Water Resources Management (2024) 38:1453–1478
https://doi.org/10.1007/s11269-023-03730-7

Comparison of Classical and Machine Learning Methods in Estimation of Missing Streamflow Data

A. B. Dariane[1] · M. I. Borhan[1]

Received: 22 April 2023 / Accepted: 27 December 2023 / Published online: 15 January 2024

Abstract

Recovering missing data and access to a complete and accurate streamflow data is of great importance in water resources management. This article aims to comparatively investigate the application of different classical and machine learning-based methods in recovering missing streamflow data in three mountainous basins in northern Iran using 26 years of data duration extending from 1991 to 2017. These include Taleghan, Karaj, and Latyan basins that provide municipal water for the capital Tehran. Two periods of artificial gaps of data were considered to avoid possible duration-based impacts that may affect the results. For this purpose, several methods are investigated including simple and multiple linear regressions (LR & MLR), artificial neural network (ANN) with five different structures, support vector regression (SVR), M5 tree and two Adaptive Neuro-Fuzzy Inference System (ANFIS) comprising Subtractive (Sub-ANFIS) and fuzzy C-means (FCM-ANFIS) classification. Although these methods have been used in different problems in the past, but the comparison of all these methods and the application of ANFIS using two clustering methods in missing data is new. Overall, it was noticed that machine learning-based methods yield better outputs. For instance, in the Taleghan basin and in the gap during 2014–2017 period it shows that the evaluation criteria of Root Mean Square Error (RMSE), Nash–Sutcliffe Index (NSE) and Coefficient of Determination (R^2) for the Sub-ANFIS method are 1.67 m^3/s, 0.96 and 0.97, respectively, while these values for the LR are 3.46 m^3/s, 0.83 and 0.87 respectively. Also, in Latyan basin during the gap of 1991–1994, FCM-ANFIS was found to be the best method to recover the missing monthly flow data with RMSE, NSE and R^2 criteria as 3.17 m^3/s, 0.88 and 0.92, respectively. In addition, results indicated that using the seasonal index in the artificial neural network model improves the estimations. Finally, a Social Choice (SC) method using the Borda count was employed to evaluate the overall performance of all methods.

Keywords Missing data · Streamflow · Recovering methods · Classical methods · Machine learning algorithms · Social choice

✉ A. B. Dariane
borhani@kntu.ac.ir

[1] Dept. of Civil Eng., K.N. Toosi University of Technology, Tehran, Iran

1 Introduction

Engineering studies in hydrology and climatology require the existence of sufficient and reliable data such as rainfall, temperature, streamflow, and evapotranspiration (Mwale et al. 2012). As such, the availability of a reliable source of complete and correct sets of hydrological data is necessary for the development of various purposes, including water supply, construction of hydropower plants, flood protection, hydraulic structure design, hydrological modeling and climate change projects (Kamwaga et al. 2018).

However, if the order of observing the data information in the time series of hydrological data is interrupted, the problem of time series missing data arises (Yozgatligil et al. 2013) which is a global problem (Dembélé et al. 2019; Lai and Kuok 2019). In developing countries, data often exhibits various deficiencies such as inadequate statistical period, poor measurement quality, and missing data, as highlighted by Ilunga and Stephenson (2005), Mwale et al. (2012), and Radi et al. (2015). Additionally, challenges like lack of awareness, insufficient staff training, and limited focus on measurement and data processing in hydrological studies further exacerbate the issue of estimating missing data in developing countries.This lack of data can be caused by various factors such as malfunction in measuring instruments and monitoring equipment, absence of supervisor and expert, human errors during data entry, manual collection instruments, limited access to measurement locations, lack of sufficient number of measuring stations, extreme weather conditions, lack of financial resources, political wars, accidental loss of data and effects of natural phenomena such as earthquakes, landslides, hurricanes, etc. (Elshorbagy et al. 2000; Harvey et al. 2010; Londhe et al. 2015; Kamwaga et al. 2018; Aguilera et al. 2020).

Missing data is a very prevalent problem in climatology and their presence affects the quality of the final results in hydrological studies and water resources management and causes unreliable analysis (Tencaliec et al. 2015; Aieb et al. 2019; Fagandini et al. 2023). As a result, data recovery and infilling the gaps in the time series of hydrological data is the essential and primary step in planning, designing and operating water resources systems and various hydrological studies.

In recent years, many studies have been carried out to demonstrate methods to recover missing data of various hydro-climatological time series including precipitation (Coulibaly and Evora 2007; Faramarzzadeh et al. 2023), temperature (Xia et al. 1999), evapotranspiration (Abudu et al. 2010) and streamflow (Giustarini et al. 2016; Baddoo et al. 2021).

Streamflow data recovery methods may consist of simple classical methods where they have been of interest to researchers for a long time. For example, Gyau-Boakye and Schultz (1994) used several techniques, including interpolation, recursive regression, autoregressive and nonlinear methods, to fill in missing streamflow data in three different catchments in Ghana. The results showed that the choice of methods can depend on several factors, such as the season, the studied area, and the length of the data gap. On average, regression models can provide good results, but in general, the simple methods yield larger deviations between the observed and predicted streamflow for long duration gaps. Harvey et al. (2012) used 15 simple techniques, based on regression, scaling and equipercentile approaches to infilling missing streamflow data in the UK. The results of this study indicated the superiority of regression-based methods over other simple methods. Tencaliec et al. (2015) proposed a hybrid method of regression and autoregressive integrated moving average (ARIMA) called Dynamic Regression Model to recover missing streamflow data. The results showed that this model provides reliable estimates for the missing data for the Durance watershed located in the South-East of France. Kamwaga et al. (2018) investigated empirical and regression methods to estimate streamflow data in the Little Ruaha basin located in Tanzania. The methods used included simple and multiple linear regression, rainfall-runoff relationship using double mass curve technique, flow duration matching and drainage-area ratio. The calibration and validation results showed that the MLR method did better than other methods in recovering missing streamflow data.

On the other hand, methods based on machine learning (ML) have gained popularity in recent decades in hydrology and water resources management and have been widely used to study the droughts (Khan et al. 2020), rainfall-runoff modeling (Mohammadi 2021), forecasting flood (Mosavi et al. 2018) and groundwater problems (Cai et al. 2021). ML methods are also used in the recovery of missing data (Zhou et al. 2023).

Ng et al. (2009) developed a hybrid model of generalized regression neural network and genetic algorithm (GRNN-GA) to recover missing streamflow data. Their results showed that this method is more successful than the k-nearest neighbor (KNN) and multiple imputation (MI) methods in infilling streamflow data of Saugeen River in Canada. Dastorani et al. (2010) estimated the missing streamflow data in four different basins in Iran using two classical methods including the normal ratio (NR) and correlation approach and two ML methods of artificial neural network and adaptive neural fuzzy inference system. The results indicated that although in some cases all four methods provide good predictions, the ANFIS model has a better ability to predict missing streamflow data, especially in the stations located in the arid region with heterogeneous data. Also, ANN model showed better performance than the classical methods for estimating missing data. Bahrami et al. (2010) in order to estimate the missing maximum annual streamflow data in the Sefidrood basin in the northern Iran used the data of 16 hydrometric stations and a 28-year period time series. They showed that the ability of ANN model is higher than nonlinear regression (NLR) in recovering missing data. Mwale et al. (2012) used the self-organizing maps (SOM) approach, which is a form of unsupervised ANN, to fill the gaps in rainfall and streamflow data in the Shire River basin of Malawi. Ergün and Demirel (2023) used a distributed hydrological model and remote sensing data to estimate missing streamflow data. The result showed that if the calibration length is appropriate, this model has a good performance in filling the data gap. Others also compared the application of some classic and machine learning methods (Souza et al. 2020; Arriagada et al. 2021).

As it can be seen, different methods have been employed in recovering the missing data. These are ranging from very simple to relatively sophisticated methods. Nevertheless, a complete evaluation of different methods is necessary which is essential in developing countries with great data limitations. Moreover, recent developments in machine learning algorithms demand a thorough evaluation and comparison of these methods with the more common classical and usually regression based methods. In addition, this article seeks to answer the question “Are ML methods more efficient than classical methods in recovering missing streamflow data?” Finally, this study investigates the effect of some parameters such as the seasonal index to determine the suitable methods with high efficiency. Moreover, a social choice (SC) approach is introduced and used to determine the superior methods among the lists of solution results.

2 Case Study and Data Set

In this article, three mountainous basins of Taleghan, Karaj, and Latyan that provide municipal water demand for the capital Tehran in northern Iran were used to evaluate the recovery methods for missing streamflow data. The mountainous terrain and complex topography of these basins often present challenges in obtaining accurate streamflow data, leading to gaps in the time series. In such situations, the proposed methodology can be applied to resolve this issue within these basins (refer to Fig. 1).

Taleghan basin is located on the southern slopes of the Alborz Mountain. This basin with an area of about 1171 km^2 has maximum and minimum heights of 4300 m and 1398 m above sea level (masl), respectively (average height of 2840 m). And it is placed between 36° 5´ to 36° 19´ N latitudes and between 50° 25´ to 51° 11´E longitudes coordinates. The existence of the Taleghan Reservoir in this basin, as one of the important sources in the supply of drinking water for Tehran and agricultural water demands in the downstream areas, has caused the importance of this basin in this region. This basin is placed in the semi-humid group with an average annual rainfall of 400 mm and an annual temperature of 11.4 °C. Almost half of this catchment area has a slope above 40%, with weak and moderate vegetation cover.

The Karaj basin is located on the southern slopes of Central Alborz Mountain, adjacent to Taleghan, between latitudes coordinates 35° 53′ N and 36° 10′ N and between longitudes coordinates 50° 3′ E and 51° 35′ E, upstream of Amirkabir Reservoir. Due to the existence of this dam, the Karaj basin is very essential in supplying water demand of Tehran and Karaj cities. This basin, has an area of about 1088 km^2, and an average annual rainfall and temperature of 247 mm and 11.4 °C, and is located in north-western Tehran. The maximum height of the basin reaches more than 4352 masl, and its lowest level is about 1295 masl in the dam site.

