



Forecasting of river water flow rate with machine learning

Akin Ilhan¹

Received: 6 December 2021 / Accepted: 27 June 2022 / Published online: 24 July 2022

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

Today, the estimation of physical parameters has become very important; for instance, the water flow rate (RWFR) estimation is one of the types that will gain considerable significance among the others performed in this way. The forecasting of RWFR plays a crucial role in planning and building of new water dams, or operating the ones that were previously built. This study proposes machine learning algorithms to estimate a one-day ahead short-term RWFR. The estimation models were developed, using historical RWFR data, in order to obtain the future RWFR values. For the purpose of RWFR predictions, long short-term memory (LSTM) neural network, adaptive neuro-fuzzy inference system (ANFIS) with fuzzy c-means (FCM), ANFIS with subtractive clustering (SC), as well as the ANFIS with grid partition (GP) were advanced. A measurement station (MS), named as Harmanli MS, located on the Maritsa River and at the border of Turkey and Bulgaria, was selected as the study region. A total of 102 models were constructed by these four algorithms. The forecasting outcomes were compared with the real measured data. The comparisons were conducted using the statistical error results obtained from mean absolute error (MAE), root mean square error (RMSE), and the correlation coefficient (R). The predictions of the daily average volumetric flow rate (VFR) data have indicated that ANFIS-FCM model had generated the best statistical error results. Namely, statistical error results of 2.54 m³/s MAE, 4.35 m³/s RMSE, and 0.9981 R have been obtained with the utilization of the ANFIS-FCM algorithm. On the other hand, when the averages of three statistical error parameters are considered, it was shown that averages of the statistical error results of the ANFIS-SC algorithm including cumulative of 48 models to be slightly better than the average statistical error results of the ANFIS-FCM. Accordingly, it was concluded and demonstrated in this study that FCM and SC tools of the ANFIS can be two useful methods in VFR predictions. Finally, as in the case of RWFR data which usually has random distributions, it has been reported and shown that both algorithms can be simply accomplished to any type of randomly distributed data.

Keywords Machine learning · Time series · Volumetric flow rate · Long short-term memory · Adaptive neuro-fuzzy inference system

1 Introduction

The utilization of the renewable energy sources (RESs) is increasing recently due to the global warming, greenhouse gases, as well as the limited fossil fuel reserves. Accordingly, RES power plants have an important spreading potential when compared with respect to the conventional power plants that are using carbon-based energy sources.

This is a general case for worldwide and many countries in the World have at least one type of RES power plants. Among whole renewable energy sources, the hydropower is still of the most important and the widespread method of renewable power generation. Figure 1 presents the comparison of World's total hydropower generation with respect to the total renewable power generation, considering years between 2010 and 2020. Similarly, the cumulative installed hydropower capacity according to the total installed renewable power of Turkey, considering the same interval of time, is indicated in Fig. 2 [1, 2].

As it is seen in Figs. 1 and 2, both for the World and the Turkey, the hydropower constitutes a huge amount of the cumulative installed renewable power between 2010 and 2020. For instance, while the cumulative installed

✉ Akin Ilhan
ailhan@ybu.edu.tr

¹ Department of Energy Systems Engineering, Faculty of Engineering and Natural Sciences, Ankara Yildirim Beyazit University, 06010 Ankara, Turkey

Fig. 1 Comparison of the installed hydropower capacity given according to the total installed renewable power of the World [1, 2]

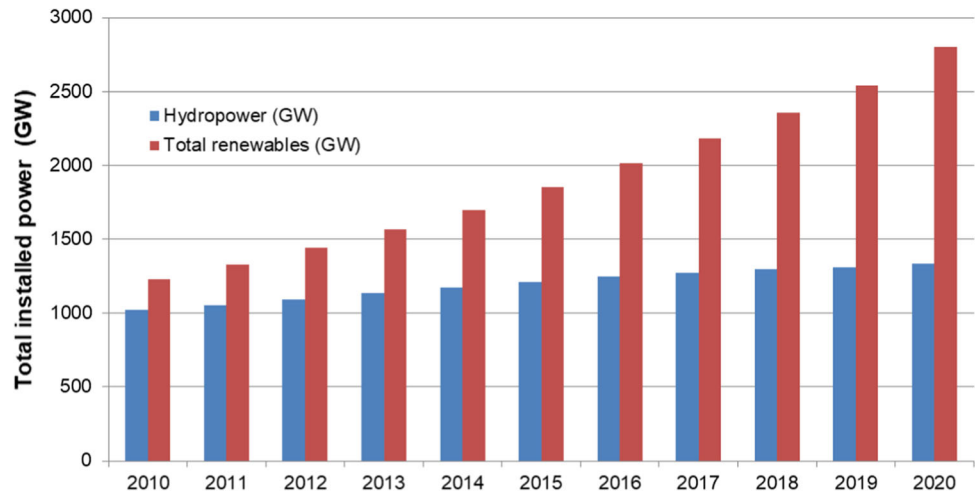
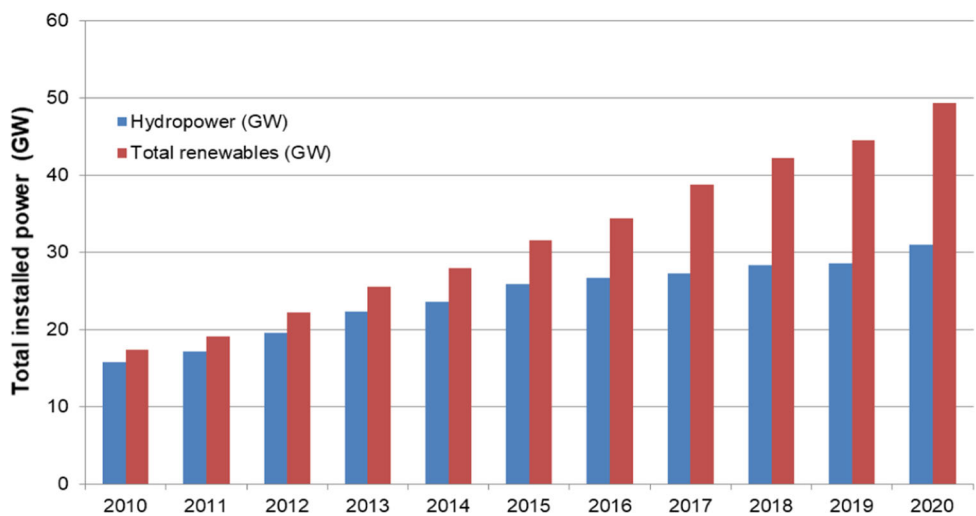


Fig. 2 Comparison of the installed hydropower capacity given according to the total installed renewable power of Turkey [1, 2]



hydropower was 1,024.833 GW by 2010 for the World, it has been increased to 1,332.885 GW of installations by 2020. In terms of the renewable power of the World, the cumulative installations were 1,226.853 GW by 2010, which has increased to 2,802.004 GW by 2020. From a different perspective, when the World's total installations are taken into account, while the ratio of total installed hydraulic power to total renewable power was 83.53% in 2010, this ratio decreased to 47.57% in 2020. Although there has been a decline in the rate, the ratio of hydro-electric power to total renewable power is quite high as of 2020. On the other hand, considering Turkey, the total hydraulic power installations were 15.831 GW by 2010. However, the amount of the hydraulic installations by 2020 was almost doubled, which was reached to 30.984 GW for this country. Besides, considering Turkey's total renewable power installations, it is observed that 17.369 GW of installations by 2010 has been increased to 49.398 GW by 2020. Similar to the situation of the World, the ratio of Turkey's hydraulic power installations on the total

renewable power installations has decreased from 91.11% to 62.72% between 2010 and 2020. Although hydraulic power installations in the World and Turkey are increasing rapidly, the relative proportional decrease in the last decade is due to the acceleration of the installations of the other renewable energy types such as wind and solar. However, even this situation could not reduce the traditional importance of hydraulic power among the whole renewable energy types. And the importance of the hydraulic power will remain this way for many years to come [1, 2].

Similar to other methods of electricity generation, in hydropower engineering, a hydro turbine is utilized to generate electricity, with the principles of the energy obtained from a falling or flowing water to turn the rotor blades. The hydraulic generator attached to the rotating blades of the turbine converts the mechanical energy to the electrical energy. The capacity of the electrical energy generated from each hydro power plant is linked to the amount of the flowing water flow rate as well as the height that the water falls from. The history of hydro engineering

includes the utilization of hydropower for mechanical milling, such as the capture of the energy from the flowing water to obtain the grinding of the grains. However, in modern time, hydro plants are used for the purpose of generating huge amounts of electricity. For this reason, the power plant facilities can be grouped according to the plant generation capacity. Large types of hydro plants include the power generation above 30 MW. Small types include power generation values in the range of 100 kW–30 MW. Finally, micro types of hydropower generation include the generation capacity below 100 kW [3].

There are many benefits and advantages of the hydropower, when compared to other power generation methods. For example, it has a clean fuel source; i.e., the fuel of hydropower generation is only water. Besides, during the production of the hydroelectric energy, air pollution is not in question. It is a domestic and national form of energy production that each country or province can easily obtain hydropower generation with its own water reserve. So, the increase of the hydropower installations in a country will reduce the dependence to the international fuel resources. It is a renewable power source being more reliable and economic than fossil fuels. It helps to minimize the greenhouse gas production and the accompanying global warming. Recreational opportunities will be offered by the hydropower applications, since hydro-engineering creates water reservoirs. Namely, the water reservoirs will be used in fishing, swimming or boating purposes. On the other hand, hydropower plants can provide energy to the grid instantly, and provide back-up power during major electricity outages. Hydropower applications also supply the control of floods, will help in irrigations of fields, etc., and provide water directly when required [3, 4].

An installed and operating small-scale hydraulic power plant is shown in Fig. 3 [3]. As can be seen, water is

lowered to the power generation facility from a high altitude and from the reservoir by the penstocks. The water that is hitting to the hydro turbine blades has initially a potential energy that is converted into mechanical energy in these blades. Subsequently, in the turbine alternator, the mechanical energy is converted into the electrical energy. Finally, in the switchyard established in the same place, the transfer and the distribution of the electrical energy to the grid is ensured.

The literature studies indicate that the forecasting applications in engineering can be applied by three methods, including physical types, statistical ones, as well as the artificial intelligence methods [5, 6]. Among those, plain methods are the physical methods that study on the physical information such as atmospheric temperature, roughness, atmospheric pressure, and obstacles [7]. For example, numerical weather prediction (NWP) is a type of physical method. A series of mathematical equations as well as physical information are required in NWP processes. Physical methods are generally applicable in medium- to long-term data estimations, such as wind power. However, one of the disadvantages of the physical methods includes complexity of the computations, as well as important computing resources are required in these methods. On the other hand, statistical models have been also applied as a forecasting method, in the literature. However, these methods have been generally limited with wind speed estimations. Data of historical wind speed is generally utilized in a proper time frame to forecast future wind speed, in statistical methods. Some of the statistical methods that are mainly available in the literature include moving average models, autoregressive models, and autoregressive moving average models. More advanced types such as autoregressive integrated moving average models and seasonal autoregressive integrated moving

Fig. 3 An installed and operating hydraulic power plant [3]



average models are also widely applied in predictions. The applied statistical models are firstly trained according to the historical data, and then, the trained algorithm will be utilized to generate forecasting of the required parameter. Statistical models have been generally applied to the forecasting of wind power of the type ultra-short-term as well as the short-term. However, their estimation precisions are pretty limited. For instance, Liu et al. [8] suggested a recent model of recursive ARIMA and EMD (empirical mode decomposition). In the study of Liu et al. [8], wind speed forecasting was implemented to the railway strong wind warning system. On the other hand, in the North Dakota region, the wind speed for one-day and two-day ahead was forecasted using the developed algorithm of the fraction-ARIMA [9].

