



Using Decision Tree and K-Means to Improve ANFIS for Predicting Missing Flow Data in Çoruh Basin

RAMAZAN ACAR¹

Abstract—Complete and continuous flow data provides contributions to the water resources projects appropriate design. However, both financial and technical factors prevent the regular collection of data and such a situation leads to missing data problems. Input variables are from five stations in the Çoruh Basin in the northeast province of Türkiye, and output variables are from one station. In the study, 75% (171 months of data) of the 228 months flow data between the years 1993–2011 are used in the training phase and 25% (57 months of data) in the testing phase. In the ANFIS models, Decision Tree (DT) and K-means (KM) methods are used in selecting the input variables and determining the number of membership functions. In addition, artificial neural network (ANN), multiple linear regression (MLR), and ANFIS models with randomly determined cluster numbers are also used to compare the model performance of these methods. The results show that using DT and KM methods, the ANFIS models generate more reliable results than other models. During the study, the regression coefficient (R^2), weighted mean square error (WMSE), and Wilcoxon (Z) values are taken into account for reliability measure. The DT–K-means–ANFIS model achieved the highest accuracy with $R^2 = 0.98$ and $WMSE = 5.89$ during the testing phase, demonstrating superior performance to the ANN and MLR models. It is shown that the most successful models can be determined in a shorter time by using DT and KM methods before generating ANFIS models.

Keywords: ANFIS, Artificial Neural Networks, Decision Tree, Hydrology modeling, K-Means, Missing flow data.

1. Introduction

In the globalizing world, population growth, industrialization, climate change and environmental problems are increasing the demand for water. This situation poses a serious risk in terms of decreasing usable water resources. In order to make good planning in the use of water resources, more research

should be done on the development of water structures (Turhan & Çağatay, 2016). Flow records are the integration part of hydro-meteorological processes in a basin and they provide useful information for the planning of projects related to water structures. In many basins in Türkiye and the world, errors may occur in flow data due to unsuitable weather conditions, failure of measurement equipment and human factors. Incomplete data can lead to wrong decisions in scientific and administrative studies in the fields of hydrology, agriculture, irrigation, design of water structures, etc. (Yılmaz & Tosunoğlu, 2023). In order to take precautions against the situations mentioned above, it is necessary to estimate the deficiencies of existing data sets. Various methods have been proposed in the literature to eliminate these deficiencies and make forward-looking predictions. The pioneers of these methods are single and multiple regression analysis (Kamwaga et al., 2018; Papailiou et al., 2022), adaptive network-based fuzzy inference systems (Dastorani et al., 2010; Saplioglu & Kucukerdem Ozturk, 2024), artificial neural networks (Dariane & Borhan, 2024; Souza et al., 2020), interpolation methodology (Ismail et al., 2017), time series analysis (Beauchamp et al., 1989; Elshorbagy et al., 2000), flow continuity line determination (Masih et al., 2010; Swain & Patra, 2017) and machine learning techniques (Sharma & Yuden, 2021). Terzi and Köse (2012) studied the estimation of Göksu River flow data in Türkiye and tried to complete the missing flow data. It has been stated that artificial neural networks (ANNs) give reliable results in estimation studies. Tfwala et al. (2013) investigated the accuracy of ANN in estimating missing flow records. Gümüş and Kavşut (2013) filled the missing monthly flow data of Zamanti River-

¹ Civil Engineering Department, Hydraulic Division, Munzur University, Tunceli, Türkiye. E-mail: ramazanacar@munzur.edu.tr

Ergenusağı station using feedforward backpropagation ANNs, radial basis and generalized regression neural network methods. Tosunoğlu et al. (2017) conducted a study to determine the missing flow data at five flow measurement stations located in the Upper and Middle Euphrates drainage basin region. Regression and flow duration curve methods were used in estimating the missing data. Although both methods provided successful results, and thus, it was stated that regression models showed more successful performance. Saplıoğlu and Küçükerdem (2018) used the ANFIS method to estimate the missing flow data of the Yeşilırmak River in Türkiye. First, models were established using ANFIS, multiple regression and normal rate methods to complete the missing data and their results were compared with each other. It is observed that the minimum number of data required for ANFIS model applications was determined using the optimum ANFIS model. As a result, the ANFIS method provides the most accurate result. It is also emphasized that a 10-year training set will be enough as a data set. Canchala-Nastar et al. (2019) managed to obtain missing monthly precipitation data with an error of approximately 9.8 mm using the ANN method in their study conducted in Colombia.

In the literature, ANFIS-based studies often rely on fixed assumptions or trial-and-error methods to determine both which stations to select as input variables and how to determine the number of membership functions for these variables. This approach causes time loss, especially when evaluating a large number of stations, and it makes it difficult to determine the most suitable model structure. In this study, decision trees were used to determine the input variables, and the K-Means method was used to select the number of membership functions, thereby reducing the fundamental uncertainties in the ANFIS model. This procedure resulted in a clearer model setup process and increased the consistency of the obtained results. In addition, while previous studies applied ANFIS together with Decision Trees or K-means as separate approaches, this study integrates both methods in a consecutive manner so that one guides the selection of input variables and the other determines the membership structure. This combined use provides a clearer distinction from existing applications and highlights the

contribution of the proposed model. The above-mentioned studies evaluated the performance of various techniques commonly used for estimating missing data. In this study, the analysis was further extended by including Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) models to provide a broader performance comparison with the proposed approach.

2. Methodology

As technology has developed over time, modern electronic devices started to be used. The first studies on artificial neurons were conducted by McCulloch and Pitts (1943), and as a result of their studies on electrical circuits, they modeled simple artificial neurons. Widrow and Hoff (1960) developed the methods called Adaline and Madaline. Madaline is the first artificial neural network (ANN) model used to solve various problems (Acar & Saplıoğlu, 2020; Saplıoğlu & Çimen, 2010). ANN can be examined in two parts concerning its structure and mathematical functions. ANN structure consists of input, hidden and output layers (Fig. 1). Hidden cells have a trigger processor to establish nonlinearity in the modeling process. The operation of ANN can be thought of as internal and external processors with two mathematical functions. The internal function of ANN is performed by hidden layer processors. In its external operation, the values between successive layer cells are first assigned. Then, with processes such as forward and feedback, the result is renewal by

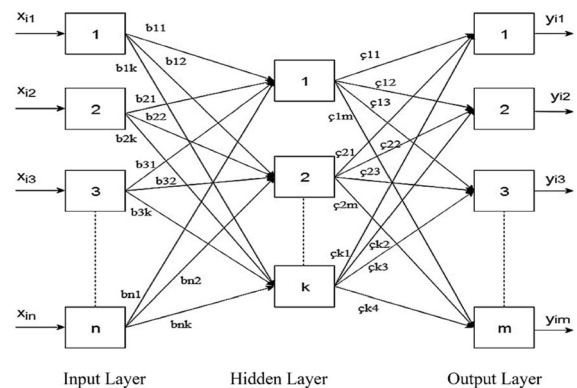


Figure 1
General structures of artificial neural networks

minimizing the error in the outputs prediction. These mathematical operations perform the tasks of learning, training, remembering, renewing network connections and frequently by recognizing new information (Şen, 2004).

Adaptive Neuro Fuzzy Inference System (ANFIS) was first developed by Jang (1993). Its aim is to obtain fuzzy rules automatically. The structure of this inference system uses the learning ability of ANNs together with fuzzy logic implementation. In this respect, the ANFIS method is more successful than running both the fuzzy logic model and the ANN model alone. When the input and output values are known, ANFIS can determine the valid fuzzy rule base. The ANFIS model consists of five layers as input membership function (MF), MF trigger layer (Π), trigger normalization layer (N), function results layer (R) and sum of the results layer (Σ) (see Fig. 2).

Fuzzy K-means is an unsupervised modeling algorithm first used by Macqueen (1967). This method works by segmenting several objects into existing categories or conditions based on the mid-point or nearest center. This method minimizes the variation among data within the same cluster. It also

maximizes the variation among data in different clusters (Al Kindhi et al., 2019). This method is a technique used in geohydrology, soil science, and vegetation mapping (Lucieer & Lucieer, 2009; Vriend et al., 1988). The K-means clustering algorithm consists of the following steps (Hartigan & Wong, 1979; Zeraatpisheh et al., 2019).

- Data points are defined by X_i and the number of clusters is denoted by K_i ($i = 1, 2, \dots, n$).
- Data points are then randomly divided into clusters. Then, all cluster centers are calculated as C_j ($j = 1, 2, \dots, k$).
- Within each cluster, the distance of each data point to the centers is calculated. This is done using the selected distance algorithm. When only the input parameter is used, the X distance is used, while when both input and output data are used together, the Euclidean distance can be used to find the spatial distance.
- Distances are used to determine which data belongs to which cluster.
- The centroids of each cluster are recalculated. Repeat the distance calculation according to the new centroids.