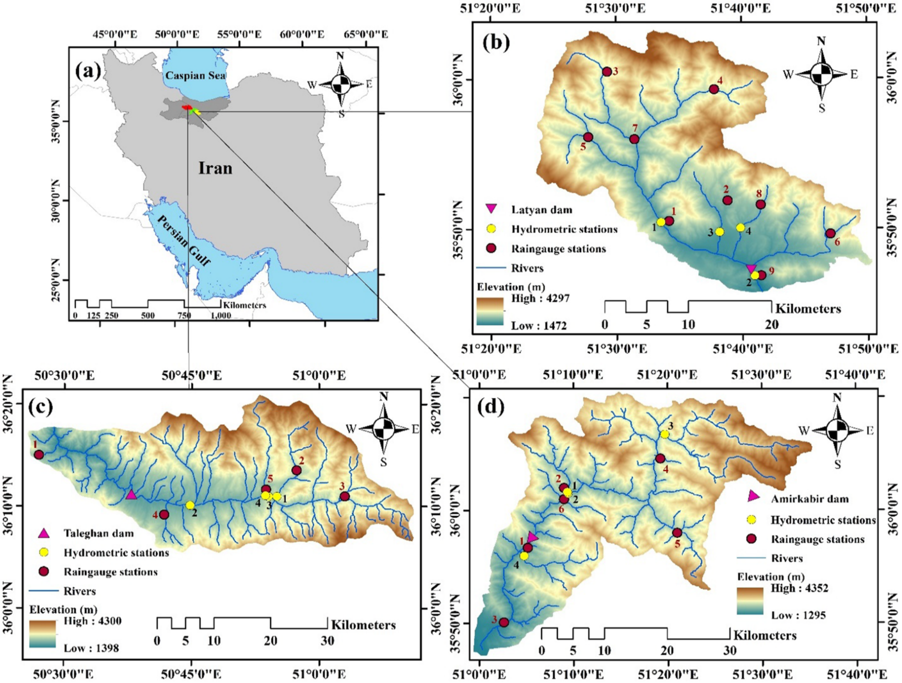

Fig. 1 Location maps showing the approximate position of the study area: **a** in Iran, and detailed maps of **b** Latyan, **c** Taleghan, and **d** Karaj basins

Latyan basin has an area of 728 km^2 and is located adjacent to Karaj basin between 35° 45′ N and 36° 6′ N latitudes and 51° 22′ E and 51° 55′ E longitudes. It has an average annual rainfall and temperature of 320 mm and 11.4 °C. The maximum and minimum altitude in this basin is 4297 and 1472 masl; respectively.

The Latyan Reservoir at the outlet of basin provides part of municipal water demand for Tehran. All the three reservoirs in the study area also provide water for the agricultural fields in the downstream regions.

In order to recover the missing streamflow data, the streamflow data of the hydrometric stations and rain gauge stations of the three basins were used as indicated in Tables 1 and 2. For this purpose, a common 26 years data duration extending from 1991–1992 to 2016–2017 water years was selected for the hydrometric and rainfall stations of the basins.

3 Methodology

This section contains descriptions of streamflow data quality control tests including Standard Normal Homogeneity (SNH) Test and Mann–Kendall (M–K) Test plus recovering models for missing monthly streamflow data such as LR, MLR, ANN, SVR, M5, ANFIS models. In addition, the evaluation criteria, the SC method to determine the best model in recovering the missing data are also presented.

3.1 Statistics Quality Control

3.1.1 Standard Normal Homogeneity Test

The SNH test is one of the most widely used homogeneity tests in hydrologic research, which was developed by Alexandersson (1986). The first step in evaluating the effects of climate change and human activities on the streamflow is to find a natural, reliable and

Table 1 The specifications of hydrometric stations

Station Number	Station Name	Latitude (N)	Longitude (E)	Elevation (m)	Mean Monthly Streamflow (m^3/s)	Basin
1	Mehran	36° 11′	50° 55′	2010	2.16	Taleghan
2	Galinak	36° 10′	55° 44′	1789	12.94	
3	Jostan	36° 11′	50° 54′	1981	8.01	
4	Alizan	36° 12′	50° 54′	1944	1.60	
1	Sira1	36° 01′	51° 09′	1819	12.4	Karaj
2	Sira2	36° 02′	51° 09′	1833	1.70	
3	Gachsar	36° 07′	51° 20′	2249	3.98	
4	Haft Cheshmeh	35° 56′	51° 05′	1668	1.01	
1	Rodak	35° 51′	51° 33′	1795	10.30	Latyan
2	Latyan	35° 47′	51° 41′	1547	6.13	
3	Najarkela	35° 49′	51° 38′	1703	0.78	
4	Naroon	35° 50′	51° 00′	1786	0.36	

Table 2 The specifications of rain gauge stations

Station Number	Station Name	Latitude (N)	Longitude (E)	Elevation (m)	Mean Monthly Rainfall (mm)	Basin
1	Angeh	36° 15′	50° 27′	1945	35.10	Taleghan
2	Dizan	36° 13′	50° 57′	2229	68.51	
3	Gatedeh	36° 10′	51° 03′	2497	89.10	
4	Jostan	36° 11′	50° 53′	1989	45.25	
5	Zidasht	36° 09′	50° 41′	1962	38.63	
1	Sade Amirkabir	36° 56′	51° 05′	1612	34.10	Karaj
2	Sira	36° 02′	51° 09′	2130	49.53	
3	Bilaghan	36° 50′	51° 02′	1404	23.52	
4	Nesa	36° 04′	51° 19′	2191	56.87	
5	Shahrestanak	35° 58′	51° 21′	2276	52.71	
6	Morood	36° 01′	51° 09′	1827	46.89	
1	Rodak	35° 50′	51° 33′	1705	49.04	Latyan
2	Kandsofla	35° 51′	51° 36′	1945	42.89	
3	Shemshak	36° 00′	51° 29′	2494	59.09	
4	Garmabdar	35° 59′	51° 37′	2435	56.23	
5	Ahar	35° 56′	51° 27′	2087	54.92	
6	Lavasanbozog	35° 49′	51° 47′	2195	45.01	
7	Fasham	35° 55′	51° 31′	1966	60.31	
8	Afjeh	35° 51′	51° 41′	2048	55.23	
9	Latyan	35° 46′	51° 41′	1580	34.47	

trend-free period in the data time series, so that there are minimal human activity and artificial changes (Mahmood and Jia 2019). SNH test can find and report the time of discontinuity and occurrence of heterogeneity in the data series. Equation (1) is employed to discover breaking or change points in the time series $x_1, x_2, \ldots, x_n$:

$$T_y = y\overline{z}_1 + (n-y)\overline{z}_2 \, for \; y = 1, 2, \ldots, n \tag{1}$$

where,

$$\overline{z}_1 = \frac{1}{y}\sum\nolimits_{i=1}^{n} \frac{(x_i - \overline{x})}{s} \; and \; \overline{z}_2 = \frac{1}{n-y}\sum\nolimits_{i=y+1}^{n} \frac{(x_i - \overline{x})}{s} \tag{2}$$

The statistic T_y is obtained to compare the first y observations average with the average of n-y observations. The maximum value of T_y is the breaking point in the time series.

The null hypothesis (H_0) (no change point) is rejected if the test statistic (T_{max}) is greater than the critical value (which dependents on the numbers of sample).

$$T_{max} = \max_{1 \le y \le n} T_y \tag{3}$$

Here, x_i represents the test variable for the year i between 1 and n. $\overline{x}$ and s are referred to mean and standard deviation of a time series, respectively.

3.1.2 Mann-Kendall Test

Mann–Kendall test (Mann 1945; Kendall 1948) recommended by the World Meteorological Organization, is widely used to determine the time trend of hydrological and meteorological data (Abghari et al. 2013; Gebremicael et al. 2017; Ali et al. 2019).

The M–K test statistic (S) for streamflow can be calculated using the Eqs. (4) and (5):

$$S = \sum_{i=1}^{n-1} \cdot \sum_{j=i+1}^{n} sign(x_j - x_i) \tag{4}$$

where,

$$sign(x_j - x_i) = \begin{cases} +1 & if\ (x_j - x_i) > 0 \\ 0 & if\ (x_j - x_i) = 0 \\ -1 & if\ (x_j - x_i) < 0 \end{cases} \tag{5}$$

Here, x_i and x_j are the data values at time i and, respectively, and n represents the length of the data set. The positive value of S indicates an increasing trend, and vice versa.

The variance of S is calculated by

$$Var(S) = \frac{s(n-1)(2n+5) - \sum_{i=1}^{P} t_i(t_i - 1)(2t_i + 5)}{18} \tag{6}$$

where, P is the number of tied groups, t_i is the number of data value in the P^{th} group.

Then the standard Z value is calculated according to Eq. (7).

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \frac{S+1}{\sqrt{Var(S)}} & if\ S < 0 \end{cases} \tag{7}$$

The calculated standardized Z value is compared with the standard normal distribution table with two-tailed confidence levels $\alpha = 0.05$. Null hypothesis (H_0) is rejected if $|Z| > \left|Z_{1-\alpha/2}\right|$, otherwise, H_0 is accepted meaning that there is no trend in the time series.