Apart from physical and statistical methods, artificial neural network (ANN) tool is the most common applied method to have a proper forecasting. Especially this tool is generally used in wind speed predictions with a large extent. The networks of the ANN can be obtained using back-propagation neural networks, radial basis functions neural networks, extreme learning machines, multilayer perceptron neural networks, and besides Bayesian neural networks. For instance as an artificial tool, long short-term memory network (LSTM) algorithm is utilized in the estimation of wind turbine output power [5]. In this study, the analysis of the error distribution characteristics for wind turbine power forecasting was conducted by the Gaussian mixture model (GMM). Machine learning predictions are also considered in predictions of solar energy. For example, to obtain the hourly solar irradiance forecasting, Khosravi et al. [10] applied a machine learning algorithm. Artificial learning was also considered in predictions of experimental data. In this regards, in the study of Shi et al. [11], LSTM method was applied to the wind power experimental results, which revealed more accurate forecasting outcomes when compared with respect to the other conventional neural networks. Artificial learning techniques including hybrid mechanisms were also considered in the literature. As an example, Yu et al. [12] had applied an innovative hybrid model consisting of support vector machine (SVM), long short-term memory neural network (LSTM), the standard recurrent neural network (RNN), as well as the gated recurrent unit neural network (GRU). In this study, they demonstrated the accuracy of the prediction results obtained by the hybrid model exceeded the precision of the forecasting outcomes obtained by the single type conventional methods. In the literature, within the scope of the same study, estimations were also performed by applying the same predictions methods to different disciplines. For instance, Han et al. [13] had performed wind and photovoltaic power production forecasting using copula function and LSTM. Besides, Zhang

et al. [14] had studied a hybrid type of model including a decomposition virtual node pruning and LSTM. In the study, they had obtained the forecasting of wind speed with a high precision.

The artificial learning in the literature also includes the control of some physical parameters by some other optimized physical parameters and estimations in this direction are also realized. In this context, a shared weight long short-term memory network (SWLSTM) was proffered by Zhang et al. [15]. In the study, the number of the optimized variables and the time required for the training were reduced. Besides, in the study, the increase of the forecasting precision was observed and reported to be improved with the SWLSTM utilizations [15]. Similarly, a model of multi-variable stacked LSTMs (MSLSTM) was utilized by Liang et al. [16]. The values of the wind speed data cluster in this model were forecasted according to the meteorological data of wind speed, pressure, temperature, humidity, dew point, and the solar radiation. Their outcomes demonstrated that the algorithm of the MSLSTM is suitable to capture and learn the uncertainties and propagate a competitive output performance [16]. Studies in the literature also include predictions depending on the time series formed of historical data of a certain physical parameter. For example, time series weather prediction was performed by Zaytar and Amrani, using deep neural network structure [17]. The results of the LSTM computations depending on the time series were demonstrated to be rather improved according to the results captured by the conventional methods [17]. Besides, time-series predictions of wind speed considering İstanbul, İzmir, Muğla, Tekirdağ, and Eskişehir cities of Turkey were studied by Ozen et al. [18]. The method of LSTM was utilized in order to obtain the forecasting based on the real historical wind speed data of these provinces. With the help of the trained algorithm with the events occurring in certain time periods, studies on the estimation of the event in the following but not yet realized next time period have also been carried out in the literature. For instance, by the utilization of one hour ahead clustering, LSTM was used by Prabha et al. [19].

Different algorithms of ANFIS consisting of fuzzy c-means (FCM), grid partitioning (GP), and subtractive clustering (SC) were employed in the study of Benmouiza and Cheknane [20]. The prediction outcomes of the study have indicated that ANFIS-FCM had provided superior results over the other two methods [20]. The literature includes also the comparison of the artificial learning in predictions according to the conventional linear and nonlinear regressions. In this regards, linear and nonlinear regressions were compared according to the ANN in the study of Bilgili and Sahin [21]. Meteorological data including wind speed, rainfall, ambient temperature, atmospheric pressure, and the relative humidity was

considered in this developed model of Bilgili and Sahin to obtain the prediction of the wind speed [21]. It was reported that machine learning of ANN was superior with respect to the conventional methods of linear and nonlinear regressions.

Predictions about the hydraulic events are also presented in the literature. In this aspect, a machine learning approach was applied in prediction of regional water flows of Lake Okeechobee [22]. This lake is the primary source which supplies water to south of the Florida. In this study, a computer algorithm was developed as an inductive technique. The technique utilized a learning system that involved a second-order table compression [22]. On the other hand, artificial neural network (ANN), the ant colony optimization (ACO), the genetic algorithm (GA), and the particle swarm optimization (PSO) were considered in the predictions of flood hydrograph [23]. Flow data of Tiber River found in central Italy was taken into account in training of the algorithms of these methods. The results obtained from the study showed that these four approaches were robust methods in forecasting of hydrographical incidents [23].

River water flow prediction was performed with the developed forecasting models using relevance vector machine (RVM) [24]. The approach included a probabilistic kernel-based learning machine. The results of the study indicated that a high accuracy would be obtained in river water flow predictions using the approach of RVM [24]. Besides, the support vector machine (SVM) was proposed as a technical approach that could be safely applied in the forecasting of flow in straight compound channels [25]. The statistical results of the study demonstrated that correlation coefficient (R^2), root-mean square error (RMSE), and mean absolute percentage error (MAPE) of the SVM were so far better with respect to the statistical error results of the other conventional methods [25]. Artificial neural network (ANN) was implemented using proper hidden nodes and a proper activation function in stream water flow guesses [26]. In this study, under these conditions, the model performance was shown to be enhanced in ANN utilizations for river water flow predictions [26]. A flood hazard rating forecasting model was suggested using long short-term memory (LSTM). Samseong District was chosen as the study location, since many floods have occurred in this region. The amount of the rainfall was determined using machine learning algorithms in this study [27]. On the predictions of discharge rates of stream flow of Tigris River of Iraq, comparison of the extreme learning machine (ELM), support vector machine (SVM), and generalized regression neural network (GRNN) was performed. The results of the study revealed that predictions of the ELM model outperformed the results of the other algorithms [28]. Runoff predictions were

performed for flood detections, using backpropagation (BP) neural network, extreme learning machine (ELM), generalized regression neural network (GRNN), and wavelet neural network (WNN) models [29]. In this study, the water flow rate and level of the Xijiang River was predicted at Wuzhou station. The best performance of the predictions was obtained by GRNN models in terms of the stream flow, whereas the best performance of the forecasting was obtained by WNN in terms of the water level [29].

The novelty of the current study includes machine learning algorithms of LSTM, ANFIS-FCM, ANFIS-SC, and ANFIS-GP in the forecasting of one-day ahead average RWFR. The predictions of future water flow rate were obtained using the past time series. Initially, four algorithms were trained using the real observed and measured water flow rate data of Harmanli measurement station located on the Maritsa River, in the vicinity of Haskovo province of Bulgaria, which is located on the border of Turkey and Bulgaria. When the training of the developed algorithms was completed, the algorithms were tested to demonstrate how the models were superior to each other. However, the novelty of this study consists of the application of the machine learning algorithms to the hydraulic water flow rate data. Besides, the novelty of the study also includes the demonstration of the influence of the selected number of historical data used for training of the algorithms. The prediction of RWFR plays a significant role of planning and building of new water dams, or operating the ones that were formerly built. In addition, the analysis of the water VFR of the rivers will help in monitoring of natural events, such as flooding in the river. By this way, possible damages of the natural events to surroundings may be prevented.

2 Materials and methods

2.1 Proposed methods

2.1.1 ANFIS systems

Adaptive neuro-fuzzy inference system (ANFIS) is a well-known and widely used universal predictor which can be applied to different branches of engineering applications. It is utilized for a true continuous function in a compact setup to any degree of accuracy. ANFIS is basically expressed as a network formed of Sugeno-type fuzzy systems. The network structure is, however, equipped with neural learning capabilities. If-then rules are created using appropriate membership functions (MF) from input and output with the employment of neural network learning algorithm. The FIS development operation that uses the

adaptive neural network is known as ANFIS [30]. There are many scientific studies available in the literature covering a general or detailed information on the ANFIS structure [31–34].

Fuzzy c-means (FMC) which is a clustering method of ANFIS allows each data point to have multiple clusters and enable the data point to be belonging to different degrees of membership. The minimization of the objective function gives the basis of the FCM algorithm. General and detailed information on ANFIS-FCM is available in the literature [34].

Each data point is considered as a candidate cluster center in subtractive clustering algorithm (SC). The calculation of the potential of each data point is performed through the measurement of the density of the data point surrounding the cluster center. An iterative process is utilized in the algorithm of the ANFIS-SC; in which each point is assumed to be potentially a cluster center, and the location of the points are defined according to the other data points. Literature also includes studies having the modeling of ANFIS-SC or the engineering applications using ANFIS-SC [20].

Finally, grid partitioning (GP) method of the ANFIS splits the input data space to a rectangular subspace by the utilization of the axis-paralleled partition. The split of each input is obtained by partitioning the inputs into membership functions of identical shape. The amount of the fuzzy if-then rules is calculated by the expression of M^n . In this expression, the input dimension and the amount of partitioned fuzzy subsets for each input variable are, respectively, demonstrated by n and M [20]. In this method, the functional decomposition is the approach to solve a problem. The literature also includes broad explanations of this method, demonstrated with the studies performed in this direction [35].

2.1.2 System of LSTM neural network

LSTM method was initially introduced in 1997 [36]. LSTM is a form of RNN that obtains the solutions of the problems in RNN by adding cell states or memory cells with constant errors. In this way, errors can be reconstituted without disappearing gradients. In an LSTM algorithm, three different gates are existed, namely those include input, output, and the forget gates. These gates control different learning activities of the algorithm. The protection of the constant error flow in the memory cell from unrelated inputs is learned by an input gate, whereas the protection of the other units from irrelevant memory content stored in the memory cell is learned by the output gate. Lastly, the way to control for how long the value will be in the memory cell is learned by the forget gate [37, 38].

The LSTM layer structure is presented in Fig. 4. As shown in this figure that the flow of an X time series is occurring with S -length and C properties (i.e., channels) across the considered LSTM layer. The output, which is also known as the hidden state as well as the cell state at time step t are indicated in this diagram with the designations of h_t and c_t , respectively. To have the calculation of the first output and to form the updated cell state, the first part of the LSTM is used for the initial state of the network and for the early time step of the series. The current state of the network, i.e., the network state at c_{t-1} and h_{t-1} , is utilized by the block at the time step of t . In this way, the next time step of the sequence is used by the block for the computation of the output and the updated cell state of c_t .

The state of the layer includes the hidden state as well as the cell state. The hidden state is also named as the output state. Accordingly, the output of the LSTM layer is existed at the hidden layer at the time step of t . The state of the cell includes information emerging from former time steps. The addition or removal of the information to or from the cell state is implemented by the layer, at each time step. The gates provide the control of these updates. The control of the cell state and the hidden state of the layer is actualized by some components of the LSTM layer structure. For instance, the input gate (i) audits the level of cell state update, whereas the output gate (o) audits the level of cell state supplemented to the hidden state. The reset of the level of the cell state, i.e., the forget of the cell state, is audited by the forget gate (f). Finally, the information is inserted to the cell state by the cell candidate (g).

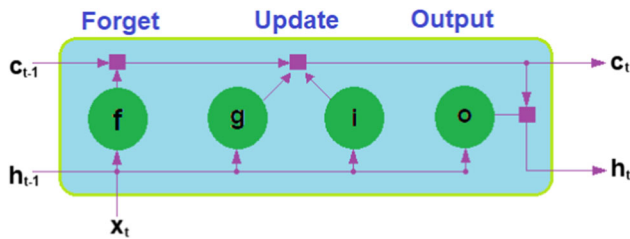
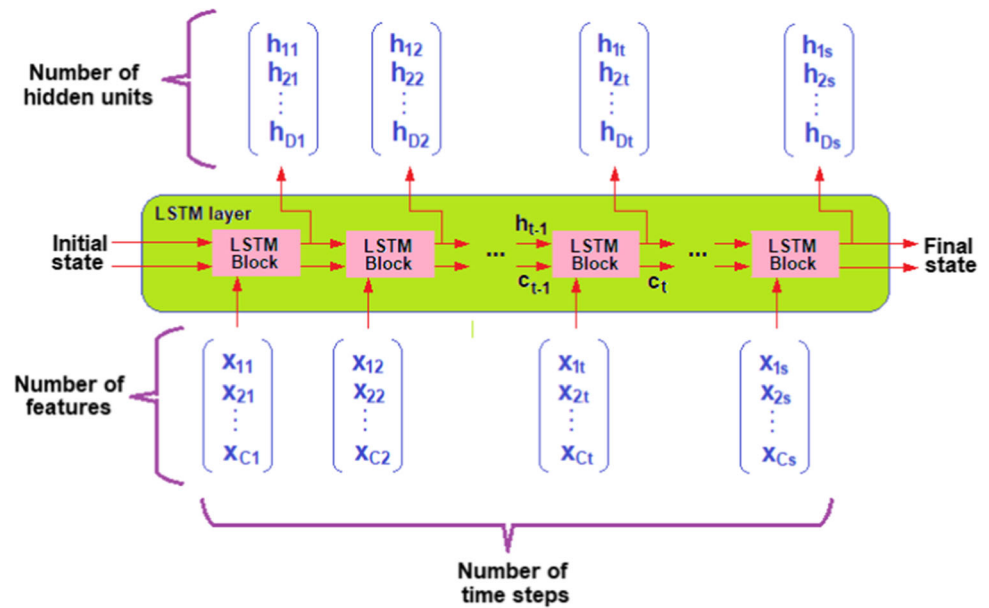
The flow of data at a time step of t is shown in Fig. 5. How the gates forget, how they update, or how they output the cell state and how they output the hidden states are displayed in this diagram of Fig. 5.

