University of Birjand

Water Harvesting Research Vol. 6, No. 2, Autumn & Winter 2023, Case Study p. 213-225

10.22077/JWHR.2024.6969.1117

Implementation of a Machine-Learning-Based Approach for Forecasting Watershed Stream Flow (Case Study: Chehel Chai Watershed, Iran)

Hamed Sahranavard[a], Mahdi Naseri[b*], Abolfazl Akbarpour[c], Farshad Ahmadi[d]

[a]MSc Student of Water Resources Management, Department of Civil Engineering, University of Birjand, Birjand, Iran.
[b]Assistant Professor, Department of Civil Engineering, University of Birjand, Birjand, Iran.
[c]Professor, Department of Civil Engineering, University of Birjand, Birjand, Iran.
[d]Assistant Professor, Department of Hydrology & Water Resources Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

[*]Corresponding Author, E-mail address: mnaseri@birjand.ac.ir

Received: 19 November 2023/ **Revised**: 25 February 2024/ **Accepted**: 03 March 2024

Abstract

The importance of the optimal and efficient use of all available water resources becomes noticeable when today due to successive droughts and a decrease in rainfall, the surface water resources are running out. Runoff and surface water resources are some of the primary and vital available water resources, and hence, modeling and predicting their behavior are especially critical. In the current research, the aim was to model the stream flow of the Chehel Chai watershed in Golestan province, Iran, using the data of the stream flow and precipitation for a period of 45 years. For this reason, 4 machine learning algorithms namely, Extreme Learning Machine (ELM), Random Forest (RF), Gaussian Process Regression (GPR), and Gene Expression Programming (GEP) were used. The data were entered into the modeling in the form of different scenarios consisting of the stream flow and precipitation with varying lags of time. The results showed that scenario M2 (using only stream flow data with two time lags) in the ELM (extreme learning machine) model with the values of RMSE (root mean square error) =0.984 (m^3/s) and R^2=0.613 had the most accurate performance and predictions among all the models and scenarios.

Keywords: Forecasting, Machine learning, Modeling, Stream flow.

1. Introduction

Global warming, climate changes, consecutive droughts, and inefficient management are among the factors that jeopardized the survival of water resources (Sahranavard and Naseri, 2022). Surface water resources are among the most important water resources. So, the modeling and forecasting of these resources can help in efficient exploitation, reducing losses and wastage of water, and preventing floods. Therefore, today modeling and simulation of surface water resources have attracted the attention of experts and researchers all over the world (Sahranavard and Naseri, 2022).

She and Basketfield (2005) compared support vector machine (SVM), linear discriminant analysis, and multinomial logistic regression for forecasting the stream flow of Cascade mountain range, United States and the projections showed the best performance was belonged to the SVM method. A FIR neural network and a fuzzy clustering-based were implemented by Luna et al. (2005) to forecast a case study stream flow and they showed that the FIR neural network had the most accurate predictions. Asefa et al. (2006) applied Support Vector Machines for modeling and forecasting the stream flow and compared the results with physical-based models. Various artificial neural networks were used by Kisi (2007) for forecasting the short-term daily stream flow. Hong (2008) combined three algorithms, namely recurrent artificial neural network, support vector machine, and chaotic particle swarm optimization algorithm for forecasting the stream flow. The results showed that the hybrid model performed satisfactorily than the other algorithms. Kalra and Ahmadi (2009) applied the SVM model for modeling and forecasting the stream flow of the Colorado River catchment, United States. The predictions illustrated that the SVM algorithm overcame the linear regression and feedforward back propagation artificial neural network. The performances of statistical methods, namely autoregressive integrated moving average and seasonal ARIMA compared to the artificial intelligence approach, and Jordan-Elman artificial neural networks, for forecasting the flow rate of Kizil River were investigated by Abudu et al. (2010). They concluded that the statistical methods presented a more justifiable performance. Li et al. (2010) implemented a modified version of the SVM algorithm to simulate and forecast the input flow of the Shihmen reservoir, Taiwan. They compared the projections with the results of two versions of multiple linear regression and deduced that the modified SVM approach presented the most accurate results. Kernel SVM approach was used for modeling and forecasting the stream flow of the Mahanadi River, India. The projections compared with the results of the Box-Jenkins method illustrated that the SVR outperformed the ARIMA model. Rasouli et al. (2010) used climatic data for forecasting the daily stream flow of a coastal watershed in Canada. The methods they applied were support vector regression, Bayesian neural network, and Gaussian process and they compared the predictions with the MLR (multiple linear regression) method. According to their results, the three machine learning algorithms worked more satisfactorily than the MLR model.

A hybrid model consisting of a wavelet function, neuro, and fuzzy algorithms was used by Shiri and Kisi (2010) to model and forecast the Stream flow of the Filyos River, Turkey. They concluded that the hybrid approach increased the accuracy of single methods. Three approaches, including modified SVM, basic SVM, and ANN (artificial neural network), were applied by Guo et al. (2011) for modeling and forecasting the stream flow. Based on the results, the modified SVM had the most precise performance. Rasouli et al. (2012) used climatic parameters as input data for modeling and forecasting the stream flow of a small watershed in Canada by four various algorithms, namely Bayesian neural network (BNN), support vector regression (SVR), Gaussian process (GP), and multiple linear regression. They showed that the machine learning approaches were more reliable. The accuracy and efficiency of SVM and MLR algorithms were compared for forecasting the annual maximum stream flow in Malaysia by Zakaria and Shabri (2012). They deduced that the SVM model performed more precisely than the MLR method.