3.2 Recovery Methods

3.2.1 Classical Methods

Linear Regression LR is the simplest method to transfer hydrological information between two gauging stations (Salas 1993). In this method, the correlation coefficients between the target station and all neighboring stations are first calculated and then ranked. Finally, the missing data is estimated using the linear regression equation with the station having the highest correlation coefficient (Eq. 8).

$$Y = \beta_0 + \beta_1 x \tag{8}$$

Multiple Linear Regression Finding the correct relationship between a dependent variable and several independent variables is a problem in statistical analysis (Tabari et al. 2011). MLR, as the general form of LR, is a beneficial and accurate statistical technique that expresses the relationships between a dependent variable and several independent variables by fitting a linear equation. The linear equation of multiple linear regression appears in the form of Eq. (9).

$$Y = \beta_0 + \sum\nolimits_{i=1}^{n} \beta_i x_i \tag{9}$$

where, Y is the dependent variable, x_i are the independent variables, β_0 and β_i are the parameters of the model and n is the total number of independent variables.

3.2.2 Machine Learning Methods

Artificial Neural Network The artificial neural network is a mathematical model inspired by the neural network in the human brain and was first introduced by McCulloch and Pitts (1943). ANN has the ability to learn and recognize the nonlinear relationship between input and output parameters, solving complex problems on large scales. This ability of the ANN makes it attractive for hydrological modeling and water resources studies (Belayneh et al. 2016). There are different structures ANN including multilayer perceptron (MLP), radial basis function networks (RBF) and recurrent neural networks (RNN) (Khazaee Poul et al. 2019) but the most common ANN structure used in engineering science, especially in hydrology research, is the MLP (Mekanik et al. 2013; Ahmadi et al. 2019). MLP is a feedforward neural network consisting of three layers: an input layer, one or two hidden layers, and an output layer (Fig. 2) and information moves forward from the input layer to the output layer (Khan et al. 2021).

According to Kolmogorov theorem, the two hidden layers in the neural network can model any problem, provided that the number of neurons in the hidden layer is sufficient (MacLeod 1999). However, in most hydrological systems, it is sufficient to use a hidden layer with the appropriate number of neurons (Dariane and Karami 2014).

Equation (10) represents the MLP neural network.

$$y_j = f_2\left[\sum\nolimits_{k=1}^{K} w_j f_1\left(\sum\nolimits_{i=1}^{I} w_k x_i + b_k\right) + b_j\right] \tag{10}$$

where, x_i and y_j are the input and output of the neural network, respectively. Indexes i, k and j refer to the input, hidden and output layers, respectively. w_k is the weight between neurons in the input and hidden layers and w_j is the weight between neurons in the hidden and output layers. b_k and b_j are the biases associated with the neurons of the hidden and output layers, respectively. f_1 and f_2 are the activation functions of hidden and output layers, respectively. In this study, due to nonlinear relationships in hydrology, the sigmoid activation function (Eq. 11) was used in the hidden layer (Uysal and Şorman 2017) and the linear transfer function (Eq. 12) in the output layer (Tongal and Booij 2018).

$$f_1(x) = \frac{2}{1 + \exp(-2x)} - 1 \tag{11}$$

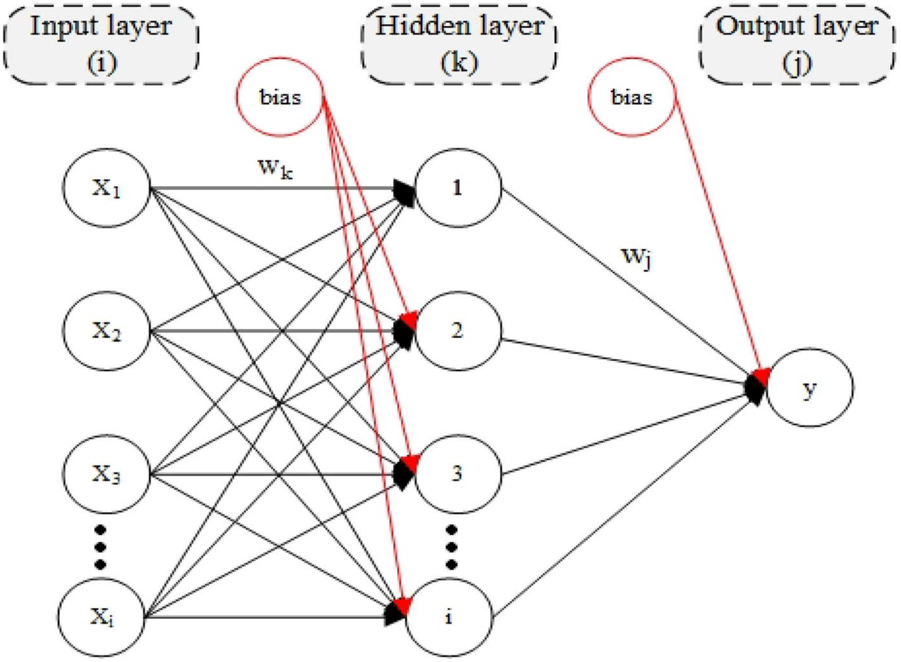

Fig. 2 A three layer ANN structure

$$f_2(x) = x \tag{12}$$

Also, the Levenberg–Marquardt (LM) algorithm is used to train the ANN. This algorithm is the most common optimization algorithm used in ANN, which is suitable for nonlinear and dynamic relationships of hydrological processes and can perform better than the gradient back-propagation algorithm (Asadi et al. 2013; Tongal and Booij 2018).

In this study, in order to recover streamflow missing data by ANN, a three-layer MLP network was built consisting of one hidden layer. The number of neurons in the hidden layer is determined by the trial and error method. Also, five different structures were considered in the input data to the neural network. The purpose of using these different structures is to show the effect of rainfall and seasonal index on infilling missing streamflow data.

Flow in a basin has an annual and seasonal cycle. Entering the information related to this cycle in the input of the neural network can improve its performance by providing more information to the model. This information was done by entering two time series (which represent 12 months of the annual cycle) according to the oscillation of two sine and cosine curves (Fig. 3) in the neural network (Nilsson et al. 2006).

Finally, the five input structures of the neural network are as it follows:

ANN(1): Using of monthly streamflow data of neighboring hydrometric stations in the basin.
ANN(2): Adding the seasonal index to ANN(1).
ANN(3): Using the monthly rainfall data of all stations in the basin.
ANN(4): Adding the seasonal index to ANN(3).

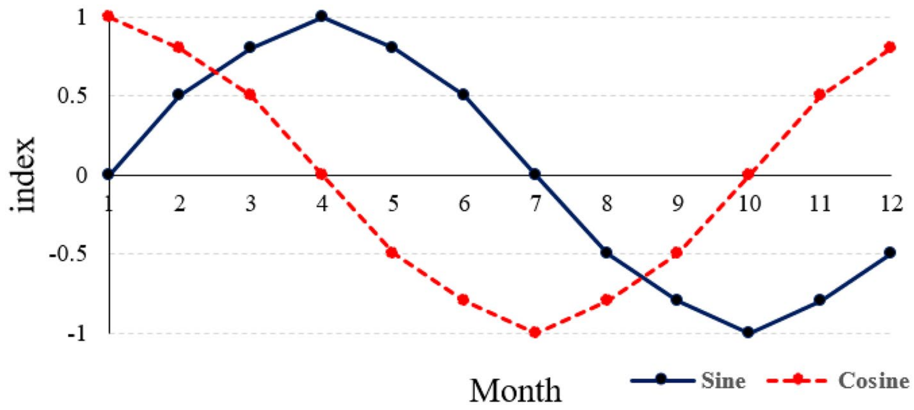

Fig. 3 Cyclic seasonal index

ANN(5): Using the monthly rainfall data of all stations and monthly streamflow data of neighboring stations in the basin along with the seasonal index (i.e., a combination of ANN2&4).

Support Vector Regression SVM is one of the most popular machine learning algorithms developed by Vapnik (1998) and has wide applications in hydrological research (Raghavendra and Deka 2014). Support Vector Regression which is an observation-based modeling technique developed based on statistical learning theory uses the principle of SVM to solve regression problems. In other words, SVR uses the principle of structural risk minimization to describe the pattern between the predictor and predicted values.

If the data set is$X = \{x_i, y_i : i = 1, \dots, n\}$, where x_i are the input vector,y_i is the target vector and n is the size of the data set. Then the general function of SVR is according to Eq. (13).

$$f(x_i) = w.\phi(x_i) + b \tag{13}$$

where, w is the weight vector, b is the bias, and $\phi(.)$ is a non-linear transformation function to map the input space into the feature space. The target of SVR is to find the values of w and b so that the values of $f(x_i)$ can be determined by minimizing the empirical risk for regression efficiency. For this purpose, it uses the loss function $L_\varepsilon(y_i, f(x_i))$, where L_ε is defined as Vapnik's ε—insensitive loss function (Vapnik 1998, 1999).

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} 0 & \text{if } \left|(y_i - f(x_i))\right| \le \varepsilon \\ \left|(y_i - f(x_i))\right| - \varepsilon & \textit{otherwise} \end{cases} \tag{14}$$

Therefore, the regression problem can be expressed as an optimization problem according to Eq. (15).

$$\begin{array}{ll} \min\limits_{w,b,\xi_i,\xi_i^*} & \frac{1}{2}||w||^2 + C\sum\limits_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{Subject to} & y_i - f(x_i) \le \varepsilon + \xi_i \\ & f(x_i) - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n \end{array} \tag{15}$$

ξ_i and ξ_i^* are slack variables that are used to measure the deviation of training samples with an error greater than ε. In the above Equation, the constant C is an integer and positive number that determines a penalty when a training error occurs, and its values are between zero and infinity. For example, if the constant C tends towards infinity, irrespective of the penalty, the result will be a complex model (Cherkassky and Ma 2004). The schematic of SVR structure is presented in Fig. 4. The above optimization formula can be written as a dual problem and solved by Eq. (16)

$$f(x_i) = \sum\nolimits_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x, x_i) + b \tag{16}$$

where, α_i and α_i^* are Lagrange multipliers, which are positive real constants and $k(x, x_i)$ is kernel function.