The learnable weights of an LSTM layer include the input weights, the recurrent weights, as well as the bias. These weights and the bias are, respectively, indicated with denotations of W , R , and b , as presented in Eq. (1). The matrices given in this equation with the cited expressions are the conjunctions of the input weights, recurrent weights as well as the bias, of each component. Namely, the concatenations of these matrices are provided as below;

$$W = \begin{bmatrix} W_i \\ W_f \\ W_g \\ W_o \end{bmatrix}, R = \begin{bmatrix} R_i \\ R_f \\ R_g \\ R_o \end{bmatrix}, b = \begin{bmatrix} b_i \\ b_f \\ b_g \\ b_o \end{bmatrix} \quad (1)$$

The subscripts given in the matrices of Eq. (1) with i , f , g , and o remark the input gate, the forget gate, the cell candidate, and the output gate, respectively. Therefore, the cell state has to be given as the function of time step of t , as exhibited in Eq. (2);

Fig. 4 LSTM layer structure

Fig. 5 The flow of data at a time step of t [39]

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (2)$$

Element-wise multiplication of the vectors, i.e., the Hadamard product, in Eqs. (2) and (3) is denoted by the designation of \odot . On the other hand, the hidden state at a time step of t has to be obtained with the Hadamard product of the output gate function at time step of t , multiplied on the state activation function (σ_c) performed on the cell state at time step of t . Namely, Eq. (3) will be defined as shown below;

$$h_t = o_t \odot \sigma_c(c_t) \quad (3)$$

The state activation function (σ_c) of the LSTM layer is computed by the utilization of the hyperbolic tangent function, by default. Formulas indicated in Eqs. (4), (5), (6), and (7) are used to describe the components at time step of t , as indicated below [39]:

$$i_t = \sigma_g(W_i x_t + R_i h_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma_g(W_f x_t + R_f h_{t-1} + b_f) \quad (5)$$

$$g_t = \sigma_c(W_g x_t + R_g h_{t-1} + b_g) \quad (6)$$

$$o_t = \sigma_g(W_o x_t + R_o h_{t-1} + b_o) \quad (7)$$

On the other hand, the gate activation function is shown with the abbreviation of σ_g . This function is utilized in Eqs. (4), (5), and (7). Besides, the sigmoid function is utilized by the LSTM layer for the computation of the gate activation function. Herewith, the gate activation function described with respect to the sigmoid function is presented in Eq. (8);

$$\sigma(x) = (1 + e^{-x})^{-1} \quad (8)$$

2.2 The statistical error analysis

In the current study, the goodness of the model was assessed using three statistical error parameters. Namely, mean absolute error (MAE), correlation coefficient (R), and root mean square error (RMSE) were used in the evaluation of the tested models. The accuracy of the VFR forecasting results was determined with respect to the real measured VFR data, using results of three statistical error parameters. The error calculation equations of these statistical methods have been shown in Eqs. (9), (10), and (11), respectively:

$$MAE = \frac{1}{N} \sum_{i=1}^N |p(i) - o(i)| \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N [p(i) - o(i)]^2} \quad (10)$$

$$R = \left(\sum_{i=1}^N [p(i) - \bar{p}][o(i) - \bar{o}] / \left(\sqrt{\sum_{i=1}^N [p(i) - \bar{p}]^2} \sqrt{\sum_{i=1}^N [o(i) - \bar{o}]^2} \right) \right) \quad (11)$$

In these equations, the anticipated and real observed values at i time are indicated, respectively, by the designations of $p(i)$ and $o(i)$. Besides, the average values of the forecasted and the real measured (observed) data are, respectively, demonstrated by the abbreviations of \bar{p} and \bar{o} in Eq. (11). Finally, the symbol N is utilized to exhibit the cumulative amount of the members of the data cluster.

3 Results and discussion

3.1 The interested area location

The study was carried out based on the hydraulic data of the Maritsa River, which is located in the northwest of Turkey. The river originates in Bulgaria, enters Turkey via Bulgaria, continues in the direction of Edirne, and flows toward the Aegean Sea to be finally poured there. To give a little more detail on the river route; the river first mixes with Arda River in the north of Edirne within the borders of Turkey. Later, Tunca River joins with Maritsa in the south of Edirne. Finally, the Ergene River mixes with the Maritsa River, which has been flowing by drawing the Turkish–Greek border for a long time, in the village of Adasarhanlı, thirty (30) kilometers away from İpsala district. The hydraulic data used in the current study has been obtained from the website of the 11th regional directorate of DSI [40]. DSI is a government institution related with the water works of the state. The hydraulic data of the Maritsa River has been obtained by the measurements performed in the Harmanli measurement station (MS). Harmanli MS is located in the Harmanli town of Haskovo province of Bulgaria, found on the border of Turkey and Bulgaria. Figure 6 demonstrates this interested area location, considered in this study. This figure involves location of Bulgaria in the World, also presents further details of the map consisting of the country and its neighbors, shows the close view of the location including the Maritsa (Meriç) river and three other rivers in this location, and demonstrates the detailed view of the VFR measurement station in the Harmanli town. In this map, the Bulgaria country, Haskovo province, Harmanli town, Maritsa River (Meriç), and Harmanli VFR measurement station found on this river are demonstrated using red coloring.

A total of 1391 daily average measurements have been considered in this study. Whole data was belonging to the time period between 01.12.2017 and 21.09.2021 dates. The data is the water volumetric flow rate (VFR) of the Maritsa River. In this station, the water flow was generally measured twice a day, namely at 08.00 and 16.00 o'clock. The study was initiated by the procedure of taking the arithmetic average of the flow data, applied for each day. The

physical unit of the instantaneous real data was m^3/s , and the considered unit in the simulations was also accordingly. Among 1391 data, 80% and 20% of the cumulative were used, respectively, for training and testing of the LSTM and the ANFIS algorithms.

3.2 About the security and privacy of the learning algorithms and the data analysis and the structure of the model

In the last decade, machine learning has also received a considerable attention in terms of the security and privacy related applications. The forecasted information is so big, and manual study on the data could be so hard as well as makes it more complicated to estimate the future. To overcome this problem, machine learning has been presented and implemented in many engineering and technical applications as well as in the computer science. During working or processing with a big data cluster, at an efficient time interval, a decisive solution will be reached to the technical problem with the utilization of a proper machine learning algorithm. Besides, the most accuracy will be obtained in the classifications of the data when a proper machine learning is applied. There is a wide range of machine learning algorithm applications or techniques in practice. But, the current researches indicate that machine learning applications or the data used in machine learning algorithms can be undefended to security attacks, malware detection, or even more. For this reason, privacy and security attacks over the past few years have been significantly taken into account including machine learning techniques, tools, applications, and systematic analyses in many scientific studies [41, 42]. In those studies, adversarial capabilities that may form against the training and testing stages of predictions were handled. In this regards, at the training stage, adversarial attempts such as data injection, data modification, and logic corruption may occur. Those attempts will probably result corruption of the utilized algorithm or will result badly trained model, since wrong data may be injected and distributed inside the training data or the training data may be polluted completely. All of such situations will unfortunately deteriorate the performance of forecasting [41].

On the other hand, adversarial attempts may also be actualized at the testing stage. For instance, those include white-box, black-box, and gray-box attempts. In white-box attempt, the used target model is completely known by the adversary including the model architecture, training data, and the model parameters, whereas, in the black-box attempt, an attack will be performed on the target model by sending a series of queries to the target model. In this context, model extraction attacks, adversarial attacks, as well as the model inversion attacks may be applied.



**Harmanli
VFR MS**

◀**Fig. 6** Detailed view of Harmanli VFR measurement station of Haskovo province of Bulgaria, located on the Maritsa River in the border of Turkey and Bulgaria

Finally, in gray-box attempt, the used target model is also completely known by the adversary as in the case of the white-box attempt; however, against the adversarial attack, the defense mechanism is not known by the adversary [41].

The malignant activity that may occur as a threat against the used algorithm in the current study was prevented with the Network Intrusion Detection (NID) systems. Since a NID system rapidly identifies any malignant web activity deteriorating the performance of the forecasting model, it demonstrates a high protecting capability. Moreover, anti-virus programs of the malware detecting types were also used. Since a malware detecting program is utilized to identify all types of defects in the system, this method had also showed a high protecting capability. It has been tried to protect the models used in this way in the current study and to ensure their confidentiality and security [42].

In statistical methods and estimations, physical data are sometimes defined as mathematical numbers and within a certain narrow range, independent of their physical properties or units. This method is referred as the normalization. Following equation is valid for the normalization of physical quantities:

$$X_N = \frac{X_R - X_{\min}}{X_{\max} - X_{\min}} \quad (12)$$

When a data cluster of any type of physical quantity that involves a data series to be in a range of $X_{\min} \leq X_R \leq X_{\max}$ is considered; minimum, maximum, and any data in this range are, respectively, indicated in Eq. (12) above, by X_{\min} , X_{\max} , and X_R . The outputs of this normalization function turn physical data to normalized dimensionless data as well as the data is now distributed in a narrow range of $0 \leq X_N \leq 1$. Thus, in this way, it will be possible to observe the distributions of the data more easily, and the tendency of the data will be easily followed.

It has been primarily detected that the methods considered in this study can be used in cases initially where the amplitude of the data is high in the range of 0–1, and secondly if there is a large amount of fluctuations. To test the prediction ability of the algorithms, without the application of any arithmetic mean, the raw data of volumetric flow rate which formed a data cluster, was investigated considering the total of 102 models. But initially, the data has been normalized according to Eq. (12). Afterward, MAE, RMSE, and R statistical error results of the tested model have been obtained. In this regards, the performance of 102 models is shown in Fig. 7. Namely, Fig. 7a indicates that the MAE among the tested normalized models

was in a range of $0.005881 \leq \text{MAE} \leq 0.064217$. Besides, the RMSE among the tested normalized models was reported to be in a range of $0.009639 \leq \text{RMSE} \leq 0.211213$. Finally, the R values were shown to be in a range of $0.550478 \leq R \leq 0.997667$. The range of the statistical error results of the MAE has indicated a high precision has been obtained in the predictions. Besides, the average value of the MAE considering all tested models was reported to be 0.029944. In statistical error analysis of the normalized values, although the upper limit of the value range of the RMSE parameter and the lower limit of the value range of the R parameter for all tested models seem, respectively, high and low, these were obtained only in a few model trials, as shown in Fig. 7b, c. In other words, low RMSEs and high R 's were obtained at normalized dimensionless forecasted values with a large percentage of the trials. In this regards, the average of the RMSE of all models was determined to be 0.029944 and the average of R of all models was determined to be 0.973823. These results showed that high success was achieved in the normalized dimensionless forecasting values.