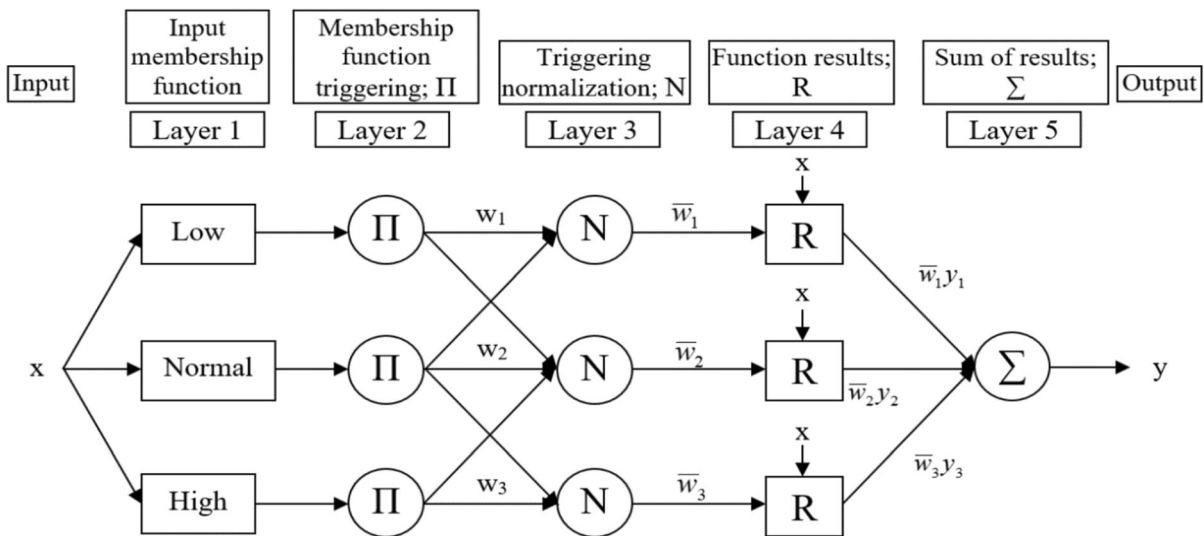


Figure 2
ANFIS architecture with one input-one output (Güçlü et al., 2014)

- When there is no transition between clusters, i.e. when the centroids do not change, the processes are terminated (Saplıoğlu & Acar, 2020).

Decision trees (DT) usage in data analysis helps for classification and identification algorithms. In many modeling studies, the number of input parameters is large. It is difficult to generate models that use all of them. In such studies, DTs help to choose between parameters. In the literature, DTs are widely used due to their ability to identify and separate statistically compatible classes in data sets and to present the results in easily readable, simple and understandable tree diagrams (Saplıoğlu & Küçükderem Öztürk, 2024). Nodes and branches constitute each tree. Each node represents the features in a category to be classified and each subset defines a value that can be taken by the node (Swain & Hauska, 1977). Due to their simple analysis and sensitivity to multiple data forms, DTs have found many applications. Figure 3 shows a structure of the DTs, which are constructed using various algorithms such as Iterative Binary Decisions 3 (ID3), Successor of ID3 (C4.5), Conditional Inference Trees (CTREE), Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE), Chi-squared Automatic Interaction Detector (CHAID), Multivariate Adaptive Regression Splines (MARS), Classification and Regression Tree (CART), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), Fast, Unbiased and Efficient Statistical Tree (QUEST) (Charbuty & Abdulazeez, 2021).

In this article, CHAID algorithm is used the DT method, which is proposed by Kass (1980). This algorithm provides a visual representation of a set of

decision rules and subdivides a group based on the statistical significance of the differences using Bonferroni measures. It continues until there are no significant divisions. The Chi-square test is used for categorical data analysis. The “best” division is the one that results from a variable that presents the highest level of significance. Once a predictor is entered, it remains in the list of predictor variables for all subsequent calculations. This allows the predictor variable to have a significant interaction with two or more other predictors, which allows for a more precise determination of the predictor variable. CHAID algorithm is a method that considers all possible combinations of predictors and their categories. In addition, it finds which factors are divided into different groups and gives the most accurate estimate of the dependent variable (increasing the Chi-square). A group of cases is obtained at the terminal nodes representing the “leaves” of the tree. In terms of the dependent variable value, these cases are as homogeneous as possible (Avilés-Jurado & León, 2013; Biggs et al., 1991; Stevenson et al., 2009; Trujillano et al., 2008).

The decision tree analysis was performed using IBM SPSS Statistics 22 software. The CHAID algorithm, which allows for multiple splits, was selected as the growth method. The significance level, $\alpha = 0.05$, was defined for splitting and merging operations. The Bonferroni correction was automatically applied by the software for multiple comparisons. To ensure the model formed reliable subgroups, a minimum of 40 observations for the parent node and a minimum of 20 observations for the child node were specified. Since the CHAID algorithm grows until there are no more statistically significant splits, no additional pruning was performed. To prevent the decision tree from becoming overly complex, the maximum depth was limited to three levels.

On the other hand, multiple regression analysis (MLR) is a statistical technique used to determine the mathematical dimension of the relationship between variables that affect each other. This is formulated based on multiple regression analysis, where the value to be predicted is written as a function of all the values that affect it as in Eq. (1) (Sun & Trevor, 2018).

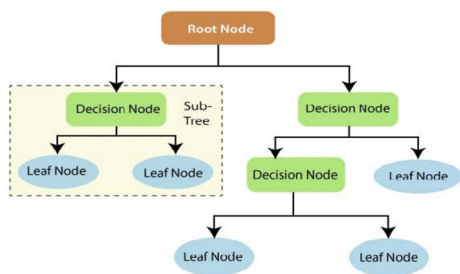


Figure 3
Schematic presentation of a decision tree (Charbuty & Abdulazeez, 2021)