The kernel functions that are implemented by the SVR includes linear, polynomial, sigmoid and radial basis function (RBF) (Mohammadi and Mehdizadeh 2020). In this study, the RBF type is used, and its mathematical relationship is according to Eq. (17). Where x_i and x_j display the vectors in the input space and σ shows the Gaussian noise level of standard deviation.

$$k(x_i, x_j) = \exp\left[-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right] \tag{17}$$

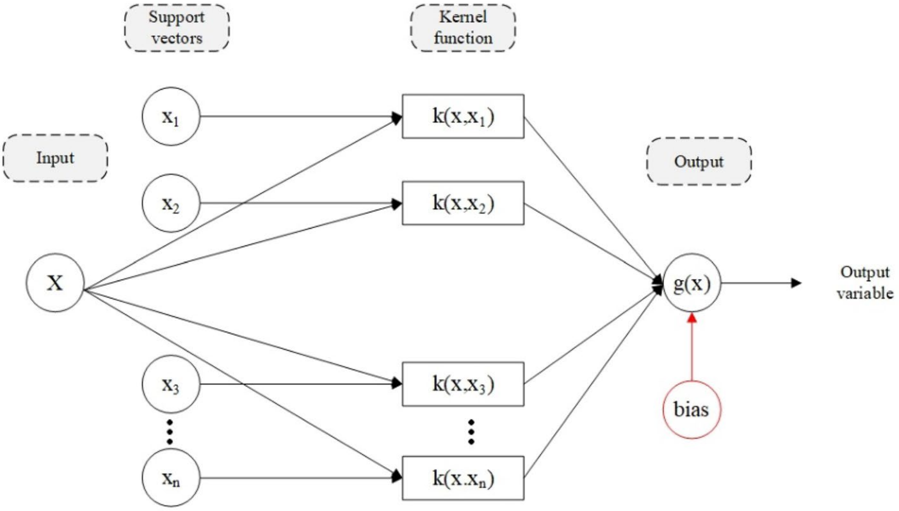

Fig. 4 Support vector regression structure

M5 Tree The M5 tree model was first proposed by Quinlan (1992). This model is based on a decision tree, but unlike the decision tree used for classification, M5 tree has linear regression functions that can be used for quantitative data (Rahimikhoob et al. 2013). The structure of this model is similar to that of an inverted tree, so that the root is at the top and the leaves are at the bottom (Keshtegar and Kisi 2018). Linear regressions in M5 model are relationships between independent and dependent variables that produce the regression bonds in its leaves.

The M5 model divides the data space into smaller sub-spaces using the divide-and-conquer method (Rezaie-balf et al. 2017). The production of the M5 model tree consists of two stages; 1. Growth and 2. Pruning (see Fig. 5).

The Growth stage, also called the Splitting stage, divides the input space into several classes using linear regression models and minimizing the errors between the measured and predicted values (Heddam and Kisi 2018). The process of splitting in each node is repeated many times until it reaches the leaf. This process stops in this model when the class values of all samples reaching a node change slightly, or only a few samples remain (Singh et al. 2010). This division criterion is based on Standard Deviation Reduction (SDR), according to the following Equation.

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \tag{18}$$

where, T shows a set of examples that reach the node; T_i denotes the subset of examples after splitting, and sd is the standard deviation. Finally, after checking all the splits, the one that maximizes the expected error reduction is selected for split in the node by the M5 model (Quinlan 1992).

The splitting process sometimes results in a large tree that looks like a tree that needs to be pruned. In the pruning stage, sub-tree nodes are replaced by linear regression functions and transformed into leaf nodes (Ghaemi et al. 2019).

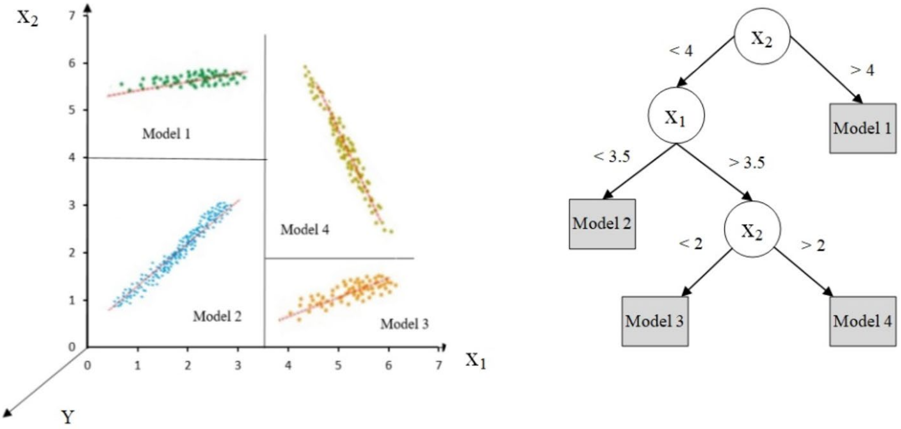

Fig. 5 M5 model tree with four linear regression models at the leaves

Adaptive Neuro-Fuzzy Inference System Adaptive Neuro-Fuzzy Inference System or ANFIS for short was first introduced by Jang (1993). ANFIS is a powerful combination of artificial neural networks with fuzzy logic. For this reason, ANFIS has advantages such as the ability to manage large amounts of input data with high uncertainty (Anusree and Varghese 2016), the potential for modeling nonlinear systems such as hydrological processes (Mosavi et al. 2018) and increasing the accuracy of estimation and forecast (Zare and Koch 2018). On the contrary, a drawback of ANFIS is the significant amount of time required for training and determining parameters for its structure (Chang and Chang 2006).

Among the different types of fuzzy models, the Takagi–Sugeno (Takagi and Sugeno 1985) model is the most widely used due to its high computational efficiency. The fuzzy model based on the first-order Takagi–Sugeno model with two fuzzy IF–THEN rules can be expressed as

$$Rule\ 1:\ if\ x\ is\ A_1\ and\ y\ is\ B_1\ then\ f_1 = p_1x + q_1y + r_1 \tag{19}$$

$$Rule\ 2:\ if\ x\ is\ A_2\ and\ y\ is\ B_2\ then\ f_1 = p_2x + q_2y + r_2 \tag{20}$$

This method consists of two inputs, two rules and one output. Where, x and y are inputs, A_i and B_i are fuzzy sets and p_i, q_i i and r_i are design parameters. This system has five layers as shown in Fig. 6.

In the system, the inputs are expressed in a fuzzy form. For this purpose, membership functions (MFs) are defined for each entry. The number and type of MFs in the construction section of the ANFIS model, are determined by clustering methods. Therefore, clustering methods are a powerful tools for classify the inputs into groups in train section of the ANFIS model and establish relationships between inputs and output space (Benmouiza and Cheknane 2019). The clustering methods include K-means, mountain, subtractive and Fuzzy C-means clustering. In this study, subtractive and FCM clustering methods were used.

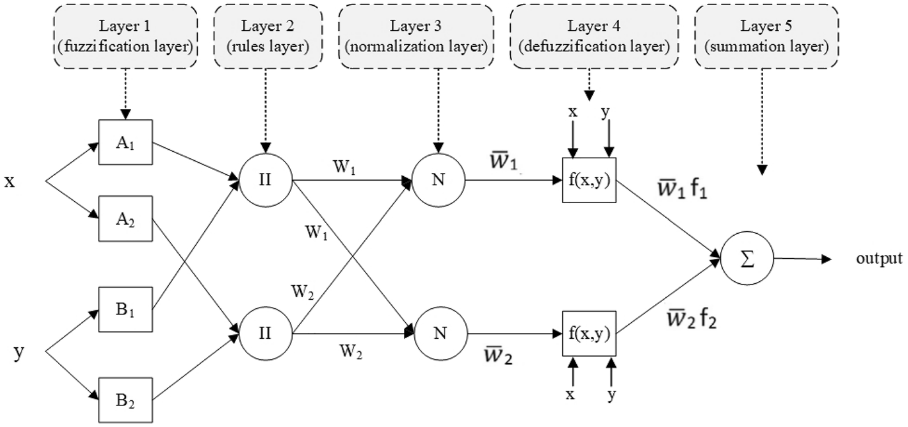

Fig. 6 ANFIS structure

Subtractive Clustering Method Subtractive clustering method is a modification of the mountain method introduced by Yager and Filev (1994). Then, (Chiu 1994) proposed Subtractive clustering method to reduce the complications of the mountain clustering method to determine the number and cluster center. This algorithm is an iterative process that assumes that each data point has the potential to be the cluster center. So that by measuring the data potential of the neighboring data points that point, the potential of each data point is calculated. Finally, the data that have the highest potential values among all the data are selected as cluster center. Then, the number of cluster is determined by determining the optimal value of the radius.

Choosing a suitable effective radius is crucial to determining the number of clusters. If the radius is considered short, a large number of clusters will be created, and the rules will increase accordingly (Cobaner 2011).

Fuzzy C-Means Clustering FCM clustering algorithm is a modified version of k-means clustering and was first introduced by Bezdek† (1973). This clustering method is used to produce less fuzzy rules and avoid the "curse of dimensionality" problem in ANFIS model (Zare and Koch 2018). According to this algorithm, after determining the cluster centers, each data point with a certain degree of membership belongs to a specific cluster, which degree of membership can be between zero and one.

3.3 Evaluation Criteria

In this study, to compare the performance of missing streamflow data estimation methods, three evaluation criteria, including Root Mean Square Error, Nash–Sutcliffe index (Nash and Sutcliffe 1970) and Coefficient of determination (Legates and McCabe Jr. 1999) are used. The details of these evaluation criteria are described in Table 3.