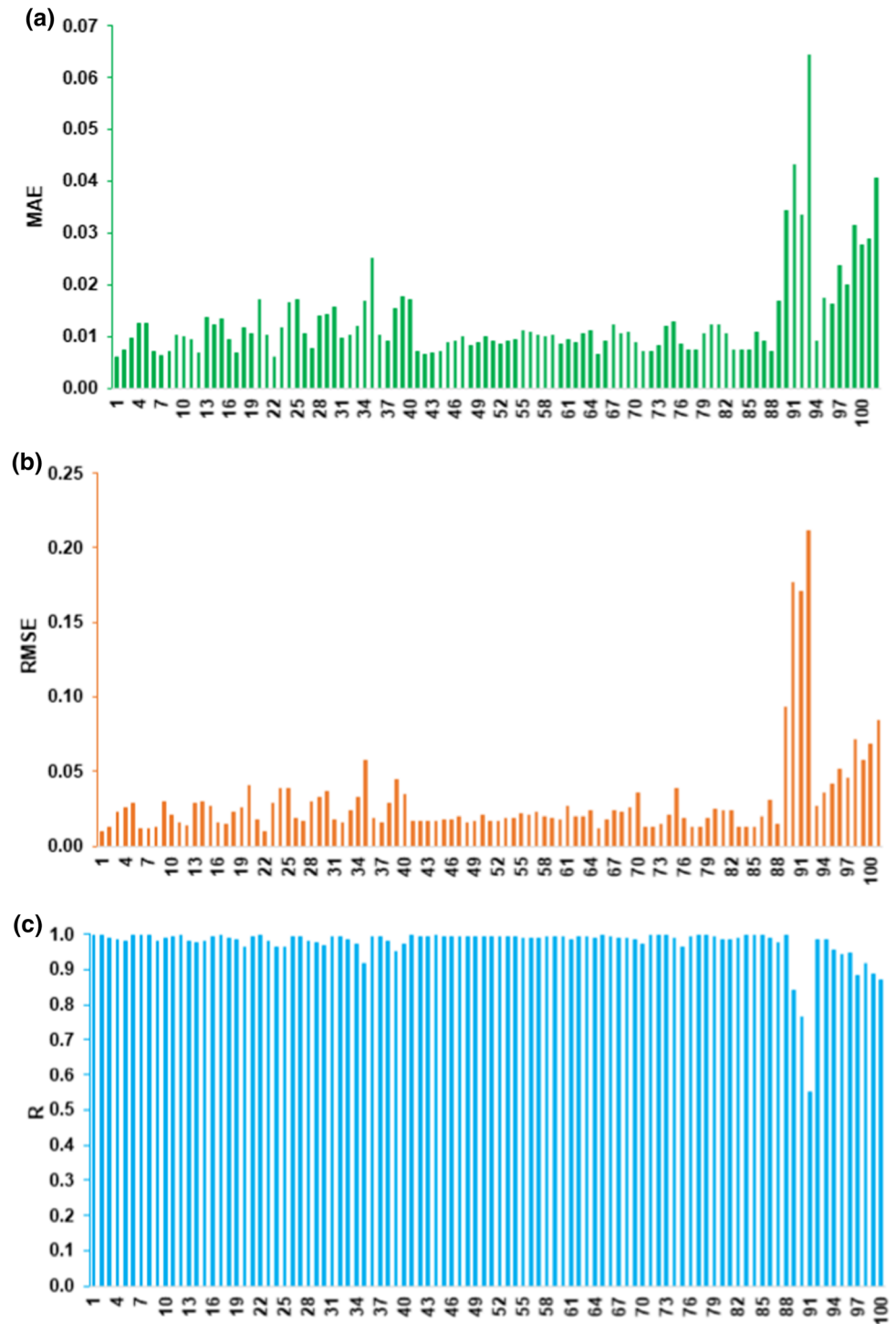
Besides, considering the best result, the normalized distributions of the real observed and the predicted VFR data are demonstrated in Fig. 8.

Thus, considering Fig. 7, it was decided that the prediction methods could be applied independently of the physical dimensions of the data; since in general the results of the MAE and RMSE of the normalized data approach to zero for most of the models and the results of the R of the normalized data approach to 1, again in most of the models. Moreover, when Fig. 8 is examined, it is observed that the normalized real and normalized forecasted VFR data have a high overlap with each other.

Since normalized flow data also shows random distributions in the range of 0–1, and because of their high amplitudes and high fluctuations, it was decided that these methods can be easily applied in the volumetric flow rate (VFR) estimations. Namely, while the river sometimes flows at very high flow rates in the winter when the snow melts, whereas sometimes due to the climate conditions, the river has a very low water level, i.e., it may be almost close to the dryness; so, there may be situations where almost no water flows in the riverbed. This situation of the nature causes a high amplitude and high fluctuation of the VFR data. Similarly, it is concluded that these methods would also be appropriate in estimating other types of data that show a feature this way.

Since these estimation algorithms give high performance in random distributions of data and since river flow rates generally have fluctuating, high amplitude, and random distributions of data; that is, sometimes very high flow rate values occur during the year and sometimes very low flow rate values occur during the year, it has been

Fig. 7 MAE, RMSE, and R results of forecasted normalized dimensionless VFR data

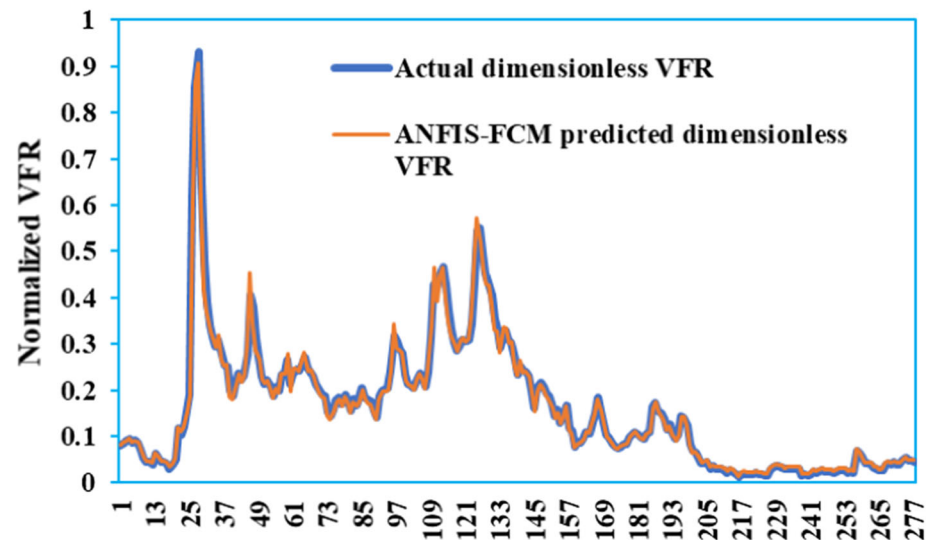


concluded that these algorithms are quite suitable for estimating river volumetric flow rates (VFR).

In the present study, VFR forecasting has been performed using the measured data of an installed hydraulic flow rate MS on the Harmanli location of the Maritsa River. The predictions of the VFR were initially actualized based on the time-series analysis considering fuzzy

c-means (FCM), subtractive clustering (SC), and grid partition (GP) algorithms of adaptive neuro-fuzzy inference system (ANFIS). This was followed by the VFR predictions executed by the long short-term memory (LSTM) neural network. Considering four types of data clusters obtained from the applied models, the aim of the study was to demonstrate the testing model validation on

Fig. 8 Dimensionless normalized VFR and its counterpart ANFIS-FCM predicted data



the measured VFR, according to the algorithms obtained on the training stage. Namely, the real and the predicted volumetric flow data cloud were compared on the statistical methods, shown in Eqs. (9), (10), and (11).

In the LSTM approach, the range of the cumulative number of the hidden layers was taken into account in between 5 and 300. The epoch number was set a constant value of 300. On the other hand, ANFIS structures involved the derivation of the forecast outputs from the inputs based on the principle of the Sugeno's fuzzy approach. Trial and error method was used in the adjusting of the selected number of historical data for training, optimal model structures, the amount of the membership functions (MFs), as well as the cumulative number of the iterations; thus, a set of ANFIS structures were tested. In this context, a total of 40 ANFIS-FCM models, 48 ANFIS-SC models, 5 ANFIS-GP models, and 9 LSTM models were tested for the comparison of the superiority of the prediction results. In the utilized ANFIS-FCM approach, the selected number of historical data, which was used for training, was considered in a range between 3 and 10, with an increment of 1. Besides, the number of the MFs was taken into account with an increment of 2, in the range of 2–10. Namely, a total of 5 MFs were used for each selected number of historical data, which was used for training. Furthermore, the max number of the iterations as indicated by the *max epoch* column of Table 2 was set to 100. Similarly, in the ANFIS-SC algorithm approach, the same procedure was followed for the selected number of historical data for training. Namely, one incremental increase for the range considering 3–10 was applied. During training of the ANFIS-SC models, in order to obtain the best clustering, the simulations were started with two different radii values for each historical data. Namely, the simulations were initiated at radii values of $R = 0.20$ and $R = 0.21$. After

$R = 0.21$, the simulations were continued with two increments on the second significant digit after the comma, on each trial, up to $R = 0.29$. The max epoch number was considered as in the case of ANFIS-FCM; i.e., kept constant of 100, so as not to prolong the computation time. Finally in the GP model of the ANFIS, only 3 and 4 were used as the selected number of historical data used for training. The number of the MFs was considered 2, 3, and 4 for ANFIS-GP algorithm. Since learning of the algorithm in this model takes a lot of time due to the complexity of the equations, the max epoch number was set to 30, instead of 100.

A general knowledge on the VFR data cluster of the Harmanli MS is provided in Table 1. As demonstrated in this table, the VFR values of the MS include a data set in the range of 01.12.2017 to 21.09.2021. According to the measurements, the VFR values were to occur in a data set of the range $23.50 \text{ m}^3/\text{s} \leq \text{VFR} \leq 495.50 \text{ m}^3/\text{s}$. Namely, a cumulative of 1,391 VFR measurements were performed in this time range and data value range, in which it formed the

Table 1 General information on the measured VFR data of Harmanli MS

Data	Harmanli MS
Minimum value (m^3/s)	23.50
Maximum value (m^3/s)	495.50
Arithmetic mean value (m^3/s)	81.74
Standard deviation (m^3/s)	60.02
Data date range	01.12.2017 to 21.09.2021
Number of cumulative data (100%)	1391
Number of training data (80%)	1114
Number of testing data (20%)	277

data cloud of this study. The analysis of the data has revealed that the arithmetic mean value of the 1391 data corresponds to $81.74 \text{ m}^3/\text{s}$. This situation means that this river has an average flow regime of $81.74 \text{ m}^3/\text{s}$, during this considered time period. The training and testing portions of the VFR data cluster have been indicated, respectively, with blue and red colors, as shown in Fig. 9.

Initially, the training of the utilized algorithms was performed with the usage of the training data cluster, and it was followed by the testing of the algorithms with the testing cluster. Whole properties of the measured data of the Harmanli MS are demonstrated in Table 1, including the standard deviation of the analyzed data, and the amount of the training and testing data.

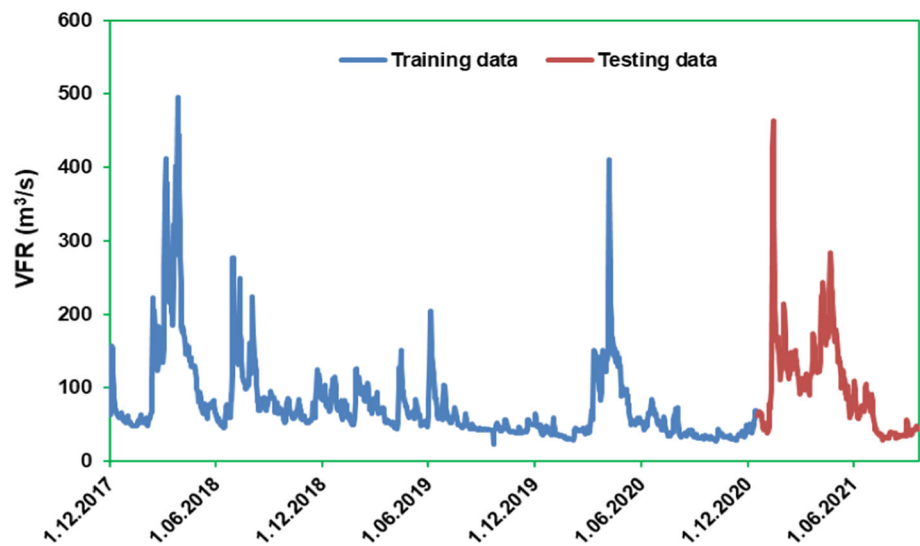
In Fig. 10, the analyzed data shown in Table 1 has been presented with different VFR data classes. This figure involves the analysis of 8 different data classes. These data classes involve the proportional cumulative frequency analysis in a data range in between the minimum and maximum VFR values of $23.5 \text{ m}^3/\text{s}$ and $495.5 \text{ m}^3/\text{s}$. The data classes have been discretized by data steps of $59 \text{ m}^3/\text{s}$. Frequency analysis showed that the Maritsa River generally flows between $23.5 \text{ m}^3/\text{s}$ and $82.5 \text{ m}^3/\text{s}$ volumetric flow rate. That is, the water flow in this range of volumetric flow rate corresponds to 68.5% of the total data. Besides, the proportional frequency indicated that the frequency in the data class of 82.6–141.5, corresponded to 19.8%. In brief, the cumulative frequency considering the data classes in the range of $23.5\text{--}200.5 \text{ m}^3/\text{s}$, corresponds to 95.6% of the total data. Flow values above $200.5 \text{ m}^3/\text{s}$ were not observed or measured much, and water flows equal to or over $200.6 \text{ m}^3/\text{s}$ up to $495.5 \text{ m}^3/\text{s}$ were generally occurred due to torrent and flood cases, which were rarely observed.