Sun et al. (2014) applied the GPR algorithm for simulating and forecasting the watershed stream flow and compared the projections with the linear regression and ANN methods. According to their achievements, the GPR had the best performance. Garsole and Rajurkar (2015) used the SVR algorithm to forecast the stream flow of the catchment area upstream of the Jayakwadi dam in India and claimed that the SVR method is a trustable model for reliability and prediction. Tongal and Booij (2018) implemented SVR, ANN, and random forest (RF) algorithms to evaluate the performance of a simulation framework and concluded that the use of this simulation framework has increased the prediction accuracy of the models. Tyralis et al. (2021) applied 10 machine learning methods to model the watershed stream flow and compared the projections with the linear regression model and concluded that all 10 machine learning algorithms improved the results compared to the linear regression.

The algorithms, namely RF, KNN, AdaBoost, and SVM were applied by Tosunoğlu et al. (2020) to model and forecast the stream flow of the Coruh river catchment, Turkey. According to their evaluation indexes, the RF approach produced the most accurate projections. Adnan et al. (2020) investigated the precision of three algorithms, including the artificial neural network, genetic algorithm, and the adaptive neuro-fuzzy inference system in modeling and forecasting the stream flow of Neelum and Kunhar Rivers, Pakistan, and compared the projections with the M5 regression tree algorithm. Based on the results, the performances of the first two models were more satisfying than the M5RT technique.

Cheng et al. (2020) applied artificial neural network (ANN) and long short-term memory (LSTM) for modeling and forecasting stream flow of the Nan and Ping Rivers catchments, Thailand. They concluded that the LSTM model produced more accurate projections.

Ikram et al. (2022) applied various machine-learning approaches such as extreme learning machine (ELM), Gaussian processes regression (GPR), support vector regression (SVR), least square SVR (LSSVR), radial basis function neural network (RBFNN) in predicting stream flow and based on six evaluation criteria, among all methods, the SVR model presented the most precise results. Singh et al. (2022) used four models, namely multiple linear regression (MLR), multiple adaptive regression splines (MARS), support vector machine (SVM), and random forest (RF) to model the stream flow of the Gola watershed, Uttarakhand based on rainfall-runoff models. According to the results of modeling and calculated evaluation criteria, the random forest algorithm has the most accurate performance.

Support vector machine (SVM), artificial neural network (ANN), and long short-term memory (LSTM) were used in the form of 3 scenarios to model the flow of 11 rivers across Malaysia by Essam et al. (2022). The findings indicated that the performance of model AAN was more acceptable than other algorithms in all 11 rivers and 3 scenarios. Adnan et al. (2023) combined the ELM method with a variety of optimization algorithms and showed that the combined models showed more justifiable projections and performance than the independent model in predicting stream flow. Kumar et al. (2023) compared a large number of machine learning algorithms such as CatBoost, ElasticNet, k-Nearest Neighbors (KNN), etc, to check the accuracy of stream flow prediction and claimed that the CatBoost algorithm performed best. The performance of Random Forest Regression (RFR) and the soil and water assessment tool (SWAT) models for predicting the stream flow of the Rio Grande Headwaters near Del Norte was evaluated and analyzed by Islam et al. (2023) and based on the results of the evaluation criteria, the RFR model Introduced as the top model.

In this research, the modeling and forecasting of the stream flow of the Chehel Chai watershed located in the Golestan province, Iran were done. For this purpose, four machine learning approaches, including the extreme learning machine (ELM), random forest (RF), gene expression programming (GEP), and Gaussian process regression (GPR) were used. The purpose of this research was to compare and evaluate the accuracy and efficiency of these algorithms in forecasting the stream flow in the climatic conditions of the studied watershed. The innovation and originality of this research are that it uses four well-known machine learning models on a case study with unique physical and geographical characteristics in six different scenarios as independent models and hybrid rainfall-runoff models and the superior and more efficient model has been determined.

2. Materials and Methods

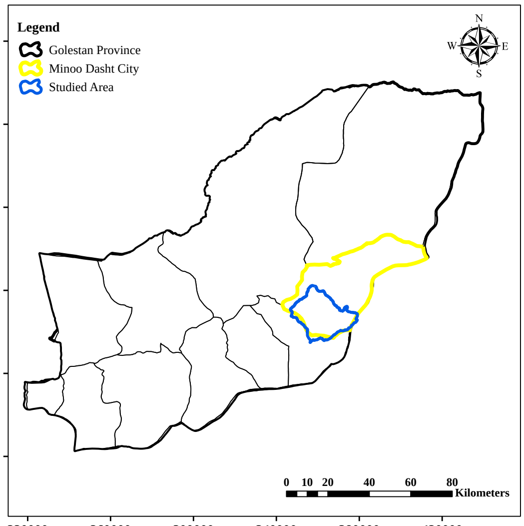
2.1. Study area

In this research, the stream flow data of the Chehel Chai watershed, which were recorded monthly from 1971 to 2017 (a 46-year period) were used. Moreover, the precipitation data of the Chehel Chai watershed were collected at the same period, which was recorded monthly and used as input data in this research. The Chehel Chai watershed is located in the Golestan province, situated in the north of Iran. The area of this watershed is 470 square kilometers. According to the received statistics, the long-term average rainfall of this catchment area is 460 mm, and in terms of temperature, the average temperature of this catchment area is reported to be 18.2 degrees Celsius. Fig. 1 presents the location of the Chehel Chai watershed. Due to the proximity of the Chehel Chai watershed to the sea level, this altitude watershed experiences a severe height difference, so its highest peak has 2,900 meters above the sea level and its lowest area reaches 22 meters. Fig. 1 shows the elevation of the studied watershed. Figures 2 and 3, present the time series plots of the stream flow and precipitation for the Chehel Chai watershed.