The Root Mean Square Error (RMSE) is a measure used to assess the level of agreement between a model's predictions and the actual observed data. When the RMSE value is zero, it indicates a perfect match between the model's output and the observed data. Conversely, as the RMSE value approaches infinity, it signifies a significant disparity between the model's output and the observed data, indicating poor performance of the model. If the value of the NSE index is equal 1, the model has the best performance and it means that the output of the model matches the observed data. If the value of the NSE index is equal to or zero, the model has the accuracy of the average

Table 3 Recovery methods accuracy criteria

Statistic Metric	Equation	Values Range	Perfect Score
Root Mean Square Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}}$	$0 < \mathrm{RMSE} < \infty$	0
Nash–Sutcliffe Efficiency (NSE)	$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2}$	$-\infty < \mathrm{NSE} < 1$	1
Coefficient of determination (R^2)	$R^2 = [\frac{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)\left(P_i - \overline{P}\right)}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2}}]^2$	$0 < R^2 < 1$	1

Where, N represents the number of data, O_i and P_i represents monthly observed flow and the predicted flow at time i respectively. $\overline{O}$ mean observed values and $\overline{P}$ mean predicted values of the observation data. Negative NSE index values occur when the performance of the average observed data is better than the performance of the desired model.

R^2, or the coefficient of determination, is a statistical measure used to assess the goodness of fit of a regression model. It indicates the proportion of variance in the dependent variable that is explained by the independent variables. A value of 0 suggests that the model does not explain any variance, indicating a poor fit, while a value of 1 indicates a perfect fit where the model explains all the variance. Therefore, R^2 is a useful tool for evaluating the effectiveness of a regression model in explaining the variability of the dependent variable.

3.4 Social Choice

Comparing data recovery methods and identifying the best performing one among several different methods may cause confusion and errors. In such problems, considering the improvement of the performance of only one criterion among all data recovery methods, it is possible to distance ourselves from the improvement of other evaluation criteria. Social choice methods can help solve this problem. So that by applying the calculated values of all the evaluation criteria in the process of comparing the data recovery methods, the final result is error-free.

The main idea of the SC approach was first introduced during the French Revolution by two French mathematicians and scientists, Jean-Charles de Borda count and Condorcet. Two centuries later, it was revived by the winner Nobel Prize in Economics in 1972 named Arrow. So that by giving priority to the candidates, it turned it into a democratic voting system. The SC approach seeks the best choice and, as far as possible, applies the preferences of decision makers equally in the final decision making process (Arrow 1951; Arrow et al. 2010). SC theory includes five approaches called Plurality voting, Hare system, Borda count, Pairwise comparisons voting and Approval voting (Srdjevic 2007).

Among these methods, the Borda count method is an efficient method for solving water resources management and hydrology problems and in various fields including water resource quality management (Zolfagharipoor and Ahmadi 2016), developing suitable algorithms for the optimal performance of multi-reservoir systems (Karami and Dariane 2018), determining the appropriate crop pattern for the proper management of water resources (Dariane et al. 2021), and streamflow modeling (Dariane and Behbahani 2022).

To enhance the comparison of streamflow data recovery methods and identify the superior method, the Borda count method was employed. This method assigns candidate i a score equal to the number of candidates lower than candidate i. So that if n candidates participate in the voting, the score of candidate i is equal to $n - i$. Finally, the winner is the candidate with the highest number of wins. For example, suppose five candidates participated in the election. In that case, four points will be assigned to the first-ranked candidate, three points to the second-ranked candidate, two points to the third-ranked candidate, one point to the fourth-ranked candidate and zero points to the fifth-ranked candidate (Karami and Dariane 2018).

The flowchart of the proposed method to recover the missing streamflow data is presented in Fig. 7.

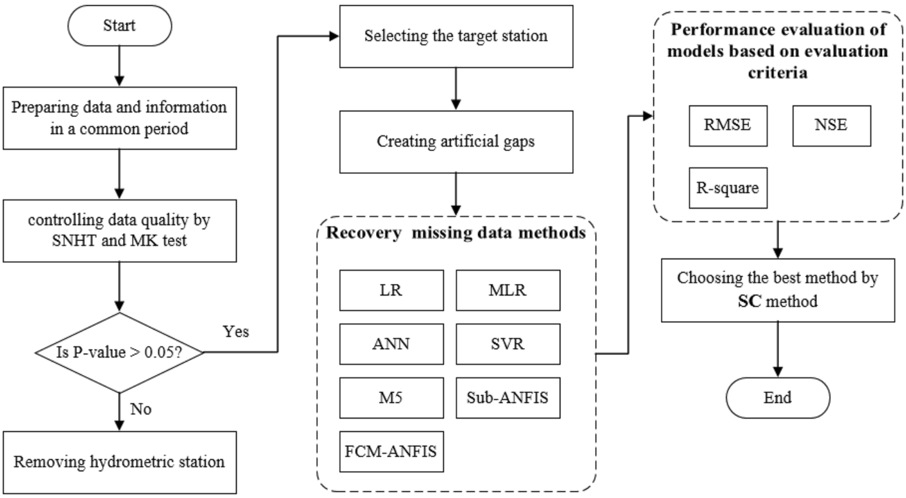
Start
Preparing data and information in a common period
controlling data quality by SNHT and MK test
Is P-value > 0.05?
Yes
No
Removing hydrometric station
Selecting the target station
Creating artificial gaps
Recovery missing data methods
LR
MLR
ANN
SVR
M5
Sub-ANFIS
FCM-ANFIS
Performance evaluation of models based on evaluation criteria
RMSE
NSE
R-square
Choosing the best method by SC method
End

Fig. 7 The flowchart of the proposed methodology

4 Results and Discussion

Monthly streamflow data in three basins of Karaj, Taleghan, and Latyan are available in a 26-year time series. SNH and M–K tests were used to determine the breaking points and trends in the time series of hydrometric stations. The null hypothesis (H_0) means the homogeneity of the data in the SNH test, and the randomness and absence of any trend or serial correlation structure among the observed data in the M–K test. These two tests were performed and the P-value was calculated for each time series of the hydrometric station. If the P-value is greater than the significance level (α), the null hypothesis is confirmed. Otherwise, the alternative hypothesis (H_1) is replaced, meaning there is heterogeneity and trend in the data. Table 4 shows the p values for each hydrometric station used in this study.

Table 4 Results of SNH test and M–K test of selected hydrometric stations at $\alpha = 0.05$

Basin	Station	SNHT p-value	MK p-value
Taleghan	Mehran	0.56	0.88
	Galinak	0.36	0.84
	Jostan	0.24	0.98
	Alizan	0.47	0.10
Karaj	Sira1	0.48	0.47
	Sira2	0.36	0.42
	Gachsar	0.50	0.48
	Haft Cheshmeh	0.14	0.35
Latyan	Rodak	0.22	0.94
	Latyan	0.60	0.05
	Najarkela	0.14	0.37
	Narvoon	0.12	0.79

The results show that the streamflow data of all hydrometric stations are without breakpoints and trends. In other words, the time series does not show any impact related to climate change and human activities. It should be noted that if the hydrometric station data has a trend and a breakpoint, it will be removed from the data set.

It should be mentioned that Jostan, Sira1, and Rodak stations are target stations in Taleghan, Karaj, and Latyan basins, respectively. The target station is the station that has missing data in its data time series.

Two artificial gaps with 36 months duration were created in the target stations. The artificial gap of the first period is between October 1991 to September 1994, and the second artificial gap is between October 2014 to September 2017. The purpose of creating two gaps is to compare the performance of methods in different periods with probably different conditions in order to decrease the impact of possible specific hydrometeorological conditions in a single period on the results. This can help to draw more accurate and reliable conclusions.

As mentioned earlier, estimation of missing data in artificial gaps was done by LR, MLR, ANN, SVR, M5, FCM-ANFIS and Sub-ANFIS models. It is worthwhile to mention that each of the five structures of the neural network model was executed 20 times, and the average of these 20 executions is presented as the representative performance for the neural network models.

According to Borda method, the results of the evaluated criteria of all the methods used to recover the streamflow data are ranked in each gap period and in each basin. The best recovery method is determined based on the total score of each method. In this way, first the results of each of the evaluation criteria, including RMSE, NSE, and R^2, were sorted separately from low to high. The lower the value of the RMSE evaluation criterion calculated for each method (in m^3/s), the higher efficiency of that method for recovering data. On the contrary, the higher the calculated value of NSE and R^2 measures, the better the performance of that method. Accordingly, the lowest RMSE criterion and the highest value of NSE and R^2 evaluation criteria were assigned the first rank when ranking the evaluation criteria. In the same way, each of the evaluation criteria of data recovery methods is ranked. In this study, 11 recovery methods were used and ranked from 1 to 11. After ranking, a value should be assigned to each rank. This value is equivalent to the number of ranks below it. For example, due to the use of 11 methods, the first rank of each evaluation criterion is given a point equal to 10. This process continues until the lowest rank, so the evaluation criterion with rank 11 is given a score equal to zero. Then the method with the highest summation of evaluation criteria points (Borda count) is the winner. For a more accurate comparison of the obtained results, after ranking the methods in each basin and in each gap, the methods were classified into three groups A, B and C. Group A includes methods 1 to 4 in the ranking according to Borda count approach, i.e. the best group, group B includes methods 5 to 8, i.e. the average group, and group C includes the last three methods, i.e. the weak group.

The described process was implemented on the results of the evaluation criteria obtained from the application of the methods on the missing data. Also for example, Table 5 shows the ranking results and Borda points for gap 1991–1994 and 2014–2017 and in Jostan station located in Taleghan. Borda count values are the rank of the total points of the evaluation indices for each method.