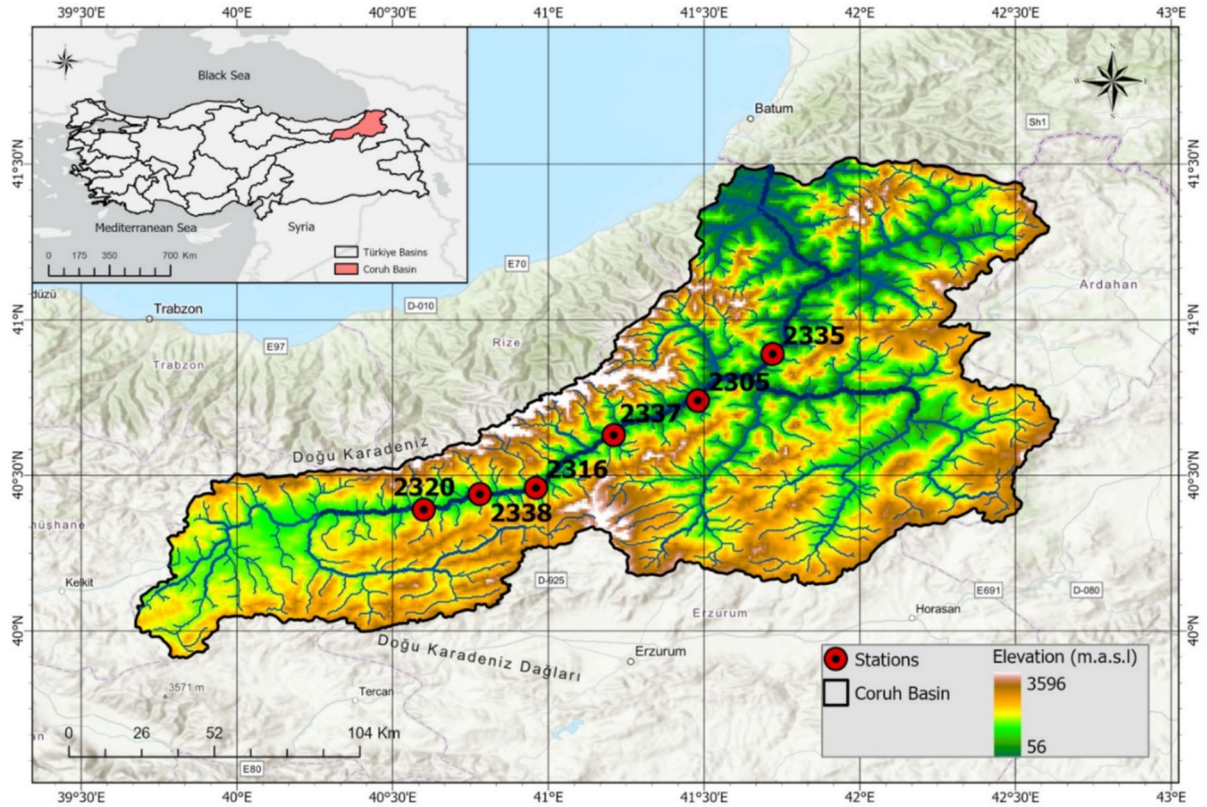


Figure 4
Locations of measurement stations in the Çoruh Basin

Table 1

Geographical information of the stations and statistical values of the data

Station No	Station Name	Latitude (N)	Longitude (E)	Max (m ³ /s)	Min (m ³ /s)	Mean (m ³ /s)	Standard Deviation	Kurtosis	Skewness
2305	Peterek	40.744	41.485	351.65	11.00	71.61	76.56	1.82	1.67
2316	Ispir	40.460	40.965	229.06	4.57	39.09	45.92	3.33	1.96
2320	Laleli	40.392	40.604	157.81	2.81	30.04	35.16	2.69	1.89
2335	Inanlı	40.889	41.716	695.39	21.10	121.86	123.99	3.22	1.88
2337	Çamlıkaya	40.629	41.206	305.48	7.96	57.35	64.93	2.43	1.80
2338	Pazaryolu	40.437	40.778	194.94	2.66	34.61	41.59	2.86	1.92

$$Y = m_1x_1 + m_2x_2 + \dots + m_nx_n + u \quad (1)$$

MLR analysis can be used when output variable, Y , is the dependent variable, x_i 's ($i = 1, 2, \dots, n$) are the independent variables, m_i 's ($i = 1, 2, \dots, n$) are the regression coefficients, and u is the error term where

the error variance for each independent variable is constant, i.e. homoscedastic (Black et al., 2010).

To evaluate the performance of the analyses, different validation models are considered such as Wilcoxon test (WT), weighted mean square error (WMSE) and coefficient of determination (R^2). WT

Table 2

Pearson correlation coefficients between streamflow data from stations

Stations	2305	2316	2320	2335	2337	2338
2305	1.000					
2316	0.972	1.000				
2320	0.977	0.982	1.000			
2335	0.996	0.989	0.984	1.000		
2337	0.977	0.984	0.998	0.985	1.000	
2338	0.991	0.956	0.961	0.987	0.961	1.000

tries to test whether the scatter of two variables is the same by taking into account the differences in two half-observations. For this purpose, the differences between the half-observations are calculated according to Eq. (2) by taking absolute values as in Eq. (3) (Lee & Kang, 2015). These absolute values are ordered from the smallest to the largest with the given ordinal numbers. The sum of the signed rows is taken as T^+ , which is the sum of the positively signed rows, and T^- , which is the sum of the negatively signed rows. The sum is given in Eq. (4) as follows according to Karagöz (2019).

$$D_i = X_i - Y_i \quad (2)$$

$$|D_i| = |X_i - Y_i| \quad (3)$$

$$T = T^+ - T^- \quad (4)$$

Here D_i is the difference between the first half data X_i and the second half data Y_i . Z_w in Eq. (5) is the value of the WT statistic for $Z_{\alpha/2}$ (for two tails), which helps determine the trend conditions. The standard deviation of the distribution is σ_T and the arithmetic mean is μ_T . Since the hypothesis is based on μ_T being zero, it is taken as zero (Wilcoxon, 1992). Hence, H_0 is expressible as $T^+ = T^-$, the sum of the positive and negative differences between the trial results is equal (Saplıoğlu & Güçlü, 2022).

$$Z_w = \frac{T - \mu_T}{\sigma_T} = \frac{T}{\sigma_T} \quad (5)$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} \quad (6)$$

In this study, the Wilcoxon signed-rank test was applied to examine whether the error values of the compared models were similar to each other. Within the scope of the test, the null hypothesis (H_0) states that there is no statistically significant difference between the error distributions of the two models; the alternative hypothesis (H_1) states that there is a significant difference between these distributions. The analysis was conducted as a two-tailed test. Clearly stating the hypotheses contributes to a clearer understanding of the role of the test within the scope of the study and how the results should be interpreted.

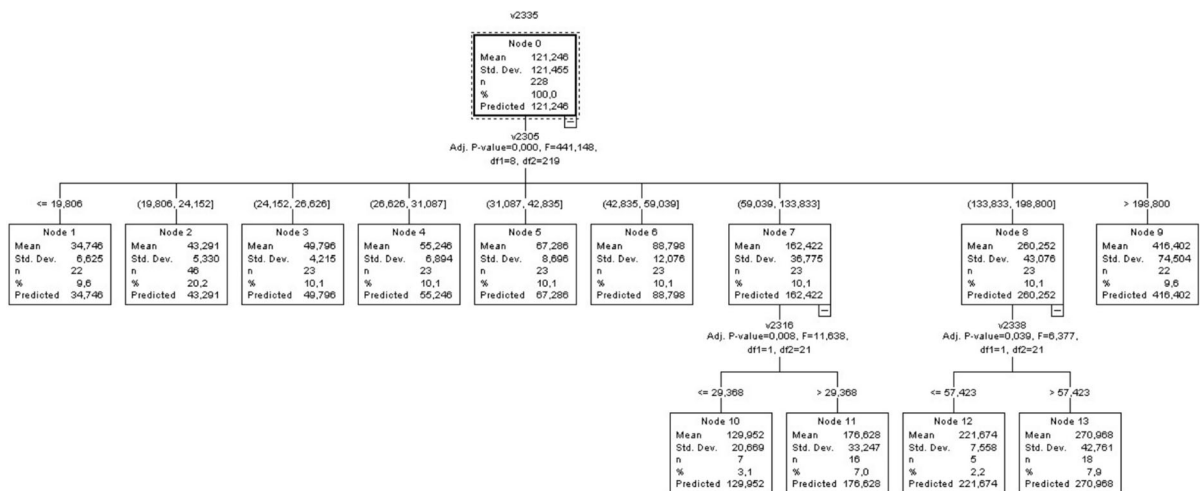


Figure 5

Implementation of a decision tree for the estimate of variables at station 2335

Table 3
Numerical characteristics of CHAID nodes and associated prediction stations

Node	Splitting variable (station)	Threshold range for splitting variable (m ³ /s)	Mean of target flow at station 2335 (m ³ /s)	Std. Dev	n	Node type
1	2305	≤ 19.806	34.746	6.625	22	Terminal
2	2305	(19.806, 24.152]	43.291	5.330	46	Terminal
3	2305	(24.152, 26.826]	49.796	4.215	23	Terminal
4	2305	(26.826, 31.087]	55.246	6.894	23	Terminal
5	2305	(31.087, 42.835]	67.286	8.696	23	Terminal
6	2305	(42.835, 59.039]	88.798	12.076	23	Terminal
7	2305 → 2316	(59.039, 133.833] (before split by 2316)	162.422	36.775	23	Internal (parent of Nodes 10–11)
8	2305 → 2338	(133.833, 198.800] (before split by 2338)	260.252	43.076	23	Internal (parent of Nodes 12–13)
9	2305	> 198.800	416.402	74.504	22	Terminal
10	2316	≤ 29.368 (within Node 7)	129.952	20.669	7	Terminal
11	2316	> 29.368 (within Node 7)	176.628	33.247	16	Terminal
12	2338	≤ 57.423 (within Node 8)	221.874	7.558	5	Terminal
13	2338	> 57.423 (within Node 8)	270.968	42.761	18	Terminal

WMSE tries to handle some of the assumptions of mean square error (MSE) in a different way. The MSE loss function treats the error incurred by the samples of all classes equally; this is valid for problems with balanced datasets. However, in the case of an imbalanced dataset, the constructed model is already built with respect to the majority classes. MSE evaluates the error incurred by the misclassification of an example in the majority class more accurately than in the minority class. However, in the proposed loss function, the error incurred by members of different classes is treated differently by a class influence term, β_n multiplied by the error of each member in the n th class. The WMSE loss is given by the following expression.

$$L = \frac{1}{M} \sum_{i=1}^M \sum_n \frac{(y_n^{(i)} - o_n^{(i)})^2}{N} \beta_n \quad (7)$$

Here, β_n is the ratio of the number of examples in the n -th class to the number of examples in the training dataset and is given by Eq. (8).

$$\beta_n = \frac{c_n}{M} \quad (8)$$

where C_n is the number of examples in class n in the training dataset. This class fitness level is higher for classes with more examples in the training dataset

than for classes with fewer examples. Therefore, classes with larger influence values contribute more error than classes with smaller influence values (Laxmi Sree & Vijava, 2018).

