Turkey's rivers. Four commonly used model selection criteria are utilized (i.e., RMSE, KGE, MAE, and R^2) to determine the best estimation procedure. Six methods are utilized to infill the missing streamflow data and are compared with each other. The results showed that no single method could be determined as the most appropriate method to reconstruct stations with missing data in Turkish river basins. However, the MARS and the SVM methods are determined as the most frequently while LR and KNN appeared as the least frequently selected. The most important contribution of this paper to the study area is the generation of continuous and longer data by structuring the missing data in the records of the flow stations. Finally, the findings of this study can provide important information and preliminary insight to engineers and decision makers in the design of water structures in any region of Turkey. We also believe that these results will provide important contributions to researchers working on physically-based hydrological models in the future.

4.Tartışma ve Sonuç

Dört model seçim kriterine dayalı olarak, altı popüler model ile verileri tamamlanan eksik seri analizinden elde edilen sonuçlar, Türkiye nehir havzalarında eksik verileri tamamlamak için en uygun yöntem olarak tek bir yöntemin belirlenemeyeceğini göstermektedir. Bununla birlikte, MARS ve SVM yöntemleri en sık tutarlı veya sağlam rekonstrüksiyon sonuçları verirken, LR ve KNN en az seçilen yöntemler olarak belirlenmiştir. Şekil 3 ve 4'te görüldüğü gibi, KGE ve R^2 açısından en başarılı model sonuçlarına bakıldığında sonuçlar genel olarak tatmin edicidir. Özellikle Türkiye'nin kuzey ve güneyinde eksik akım tahmininde oldukça başarılı sonuçlar elde edilmiştir. Bunun nedeni dağlık bölgelerdeki topoğrafya ve havzada birbirine yakın istasyon sayısının fazla olması olabilir.

ANN, ANFIS ve LR yöntemleri Türkiye'de ve dünyada eksik akım verilerini tamamlamak için genellikle veriye dayalı teknikler olarak kullanılırken, KNN, SVM ve MARS yöntemleri daha az sıklıkla kullanılmıştır. Ancak, bu çalışmada eksik akış verilerinin yeniden yapılandırılmasının doğruluğunun diğer yeniden yapılandırma çalışmalarında elde edilen sonuçlarla karşılaştırılması referans amaçlı olarak önemlidir. [9], Fırat Havzası boyunca günlük akım verilerini yeniden oluşturmak için ANFIS, ANN, genetik programlama (GP) ve LR yöntemlerini uygulamışlar ve Yukarı ve Aşağı Fırat Havzalarında eksik verileri tamamlamak için ANFIS ve ANN yöntemlerinin en uygun yöntemler olduğunu vurgulamışlardır. GP ve ANFIS modelleri ise Orta Fırat Havzası'nda en başarılı sonuçları vermiştir. Aylık akım verilerini tamamlamak için farklı ANN türleri uygulayan [36], test aşamasında 0,56 ila 0,73 arasında değişen korelasyon katsayısı bulmuştur. [12], aylık ortalama akımlardaki eksik verileri doldurmak için dört farklı ANN türünü çalışmıştır ve 0.94 arasında değişen R^2 değerleri belirlemiştir. Bu çalışma, aynı sahadaki eksik akım verilerini tahmin etmek için farklı tahmin yöntemlerinin uygun olabileceğini göstermiştir. Aynı iklim, havza ve hidrolojik özelliklere sahip bir havzada eksik akım verilerinin tahmininde farklı yöntemlerin başarılı performansı, farklı tahmin yöntemlerinin kullanılması gerekliliğini ortaya koymaktadır. Örneğin Çoruh havzasında