3.3 Systematics of the applied forecasting methods and the outputs obtained

3.3.1 Results of the ANFIS-FCM model

The results of the FCM models of the ANFIS are given in Table 2. The forecasting data was obtained according to the outcomes of the testing processes. Besides, the testing procedure was performed with the trained algorithm during the training stage. The novelty of this study includes the application of the artificial intelligence to the fluids, in particular to the volumetric flow parameter; and comprises the changing of the amount of the historical data in the estimation of the current data. A cumulative of 40 models were considered in the FCM algorithm of the ANFIS. The best model among whole 40 models was concluded to occur at the used historical data for training corresponding to 3 and at the MFs of 2. In this best model, three statistical models of MAE, RMSE, and R were, respectively, obtained as $2.54 \text{ m}^3/\text{s}$, $4.35 \text{ m}^3/\text{s}$, and 0.9981. However, it is seen in this table that 40 models gave quite acceptable statistical error results. At certain selected numbers of historical data used for training, for instance at 4 of the historical data and 2 of MFs, at 5 of the historical data and 4 of MFs, at 6 of the historical data and 4 of MFs, at 7 of the historical data and 4 of MFs, at 8 of the historical data and 4 of MFs, at 9 of the historical data and 4 of MFs, at 10 of the historical data and 4 of MFs, prediction values very close to the real data were obtained. In Table 2, the best results of the relevant set containing the selected amount of historical learning data between 3 and 10 are shown in bold. On the other hand, the best result among the best ones has been shown in bold and italics in this table for historical data corresponding to 3 and MFs corresponding to 2. The slightly poor results which were obtained in the simulations

Fig. 9 Time series of the real measured daily average VFR data of Harmanli



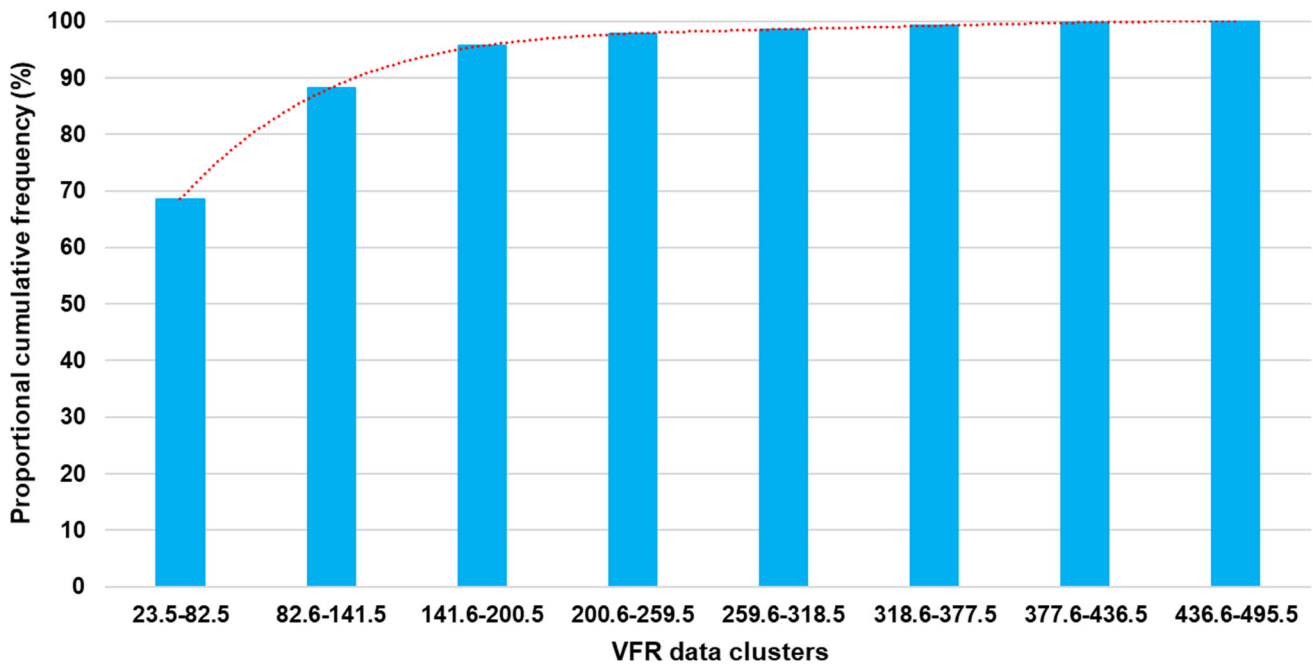


Fig. 10 Proportional cumulative frequency (%) of the real measured VFR data

and shown in this table were occurred due to not-well partitioning of the inputs, during the training stage. It was not preferred to use historical data exceeding 10, since the computation time was significantly increased when selected historical data had exceeded 10. Besides, it was observed that exceeding the number of the historical data over 10 did not cause a significant increase on the accuracy of the computations. Although the accuracy has not dropped much, it was not improved either with respect to the increase of the historical data.

Considering the comparison of the functions belonging to the actual measured and the forecasted data, the best result, among the whole simulations, obtained at the historical data of 3 and at the number of MFs of 2 is provided in Fig. 11. As presented in this figure, at the cited historical data and MFs, this suggested model in VFR prediction was concluded to match well with the real VFR data. For this best result, the regression plot of the real measured values with respect to the predicted ones obtained with the FCM algorithm of the ANFIS is shown in Fig. 12. The average correlation coefficient value (R) of the whole data set has been obtained by taking the average value of the whole daily VFR correlation coefficients; and this average value of the correlation coefficient has also been indicated in this figure. Besides, the daily correlation coefficient (R) was calculated according to the comparisons of the instantaneous real measured and the forecasted members, of the data set. The regression plot of Fig. 12 actually belongs to the obtained model at the historical data corresponding to 3 and at the number of MFs corresponding to 2. The

correlation plot and the result of the value of the correlation coefficient, not only represent the actual and forecasted value distributions, but also these results lead to understand how accurately the forecasting values of the applied model match with the data of the real observed values.

3.3.2 Results of the ANFIS-SC model

The statistical evaluation criteria for the ANFIS-SC models are demonstrated in Table 3, in order to indicate the accuracy of the prediction results. As similar to the FCM algorithm of the ANFIS, selected number of historical data for training has been also controlled in the SC algorithm of the ANFIS. It was only possible to construct the algorithms around $R = 0.20$; but at farther radii values, the training of the algorithms did not take place and during the testing phase, the algorithm gave values that were far from the measured VFR values and sometimes negative values of the VFR were obtained. For this reason, the influence of the radius was considered in the range $0.20 \leq R \leq 0.29$. The number of the iterations was kept constant at 100 for the maximum epoch. Briefly, a total of 48 models were taken into account in the SC algorithm of the ANFIS. Table 3 gives the outcomes of the model estimations which were obtained at the testing stages of the ANFIS-SC models. Namely, the forecasting outcomes of the models have been obtained at the testing stages, as soon as the training stages were completed. When three accuracy statistical comparison parameters are taken into account, it is concluded in Table 3 that the studied ANFIS-SC models generated more

Table 2 ANFIS-FCM statistical results

Selected number of historical data for training	Number of MFs	Max epoch	MAE (m ³ /s)	RMSE (m ³ /s)	<i>R</i>
3	2	100	2.54	4.35	0.9981
	4	100	3.99	6.38	0.9918
	6	100	4.01	7.56	0.9874
	8	100	4.02	8.12	0.9926
	10	100	5.79	11.13	0.9867
4	2	100	3.04	4.69	0.9977
	4	100	3.52	5.52	0.9969
	6	100	3.70	6.76	0.9948
	8	100	4.02	7.81	0.9862
	10	100	5.10	10.71	0.9872
5	2	100	3.50	5.86	0.9966
	4	100	3.35	5.48	0.9968
	6	100	3.91	6.00	0.9959
	8	100	4.73	9.74	0.9892
	10	100	4.91	9.32	0.9906
6	2	100	3.69	6.41	0.9960
	4	100	3.51	5.69	0.9964
	6	100	4.19	6.77	0.9948
	8	100	4.73	8.40	0.9919
	10	100	5.27	10.57	0.9879
7	2	100	3.89	6.64	0.9956
	4	100	3.39	5.29	0.9969
	6	100	3.98	6.33	0.9955
	8	100	5.18	10.24	0.9880
	10	100	6.06	13.67	0.9792
8	2	100	3.99	7.03	0.9950
	4	100	3.62	5.61	0.9965
	6	100	4.86	10.10	0.9883
	8	100	5.89	11.73	0.9843
	10	100	6.28	15.36	0.9729
9	2	100	4.19	7.13	0.9948
	4	100	3.67	6.02	0.9959
	6	100	4.97	10.10	0.9883
	8	100	5.28	12.28	0.9828
	10	100	6.82	18.51	0.9604
10	2	100	4.25	7.41	0.9944
	4	100	3.74	6.13	0.9958
	6	100	4.86	10.01	0.9886
	8	100	5.75	12.51	0.9820
	10	100	6.66	17.11	0.9663

The best results are shown in bold

or less similar results. However, the best anticipation result was obtained when the selected number of historical data used for training was set to 3 and the influence of the radius was adjusted to $R = 0.20$. Accordingly, the statistical accuracy measures at this simulated model were

determined to be 2.81 m³/s MAE, 5.10 m³/s RMSE, and 0.9978 *R*. On the other hand, as indicated in this table, among the tested 48 models, 8 models were observed to be superior to the others for each number of historical data. For instance, at 4 of the historical data and 0.20 and 0.21 of

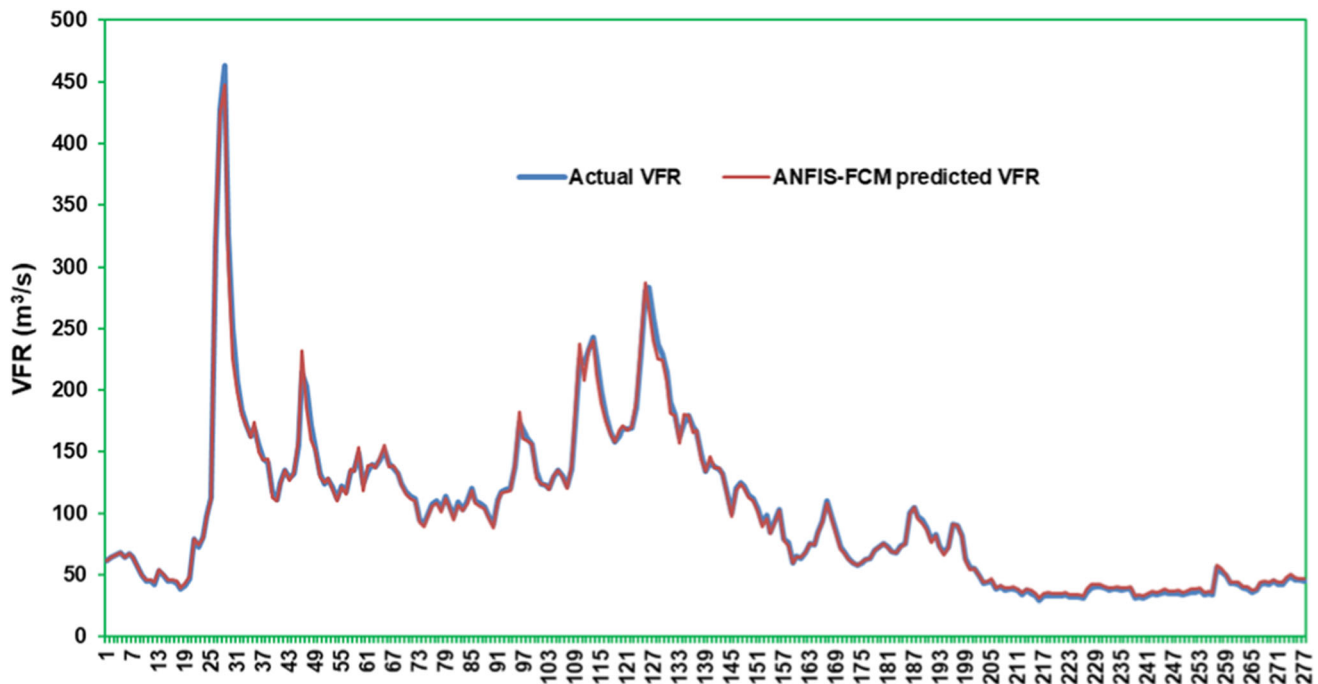


Fig. 11 The actual daily average time series data of the VFR and corresponding ANFIS-FCM predicted data