Regression analysis and analysis of variance both use R^2 concept, which is a measure of the amount of variance explained in the data. Therefore, the higher the R^2 value, the better the model explains the data (Bucchianico, 2008). The formula for the coefficient of determination is given in Eq. (9).

$$R^2 = 1 - \frac{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^n (Y_j - \bar{Y}_j)^2} \quad (9)$$

Here Y_j ($i = 1, 2, \dots, n$) represents the actual values, $\hat{Y}_j$ ($j = 1, 2, \dots, n$) the predicted values and $\bar{Y}_j$ ($j = 1, 2, \dots, n$) the mean of the data set (Acar & Saplıoğlu, 2022; Güler & Kandemir, 2022).

The proposed method introduces stations to the decision tree. The most effective stations are identified through analysis of the decision tree. Additionally, K-means clustering is used to determine the number of clusters for each parameter. Otherwise, too many alternatives arise in the ANFIS application. Modelling all these alternatives would be difficult and time-consuming. In this study, selecting five stations as input parameters or creating different

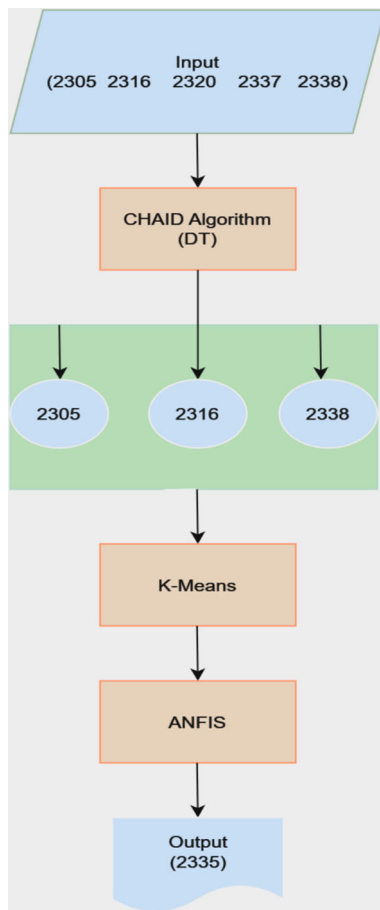


Figure 6
Flowchart of the model structure

combinations results in 120 different models. Furthermore, when different combinations are made with the number of sub-sets, the number of models increases exponentially. Therefore, it was considered necessary to use the decision tree and K-means methods to select the best model. Consequently, the accuracy of the models established based on the results obtained from these methods was examined.

3. Study Area and Data

The Çoruh basin, located on the northeastern border of Türkiye, and covers approximately 2.53% of Türkiye with an area of 19,748 km². The average annual precipitation of the basin is approximately 480 mm. In addition, the land elevation values within the

basin borders vary between 30 and 2200 m above mean sea level (Tosunoğlu, 2017; Yerdelen et al., 2010). In this study, data from 6 different stations on the Çoruh River were analyzed (Fig. 4).

To obtain the data for station number 2335, monthly flow data from stations number 2305, 2316, 2320, 2337 and 2338 are used for the water years 1993–2011. In time series-based hydrological artificial intelligence applications, it is common practice to use around 70–80% of the dataset for training and the rest for testing. For instance, Similarly, Saphioğlu and Küçükerdem (2018) evaluated 75% of the 540 observations for training and the remainder for testing. Dalkılıç and Hashimi (2020) and Kiliç and Yurtsever (2022) divided their data sets into training and testing at ratios of 70%-30% and 75%-25%, respectively. Furthermore, Latifoğlu (2022) stated that 70% of the data was used for training and 30% for testing, and Çırağ et al., (2025) stated that daily flow data was modeled using a 70%-30% training–testing split. Therefore, consistent with the literature, this study divided the 228-month dataset into 75% (171 months) for training and 25% (57 months) for testing. Descriptive statistics of the stations are presented in Table 1. The Pearson correlation coefficients of the flow data of these stations are summarized in Table 2.

Inspection of the standard deviation values of the monthly flow data in Table 1 shows that the standard deviation exceeds the average value in the time interval established for all stations. Therefore, it can be interpreted that the measurements are distributed far from the calculated mean. The skewness values of the monthly flow data in the table are between 1.67–1.96 and skewness coefficients are greater than 1. In addition, all station data are skewed to the right. Kurtosis values of all stations are positive, and a leptokurtic distribution is observed. In addition, it can be interpreted that the monthly flow data of all stations have very thick tails and many outliers due to their kurtosis values.

4. Results and Discussion

In this study, ANFIS, ANN and MLR methods are used to generate models to estimate the missing flow

Table 4
Details of the clustering performed with the K-means algorithm

Station No	Number of Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
		Number of members	Number of members	Number of members	Number of members	Number of members	Number of members
2305	2	178	50	–	–	–	–
	3	36	22	170	–	–	–
	4	161	20	18	29	–	–
	5	161	29	4	16	18	–
	6	35	26	136	4	13	14
2316	2	189	39	–	–	–	–
	3	171	38	19	–	–	–
	4	171	3	18	38	–	–
	5	40	11	30	131	16	–
	6	40	6	13	21	131	17
2338	2	42	186	–	–	–	–
	3	171	20	37	–	–	–
	4	23	169	15	21	–	–
	5	169	15	19	2	23	–
	6	134	15	22	12	36	9

data from station number 2335. Using five stations in the model increases the number of possible models and renders ANFIS models useless. Therefore, the modelling procedure is divided into two stages. In the first stage, the combination of five stations are determined using a DT to be used in the ANFIS model to detect missing flow data from station 2335. The purpose of performing this analysis using a DT is to minimize the number of scenarios that need for the generation of input variables. Otherwise, 31 scenarios may need to be generated for prediction models with five input variables. Because all possible input combinations that can be created from five station inputs require 31 (2^5-1) different subsets, excluding the null set. Therefore, if no pre-selection is made, at least 31 different input scenarios must be tested for ANFIS. Furthermore, the aforementioned number increases exponentially if different MF types are used for each subset scenario or if the number of MFs increases to higher values such as 3, 4, 5, etc. The stations are selected using a DT method generated by the CHAID algorithm. This method was used to determine the most appropriate station combinations by analyzing the relationship between the data. The CHAID algorithm allowed for the best choices to be made by considering the differences and similarities in the

flow data of the stations (Fig. 5). Additionally, the terminal nodes presented in Fig. 5 have been transferred to a table and presented numerically (Table 3).

The results show that station 2305 represents data with nine different clusters that can be used to model the data of station 2335. Station 2316 is needed to make decisions for values between 59.039 and 133.833 m³/s. In addition, for values between 133,833 and 198,800 m³/s, the decision was made using the data from station number 2338. In order to strengthen the decision structure, it is enough to analyze the values of stations number 2305, 2316 and 2338 together with the 9-2-2 subset. According to this DT, it was decided to use stations number 2305, 2316 and 2338 in the ANFIS model to recover the missing flow data of station number 2335.

In the second stage of the study, the data belonging to stations 2305, 2316 and 2335, which are used in the DT method, are first clustered using the K-means algorithm (Fig. 6). Clustering was done in two ways.