Table 5 shows that in the first artificial gap (1991–1994) in Taleghan basin, the FCM-ANFIS method with RMSE, NSE and R^2 criteria equal to 2.500, 0.945 and 0.952, respectively, is the best method to estimate the missing flow data. According to Table 5, in the second artificial gap in this basin, the sub-ANFIS method is the best method with RMSE,

Table 5 Results of Borda count for data estimation methods in Taleghan basin

Gaps	Method	Test period			point			Sum of points	Borda count	Group
periods		RMSE	NSE	R2	RMSE	NSE	R2			
1991 to 1994	LR	4.02	0.86	0.90	3	3	4	10	7	B
	MLR	5.26	0.75	0.86	2	2	2	6	8	B
	ANN(1)	3.40	0.89	0.90	4	4	3	11	6	B
	ANN(2)	2.92	0.92	0.93	7	7	7	21	3	A
	ANN(3)	10.15	0.08	0.13	0	0	0	0	10	C
	ANN(4)	5.36	0.74	0.78	1	1	1	3	9	C
	ANN(5)	2.91	0.92	0.93	8	6	6	20	4	A
	SVR	3.08	0.91	0.95	5	5	9	19	5	B
	M5	2.69	0.93	0.94	9	9	8	26	2	A
	Sub-ANFIS	2.99	0.92	0.92	6	8	5	19	5	B
	FCM-ANFIS	2.50	0.95	0.95	10	10	10	30	1	A
2014 to 2017	LR	3.46	0.83	0.87	2	2	2	6	7	B
	MLR	2.55	0.91	0.92	4	5	4	13	5	B
	ANN(1)	2.66	0.90	0.90	3	3	3	9	6	B
	ANN(2)	2.40	0.91	0.92	6	6	5	17	4	A
	ANN(3)	9.36	-0.43	0.18	0	0	0	0	9	C
	ANN(4)	4.14	0.76	0.78	1	1	1	3	8	B
	ANN(5)	2.52	0.91	0.91	5	4	8	17	4	A
	SVR	2.38	0.93	0.94	8	8	9	25	2	A
	M5	2.40	0.92	0.93	7	7	6	20	3	A
	Sub-ANFIS	1.67	0.96	0.97	10	10	10	30	1	A
	FCM-ANFIS	2.10	0.94	0.94	9	9	7	25	2	A

NSE and R^2 criteria of 1.666, 0.964, and 0.967 respectively. As a result, ANFIS model is the best method to recover the monthly flow data of the Taleghan basin compared to other methods.

It is worthwhile to mention that all methods use streamflow data from surrounding hydrometric stations except the ANN(3) and ANN(4) where only surrounding precipitation data are used. Also, ANN(5) uses both precipitation and streamflow data from surrounding stations. Therefore, it is not surprising to see that LR and MLR performs better than ANN(3) and ANN(4). ANN(5) suffers from precipitation inputs that not only do not help effectively the model performance but also introduce excessive parameters resulting in worse outputs than the ANN(2) that only uses streamflow data.

The importance of utilizing the seasonal index becomes evident when comparing models with and without it. For example, when comparing the results of ANN(1) with ANN(2), and ANN(3) with ANN(4), the effectiveness of the seasonal index becomes apparent. The seasonal nature of streamflow in basins strongly impacts the accuracy of peak streamflow estimation by the ANN(1) and ANN(3) models. However, the addition of seasonal index as input to the neural network resulted in a significant improvement in estimating missing values. In Fig. 8, the peak streamflow in Karaj basin is depicted, showing that the data estimated by ANN(2) and ANN(4) align more closely with the observed peak streamflow data compared to the other two neural network models.

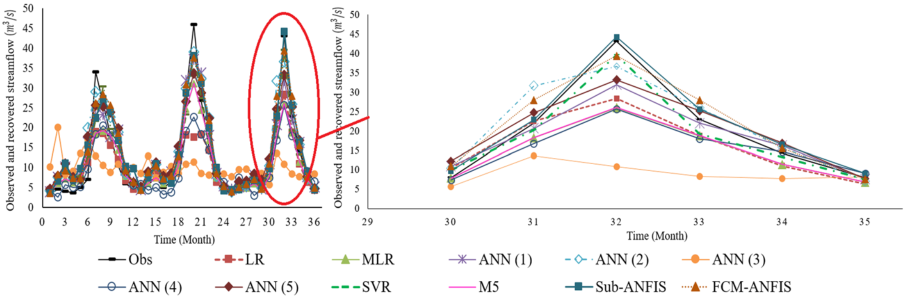

Fig. 8 Observed and estimated streamflow in Sira1 station Karaj basin, in the period 2014–2017

As it can be seen from Table 6, there are similarities and some differences among the performance of methods in different basins and different periods. However, it is noticeable that ANFIS methods as well as SVR are superior in most cases. Also, it is interesting to note that simple classic methods of LR and MLR are in some cases and overall better than some machine learning methods when proper inputs are used. In other words, selecting proper inputs (i.e., streamflow here) is more important than using more advanced method

Table 6 Results of Borda count-based grouping for all basins

Gap periods	**Method**	**Taleghan**	**Karaj**	**Latyan**	**Taleghan**	**Karaj**	**Latyan**	**Overall rating**
		Borda count			**Group**			
1991 to 1994	LR	7	3	4	B	A	A	B
	MLR	8	4	7	B	A	B	B
	ANN(1)	6	7	9	B	B	C	B
	ANN(2)	3	5	3	A	B	A	A
	ANN(3)	10	11	11	C	C	C	C
	ANN(4)	9	10	10	C	C	C	C
	ANN(5)	4	8	5	A	B	B	B
	SVR	5	2	2	B	A	A	A
	M5	2	6	8	A	B	B	B
	Sub-ANFIS	5	9	6	B	C	B	B
	FCM-ANFIS	**1**	**1**	**1**	**A***	**A***	**A***	A
2014 to 2017	LR	7	9	9	B	C	C	B
	MLR	5	7	7	B	B	B	B
	ANN(1)	6	5	6	B	B	B	B
	ANN(2)	4	4	3	A	A	A	A
	ANN(3)	9	10	11	C	C	C	C
	ANN(4)	8	8	8	B	B	B	B
	ANN(5)	4	7	4	A	B	A	B
	SVR	2	**1**	**1**	A	**A***	**A***	A
	M5	3	6	10	A	B	C	B
	Sub-ANFIS	**1**	2	5	**A***	A	B	A
	FCM-ANFIS	2	3	2	A	A	A	A

A* is the first rank among recovery methods in each basin and gap (i.e., ANN compared to LR or MLR). LR does well in the first gap period in all three basins but it gives poor results in the second period in two basins, which is an indication that it is unable to handle data variations during specific periods (i.e., wet or dry conditions). MLR overcomes this problem by using more streamflow variables. M5 behaves differently not only in different periods but also in different basins. It shows good results in Taleghan but performs poorly in other two basins.

Figures 8 and 9 clearly show that method ANN(3) has very little accuracy in recovery the base flow and peak flow data. It also supports the finding in evaluation figures estimated earlier. This issue was clearly identified by comparing structures 1 and 2. Additionally, ANN(5) exhibits poorer results compared to ANN(2) in the recovery of peak streamflow data. This is due to the rainfall variable contributing to an increase in network error, similar to the conditions observed for structures 3 and 4.

On the other hand, ANN(5) works better than ANN(3) and ANN(4), because, in this situation, the presence of the flow variable improves the performance of the network and helps the learning process ANN. In general, the ANN(2) model performs better than other ANN models, and the two input variables of flow and seasonal index make the neural network perform well. Therefore, the use of ANN(5), in addition to increasing the volume of calculations, also leads to weaker results.

Figure 8 shows the data estimated by the classical methods have less agreement with the data observed in the peak streamflows. This issue shows the uncertainty of classical methods in peak streamflow data recovering.

Based on the results from Table 5, it is evident that the FCM-ANFIS method outperforms the classic LR (MLR) method in estimating missing data during the first artificial gap (1991–1994) in the Taleghan basin. The FCM-ANFIS method achieved RMSE, NSE, and R^2 criteria of 2.50, 0.95, and 0.95 respectively, while the classic LR (MLR) method showed inferior performance with RMSE, NSE, and R^2 criteria of 4.02 (5.26), 0.86 (0.75), and 0.90 (0.86) respectively.

Similarly, during the second artificial gap (2014–2017), the sub-ANFIS method demonstrated superior performance with RMSE, NSE, and R^2 criteria of 1.67, 0.96, and 0.97 respectively. On the other hand, the LR and MLR methods performed worse with NSE values of 0.83 and 0.91 in estimating missing streamflow data.

These results reinforce the overall ranking in Table 6, indicating that machine learning methods, particularly FCM-ANFIS and sub-ANFIS, are recommended for estimating missing data in both the first and second artificial gaps. Therefore, it is advisable to utilize machine learning methods in such situations based on the findings.

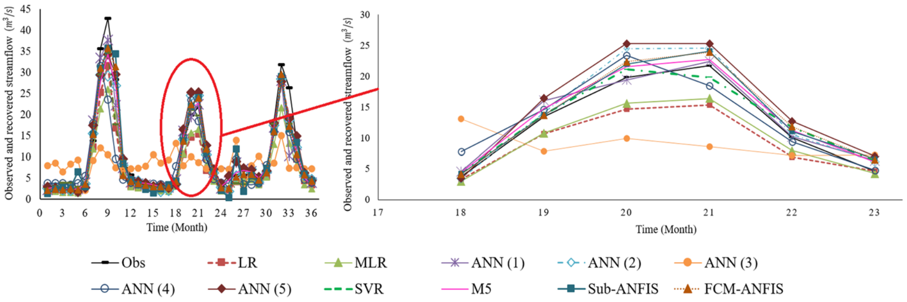

Fig. 9 Observed and estimated streamflow in Jostan station Taleghan basin, in the period 1991–1994

In addition, Figs. 8 and 9 shows that the data estimated by sub-ANFIS and FCM-ANFIS models have a good match with the observed peak streamflow data. The performance of the methods based on machine learning, especially the ANFIS model, is very satisfactory when the river flow increases.

Figure 10 shows the membership percentage of recovery methods in groups A, B and C in this study. These results are based on the outputs obtained in three areas and two gap periods.