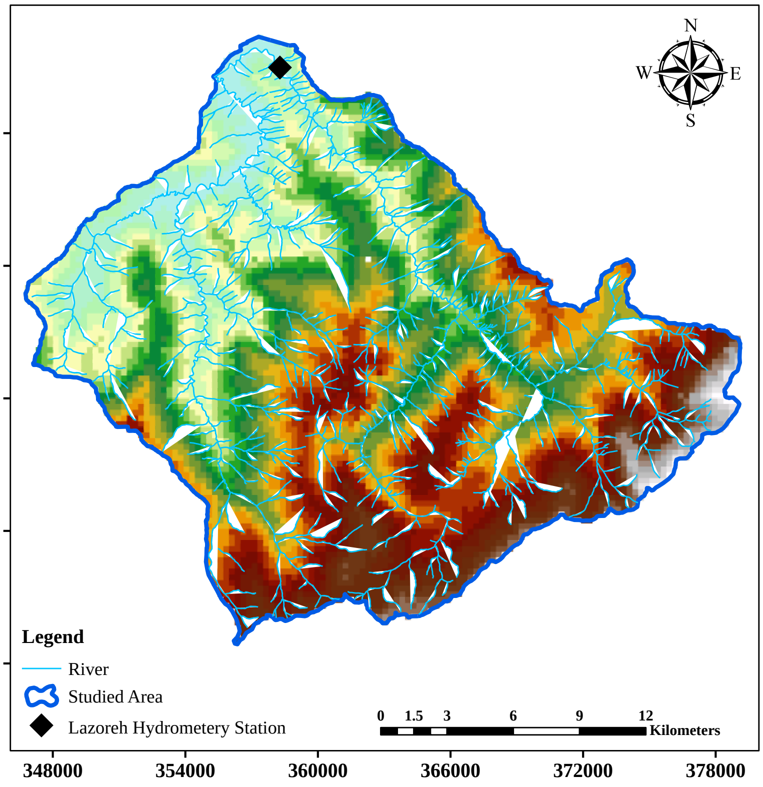
2.2. Extreme Learning Machine (ELM)

A new algorithm that has been proposed in recent years and used and applied in many researches is called the ELM, which was first introduced by Huang et al. (2006). This algorithm is very similar to feed-forward neural networks in the implementation process and the only difference is that the ELM determines the weight of hidden neurons and as a result, it can be said to be among the simplest artificial neural networks. One of the most important advantages of this simple method, which makes the implementation easier for processing, is to reduce the computational time required in calculations and reduce the number of structures (Miche et al. 2008).