The reason for clustering in two ways is that the input data directly or indirectly affects the output data. The first is the clustering of the data itself. The other is the clustering of the output positions corresponding to the input values. The results of the K-

Table 5
Comparison of DT, K-means, and trial-and-error models

Method	Model (Input)	MF _s and Neuron number	Training Data			Testing Data		
			R ²	WMSE	Wilcoxon (Z)	R ²	WMSE	Wilcoxon (Z)
ANFIS	2305	5	0.885	11.09	−1.73	0.827	54.71	−0.92
	2305	3	0.872	11.84	−2.07	0.822	55.60	−0.80
	2316	5	0.799	20.15	−1.33	0.724	85.24	−0.74
	2316	3	0.775	21.82	−1.18	0.733	82.30	−0.66
	2320	5	0.751	23.49	−2.14	0.752	67.69	−0.17
	2320	3	0.694	24.75	−1.52	0.694	86.51	−0.47
	2337	5	0.738	25.05	−1.75	0.713	81.26	−2.11
	2337	3	0.680	31.77	−2.26	0.675	100.07	−1.39
	2338	5	0.843	14.28	−2.14	0.818	49.64	−1.08
	2338	3	0.822	16.38	−1.91	0.789	57.89	−1.15
	2305-2337	5-3	0.885	11.27	−1.29	0.802	65.69	−0.39
	2316-2337	5-2	0.870	12.40	−0.35	0.790	65.36	−3.23
	2320-2338	4-3	0.778	20.63	−1.96	0.818	51.72	−1.81
	2316-2338	4-5	0.866	12.41	−1.26	0.813	60.79	−0.53
	2337-2338	3-5	0.792	23.77	−1.70	0.771	93.68	−2.24
	2305-2316-2337	5-2-3	0.920	6.84	−0.03	0.840	60.78	−1.40
	2316-2320-2337	5-4-5	0.847	13.80	−0.60	0.815	55.01	−2.93
	2320-2337-2338	3-4-5	0.828	17.79	−0.77	0.738	116.44	−2.34
	2305-2337-2338	3-3-3	0.844	14.49	−0.22	0.703	88.27	−1.44
	2305-2320-2337	4-4-4	0.855	13.75	−0.28	0.684	92.81	−0.85
	2305-2316-2320-2338	3-3-3-3	0.943	4.73	−0.65	0.844	54.54	−1.49
	2316-2320-2337-2338	4-4-4-4	0.898	10.38	−1.27	0.761	88.43	−3.38
	2305-2320-2337-2338	5-5-5-5	0.919	8.32	−0.47	0.841	58.94	−0.39
	2305-2316-2320-2337	5-3-4-3	0.904	7.73	−0.33	0.589	144.87	−1.35
	2305-2316-2338	4-3-5-5	0.917	8.51	−0.80	0.827	57.95	−0.78
	2305-2316-2320-2337-2338	4-4-4-4-4	0.975	2.12	−0.05	0.785	80.05	−1.64
	2305-2316-2320-2337-2338	3-3-3-3-3	0.939	5.91	−0.54	0.801	67.45	−0.77
	2305-2316-2320-2337-2338	4-5-3-5-4	0.960	3.21	−0.32	0.862	50.37	−0.76
	2305-2316-2338 (DT)	9-2-2	0.979	1.98	−0.25	0.948	16.39	−2.12
	2305-2316-2338 (DT-KM)	4-5-4	0.996	0.37	−0.49	0.980	5.89	−0.94
ANN	2305	10	0.864	12.55	−1.88	0.796	71.15	−0.54
	2305	12	0.861	13.61	−1.46	0.780	62.81	−0.79
	2316	10	0.753	24.06	−1.28	0.734	82.94	−1.08
	2316	12	0.791	19.82	−1.22	0.744	79.37	−0.95
	2320	10	0.745	24.05	−1.89	0.685	88.71	−0.16
	2320	12	0.731	25.33	−1.65	0.696	86.14	−0.64
	2337	10	0.697	28.87	−1.15	0.630	117.96	−2.37
	2337	12	0.680	32.16	−2.37	0.667	102.58	−1.31
	2338	10	0.800	19.21	−1.98	0.745	71.37	−1.32
	2338	12	0.818	16.78	−1.96	0.765	67.48	−1.23
	2305-2337	14	0.847	14.96	−1.32	0.748	87.38	−0.65
	2316-2337	12	0.834	15.82	−0.41	0.774	67.55	−3.61
	2320-2338	16	0.767	21.70	−1.57	0.772	66.63	−1.91
	2316-2338	10	0.830	16.09	−1.11	0.774	82.56	−0.62
	2337-2338	14	0.772	26.08	−1.59	0.763	99.87	−1.66
	2305-2316-2337	6	0.882	10.42	−0.16	0.809	69.67	−1.50
	2316-2320-2337	8	0.826	15.81	−0.28	0.757	73.66	−2.67
	2320-2337-2338	16	0.812	20.34	−0.93	0.727	120.45	−2.42
	2305-2337-2338	14	0.829	16.00	−0.24	0.737	77.48	−1.41
	2305-2320-2337	10	0.837	15.44	−0.24	0.718	82.51	−1.07
	2305-2316-2337-2338	12	0.908	8.27	−0.31	0.873	41.78	−1.65
	2316-2320-2337-2338	16	0.862	14.10	−0.97	0.754	94.69	−3.32
	2305-2320-2337-2338	12	0.897	11.03	−0.43	0.816	67.48	−0.07
	2305-2316-2320-2337	14	0.924	6.11	−0.15	0.806	65.89	−1.04
	2305-2316-2320-2338	8	0.913	8.95	−0.67	0.856	48.25	−0.77
	2305-2316-2320-2337-2338	6	0.959	3.75	−0.07	0.842	56.37	−1.64
	2305-2316-2320-2337-2338	12	0.915	8.32	−0.46	0.798	68.71	−0.95
	2305-2316-2320-2337-2338	14	0.947	4.49	−0.19	0.843	57.84	−0.75
	2305-2316-2338 (DT)	16	0.969	2.95	−0.08	0.908	28.55	−2.10

Table 5 *continued*

Method	Model (Input)	MF _s and Neuron number	Training Data			Testing Data		
			R ²	WMSE	Wilcoxon (Z)	R ²	WMSE	Wilcoxon (Z)
MLR	2305-2316-2320-2337-2338	–	0.921	11.30	–9.66	0.888	48.78	–6.10

(DT) Denotes Decision Tree, (KM) denotes K-means, (MFs) Denotes membership function

means clustering method are shown in Table 4. Table 4 shows the 4-5-4 cluster numbers only as final results. To check the suitability of these cluster numbers, the K-means algorithm was tested on each variable group within the range $k = 2-8$. Additionally, 50 different starting points were used for each value. When examining the models that yielded the lowest SSE values, it was determined that the most suitable cluster numbers for input stations 2305, 2316, and 2338 were 4-5-4, respectively. Furthermore, the Silhouette scores also reached their highest values at these cluster numbers. In all analyses, the algorithm converged stably after 300 iterations. Therefore, the selection of the 4–5–4 structure was validated both through graphical examinations and quantitative criteria.

When the obtained results are analyzed, it is seen that the homogeneously distributed elements appear in four clusters for station 2305, five clusters for station 2316 and four clusters for station 2338. These clusters are marked in the model section of Table 5.

In the second stage, in addition to the ANFIS model generated using the DT and K-means methods, other models are also generated. These are ANFIS models with random cluster numbers, MLR model using all station data and ANN models with different neuron numbers. The biggest problem with artificial intelligence (AI) methods, especially those used in modeling, is that data is memorized, and results are given accordingly. To understand whether this is due to memorization or learning, 57 of the 228 monthly flow data obtained during modeling are used in the testing phase. The remaining data are used in learning.

After determining the effective stations for the estimation of the missing flow data coming from station number 2335, analyses are performed using

ANFIS, ANN and MLR methods. The model with the MLR method is shown in Eq. (10).

$$\begin{aligned}
 Y_{2335} = & X_{2305} * 1.964 + X_{2316} * (-0.35) + X_{2320} \\
 & * (-0.274) + X_{2337} * 0.15 + X_{2338} \\
 & * (-0.362) + 5.82
 \end{aligned}
 \quad (10)$$

Random models with 1, 2, 3, 4 and 5 inputs were created for ANFIS and ANN models. Different versions with different cluster numbers (ANFIS) and neuron numbers (ANN) are derived for each model. The evaluation values of the results obtained for all models are given in Table 5. When the results were evaluated, R², WMSE and WT values are obtained for the data separation into training and testing phases. Since the results obtained after training may be misleading, it would be much more appropriate to evaluate the values from the test results. For example, while the R² and WMSE values for some models reached reasonable levels in the training phase, the test phase results remained well below reasonable levels. In addition, while the WT Z values are accepted as accurate according to the 95% confidence interval in the training phase, it is seen that the test values of the same models do not support this accuracy. This situation is due to the memorization of the methods used. Therefore, it is very important for decision makers to separate data that have never been used in training for testing, as in this study.