2304, 2321 ve 2329 numaralı istasyonlar için ANN, 2320 numaralı istasyon için SVM, 2330 numaralı istasyon için MARS, D23A003 istasyonu için ANFIS en iyi tahmin yöntemi olarak seçilmiştir. Sonuç olarak, eksik akım verilerinin tahmininde farklı tahmin tekniklerinin kullanılmasının oldukça verimli ve uygun bir yaklaşım olduğu sonucuna varılabilir. Akım verilerinin eksiksiz kayıtları, su kaynaklarının etkili bir şekilde yönetilmesi için gerekli ve kritik öneme sahiptir. Ancak, gözlemlenen verilerde boşlukların oluşabilmesinin birçok nedeni göz önüne alındığında, bu tür serileri toplamak çok zor olabilir. Geçtiğimiz yıllarda araştırmacılar, parametrik ve parametrik olmayan teknikler gibi çeşitli yaklaşımlar kullanarak bu serileri yeniden oluşturmak için yöntemler önermişlerdir. Bu çalışmanın amacı, Türkiye nehirlerindeki istasyonların akım kayıtlarındaki eksik verileri doldurmak için farklı modeller geliştirmektir. En iyi tahmin prosedürünü belirlemek için yaygın olarak kullanılan dört model seçim kriteri (yani, RMSE, KGE, MAE ve R^2) kullanılmıştır. Eksik verilerini doldurmak için altı yöntem kullanılmış ve birbirleriyle karşılaştırılmıştır. Sonuçlar, Türkiye nehir havzalarında eksik veriye sahip istasyonları yeniden oluşturmak için en uygun yöntem olarak tek bir yöntemin belirlenemeyeceğini göstermiştir. Ancak MARS ve SVM yöntemleri en sık olarak belirlenirken, LR ve KNN en az seçilen yöntemler olarak ortaya çıkmıştır. Bu çalışmanın çalışma alanına en önemli katkısı akım istasyonları kayıtlarındaki eksik verilerin yapılandırılarak sürekli ve daha uzun süreli verilerin üretilmesidir. Son olarak, bu çalışmanın bulguları, Türkiye'nin herhangi bir bölgesindeki su yapılarının tasarımında mühendislere ve karar vericilere önemli bilgiler ve ön bilgiler sağlayabilir. Bu sonuçların gelecekte fiziksel tabanlı hidrolojik modeller üzerinde çalışan araştırmacılara da önemli katkılar sağlayacağına inanıyoruz.

5. Ethics committee approval and conflicts of interest

The authors declare no need for an ethics committee approval and no conflict of interest in this paper.

Acknowledgment

The authors acknowledge the General Directorate of State Hydraulic Works, Turkey, for providing the daily streamflow data records used in this study.

References

[1] Kuriqi, A., Ali, R., Pham, QB., et al 2020. Seasonality shift and streamflow flow variability trends in central India. Acta Geophys 68:1461–1475. https://doi.org/10.1007/s11600-020-00475-4.

[2] Dikbas, F., Yasar, M. 2020. Data-Driven Modeling of Flows of Antalya Basin and Reconstruction of Missing Data. Iran J Sci Technol - Trans Civ Eng 44:1335–1344. https://doi.org/10.1007/s40996-019-00331-6

[3] Ergen, K., Kentel, E. 2016. 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. J Environ Manage 166:309–320. https://doi.org/10.1016/j.jenvman.2015.10.036

[4] Dembélé, M., Oriani, F., Tumbulto, J., et al 2019. Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. J Hydrol 569:573–586.

[5] Kim. J.W., Pachepsky, Y.A. 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J Hydrol 394:305–314. https://doi.org/10.1016/j.jhydrol.2010.09.005

[6] Xia, Y., Fabian, P., Stohl, A., Winterhalter, M. 1999 Forest climatology: Estimation of missing values for Bavaria, Germany. Agric For Meteorol 96:131–144. https://doi.org/10.1016/S0168-1923(99)00056-8

[7] Ng, W.W., Panu, U.S., Lennox, W.C. 2009. Comparative Studies in Problems of Missing Extreme Daily Streamflow Records. J Hydrol Eng 14:91–100. https://doi.org/10.1061/(asce)1084-0699(2009)14:1(91)

[8] Nayak, P.C., Sudheer, K.P., Rangan, D.M., Ramasastri, K.S. 2004 A neuro-fuzzy computing technique for modeling hydrological time series. J Hydrol 291:52–66. https://doi.org/10.1016/j.jhydrol.2003.12.010

[9] Yilmaz, A.G., Muttil, N. 2014. Runoff Estimation by Machine Learning Methods and Application to the Euphrates Basin in Turkey. J Hydrol Eng 19:1015–1025. https://doi.org/10.1061/(asce)he.1943-5584.0000869

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

[11] Kim, M., Baek, S., Ligaray, M., et al 2015. Comparative studies of different imputation methods for recovering streamflow observation. Water (Switzerland) 7:6847–6860. https://doi.org/10.3390/w7126663

[12] de Souza, G.R., Bello, I.P., Corrêa, F.V., de Oliveira, L.F.C. 2020. Artificial Neural Networks for Filling Missing Streamflow Data in Rio do Carmo Basin, Minas Gerais, Brazil. Brazilian Arch Biol Technol 63:1–8. https://doi.org/10.1590/1678-4324-2020180522

[13] Tabari, H., Sabziparvar, A.A., Ahmadi, M. 2011 Comparison of artificial neural network and multivariate linear regression methods for estimation of daily soil temperature in an arid region. Meteorol Atmos Phys 110:135–142. https://doi.org/10.1007/s00703-010-0110-z

[14] Uysal, G., Şorman, A.A., Şensoy, A. 2016 Streamflow Forecasting Using Different Neural Network Models With Satellite Data for a Snow Dominated Region in Turkey. Procedia Engineering 154 1185 – 1192.