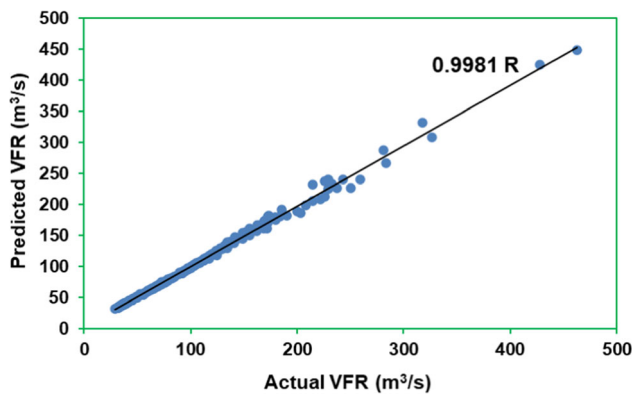


Fig. 12 Statistical regression plot for the actual daily average VFR data and corresponding ANFIS-FCM predictions

radii, at 5 of the historical data and 0.20 of radius, at 6 of the historical data and 0.20 of radius, at 7 of the historical data and 0.23 of radius, at 8 of the historical data and 0.20 of radius, at 9 of the historical data and 0.20 of radius, at 10 of the historical data and 0.20 of radius, prediction values very close to the real data were obtained. The best outcomes of the relevant set containing the selected amount of historical learning data between 3 and 10 are demonstrated in bold. On the other hand, the best outcome among the best ones has been given in bold and italics (i.e., at the selected number of historical data used for training corresponding to 3 and at the $R = 0.20$). These outcomes have indicated that the models which were not shown by bold, caused slightly lower mapping of the model. However, in

the models that were shown by bold in Table 3 permitted excellent mapping of the model. On the other hand, in the training stage, the increase of the selected number of historical data used for training over 10 had caused over-fitting or undesirable input memorization. Consequently, significant amount of errors have been obtained, in which huge differences between the real measured and estimated VFR values occurred, or forecasted VFR values were even obtained at negative values. For this reason, more than 10 historical data utilizations were not analyzed and included in this table.

ANFIS-SC forecasted VFR data presented with respect to the real measured daily average time series VFR data is shown in Fig. 13. This figure involves the best VFR predictions obtained by the SC algorithm of the ANFIS. Namely, the anticipations obtained at the selected number of historical data used for training corresponding to 3 and the influence of radius corresponding to $R = 0.20$ have been presented in this figure. The VFR predictions obtained by ANFIS-SC and the real observed time series data are reported to coincide well in this figure. Furthermore, the regression plot of the real measured and the forecasted values of the Harmanli VFR achieved with the SC tool of the ANFIS is demonstrated in Fig. 14. The mean value of the correlation coefficient (R) which is also indicated in this figure, has been obtained in the same way of ANFIS-FCM. Namely, the average value of the correlation coefficient (R) was determined using the daily correlation coefficient (R) values which were calculated according to the

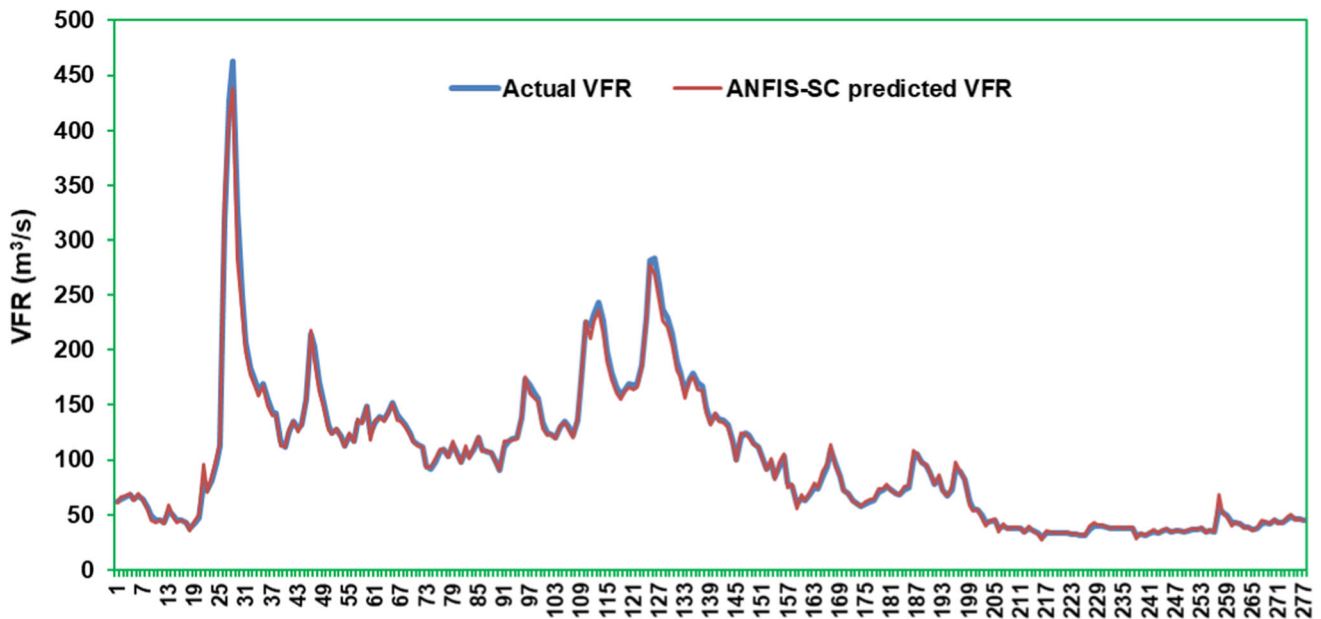
Table 3 ANFIS-SC statistical results

Selected number of historical data for training	Influence radius	Max epoch	MAE (m ³ /s)	RMSE (m ³ /s)	R
3	0.20	100	2.81	5.10	0.9978
	0.21	100	2.82	5.10	0.9978
	0.23	100	2.85	5.11	0.9978
	0.25	100	2.87	5.13	0.9977
	0.27	100	2.90	5.13	0.9977
	0.29	100	2.95	5.17	0.9977
4	0.20	100	2.99	5.05	0.9978
	0.21	100	2.99	5.05	0.9978
	0.23	100	3.02	5.06	0.9978
	0.25	100	3.08	5.06	0.9978
	0.27	100	3.11	5.08	0.9978
	0.29	100	3.12	5.09	0.9977
5	0.20	100	3.01	5.45	0.9973
	0.21	100	3.02	5.44	0.9973
	0.23	100	3.03	5.44	0.9973
	0.25	100	3.06	5.45	0.9973
	0.27	100	3.08	5.44	0.9973
	0.29	100	3.12	5.44	0.9973
6	0.20	100	3.08	5.47	0.9973
	0.21	100	3.11	5.47	0.9973
	0.23	100	3.14	5.45	0.9973
	0.25	100	3.15	5.46	0.9973
	0.27	100	3.15	5.45	0.9973
	0.29	100	3.17	5.46	0.9973
7	0.20	100	3.10	5.49	0.9973
	0.21	100	3.10	5.49	0.9973
	0.23	100	3.09	5.49	0.9973
	0.25	100	3.14	5.48	0.9973
	0.27	100	3.13	5.49	0.9973
	0.29	100	3.14	5.47	0.9973
8	0.20	100	3.28	5.84	0.9968
	0.21	100	3.31	5.87	0.9968
	0.23	100	3.32	5.88	0.9968
	0.25	100	3.32	5.87	0.9968
	0.27	100	3.35	5.88	0.9968
	0.29	100	3.36	5.88	0.9968
9	0.20	100	3.40	6.02	0.9966
	0.21	100	3.43	6.06	0.9965
	0.23	100	3.47	6.08	0.9965
	0.25	100	3.48	6.09	0.9965
	0.27	100	3.47	6.07	0.9965
	0.29	100	3.50	6.08	0.9965
10	0.20	100	3.45	6.12	0.9964
	0.21	100	3.46	6.12	0.9964
	0.23	100	3.49	6.14	0.9964
	0.25	100	3.51	6.15	0.9964
	0.27	100	3.52	6.15	0.9965

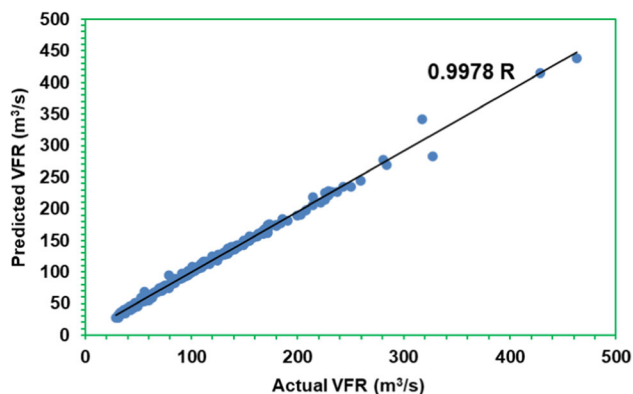
Table 3 (continued)

Selected number of historical data for training	Influence radius	Max epoch	MAE (m ³ /s)	RMSE (m ³ /s)	<i>R</i>
	0.29	100	3.52	6.14	0.9964

The best results are shown in bold

**Fig. 13** The actual daily average time series data of the VFR and corresponding ANFIS-SC predicted data

comparisons of the instantaneous real measured and the forecasted members, of the data set. Similarly in this model, the correlation plot and the average value of the *R* of the data set indicate actual and predicted value distributions, as well as give an understanding of the accuracy of the applied ANFIS-SC model to the real VFR time series.