The absolute value of the Z statistic was used in the Wilcoxon signed-in the testing phase, the WMSErank test. Values with $|Z| \geq 1.96$ were considered significant differences at the 5% significance level; values with $|Z| < 1.96$ were considered non-significant differences. When examining the results in Table 5, it is seen that in most ANFIS models, the $|Z|$ values are below 1.96 and the error distributions are close to each other. In contrast, exceeding the

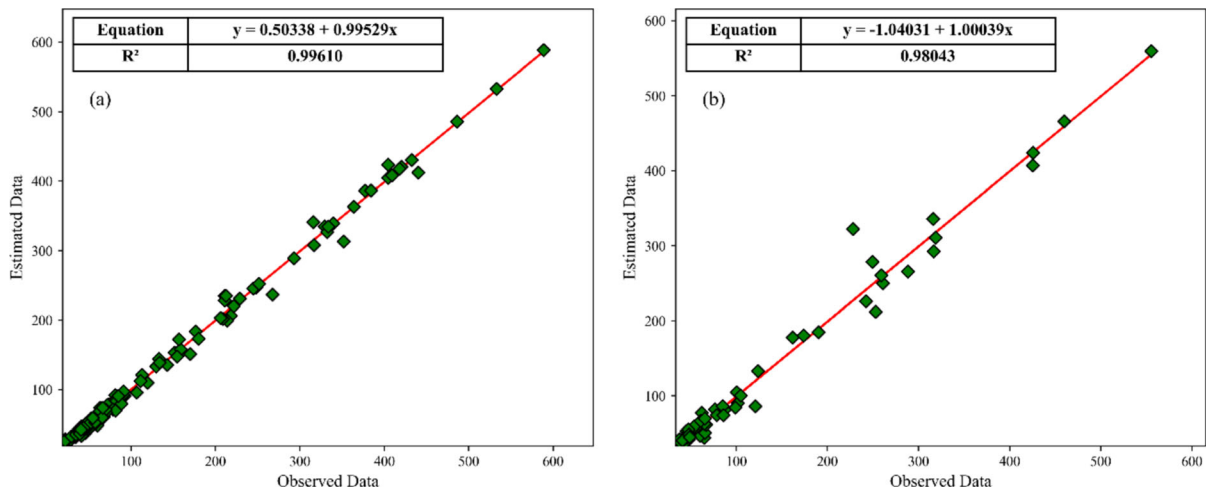


Figure 7
ANFIS (4-5-4) model scatter diagrams: **a** training, **b** test

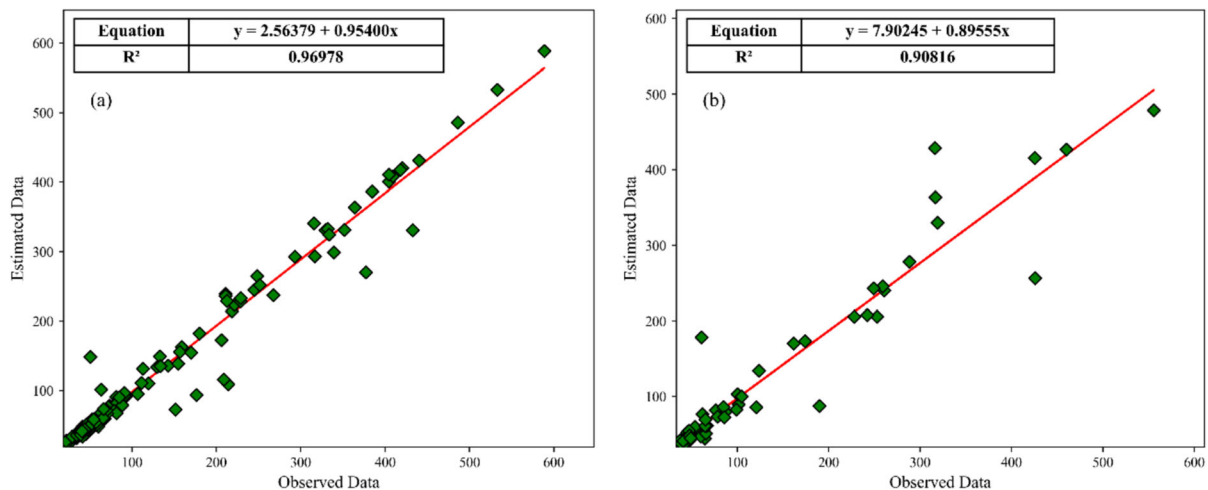


Figure 8
ANN (16 neuron) model scatter diagrams: **a** training, **b** test

threshold value in some input combinations (e.g., 2337, 2316–2337, and 2316–2320–2337) indicates that the error values in these models are statistically different from those in other models. A similar situation exists in ANN models. While no significant difference was found in most models, higher $|Z|$ values were obtained, particularly in models containing 2337. The MLR model stood out significantly from other methods with a very high Z value. In general, most ANFIS and ANN models exhibit

similar error characteristics, but significant differences emerge in certain station combinations.

The analysis of the results shows that the MLR model can be used according to the R^2 value. However, in the testing phase, the WMSE and WT Z values do not support this result at the 95% confidence interval. There are models with reasonable R^2 values in the randomly generated ANFIS and ANN models. However, these models also show weak results in terms of the WMSE criterion. The most

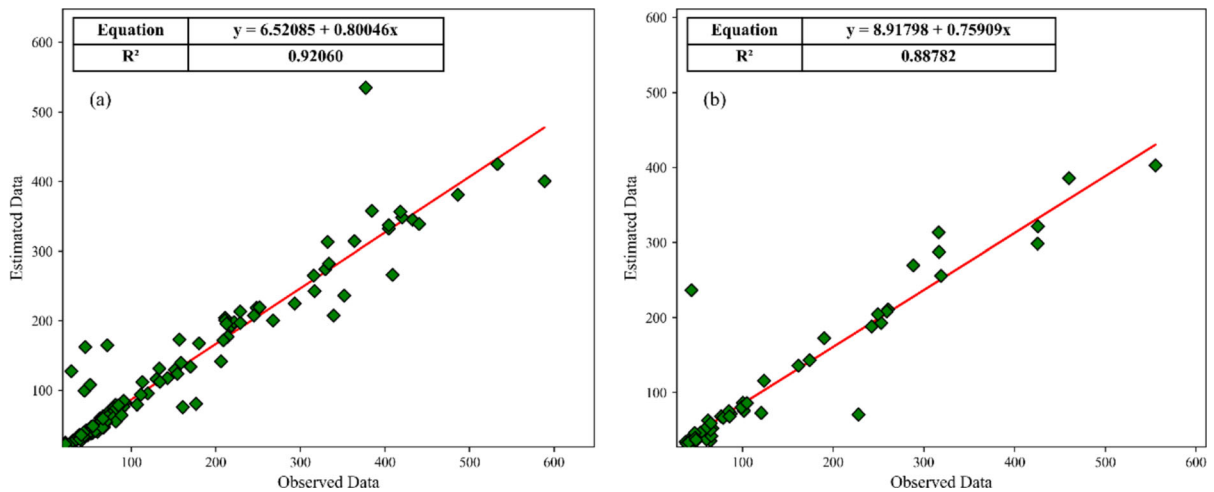


Figure 9
MLR model scatter diagrams: **a** training, **b** test

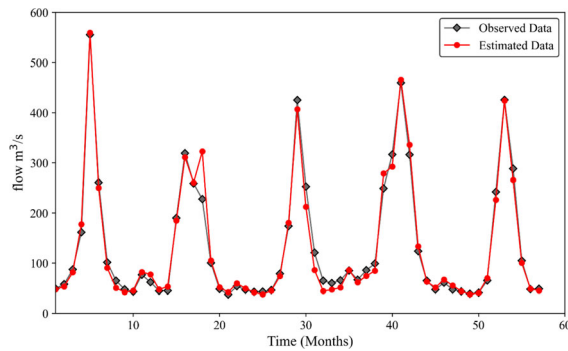


Figure 10
Comparison of observed data and estimated data of ANFIS (4-5-4) model test data