[15] Sun, Y., Niu, J., Sivakumar, B. 2019. A comparative study of models for short-term streamflow forecasting with emphasis on wavelet-based approach. Stoch Environ Res Risk Assess 33:1875–1891. https://doi.org/10.1007/s00477-019-01734-7

[16] Jang, J.R. 1993 ANFIS : Adap tive-Ne twork-Based Fuzzy Inference System. 23

[17] Karaboga, D., Kaya, E. 2019. Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey. Artif Intell Rev 52:2263–2293. https://doi.org/10.1007/s10462-017-9610-2

[18] Kisi, O., Nia, A.M., Gosheh, M.G., et al 2012. Intermittent Streamflow Forecasting by Using Several Data Driven Techniques. Water Resour Manag 26:457–474. https://doi.org/10.1007/s11269-011-9926-7

[19] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[20] Parisouj, P., Mohebzadeh, H., Lee, T. 2020. Employing Machine Learning Algorithms for Streamflow Prediction: A Case Study of Four River Basins with Different Climatic Zones in the United States. Water Resour Manag 34:4113–4131. https://doi.org/10.1007/s11269-020-02659-5

[21] Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B. 2001. Model Induction with Support Vector Machines:Introduction and Applications. J Comput Civ Eng 15:208–216. https://doi.org/10.1061/(asce)0887-3801(2001)15:3(208)

[22] Lin, J.Y., Cheng, C.T., Chau, K.W. 2006. Using support vector machines for long-term discharge prediction. Hydrol Sci J 51:599–612. https://doi.org/10.1623/hysj.51.4.599

[23] Friedman, J.H. 1991 Multivariate Adaptive Regression Splines. Annals of Statistics, 19, 1-67. https://doi.org/10.1214/aos/1176347963

[24] Mehdizadeh, S., Fathian, F., Safari, M.J.S., Adamowski, J.F. 2019. Comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: A local and external data analysis approach. J Hydrol 579:. https://doi.org/10.1016/j.jhydrol.2019.124225

[25] Zhang, W., Goh, A.T.C. 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci Front 7:45–52. https://doi.org/10.1016/j.gsf.2014.10.003

[26] Alizamir, M., Heddam, S., Kim, S., et al 2021. Prediction of daily chlorophyll-a concentration in rivers by water quality parameters using an efficient data-driven model: online sequential extreme learning machine.ActaGeophys. https://doi.org/10.1007/s11600-021-00678-3

[27] Khazaee, Poul, A., Shourian, M., Ebrahimi, H.2019. A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction. Water Resour Manag 33:2907–2923. https://doi.org/10.1007/s11269-019-02273-0

[28] Rainer, S., Kenneth, P. 1997. Differential Evolution: A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J Glob Optim 11:341

[29] Jiang ,Z., Ma, W. 2018. Integrating differential evolution optimization to cognitive diagnostic model estimation. Front Psychol 9:1–9. https://doi.org/10.3389/fpsyg.2018.02142

[30] Mullen, K.M., Ardia, D., Gil, D.L., et al 2011. DEoptim: An R package for global optimization by differential evolution. J Stat Softw 40:1–26. https://doi.org/10.18637/jss.v040.i06

[31] Sleziak, P., Holko, L., Danko, M., Parajka, J. 2020. Uncertainty in the number of calibration repetitions of a hydrologic model in varying climatic conditions. Water (Switzerland) 12:. https://doi.org/10.3390/W12092362

[32] Tang, S., Jiang, J., Zheng, Y., et al 2021. Robustness analysis of storm water quality modelling with LID infrastructures from natural event-based field monitoring. Sci Total Environ 753:142007. https://doi.org/10.1016/j.scitotenv.2020.142007

[33] Yilmaz, M., Tosunoglu, F., Demirel, M.C. 2021. Comparison of conventional and differential evolution-based parameter estimation methods on the flood frequency analysis. Acta Geophys 69:1887–1900. https://doi.org/10.1007/s11600-021-00645-y

[34] Kadiolu, M. 2000 Regional variability of seasonal precipitation over Turkey. Int J Climatol 20:1743–1760. https://doi.org/10.1002/1097-0088(20001130)20:14<1743::AID-JOC584>3.0.CO;2-G

[35] Güçlü, Y.S. 2018. Multiple Şen-innovative trend analyses and partial Mann-Kendall test. J Hydrol 566:685–704. https://doi.org/10.1016/j.jhydrol.2018.09.034

[36] Wambua, R.M., Mutua, B.M., Raude, J.M. 2016. Prediction of Missing Hydro-Meteorological Data Series Using Artificial Neural Networks (ANN) for Upper Tana River Basin , Kenya. Am J Water Resour 4:35–43. https://doi.org/10.12691/ajwr-4-2-2