According to Fig. 10 ANN(2), SVR, and FCM-ANFIS methods are placed in group A more than other models, and none of them are seen in group C in any of the cases in this research. Thus, the FCM-ANFIS method is always among the top 4 methods. Also ANN(3) is always one of the weakest recovery methods (i.e., group C).

More detailed investigations show that the new M5 method is less than the classic LR method in group C. As a result, compared to the classic LR method, it has a higher accuracy in recovering missing data. Also, this method has a better performance than MLR in recovering streamflow data in all basins, and it is mostly included in group A.

On the other hand, according to Fig. 10, the MLR method has better results than the LR method. It can be concluded that the use of data from neighboring stations compared to a neighboring station with high correlation, increases the accuracy of estimated data in regression methods. Furthermore, the FCM-ANFIS method consistently belongs to group A, unlike the sub-ANFIS method. Therefore, the selection of an appropriate clustering method can significantly impact the final results.

Although the difference between the results of ANN, SVR and FCM-ANFIS models is not significant, the results obtained by ANFIS models are mainly superior among other models. The results from three basins demonstrate the effectiveness of machine learning based methods in estimating missing streamflow data. This finding is consistent with previous studies by Jing et al. (2022), Zhou et al. (2023), and Kim et al. (2015). It is important to note that there is no single best model for all situations, as the selection

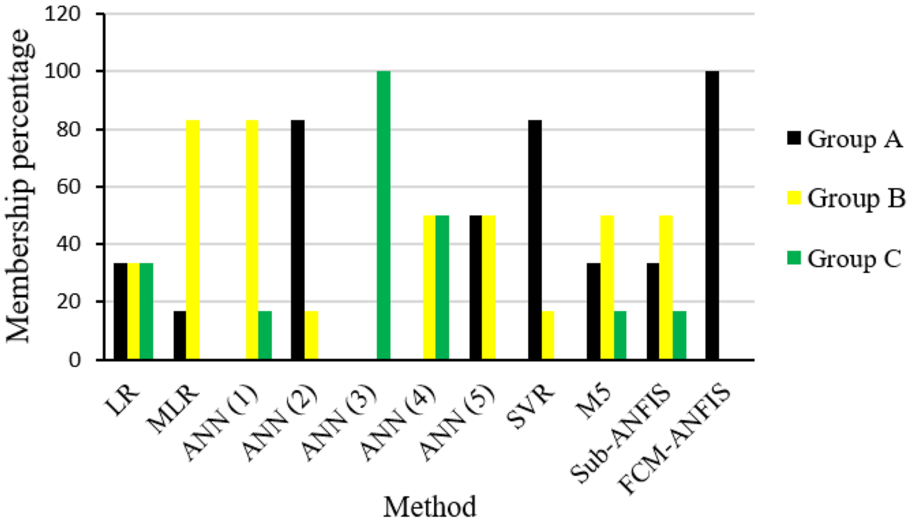

Fig. 10 Membership percentage of missing data recovery methods in A, B and C groups

of appropriate infilling methods depends on various factors such as the length of the gap, available data length, and the topographical and climatic conditions of the region.

5 Conclusion

In the present study, the performance of 11 models, including LR, MLR, ANN, SVR, M5, FCM-ANFIS and Sub-ANFIS, was evaluated in retrieving monthly streamflow data in three basins in Alborz mountainous regions in northern Iran. Models were evaluated using 26 years of data extending from 1991 to 2017, two periods of artificial gaps of data were considered to overcome possible duration-based climate conditions that may affect the results. Overall, as expected it was noticed that machine learning-based methods yield better outputs compared to classical methods.

Also, it is interesting to note that simple classic methods of LR and MLR were better than some machine learning methods when proper inputs are used. In other words, it was shown that selecting proper inputs (i.e., streamflow here) is more important than using more advanced method (i.e., ANN compared to LR or MLR). Additionally, the significance of using the seasonal index was demonstrated by comparing the results of similar models with and without the seasonal index. For instance, during the first artificial gap (1991–1994) in the Taleghan basin, the values of RMSE, NSE, and R^2 in estimating the missing data using the ANN(1) model were 3.40, 0.89, and 0.90, respectively. However, these values improved when the seasonal index was added to the artificial neural network (ANN (2)), resulting in values of 2.92, 0.91, and 0.93, respectively. This improvement was also observed in other basins.

Additionally, it has been observed that methods utilizing streamflow data from surrounding stations outperform those using rainfall data for estimating streamflow at the target station. For example, during the first artificial gap (1991–1994) in the Taleghan basin, the performance metrics for estimating streamflow using rainfall data (ANN(3)) resulted in RMSE of 10.15, NSE of 0.08, and R^2 of 0.13, which are inferior to the performance of ANN(1) utilizing streamflow data from surrounding stations.

In order to compare the recovery methods of streamflow data and determine the methods with superior performance, Borda count method was used. Due to the large number of models and stations investigated, Borda count method was used to summarize the general results. For a more accurate comparison of the obtained results, after ranking the methods in each basin and in each gap, the methods were classified into three groups A, B and C. It was found that ANFIS methods as well as SVR are superior in most cases. The ANFIS method with FCM clustering consistently ranks in group A across all basins, indicating the significance of selecting the right clustering approach for the ANFIS model. 5M behaves differently in different basins and thus is not a reliable method for the area.

Author Contributions Both authors contributed to the study in all levels and original draft preparation. The study is the result of a graduate level thesis and was guided by **Alireza Borhani Dariane** as the advisor of **Matineh Imani Borhan** (student).

Funding No funding was used in this research.

Availability of Data and Materials All authors made sure that all data and materials support our published claims and comply with field standards.

Declarations

Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abghari H, Tabari H, Hosseinzadeh Talaee P (2013) River flow trends in the west of Iran during the past 40years: Impact of precipitation variability. Glob Planet Change 101:52–60. https://doi.org/10.1016/j.gloplacha.2012.12.003

Abudu S, Bawazir AS, King JP (2010) Infilling missing daily evapotranspiration data using neural networks. J Irrig Drain Eng 136:317–325

Aguilera H, Guardiola-Albert C, Serrano-Hidalgo C (2020) Estimating extremely large amounts of missing precipitation data. J Hydroinformatics 22:578–592. https://doi.org/10.2166/hydro.2020.127

Ahmadi M, Moeini A, Ahmadi H et al (2019) Comparison of the performance of SWAT, IHACRES and artificial neural networks models in rainfall-runoff simulation (case study: Kan watershed, Iran). Phys Chem Earth Parts a/b/c 111:65–77. https://doi.org/10.1016/j.pce.2019.05.002

Aieb A, Madani K, Scarpa M et al (2019) A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed. Algeria. Heliyon 5:e01247. https://doi.org/10.1016/j.heliyon.2019.e01247

Alexandersson H (1986) A homogeneity test applied to precipitation data. J Climatol 6:661–675. https://doi.org/10.1002/joc.3370060607

Ali R, Kuriqi A, Abubaker S, Kisi O (2019) Long-term trends and seasonality detection of the observed flow in Yangtze River using Mann-Kendall and Sen's innovative trend method. Water 11

Anusree K, Varghese KO (2016) Streamflow prediction of karuvannur river basin using ANFIS, ANN and MNLR models. Procedia Technol 24:101–108. https://doi.org/10.1016/j.protcy.2016.05.015

Arriagada P, Karelovic B, Link O (2021) Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. J Hydrol 598:126454. https://doi.org/10.1016/j.jhydrol.2021.126454

Arrow KJ (1951) Social Choice and Individual Values. John Wiley Sons Inc, Nueva York

Arrow KJ, Sen A, Suzumura K (2010) Handbook of social choice and welfare. Elsevier

Asadi S, Shahrabi J, Abbaszadeh P, Tabanmehr S (2013) A new hybrid artificial neural networks for rainfall–runoff process modeling. Neurocomputing 121:470–480. https://doi.org/10.1016/j.neucom.2013.05.023

Baddoo TD, Li Z, Odai SN et al (2021) Comparison of missing data infilling mechanisms for recovering a real-world single station streamflow observation. Int J Environ Res Public Health 18

Bahrami J, Kavianpour MR, Abdi MS et al (2010) A comparison between artificial neural network method and nonlinear regression method to estimate the missing hydrometric data. J Hydroinformatics 13:245–254. https://doi.org/10.2166/hydro.2010.069

Belayneh A, Adamowski J, Khalil B, Quilty J (2016) Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmos Res 172–173:37–47. https://doi.org/10.1016/j.atmosres.2015.12.017

Benmouiza K, Cheknane A (2019) Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. Theor Appl Climatol 137:31–43. https://doi.org/10.1007/s00704-018-2576-4

Bezdek† JC (1973) Cluster Validity with Fuzzy Sets. J Cybern 3:58–73. https://doi.org/10.1080/01969727308546047

Cai H, Shi H, Liu S, Babovic V (2021) Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States. J Hydrol Reg Stud 37:100930. https://doi.org/10.1016/j.ejrh.2021.100930

Chang F-J, Chang Y-T (2006) Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. Adv Water Resour 29:1–10. https://doi.org/10.1016/j.advwatres.2005.04.015

Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17:113–126. https://doi.org/10.1016/S0893-6080(03)00169-2

Chiu SL (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:267–278. https://doi.org/10.3233/IFS-1994-2306

Cobaner M (2011) Evapotranspiration estimation by two different neuro-fuzzy inference systems. J Hydrol 398:292–302. https://doi.org/10.1016/j.jhydrol.2010.12.030

Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. J Hydrol 341:27–41. https://doi.org/10.1016/j.jhydrol.2007.04.020

Dariane AB, Behbahani MM (2022) Development of an efficient input selection method for NN based streamflow model. J Appl Water Eng Res 11:127–140. https://doi.org/10.1080/23249676.2022.2088631