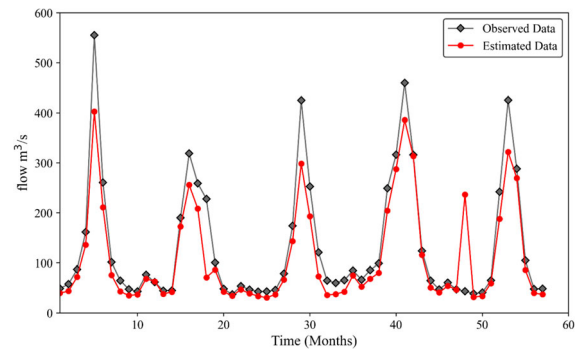


Figure 12
Comparison of observed data and estimated data of MLR model test data

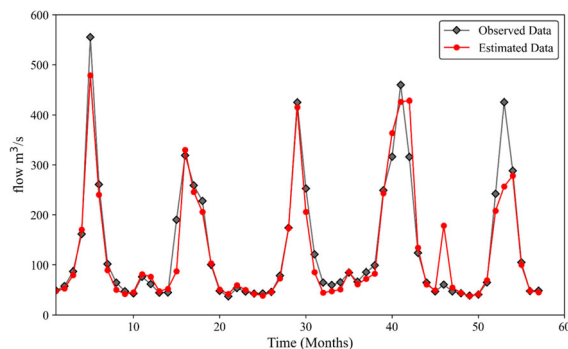


Figure 11
Comparison of observed data and estimated data of ANN (16 neuron) model test data

successful model is determined according to the highest R^2 value, lowest WMSE and WT Z values with a 95% confidence interval. According to the test-phase results in Table 5, the ANFIS-DT-KM model shows a clear superiority compared with the ANN and MLR approaches. The ANN models produced test R^2 values between 0.630 and 0.908, with WMSE values generally ranging between 28.55 and 120.45, indicating unstable performance and higher prediction errors. Although the MLR model performed more consistently than ANN, its test R^2 value (0.888) and WMSE (48.78) showed that it could not reach the accuracy of the ANFIS-DT-KM model. In contrast, the proposed 4-5-4 ANFIS-DT-KM model yielded the highest test R^2 (0.980), the lowest WMSE (5.89),

and an acceptable WT Z value (-0.94) at the 95% confidence level. These findings highlight that selecting effective inputs with DT and determining the MF numbers using K-means significantly enhance the model structure and the reliability of flow estimation. The scatter plots of the training and testing phases of the most successful model are shown in Fig. 7. Additionally, the scatter plots of the most successful models using the ANN and MLR methods are presented in Figs. 8 and 9, respectively.

In addition, Fig. 10 shows the observed and estimated flow lines for the 57 data points assigned to the testing phase. In this graph, the errors in the maximum and minimum current values are very small. In general, it can be said that a reasonable estimate is made for the entire time series. Additionally, the the observed and estimated flow lines for the most successful models using the ANN and MLR methods are presented in Figs. 11 and 12, respectively.

In this study, the effects of DT and K-means methods on the ANFIS model are investigated. In addition, ANN, MLR and ANFIS models with random number of clusters were used to compare the model performances of these methods. ANN, MLR and ANFIS models are applied to many missing data studies in the literature. Some of these studies are presented and explained in the Introduction section. The difficulties of the model in the ANFIS method are both the number of stations to be used in the analysis and the number of subsets to be selected from an infinite number of possibilities. The main purpose here is to overcome this difficulty by using both DT and K-means methods. ANFIS, ANN and MLR methods are used in many studies on missing data and hydrometeorological prediction in the literature (Rezaeianzadeh et al., 2014).

Dastorani et al. (2010) compared the normal ratio, correlation, ANN, and ANFIS methods for estimating missing flow data and found that the ANFIS method demonstrated the best performance in their study. The fact that R^2 values in the range of ≈ 0.961 – 0.997 were obtained during the testing phase shows that ANFIS is significantly more successful than both ANN and classical statistical methods. Dariane and Borhan (2024) evaluated various methods, including LR, MLR, ANN, SVM, M5 Tree, Sub-ANFIS, and FCM-ANFIS, for the estimation of missing flow data.

The study indicates that the Sub-ANFIS model provided the highest accuracy with $R^2 = 0.97$ in the Taleghan basin, while the FCM-ANFIS model yielded the most successful results with an R^2 value of 0.92 in the Latyan basin. Saplıoğlu and Küçükerdem (2018) conducted a study for the estimation of missing flow data using ANFIS, MLR and normal rate methods. In this study, the authors evaluated all stations by generating suitable models. In addition, the number of clusters required for ANFIS models is determined randomly. Acar and Saplıoğlu (2020) used ANFIS and ANN methods to estimate sediment concentrations. The three stations used in the study are also evaluated in the analyses. In addition, the number of clusters and neurons in the models of ANFIS and ANN are generated randomly. Yılmaz and Tosunoğlu (2023) used different statistical methods such as MLR, ANN, ANFIS, support vector machine (SVM), multivariate adaptive regression splines and K-nearest neighbor to estimate the missing flow data of 54 stations across Türkiye. In the study, data from nearby stations are used to estimate the missing flow data of a station and the input variables of the selective analysis models.

The proposed model also has practical significance for water management in Türkiye. In some river basins in Türkiye, flow measurements may be interrupted due to instrument malfunctions, human factors, and extreme weather events, making it difficult to obtain long-term flow series on a regular basis (Yılmaz & Tosunoğlu, 2023). This situation creates significant uncertainties, particularly in processes such as flood analysis, drought assessment, and water allocation planning. The model proposed in this study ensures that missing flow data is completed in a systematic and reliable manner. Thus, it provides a practical contribution to water resources management in basins where data deficiencies are common. The DT and K-means methods proposed in this study help to obtain more successful results in a short time by determining the effective input variables and determining the optimum number of clusters, especially for ANFIS models. These results also indicate that the DT–K-Means procedure reduces the number of ANFIS scenarios considerably, thereby accelerating the model-building process.

5. Conclusion

Among the methods used in the study, especially in ANFIS models, both the selection of input variables and the number of membership functions of these variables are quite important. If the number of selected variables and membership functions is high, many models need to be established. Trying all models is time-consuming and the best model may be missed. Therefore, before starting ANFIS modeling, a decision tree (DT) and K-means method are used for the number of input variables and membership functions.

It is shown that 31 models could be established with five input variables and the number of models could be increased by changing the number of membership functions. In addition, instead of selecting the most successful model by trial and error, the DT and K-means method are used. Furthermore, R^2 , WMSE and Wilcoxon test (WT) values are used to compare the results. In order to eliminate the error of remembering data from the training phase of artificial intelligence (AI) methods, the results obtained from the data allocated to the test phase are used in the comparisons.

In this study, data from stations 2305, 2316, 2320, 2337 and 2338 are used to estimate the missing flow data of station numbered 2335 from Çoruh Basin stations. ANFIS, ANN and MLR methods are used to estimate the missing flow in the study. When the results obtained in the study are analyzed, the most successful performance is obtained with the ANFIS model (4-5-4) based on the DT and K-means methods. Thus, it is shown that the use of the DT and K-means methods in the ANFIS model is effective, especially when there are many input variables and membership functions. In line with this result, it can be stated that the DT and K-means methods can be used to select the best model or the model closest to the best among the different models that can be created. In different ANFIS model generations in the future, the accuracy of the model can be strengthened by using the DT and K-means methods together.

Author Contributions Ramazan ACAR: Conceptualization, Data curation, Methodology, Formal analysis, Software,

Investigation, Methodology supervision, Writing-Review & editing.

Funding

Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support was received during the preparation of this manuscript.

Data Availability

The datasets used and/or analyzed during the current study are available from the authors on reasonable request.

Declarations

Conflict of Interests The authors declare no competing interests.

Ethical Approval The manuscript has not been submitted to any other journal. The proposed work is original and has not been published anywhere else.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Acar, R., & Saphioğlu, K. (2020). Detection of Sediment Transport in Streams by Using Artificial Neural Networks and Anfis