a) Location of Golestan Province



b) Location of the studied area



c) Situation of Elevation in the studies watershed

Fig. 1. Location of the Chehel Chai Watershed

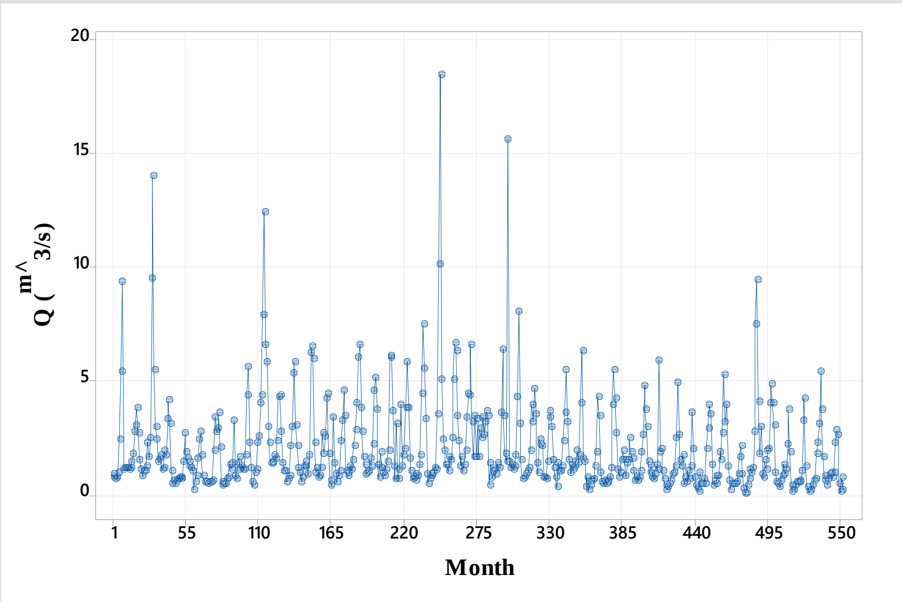


Fig. 2. Time series plot of the stream flow

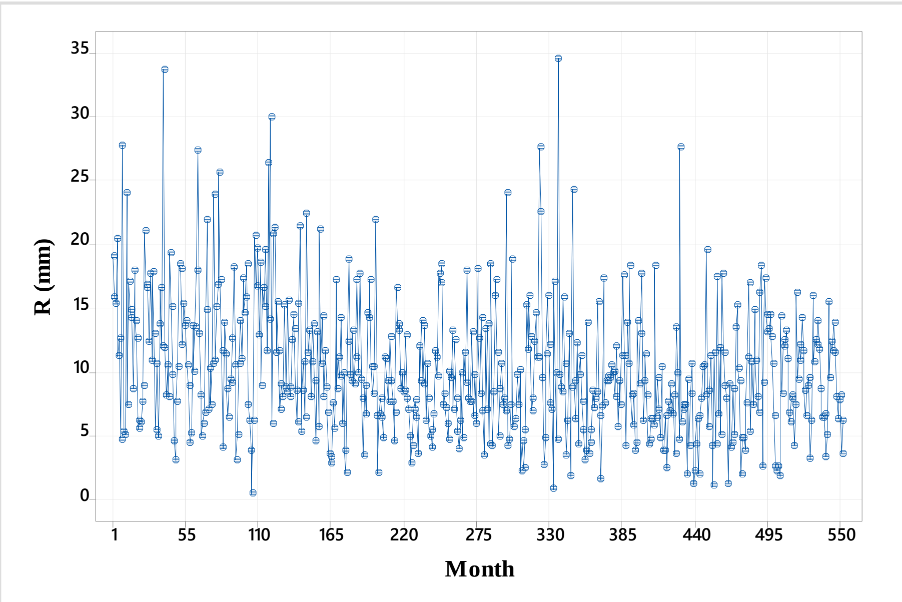


Fig. 3. Time series plot of the precipitation

In general, this algorithm has two main equations:

$$XW=H \quad (1)$$

$$H\beta = Y \quad (2)$$

In these equations, X is a complete set of input data, Y is the target vector, β represents the vector of output weights between the hidden layer and the output node, H is the hidden layer, and finally, W is the random weight matrix and play the role of connecting the output layer to the hidden layer (Huang et al. 2006).

2.3. Random Forest (RF)

Random Forest, a method that was first proposed by Breiman (2001) and later was implemented in many researches in different fields due to its good advantages such as high accuracy, efficiency, and convenience (Were et al. 2015). In general, there are many trees in this method, which each of them has a random vector that has been sampled separately and the values of each tree depend on this vector as well. Finally, the random forest is the result of the collection of all these tree predictors. The main task of the random vectors that each tree has separately is to monitor the growth of each tree that is produced in this collection (Ahmadi et al. 2022). In this method, the random vector X_i for the nth tree is generated separately and is not dependent on the vectors generated for other trees. The construction of the random vector for each tree is as follows:

$$x_n = \{h_1(x), h_2(x), \ldots, h_n(x)\} \quad (3)$$

$$h_n = h(x, X_n), x = \{x_1, x_2, \ldots, x_p\} \quad (4)$$

where x and h are the feature and the threshold of each tree, respectively. In equation 4, the next forest is formed by the next p vector, and finally, the results of each tree are stated below:

$$y_1 = h_1(x), y_2 = h_2(x), \ldots, y_n = h_n(x) \quad (5)$$

In equation 5, the obtained output corresponds to tree i, in which y_i represents this. Ultimately, all the obtained results are compared and the best result is selected (Orellana-Alvear et al. 2020).

2.4. Gaussian Process Regression (GRP)

Through the combination of Bayesian and statistical theory and developing them, the GPR method was created, which is very well known and popular until now and has been used in extensive researches (Schulz et al. 2018). One of the important applications of this method is to solve complex non-linear problems where the samples have large dimensions and small numbers. Another advantage of this algorithm is the ability of flexible deductive reasoning, which is also present in the Bayesian method. Self-organization, adaptation, parallel processing, and self-learning are among the other privileges of this approach (Pustokhina et al. 2021). In general, the GPR algorithm consists of two parts: regression and Gaussian process. The regression section is in charge of forecasting. Gaussian process is a Gaussian distribution between functions, which is defined as follows:

$$g(X) \sim GP(E(X), K(X, X)) \quad (6)$$

where $K(X,X)$ is the covariance function matrix and $E(X)$ indicates the mean function. If we consider the set $S_n^{test} = \{(x_{test}, y_{test}) | x_{test} \in R^m, y_{test} \in R\}$ as the input test data set to the model, the Gaussian process distribution is defined as follows:

$$\begin{bmatrix} y \\ y_{test} \end{bmatrix} \sim N\left(0, \begin{pmatrix} K(X,X) + \partial_n^2 I_n & K(X, x_{test}) \\ K(x_{test}, X) & K(x_{test}, x_{test}) \end{pmatrix}\right) \quad (7)$$

where in the above equation, $K(X,x_{test})$ is the covariance of the test data set x_{test} and the variables of the learning input data X.

2.5. Gene Expression Programming (GEP)

Gene Expression Programming (GEP) is a machine learning algorithm that was first introduced by Ferreira (2001), which was developed from the development of Genetic Algorithm (GA) and Genetic Programming (GP) methods (Ferreira 2001). In general, this method works on the chromosomes of the population and evaluates them according to the criteria of fitness conditions, and performs genetic changes using one or more genetic operators (Khan et al. 2021). One of the important differences of this method compared to classical regression is that some pre-defined functions are used in classical regression, but in this method, without using these functions, it reflects the primary non-linear equations (Khan et al. 2021). ET (Expression Tree) in the GEP algorithm represents various complexities, including constants, functions, operators, and variables. A single ET consists of a root node, a functional node, and an end node. The mathematical expression of ET has two genes with multiplication as the link function and can be written as the following equations:

$$\text{Gene } 1 = \sqrt[3]{\frac{a}{b} + a} \quad (8)$$

$$\text{Gene } 2 = \log(a \times b) \quad (9)$$

$$\text{Prediction} = \text{Gene } 1 \times \text{Gene } 2 \quad (10)$$

In general, this algorithm has five main steps as follows:

The first step is to choose a fitness function, which can be said to be one of the features of this function that gives this algorithm the ability to reach an optimal solution. In the second step, to forecast the target, a set of terminals consisting of explanatory variables should be selected. In the third step, some functions should be selected in such a way that they have the ability to solve the simple equation of this algorithm. In the fourth step, the size of the vertex, the function and the gene, that generally called the chromosome architecture, have to be chosen. In the last step, the genetic operator must be determined, which is done by choosing a set of genetic operators that consists of transposition, crossover, and mutation (Khan et al. 2021).