**Fig. 14** Statistical regression plot for the actual daily average VFR data and corresponding ANFIS-SC predictions

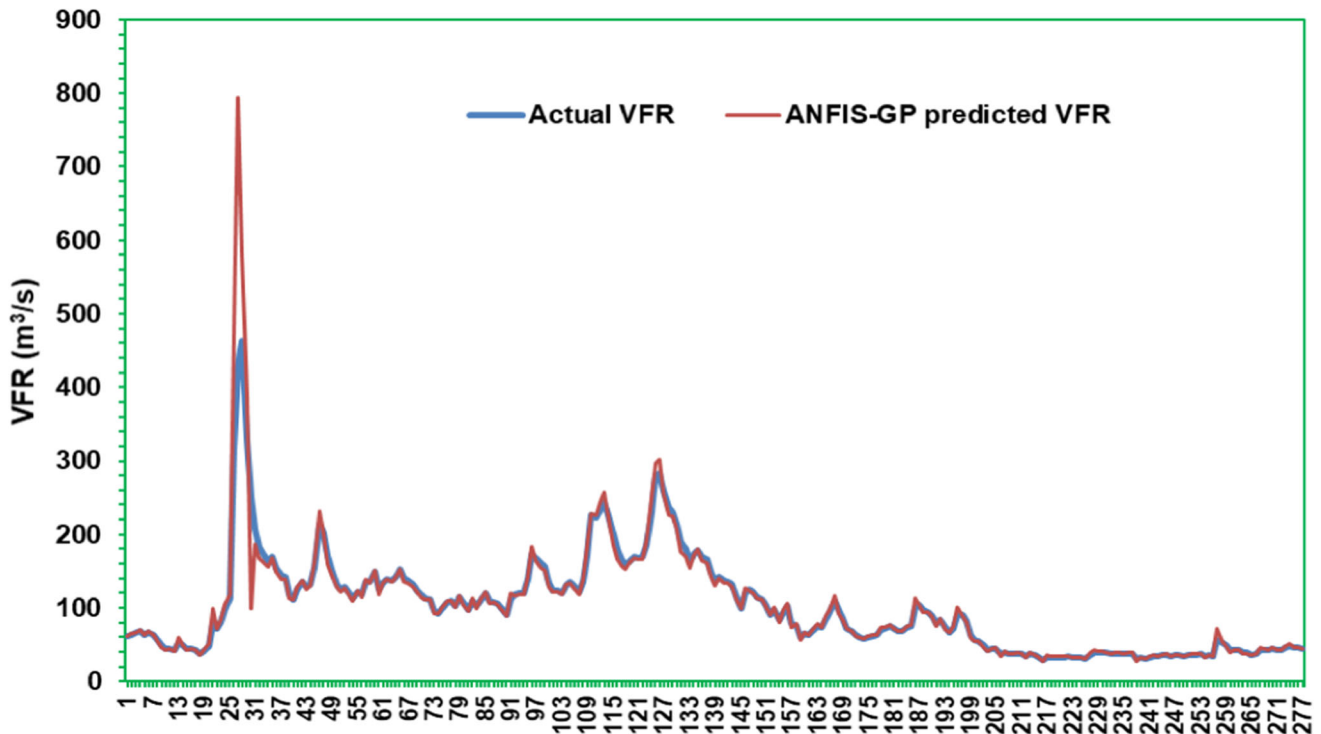
3.3.3 Results of the ANFIS-GP model

In the GP predictions, a similar forecasting methodology was applied. In the GP algorithm of the ANFIS, Gaussian membership and linear membership functions were utilized on the input and the output data. Due to the increase of the computation time, and not good grid-partitioning of the data, selected number of historical data used for training higher than 4, number of MFs exceeding 4, and the max. epoch number over 30 were not considered in the analyses. The most accurate and successful model was obtained when the number of the historical data was set to 3 and the number of MFs was adjusted to 2, as indicated in Table 4. Accordingly, Fig. 15 demonstrates the variation of the predicted VFR data, obtained by this best model of GP, with respect to the real measured VFR time series. When the outcomes obtained by the other two algorithms of the ANFIS are considered, it has been concluded that ANFIS-GP algorithm to be not so successful in predictions of VFR fluctuating data. However, the best result, obtained at the historical data of 3 and at the number of MFs of 2, is given in Table 4 with the bold; and it has been conducted that this

Table 4 ANFIS-GP statistical results

Selected number of historical data for training	Number of MFs	Max epoch	MAE (m ³ /s)	RMSE (m ³ /s)	<i>R</i>
3	2	30	6.11	27.30	0.9487
	3	30	900.37	9570.10	0.1062
	4	30	1848.69	17,774.48	0.3855
4	2	30	35.43	273.78	0.1525
	3	30	4370.79	36,194.18	0.0728

The best result is shown in bold

**Fig. 15** The actual daily average time series data of the VFR and corresponding ANFIS-GP predicted data

model is significantly sufficient than the other four models presented in the table. In this model, the statistical comparison outputs of MAE, RMSE, and *R* were computed as 6.11 m³/s, 27.30 m³/s, and 0.9487, respectively. In this best model, testing part of the GP algorithm of the ANFIS displays that the guesses belonging to the VFR time series coincide sufficiently well with the real measured values of the VFR. If to analyze the outputs of the ANFIS-GP algorithm in a closer look; Fig. 15 can be said to present the detailed analysis of the testing part of Fig. 9, which was shown with the red curves. So, while Fig. 15 gives the testing part of Fig. 9, it also exhibits the corresponding estimates at each instantaneous daily average data.

The regression plot for the estimates obtained by the ANFIS-GP algorithm, including the comparison of the real measured VFR and the forecasted VFR data, is presented in

Fig. 16. The mean value of the correlation coefficient (*R*) for the best result of GP is also indicated in this figure. This value of the *R* has been obtained in the same way of the procedure followed in the previous ANFIS tools. Namely, the ultimate *R* value was computed utilizing the measured and the forecasted daily average VFR correlation coefficient (*R*) values. The averages of the instantaneous *R* values determined the ultimate average *R* value of the whole data set. The scatter of the correlation plot and the ultimate value of the *R* of the whole data, given at the best result present the distributions of the real measured and the forecasted VFR values. Although the figure displays the degree of the accuracy of the ANFIS-GP model that was applied for VFR predictions, it is certain that forecasting capability was lower in this model. This is especially well seen at high water flow rates, by observing the actual

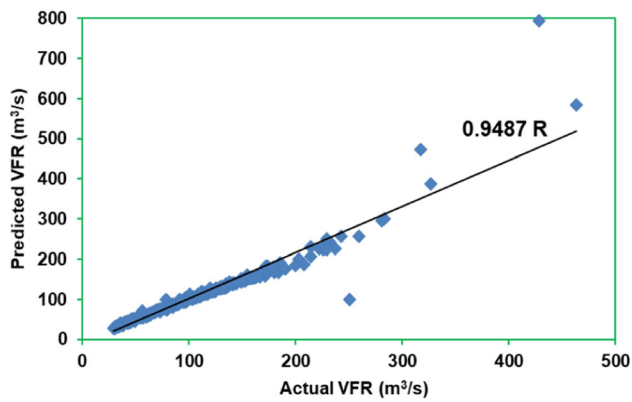


Fig. 16 Statistical regression plot for the actual daily average VFR data and corresponding ANFIS-GP predictions

measured values, which significantly diverge from the regression curve. Eventually, it has been concluded that although this model is slightly poor in predictions; it was decided to be still sufficient in fitting of the predictions to the real observed VFR time series.

3.3.4 Results of the LSTM

The VFR forecasting results obtained by the LSTM algorithm are given in Table 5. Similarly, statistical evaluation criteria such as MAE, RMSE, and R have been used to compare the prediction outcomes at the testing stage.

In LSTM modeling, a total of 9 models have been tested and compared. The number of the hidden layers was altered between 5 and 300. However, as indicated by bold in Table 5, the best result among the tested models was obtained at the hidden layer number corresponding to 5. In this number of hidden layer, the VFR statistical error results of 4.02 m³/s, 12.14 m³/s, and 0.9859 were obtained, respectively, for MAE, RMSE, and R . The obtained results have demonstrated that the suggested LSTM model in VFR forecasting performed a satisfactory performance. It has been reported that this result is very good compared to

ANFIS-GP and approximates the other two methods of ANFIS.

The LSTM predicted VFR data with respect to the VFR time series including real observed data is shown in Fig. 17. This figure shows the results of the model having the best predictions, i.e., the model which performed the predictions at the hidden layer number of 5 and epoch number of 300. The number and the row of the daily average VFR samples and the corresponding VFR values are, respectively, located on the x and y axes of this figure. As shown in this figure, although there is a significant fluctuation in the average values of the daily measured samples of the VFR time series data, the LSTM technique which was applied to the VFR data of the Harmanli MS has revealed that the anticipated outcomes overlapped with the actual observed values. Besides as given in Table 5, the sufficiency of the overlapping incidents, which are obtained in Fig. 17, was as well proved based on the statistical calculations. So, it is concluded and reported that the usage of the LSTM neural network in such sinusoidal data clusters results satisfactory prediction outcomes. On the other hand, to show the relevance and the accuracy of the predictions according to the observed data, the regression plot of correlation coefficient which was obtained by the LSTM algorithm is presented in Fig. 18. In this figure, similarly, the measured and the predicted values of the Harmanli VFR measurement station are, respectively, denoted in x-axis and y-axis. The volumetric water flow rate values are provided with the unit of m³/s on these axes. The best correlation coefficient outcome of 0.9859 R among the tested models given in Table 5, which is obtained at the hidden layer number of 5, is also indicated in this figure as accompanying the actual and the predicted value distributions. To briefly explain the success of the LSTM estimation method, the water flow rate, which reaches a maximum of 495 m³/s, is calculated differently with an error of 4 m³/s and 12 m³/s. That is, the estimation of the flow rate values of this size with such low errors reveals the success and the accuracy of the LSTM method.

Table 5 LSTM statistical results

Hidden layer number	Epoch number	MAE (m ³ /s)	RMSE (m ³ /s)	R
5	300	4.02	12.14	0.9859
10	300	8.18	16.87	0.9857
25	300	7.64	18.63	0.9596
50	300	9.02	20.80	0.9534
75	300	8.49	20.34	0.9517
100	300	15.23	34.26	0.8657
125	300	12.67	31.28	0.8958
150	300	18.27	39.69	0.8715
300	300	7.60	19.71	0.9582

The best result is shown in bold

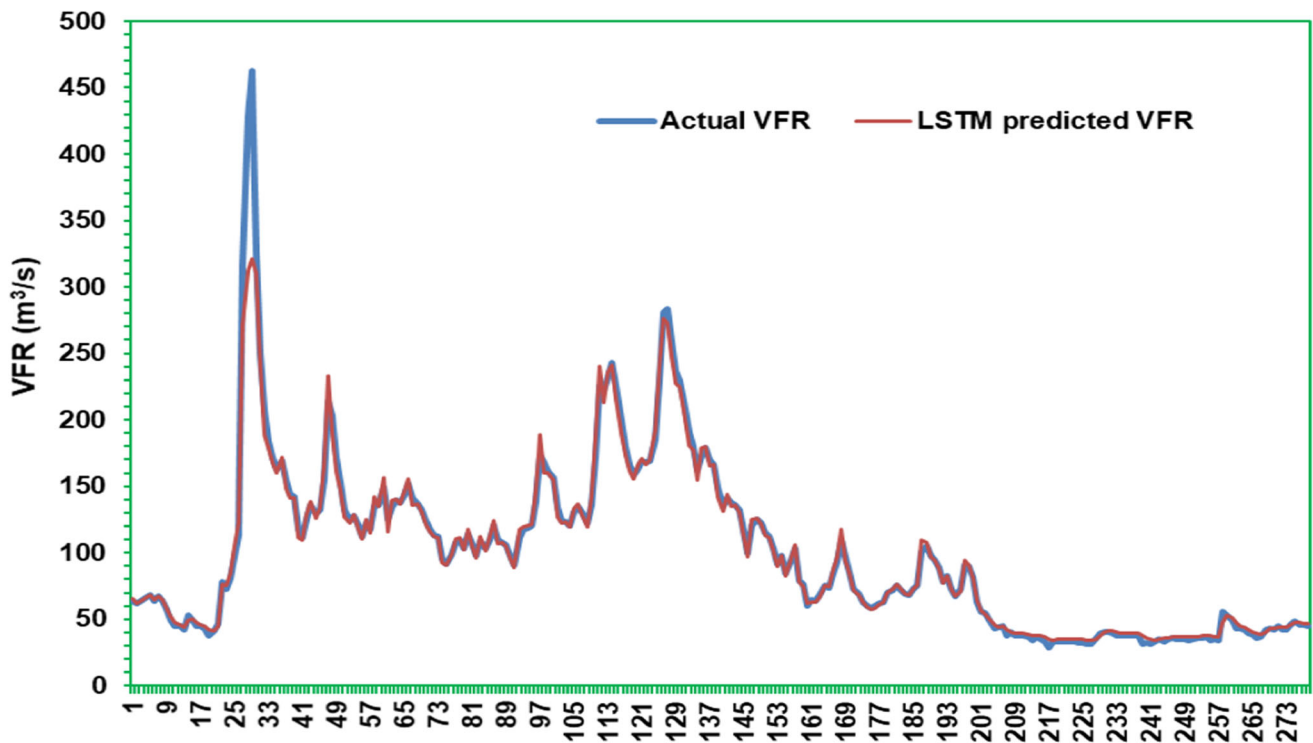


Fig. 17 The actual daily average time series data of the VFR and corresponding LSTM predicted data

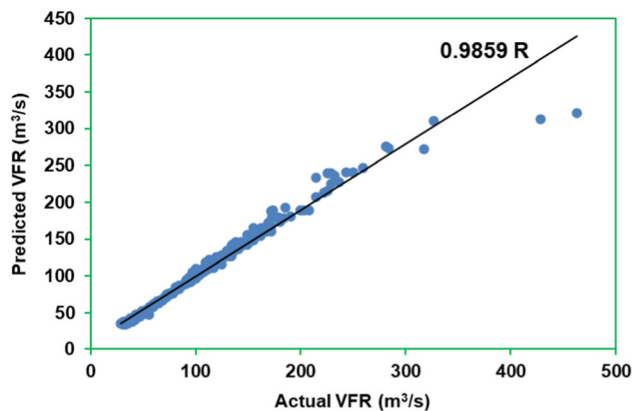


Fig. 18 Statistical regression plot for the actual daily average VFR data and corresponding LSTM predictions

3.3.5 Comparison of the results

Table 6 gives the statistical accuracy outcomes of the VFR forecasting of Harmanli MS, considering the best results obtained from ANFIS-FCM, ANFIS-SC, ANFIS-GP, and LSTM. The statistical accuracy outcomes of the whole data set have included MAE, RMSE, and R results of the comparisons of the predictions according to the real measured data. On the other hand, the table indicates that the best forecasting has been captured with the FCM algorithm of the ANFIS. This best result is indicated in this table by bold. Namely, this method of the ANFIS generated an

Table 6 Comparisons of the statistical accuracy measures of VFR predictions

Station	Method	MAE (m ³ /s)	RMSE (m ³ /s)	R
Harmanli	ANFIS-FCM	2.54	4.35	0.9981
	ANFIS-SC	2.81	5.10	0.9978
	ANFIS-GP	6.11	27.30	0.9487
	LSTM	4.02	12.14	0.9859

average statistical error output value of 2.54 m³/s MAE, 4.35 m³/s RMSE, and 0.9981 R . When the whole of the data cloud which has a VFR value of 495.50 m³/s as the maximum value is considered, these MAE and RMSE values can be regarded to be significantly small. Besides, the correlation coefficient (R) value corresponding to 0.9981 indicates that the regressions of the ANFIS-FCM consisting of the predictions and the real measured data are compatible with each other. Finally, it has been concluded that ANFIS-SC tool can be a beneficial tool in terms of forecasting the VFR data showing random distributions.