- Methods (in Turkish). *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 9(1), 437–450.
- Acar, R., & Saplıoğlu, K. (2022). Estimation of Water Hardness by ANFIS Method in which Effective Input are Determined by Multiple Regression (in Turkish). *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, 22(6), 1413–1424.
- Al Kindhi, B., Sardjono, T., Purnomo, M., & Verkerke, G. (2019). Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis. *Expert Systems with Applications*, 121, 373–381.
- Avilés-Jurado, F., & León, X. (2013). Prognostic factors in head and neck squamous cell carcinoma: Comparison of CHAID decision trees technology and Cox analysis. *Head & Neck*, 35(6), 877–883.
- Beauchamp, J., Downing, D., & Railsback, S. (1989). Comparison of Regression and Time-Series Methods for Synthesizing Missing Streamflow Records. *JAWRA Journal of the American Water Resources Association*, 25(5), 961–975.
- Biggs, D., De Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18(1), 49–62.
- Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis: A global perspective*. Pearson.
- Canchala-Nastar, T., Carvajal-Escobar, Y., Alfonso-Morales, W., Cerón, W., & Caicedo, E. (2019). Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. *Data in Brief*, 26, Article 104517.
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28.
- Çırağ, B., Acar, R., Katipoğlu, O. M., Yağanoğlu, M., Şengül, S., 2025. Reconstructing Missing Streamflow Data in Mountainous, Snow-Dominated Regions Using Optimized Hybrid Machine Learning. *Water Resources Management*, 1–23.
- Dalkılıç, H. Y., & Hashimi, S. A. (2020). Prediction of daily streamflow using artificial neural networks (ANNs), wavelet neural networks (WNNs), and adaptive neuro-fuzzy inference system (ANFIS) models. *Water Supply*, 20(4), 1396–1408.
- Dariane, A., & Borhan, M. (2024). Comparison of Classical and Machine Learning Methods in Estimation of Missing Streamflow Data. *Water Resources Management*, 38(4), 1453–1478.
- Dastorani, M., Moghadamnia, A., Piri, J., & Rico-Ramirez, M. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment*, 166, 421–434.
- Di Bucchianico, A., 2008. Coefficient of determination (R²). *n Encyclopedia of Statistics in Quality and Reliability*; Wiley: Hoboken, NJ, USA
- Elshorbagy, A., Panu, U., & Simonovic, S. (2000). Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrological Sciences Journal*, 45(6), 849–866.
- Güçlü, Y., Yeleğen, M., Dabanlı, İ., & Şişman, E. (2014). Solar irradiation estimations and comparisons by ANFIS, Angström-Prescott and dependency models. *Solar Energy*, 109, 118–124.
- Güler, E., & Kandemir, S. (2022). Estimation of CO2 Emissions by OECD Countries Using Linear and Cubic Regression Analyzes (in Turkish). *Avrupa Bilim Ve Teknoloji Dergisi*, 34, 175–180.
- Gümüş, V., & Kavşut, M. (2013). Zamanti Nehri-Ergenusağı İstasyonu Eksik Aylık Akım Verilerinin Tahmini. *Gazi University Journal of Science Part c: Design and Technology*, 1(2), 81–91.
- Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series c (applied Statistics)*, 28(1), 100–108.
- Ismail, W., Zin, W., & Ibrahim, W. (2017). Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods. *Malaysian Journal of Fundamental and Applied Sciences*, 13, 214–218.
- Jang, J. S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3), 665–685.
- Kamwaga, S., Mulungu, D., & Valimba, P. (2018). Assessment of empirical and regression methods for infilling missing stream-flow data in Little Ruaha catchment Tanzania. *Physics and Chemistry of the Earth, Parts a/b/c*, 106, 17–28.
- Karagöz, Y., 2019. SPSS-AMOS-META uygulamalı istatistiksel analizler (in Turkish). *Ankara: Nobel Yayıncılık* 1043–1044.
- Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), 119–127.
- Kilinc, H. C., & Yurtsever, A. (2022). Short-term streamflow forecasting using hybrid deep learning model based on grey wolf algorithm for hydrological time series. *Sustainability*, 14(6), 3352.
- Latifoğlu, L. (2022). A novel approach for prediction of daily streamflow discharge data using correlation based feature selection and random forest method. *International Advanced Researches and Engineering Journal*, 6(1), 1–7.
- Laxmi Sree, B., & Vijaya, M. (2018). A weighted mean square error technique to train deep belief networks for imbalanced data. *International Journal of Simulation: Systems, Science & Technology*, 19(6), 141–148.
- Lee, H., & Kang, K. (2015). Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Advances in Meteorology*, 2015(1), Article 935868.
- Lucieer, V., & Lucieer, A. (2009). Fuzzy clustering for seafloor classification. *Marine Geology*, 264(3–4), 230–241.
- Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. *In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.
- Masih, I., Uhlenbrook, S., Maskey, S., & Ahmad, M. (2010). Regionalization of a conceptual rainfall–runoff model based on similarity of the flow duration curve: A case study from the semi-arid Karkheh basin, Iran. *Journal of Hydrology*, 391(1–2), 188–201.
- McCulloch Warren, S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5, 115–133.
- Papailiou, I., Spyropoulos, F., Trichakis, I., & Karatzas, G. (2022). Artificial neural networks and multiple linear regression for filling in missing daily rainfall data. *Water*, 14(18), 2892.
- Rezaeianzadeh, M., Tabari, H., Arabi Yazdi, A., Isik, S., & Kalin, L. (2014). Flood flow forecasting using ANN, ANFIS and regression models. *Neural Computing and Applications*, 25, 25–37.
- Saplıoğlu, K., & Kucukerdem Ozturk, T. (2024). Effect of decision tree in the ANFIS models: an example of completing missing data. *Russian Meteorology and Hydrology*, 49(5), 435–445.
- Saplıoğlu, K., & Kucukerdem, T. (2018). Estimation of missing streamflow data using ANFIS models and determination of the number of datasets for ANFIS: The case of Yeşilirmak River.

- Applied Ecology and Environmental Research*, 16(3), 3583–3594.
- Saplıoğlu, K., & Acar, R. (2020). Detection of Sediment Transport with Artificial Intelligence Models Formed by Using K-Means Clustering Algorithm (in Turkish). *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 9(1), 306–322.
- Saplıoğlu, K., & Çimen, M. (2010). Predicting of Daily Precipitation Using Artificial Neural Network (in Turkish). *Mühendislik Bilimleri Ve Tasarım Dergisi*, 1(1), 14–21.
- Saplıoğlu, K., & Güçlü, Y. (2022). Combination of Wilcoxon test and scatter diagram for trend analysis of hydrological data. *Journal of Hydrology*, 612, Article 128132.
- Sharma, V., & Yuden, K. (2021). Imputing missing data in hydrology using machine learning models. *International Journal of Engineering Resources and Technology*, 10(2021), 78–82.
- Souza, G., Bello, I., Corrêa, F., & Oliveira, L. (2020). Artificial neural networks for filling missing streamflow data in Rio do carmo basin, minas gerais, Brazil. *Brazilian Archives of Biology and Technology*, 63, Article e20180522.
- Stevenson, J., Mahoney, E., Walker, P., & Everson, P. (2009). Prediction of sex based on five skull traits using decision analysis (CHAID). *American Journal of Physical Anthropology: THE Official Publication of the American Association of Physical Anthropologists*, 139(3), 434–441.
- Sun, W., & Trevor, B. (2018). Multiple model combination methods for annual maximum water level prediction during river ice breakup. *Hydrological Processes*, 32(3), 421–435.
- Swain, J., & Patra, K. (2017). Streamflow estimation in ungauged catchments using regional flow duration curve: Comparative study. *Journal of Hydrologic Engineering*, 22(7), 04017010.
- Swain, P., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3), 142–147.
- Şen, Z. (2004). *Principals of artificial neural networks*. Turkey Water Foundation Press.
- Terzi, Ö., & Köse, M. (2012). Flow Forecasting of Göksu River With Artificial Neural Networks Method (in Turkish). *International Journal of Technological Sciences*, 4(3), 1–7.
- Tfwala, S., Wang, Y., & Lin, Y. (2013). Prediction of missing flow records using multilayer perceptron and coactive neurofuzzy inference system. *The Scientific World Journal*, 2013(1), Article 584516.
- Tosunoğlu, F. (2017). Trend analysis of daily maximum rainfall series in Çoruh Basin, Turkey. *Journal of the Institute of Science and Technology*, 7(1), 195–205.
- Tosunoğlu, F., İspirli, M., Gürbüz, F., & Şengül, S. (2017). Estimation of Missing Streamflow Records in the Euphrates Basin using Flow Duration Curves and Regression Models (in Turkish). *Journal of the Institute of Science and Technology*, 7(4), 85–94.
- Trujillano, J., Sarria-Santamera, A., Esquerda, A., Badia, M., Palma, M., & March, J. (2008). Approach to the methodology of classification and regression trees. *Gaceta Sanitaria*, 22(1), 65–72.
- Turhan, E., & Çağatay, H. (2016). Using of Artificial Neural Network (ANN) for Setting Estimation Model of Missing Flow Data: Asi River-Demirköprü Flow Observation Station (FOS) (in Turkish). *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 31(1), 93–106.
- Vriend, S., Van Gaans, P., Middelburg, J., & De Nijs, A. (1988). The application of fuzzy c-means cluster analysis and non-linear mapping to geochemical datasets: Examples from Portugal. *Applied Geochemistry*, 3(2), 213–224.
- Widrow, B., & Hoff, M. (1960). *IRE WESCON convention record* (pp. 96–104). IRE.
- Wilcoxon, F. (1992). Individual comparisons by ranking methods. *Breakthroughs in statistics Methodology and distribution* (pp. 196–202). Springer New York.
- Yerdelen, C., Karimi, Y., & Kahya, E. (2010). Frequency analysis of mean monthly streamflow in Çoruh Basin. Turkey. *Fresenius Environmental Bulletin*, 19(7), 1300–1311.
- Yılmaz, M., & Tosunoğlu, F. (2023). Performance evaluation of various data driven techniques for infilling missing streamflow data across Turkey's rivers (in Turkish). *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen Ve Mühendislik Dergisi*, 25(74), 317–328.
- Zeraatpisheh, M., Ayoubi, S., Brungard, C., & Finke, P. (2019). Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran. *Geoderma*, 340, 249–258.