3. Results and Discussion

As mentioned above, the stream flow data of the Chehel Chai watershed, which was recorded monthly for a period of 46 years, were used in this research. The correct selection of influential hydrological parameters is one of the most important actions in modeling hydrological phenomena. Since in the current research, the modeled parameter was stream flow, the most important hydrological parameters are precipitation, evaporation, transpiration, and temperature.

Ahmadi et al. (2022) proved that the most influential hydrological parameter in stream flow modeling is the precipitation statistics of the studied area. Therefore, in this research, only the precipitation parameter was selected as the influential hydrological parameter. Based on autocorrelation and partial autocorrelation plots, which are presented in Fig.4, six scenarios were determined for entering data into the selected algorithms so that in the first three scenarios, the input data entered into the models were only the stream flow data, and in the next three scenarios, a combination of the recorded stream flow and precipitation data were entered into the algorithms.

In scenario M1, the data were entered into the modeling process with a time correlation lag. The M2 scenario was defined in such a way that the stream flow data entered into the models should be accompanied by two time correlation lags and this time correlation lag was considered equal to 3 in the M3 scenario. Moreover, MM1 to MM3 scenarios are the combination of the stream flow and precipitation data for inputting into the models. For each of the scenarios, different patterns of data combination were considered, which are shown in Table 1.

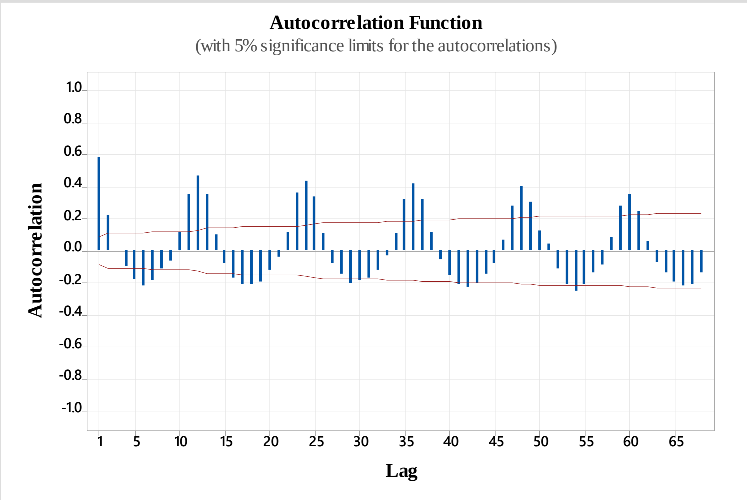
In this research, the collected stream flow and precipitation data were divided into two parts: the training data and the test data. 80% of the data were included in the training data section and 20% of the data were selected as the test data. Fig.5 shows this division in the stream flow data completely.

After entering the data into the models and based on all the planned scenarios, as well as the considered error and performance criteria, the best model was selected. Table 2 shows the values of the calculated evaluation criteria.

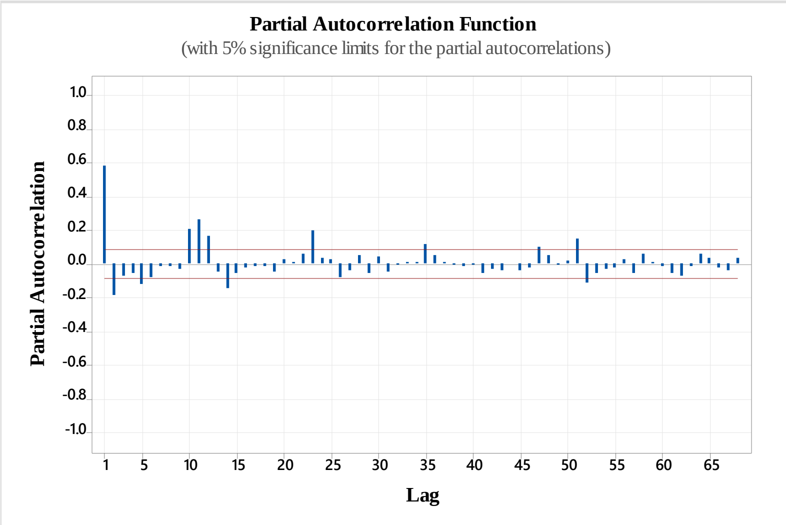
If we pay attention to the values of the evaluation criteria calculated for the Chehel Chai catchment area, it can be seen that in the ELM model, the M3 scenario with RMSE = 0.984 and MAE = 0.668 has the best results compared to other models. In relation to the GEP model, the MM2 scenario has the best predictions with RMSE = 1.017 and MAE = 0.685. Regarding to the GPR model, the MM3 scenario has more satisfactory results than the other scenarios with values of RMSE = 1.336 and MAE = 1.072.

In the RF algorithm, the MM3 scenario has a more acceptable performance with values of RMSE = 1.139 and MAE = 0.813. In general, since the M3 scenario is a model without the involvement of the precipitation parameter, it can be concluded that for modeling and forecasting the flow rate in the watershed with similar climatic conditions to the Chehel Chai watershed using the ELM algorithm, the flow parameter should be used as the only input parameter of the model and it will provide the most accurate results. But if the goal is to model with GPR, GEP and RF algorithms, according to the structure of MM2 and MM3 models, the use of flow rate and precipitation parameters with longer time delays provides more accurate and reliable results.