In ANFIS-FCM and ANFIS-SC algorithms, the machine learning is performed considering each point individually in the data cloud. Namely, while in ANFIS-FCM, each data point is allowed to have multiple clusters, whereas in ANFIS-SC, each data point is allowed to be the cluster center. In these two methods, estimates are obtained with a

high precision, since each point is studied separately by machine learning. On the other hand, in ANFIS-GP, rectangular partitioning of the inputs is done, which may result the user to do incorrect or insensitive partitioning, and finally high errors may be obtained in the predictions. In LSTM machine learning, fixed errors will be added to the cell states or memory cells, and continuous reproduction of the errors will occur. Eventually, these errors accumulate and increase throughout the machine learning. For these reasons, ANFIS-FCM and ANFIS-SC machine learnings are successful compared to the ANFIS-GP and LSTM.

4 Conclusions

In this study, water VFR forecasting of the real measured data was performed, using the data of the Harmanli MS located on the Maritsa River in Bulgaria. In the conducted river VFR predictions, LSTM as well as the FCM, SC, and GP algorithms of the ANFIS were used. Considering these four algorithms, a total of 102 models were designed and constructed. A cumulative of 1,391 instantaneous daily average VFR data was utilized in the VFR anticipations. Among those cumulative data, 80% and 20% of the total data were used, respectively, for training the algorithm and testing of the algorithm. The outcomes of the predictions at the testing stages of the models were compared, using 3 statistical parameters including average values of the MAE (m^3/s), RMSE (m^3/s), and R , of the whole data set. In this study, it has been reported that VFR predictions performed by four tools have generally generated high accuracies. On the other hand, it is concluded that the best river VFR predictions were obtained during the utilization of the ANFIS-FCM. The statistical error analyses of the VFR forecasting have indicated that this model out came with the best results in terms of these three statistical error criteria. Namely, in the FCM algorithm of the ANFIS, at the selected number of historical data used for training corresponding to 3, at the number of MFs of 2, and at the max epoch number of 100, the best prediction result having MAE, RMSE, and R corresponding to $2.54 \text{ m}^3/\text{s}$, $4.35 \text{ m}^3/\text{s}$, and 0.9981 was obtained, respectively.

On the other hand, when whole 40 models tested under ANFIS-FCM and all 48 models tested under ANFIS-SC are evaluated, as well as when the averages of their three statistical error parameters are taken into account, it was concluded and shown that ANFIS-SC was to be slightly better than ANFIS-FCM in all results. To sum up, it has been concluded that the FCM and the SC algorithms of the ANFIS can be two beneficial tools in terms of forecasting of river VFR data showing random distributions. Besides, it has been reported that these two algorithms can be easily

and successfully applied to obtain predictions to any type of such data having random distributions.

It was concluded that ANFIS and LSTM models performed better predictions in randomly distributed data which has bumpy characteristics. Namely, these models were applied safely to the data having a high amplitude and high fluctuation. It was also concluded that they performed a high success in volumetric flow rate predictions. However, it was also shown that the algorithms would be also proper in predictions of other types of data that demonstrate a feature in this way.

Funding No funding was received for conducting this study.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

- IRENA (2020) Renewable energy statistics 2020. International Renewable Energy Agency. <https://www.irena.org/>. Accessed 28 Nov 2021
- IRENA (2021) Renewable energy statistics 2021. International Renewable Energy Agency. <https://www.irena.org/>. Accessed 28 Nov 2021
- N-Sci Technologies (2021) The importance of hydropower. <https://nsci.ca/2019/08/14/the-importance-of-hydro-power/>. Accessed 20 Nov 2021
- REN21 (2021) Global status report. Renewables 2021. <https://www.ren21.net/>. Accessed 15 Nov 2021
- Zhang J, Yan J, Infield D, Liu Y, Lien F (2019) Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and Gaussian mixture model. *Appl Energy* 241:229–244. <https://doi.org/10.1016/j.apenergy.2019.03.044>
- Jung J, Broadwater RP (2014) Current status and future advances for wind speed and power forecasting. *Renew Sustain Energy Rev* 31:762–777. <https://doi.org/10.1016/j.rser.2013.12.054>
- Tascikaraoglu A, Uzunoglu M (2014) A review of combined approaches for prediction of short-term wind speed and power. *Renew Sustain Energy Rev* 34:243–254. <https://doi.org/10.1016/j.rser.2014.03.033>
- Liu H, Tian H, Li Y (2015) An EMD-recursive ARIMA method to predict wind speed for railway strong wind warning system. *J Wind Eng Ind Aerodyn* 141:27–38. <https://doi.org/10.1016/j.jweia.2015.02.004>
- Kavasseri RG, Seetharaman K (2009) Day-ahead wind speed forecasting using f-ARIMA models. *Renew Energy* 34(5):1388–1393. <https://doi.org/10.1016/j.renene.2008.09.006>
- Khosravi A, Koury RNN, Machado L, Pabon JIG (2018) Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. *J Clean Prod* 176:63–75. <https://doi.org/10.1016/j.jclepro.2017.12.065>
- Shi X, Lei X, Huang Q, Huang S, Ren K, Hu Y (2018) Hourly day-ahead wind power prediction using the hybrid model of variational model decomposition and long short-term memory. *Energies* 11(11):1–20. <https://doi.org/10.3390/en11113227>

12. Yu C, Li Y, Bao Y, Tang H, Zhai G (2018) A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. *Energy Convers Manag* 178:137–145. <https://doi.org/10.1016/j.enconman.2018.10.008>
13. Han S, Qiao YH, Yan J, Liu YQ, Li L, Wang Z (2019) Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network. *Appl Energy* 239:181–191. <https://doi.org/10.1016/j.apenergy.2019.01.193>
14. Zhang Z, Qin H, Liu Y, Wang Y, Yao L, Li Q, Li J, Pei S (2019) Long short-term memory network based on neighborhood gates for processing complex causality in wind speed prediction. *Energy Convers Manag* 192:37–51. <https://doi.org/10.1016/j.enconman.2019.04.006>
15. Zhang Z, Ye L, Qin H, Liu Y, Wang C, Yu X, Yin X, Li J (2019) Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression. *Appl Energy* 247:270–284. <https://doi.org/10.1016/j.apenergy.2019.04.047>
16. Liang S, Nguyen L, Jin F (2018) A multi-variable stacked long-short term memory network for wind speed forecasting. In: 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp 4561–4564
17. Zaytar MA, Amrani CE (2016) Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. *Int J Comput Appl* 143(11):7–11. <https://doi.org/10.5120/ijca2016910497>
18. Özen C, Kaplan O, Özcan C, Dinç U (2019) Short term wind speed forecast by using long short term memory. In: 9th International Symposium on Atmospheric Sciences (ATMOS 2019), İstanbul, Turkey
19. Prabha PP, Vanitha V, Resmi R (2019) Wind speed forecasting using long short term memory networks. In: 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT), Kannur, Kerala, India, pp 1310–1314
20. 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>
21. Bilgili M, Sahin B (2010) Comparative analysis of regression and artificial neural network models for wind speed prediction. *Meteorol Atmos Phys* 109:61–72. <https://doi.org/10.1007/s00703-010-0093-9>
22. Hewett R, Leuchner J, Carvalho M (2001) From climate history to prediction of regional water flows with machine learning. Institute of Electrical and Electronics Engineers (IEEE), pp 292–297
23. Tayfur G, Singh VP, Moramarco T, Barbetta S (2018) Flood hydrograph prediction using machine learning methods. *Water (Multidisciplinary Digital Publishing Institute)* 986:1–13. <https://doi.org/10.3390/w10080968>
24. Flake JT (2007) Application of the relevance vector machine to canal flow prediction in the Sevier River basin. Dissertation, Utah State University
25. Farhadi H, Zahiri A, Hashemi MR, Esmaili K (2019) Incorporating a machine learning technique to improve open-channel flow computations. *Neural Comput Appl* 31:909–921. <https://doi.org/10.1007/s00521-017-3120-7>
26. Sahraei A, Chamorro A, Kraft P, Breuer L (2021) Application of machine learning models to predict maximum event water fractions in streamflow. *Front Water* 3:1–21. <https://doi.org/10.3389/frwa.2021.652100>
27. Il Kim H, Kim BH (2020) Flood hazard rating prediction for urban areas using random forest and LSTM. *KSCE J Civ Eng* 24(12):3884–3896. <https://doi.org/10.1007/s12205-020-0951-z>
28. Yaseen ZM, Jaafar O, Deo RC, Kisi O, Adamowski J, Quilty J, El-Shafie A (2016) Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. *J Hydrol* 542:603–614. <https://doi.org/10.1016/j.jhydrol.2016.09.035>
29. Xiao L, Zhong M, Zha D (2022) Runoff forecasting using machine-learning methods: case study in the middle reaches of Xijiang River. *Front Big Data* 4:1–11. <https://doi.org/10.3389/fdata.2021.752406>
30. Abyaneh HZ, Nia AM, Varkeshi MB, Marofi S, Kisi O (2011) Performance evaluation of ANN and ANFIS models for estimating garlic crop evapotranspiration. *J Irrig Drain Eng* 137:280–286. [https://doi.org/10.1061/\(ASCE\)IR.1943-4774.0000298](https://doi.org/10.1061/(ASCE)IR.1943-4774.0000298)
31. Jang JR (1993) ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans Syst Man Cybern* 23(3):665–685. <https://doi.org/10.1109/21.256541>
32. Karakuş O, Kuruoğlu EE, Altinkaya MA (2017) One-day ahead wind speed/power prediction based on polynomial autoregressive model. *IET Renew Power Gener* 11(11):1430–1439. <https://doi.org/10.1049/iet-rpg.2016.0972>
33. Tabari H, Kisi O, Ezani A, Talaei PH (2012) SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. *J Hydrol* 444–445:78–89. <https://doi.org/10.1016/j.jhydrol.2012.04.007>
34. Mathworks (2020) Long short-term memory networks. <https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html>. Accessed 1 May 2020
35. Chandy KM, Taylor S (1992) An introduction to parallel programming. Jones and Bartlett, Boston
36. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
37. Zahroh S, Hidayat Y, Pontoh RS, Santoso A, Sukono Bon AT (2019) Modeling and forecasting daily temperature in Bandung. In: Proceedings of the International Conference on Industrial Engineering and Operations Management, Riyadh, Saudi Arabia, pp 406–412
38. Salman AG, Heryadi Y, Abdurahman E, Suparta W (2018) Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. *Procedia Comput Sci* 135:89–98. <https://doi.org/10.1016/j.procs.2018.08.153>
39. Liu R, Liu L (2019) Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm. *Soft Comput* 23:11829–11838. <https://doi.org/10.1007/s00500-018-03739-w>
40. Dsi (2021) River volumetric flow rates. In: 11th Regional Directorate of DSI of Turkey. <http://edirnenehir.dsi.gov.tr/>. Accessed 30 Nov 2021
41. Liu X, Xiong J, Vasilakos AV (2021) Privacy and security issues in deep learning: a survey. *IEEE Access* 9:4566–4593. <https://doi.org/10.1109/ACCESS.2020.3045078>
42. Kuntla GS, Tian X, Li Z (2021) Security and privacy in machine learning: a survey. *Issues Inf Syst* 22(3):224–240. https://doi.org/10.48009/3_iis_2021_242-258

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