Dariane AB, Ghasemi M, Karami F et al (2021) Crop pattern optimization in a multi-reservoir system by combining many-objective and social choice methods. Agric Water Manag 257:107162. https://doi.org/10.1016/j.agwat.2021.107162

Dariane AB, Karami F (2014) Deriving hedging rules of multi-reservoir system by online evolving neural networks. Water Resour Manag 28:3651–3665. https://doi.org/10.1007/s11269-014-0693-0

Dastorani MT, 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

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. https://doi.org/10.1016/j.jhydrol.2018.11.076

Elshorbagy AA, Panu US, Simonovic SP (2000) Group-based estimation of missing hydrological data: I. Approach and general methodology. Hydrol Sci J 45:849–866. https://doi.org/10.1080/02626660009492388

Ergün E, Demirel MC (2023) On the use of distributed hydrologic model for filling large gaps at different parts of the streamflow data. Eng Sci Technol an Int J 37:101321. https://doi.org/10.1016/j.jestch.2022.101321

Fagandini C, Todaro V, Tanda MG et al (2023) Missing rainfall daily data: a comparison among gap-filling approaches. Math Geosci. https://doi.org/10.1007/s11004-023-10078-6

Faramarzzadeh M, Ehsani MR, Akbari M et al (2023) Application of machine learning and remote sensing for gap-filling daily precipitation data of a sparsely gauged basin in East Africa. Environ Process 10:8. https://doi.org/10.1007/s40710-023-00625-y

Gebremicael TG, Mohamed YA, Hagos EY (2017) Temporal and spatial changes of rainfall and streamflow in the Upper Tekezē-Atbara river basin, Ethiopia. Hydrol Earth Syst Sci 21:2127–2142

Ghaemi A, Rezaie-Balf M, Adamowski J et al (2019) On the applicability of maximum overlap discrete wavelet transform integrated with MARS and M5 model tree for monthly pan evaporation prediction. Agric for Meteorol 278:107647. https://doi.org/10.1016/j.agrformet.2019.107647

Giustarini L, Parisot O, Ghoniem M et al (2016) A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. Environ Model Softw 82:308–320. https://doi.org/10.1016/j.envsoft.2016.04.013

Gyau-Boakye P, Schultz GA (1994) Filling gaps in runoff time series in West Africa. Hydrol Sci J 39:621–636. https://doi.org/10.1080/02626669409492784

Harvey CL, Dixon H, Hannaford J (2010) Developing best practice for infilling daily river flow data. Role Hydrol Manag Consequences a Chang Glob Environ 816–823

Harvey CL, Dixon H, Hannaford J (2012) An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. Hydrol Res 43:618–636. https://doi.org/10.2166/nh.2012.110

Heddam S, Kisi O (2018) Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 559:499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061

Ilunga M, Stephenson D (2005) Infilling streamflow data using feed-forward back-propagation (BP) artificial neural networks: application of standard BP and Pseudo Mac Laurin power series BP techniques. Water SA 31:171–176

Jang J-SR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23:665–685. https://doi.org/10.1109/21.256541

Jing X, Luo J, Wang J et al (2022) A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. Water Resour Manag 36:1159–1173. https://doi.org/10.1007/s11269-021-03037-5

Kamwaga S, Mulungu DMM, Valimba P (2018) Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania. Phys Chem Earth Parts a/b/c 106:17–28. https://doi.org/10.1016/j.pce.2018.05.008

Karami F, Dariane AB (2018) Many-objective multi-scenario algorithm for optimal reservoir operation under future uncertainties. Water Resour Manag 32:3887–3902. https://doi.org/10.1007/s11269-018-2025-2

Kendall MG (1948) Rank correlation methods

Keshtegar B, Kisi O (2018) RM5Tree: Radial basis M5 model tree for accurate structural reliability analysis. Reliab Eng Syst Saf 180:49–61. https://doi.org/10.1016/j.ress.2018.06.027

Khan MT, Shoaib M, Hammad M et al (2021) Application of machine learning techniques in rainfall–runoff modelling of the soan river basin, Pakistan. Water 13

Khan N, Sachindra DA, Shahid S et al (2020) Prediction of droughts over Pakistan using machine learning algorithms. Adv Water Resour 139:103562. https://doi.org/10.1016/j.advwatres.2020.103562

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

Kim M, Baek S, Ligaray M et al (2015) Comparative studies of different imputation methods for recovering streamflow observation. Water 7:6847–6860

Lai WY, Kuok KK (2019) A study on bayesian principal component analysis for addressing missing rainfall data. Water Resour Manag 33:2615–2628. https://doi.org/10.1007/s11269-019-02209-8

Legates DR, McCabe GJ Jr (1999) Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241. https://doi.org/10.1029/1998WR900018

Londhe S, Dixit P, Shah S, Narkhede S (2015) Infilling of missing daily rainfall records using artificial neural network. ISH J Hydraul Eng 21:255–264. https://doi.org/10.1080/09715010.2015.1016126

MacLeod C (1999) The synthesis of artificial neural networks using single string evolutionary techniques. PhD Dissertation, The Robert Gordon University, Aberdeen, Scotland

Mahmood R, Jia S (2019) Assessment of hydro-climatic trends and causes of dramatically declining stream flow to Lake Chad, Africa, using a hydrological approach. Sci Total Environ 675:122–140. https://doi.org/10.1016/j.scitotenv.2019.04.219

Mann HB (1945) Nonparametric tests against trend. Econom J Econom Soc 245–259

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133. https://doi.org/10.1007/BF02478259

Mekanik F, Imteaz MA, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. J Hydrol 503:11–21. https://doi.org/10.1016/j.jhydrol.2013.08.035

Mohammadi B (2021) A review on the applications of machine learning for runoff modeling. Sustain Water Resour Manag 7:98. https://doi.org/10.1007/s40899-021-00584-y

Mohammadi B, Mehdizadeh S (2020) Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. Agric Water Manag 237:106145. https://doi.org/10.1016/j.agwat.2020.106145

Mosavi A, Ozturk P, Chau K (2018) Flood Prediction using machine learning models: Literature review. Water 10

Mwale FD, Adeloye AJ, Rustum R (2012) Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. Phys Chem Earth Parts a/b/c 50–52:34–43. https://doi.org/10.1016/j.pce.2012.09.006

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ng WW, Panu US, Lennox WC (2009) Comparative studies in problems of missing extreme daily streamflow records. J Hydrol Eng 14:91–100

Nilsson P, Uvo CB, Berndtsson R (2006) Monthly runoff simulation: Comparing and combining conceptual and neural network models. J Hydrol 321:344–363. https://doi.org/10.1016/j.jhydrol.2005.08.007

Quinlan JR (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. World Scientific, pp 343–348

Radi NFA, Zakaria R, Azman MA (2015) Estimation of missing rainfall data using spatial interpolation and imputation methods. AIP Conf Proc 1643:42–48. https://doi.org/10.1063/1.4907423

Raghavendra NS, Deka PC (2014) Support vector machine applications in the field of hydrology: A review. Appl Soft Comput 19:372–386. https://doi.org/10.1016/j.asoc.2014.02.002

Rahimikhoob A, Asadi M, Mashal M (2013) A comparison between conventional and m5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manag 27:4815–4826. https://doi.org/10.1007/s11269-013-0440-y

Rezaie-balf M, Naganna SR, Ghaemi A, Deka PC (2017) Wavelet coupled MARS and M5 Model Tree approaches for groundwater level forecasting. J Hydrol 553:356–373. https://doi.org/10.1016/j.jhydrol.2017.08.006

Salas JD (1993) Analysis and modelling of hydrological time series. Handb Hydrol 19

Singh KK, Pal M, Singh VP (2010) Estimation of mean annual flood in indian catchments using back-propagation neural network and M5 model tree. Water Resour Manag 24:2007–2019. https://doi.org/10.1007/s11269-009-9535-x

Souza GRD, Bello IP, Corrêa FV, Oliveira LFCD (2020) Artificial neural networks for filling missing streamflow data in Rio do carmo basin, minas gerais, Brazil. Braz Arch Biol Technol 63

Srdjevic B (2007) Linking analytic hierarchy process and social choice methods to support group decision-making in water management. Decis Support Syst 42:2261–2273. https://doi.org/10.1016/j.dss.2006.08.001

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

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern SMC 15:116–132. https://doi.org/10.1109/TSMC.1985.6313399

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

Tongal H, Booij MJ (2018) Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. J Hydrol 564:266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004

Uysal G, Şorman AÜ (2017) Monthly streamflow estimation using wavelet-artificial neural network model: A case study on Çamlıdere dam basin, Turkey. Procedia Comput Sci 120:237–244. https://doi.org/10.1016/j.procs.2017.11.234

Vapnik V (1998) Statistical Learning Theory Wiley New York 1:2

Vapnik V (1999) The nature of statistical learning theory. Springer science & business media

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

Yager RR, Filev DP (1994) Approximate clustering via the mountain method. IEEE Trans Syst Man Cybern 24:1279–1284. https://doi.org/10.1109/21.299710

Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 112:143–167. https://doi.org/10.1007/s00704-012-0723-x

Zare M, Koch M (2018) Groundwater level fluctuations simulation and prediction by ANFIS- and hybrid Wavelet-ANFIS/Fuzzy C-Means (FCM) clustering models: Application to the Miandarband plain. J Hydro-Environment Res 18:63–76. https://doi.org/10.1016/j.jher.2017.11.004

Zhou Y, Tang Q, Zhao G (2023) Gap infilling of daily streamflow data using a machine learning algorithm (MissForest) for impact assessment of human activities. J Hydrol 627:130404. https://doi.org/10.1016/j.jhydrol.2023.130404

Zolfagharipoor MA, Ahmadi A (2016) A decision-making framework for river water quality management under uncertainty: Application of social choice rules. J Environ Manag 183:152–163. https://doi.org/10.1016/j.jenvman.2016.07.094

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