Figures 6 to 9 show the time series related to the best scenario in each model. Generally, as shown in Fig.10, the ELM algorithm has the highest value of R^2. Therefore, in the stream flow modeling of the Chehel Chai watershed using the stream flow and precipitation data, the ELM algorithm has shown the best performance among other methods.



a) ACF plot



b) PACF plot

Fig. 4. The autocorrelation function and partial autocorrelation function plots

Table 1. The chosen patterns for entering data to the algorithms

Pattern	Scenario	Row	Watershed
$Q_t = f(Q_{t-1})$	M1	1	Chehel Chai
$Q_t = f(Q_{t-1}, Q_{t-2})$	M2	2	
$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3})$	M3	3	
$Q_t = f(Q_{t-1}, R_{t-1})$	MM1	4	
$Q_t = f(Q_{t-1}, Q_{t-2}, R_{t-1}, R_{t-2})$	MM2	5	
$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, R_{t-1}, R_{t-2}, R_{t-3})$	MM3	6	

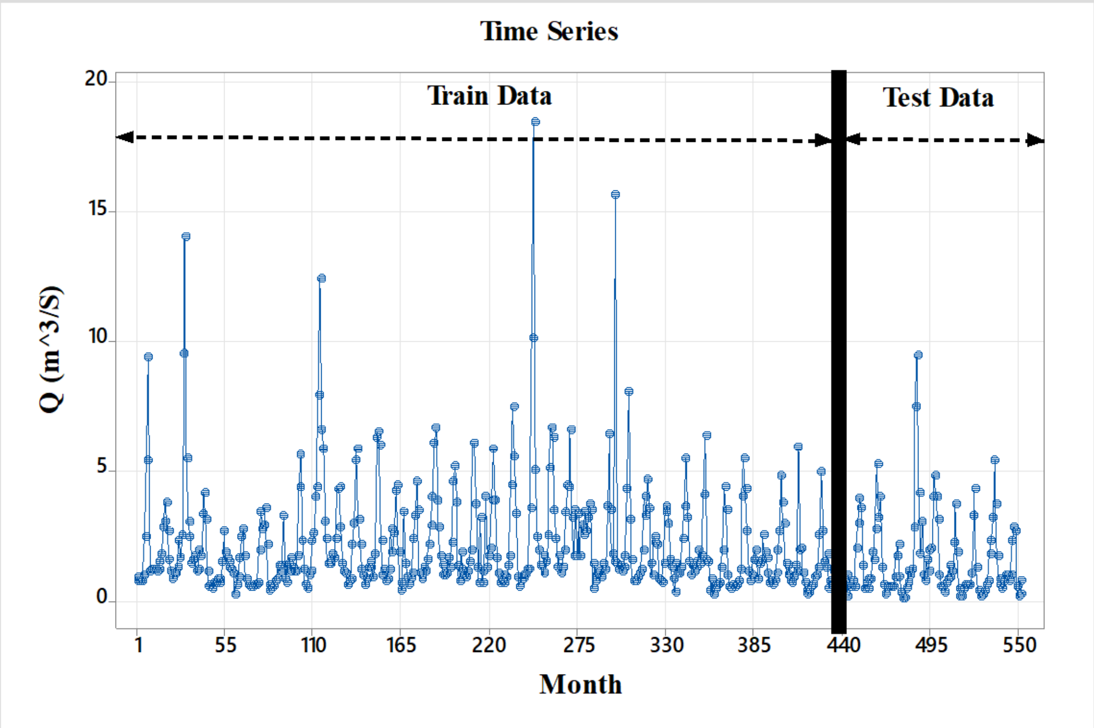


Fig. 5. The situation of the train and test data of the stream flow

Table 2. The calculated evaluation criteria for each algorithm

Model	Scenario	RMSE	MAE
ELM	M1	1.132	0.771
	M2	1.002	0.675
	M3	0.984	0.668
	MM1	1.061	0.754
	MM2	1.016	0.699
	MM3	1.005	0.689
GEP	M1	1.138	0.811
	M2	1.191	0.883
	M3	1.067	0.737
	MM1	1.146	0.857
	MM2	1.017	0.685
	MM3	1.315	0.942
GPR	M1	1.482	1.254
	M2	1.474	1.241
	M3	1.467	1.247
	MM1	1.396	1.159
	MM2	1.362	1.111
	MM3	1.336	1.072
RF	M1	1.462	0.910
	M2	1.194	0.806
	M3	1.170	0.790
	MM1	1.264	0.851
	MM2	1.155	0.818
	MM3	1.139	0.813

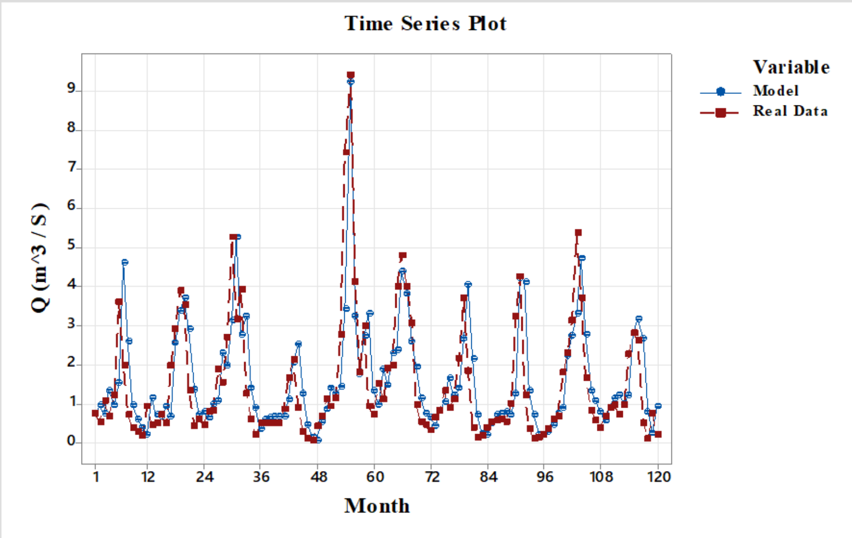


Fig. 6. Time series plot of forecasted and real data for the ELM model

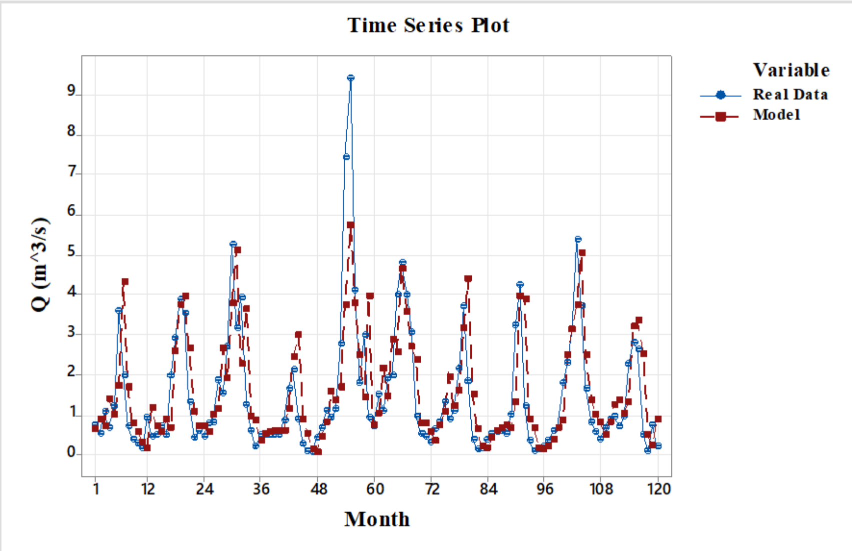


Fig. 7. Time series plot of forecasted and real data for the GEP model

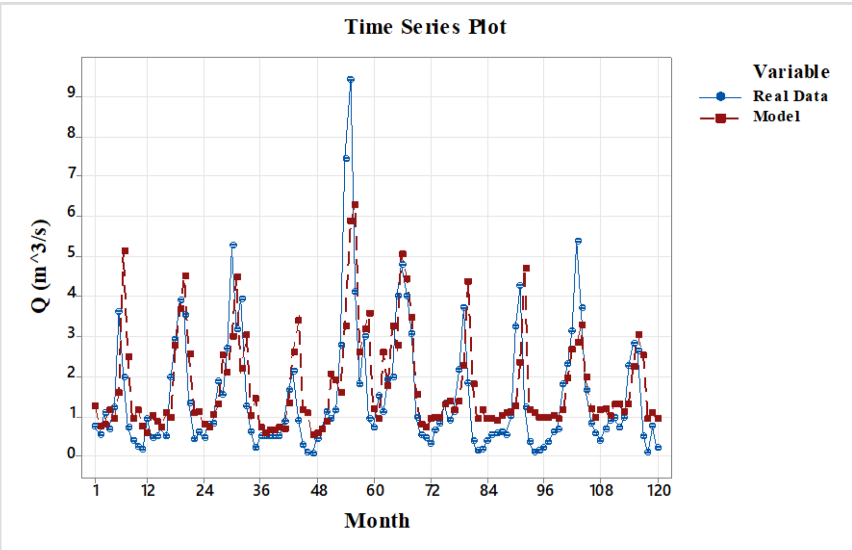


Fig. 8. Time series plot of forecasted and real data for the RF model

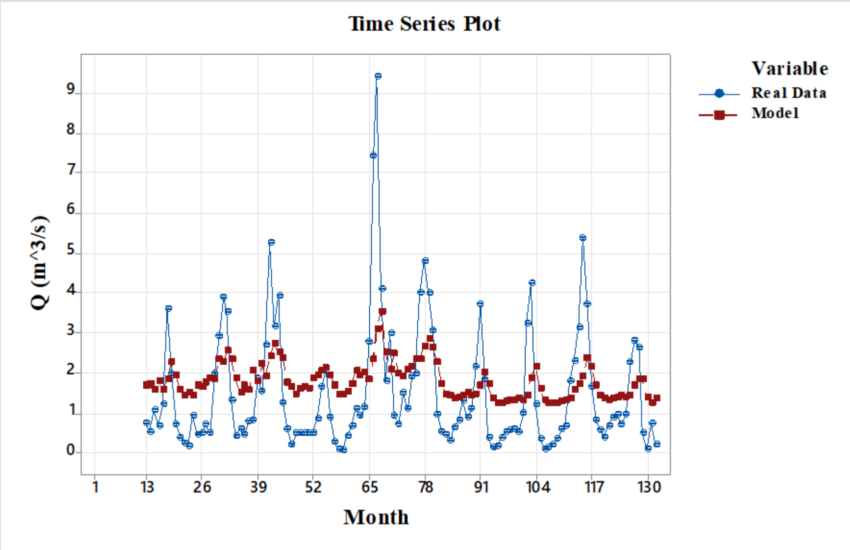


Fig. 9. Time series plot of forecasted and real data for the GPR model

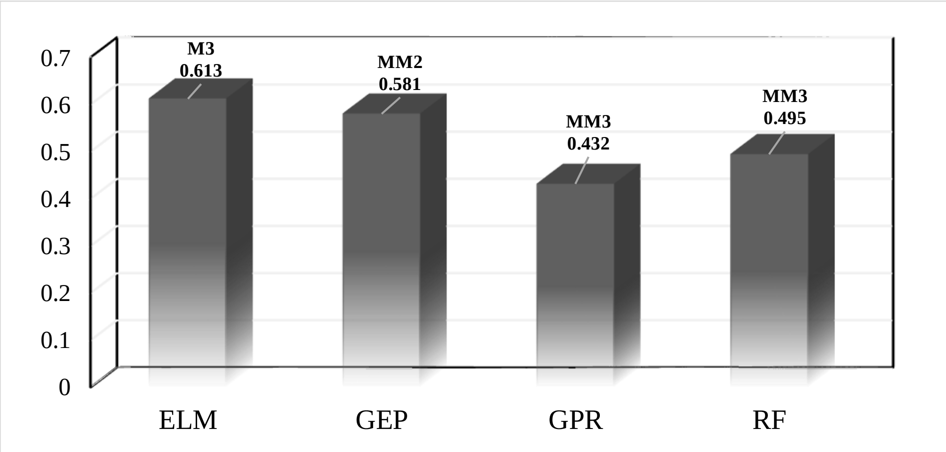


Fig. 10. The R^2 values for the best scenario in each model

4. Conclusion

In this article, 4 machine learning algorithms, namely the ELM, GEP, GPR, and RF were used. The studied area was the Chehel Chai watershed located in Golestan province, Iran. The stream flow and precipitation data of this watershed were collected during a 46-year period and the modeling process was done on the data. In order to model each algorithm, 6 scenarios were examined. The first 3 scenarios, the stream flow data were entered into the model with time lags 1, 2, and 3, and in the next three scenarios, the data were a combination of the stream flow and precipitation with 1, 2, and 3 time lags that were included in the modeling process. Finally, scenario M3 in the ELM algorithm with the value of RMSE=0.984 had the best performance. In the GEP method, the MM2 scenario with the value of RMSE=1.017 has provided the most accurate performance. In algorithms GPR and RF, scenario MM3 with values of RMSE equal to 1.336 and 1.139 have had the strongest performance, respectively. In general, scenario M3 in the ELM model with value of R^2=0.613 had the most accurate performance among all models and scenarios, and in the next place, scenario MM2 in the GEP algorithm with value of R^2=0.581 had better results than the other algorithms and scenarios. For future work, the use of hybrid models such as the use of decomposition functions such as CEEMD (complete empirical mode decomposition) or dimensionality reduction functions such as PCA (principle component analysis) are suggested to investigate the ability to increase the accuracy of modeling.

5. List of Acronyms

RMSE: Root Mean Square Error
MAE: Mean Absolute Error
GPR: Gaussian Process Regression
GEP: Gene Expression Programming
ELM: Extreme Learning Machine
ACF: Auto Correlation Function
PACF: Partial Auto Correlation Function
MLR: Multiple Linear Regression
SVR: Support Vector Machine
ARIMA: Autoregressive Moving Average
SVM: Support Vector Machine
BNN: Bayesian Neural Networks
KNN: Kernel Nearest Neighborhood
M5RT: M5 Regression Tree
ANN: Artificial Neural Networks
GA: Genetic Algorithm
RF: Random Forest
GP: Gaussian Process
LSTM: Long Short-Term Memory

6. Disclosure statement

No potential conflict of interest was reported by the authors

7. References

Abudu, S., Cui, C. L., King, J. P., & Abudukadeer, K. (2010). Comparison of performance of statistical models in forecasting monthly stream flow of Kizil River, China. *Water Science and Engineering, 3*(3), 269-281.

Adnan, R. M., Dai, H. L., Mostafa, R. R., Islam, A. R. M. T., Kisi, O., Elbeltagi, A., & Zounemat-Kermani, M. (2023). Application of novel binary optimized machine learning models for monthly streamflow prediction. *Applied Water Science, 13*(5), 110.

Adnan, R. M., Liang, Z., Kuriqi, A., Kisi, O., Malik, A., & Li, B. (2020, October). *Stream flow forecasting using heuristic machine learning methods*. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.

Ahmadi, F., Mehdizadeh, S., & Nourani, V. (2022). Improving the performance of random forest for estimating monthly reservoir inflow via complete ensemble empirical mode decomposition and wavelet analysis. *Stochastic Environmental Research and Risk Assessment, 36*(9), 2753-2768.

Asefa, T., Kemblowski, M., McKee, M., & Khalil, A. (2006). Multi-time scale stream flow predictions: The support vector machines approach. *Journal of hydrology, 318*(1-4), 7-16.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly stream flow forecasting using machine learning methods. *Journal of Hydrology, 590*, 125376.

Essam, Y., Huang, Y. F., Ng, J. L., Birima, A. H., Ahmed, A. N., & El-Shafie, A. (2022). Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms. *Scientific Reports, 12*(1), 3883.

Ferreira, C. (2001). Gene expression programming: a new adaptive algorithm for solving problems. *arXiv preprint cs/0102027*.

Garsole, P., & Rajurkar, M. (2015). *Stream flow forecasting by using support vector regression. In Proc., 20th Int*. Conf. of Hydraulics, Water Resources and River Engineering.

Guo, J., Zhou, J., Qin, H., Zou, Q., & Li, Q. (2011). Monthly stream flow forecasting based on improved support vector machine model. *Expert Systems with Applications, 38*(10), 13073-13081.

Hong, W. C. (2008). Rainfall forecasting by technological machine learning models. *Applied Mathematics and Computation, 200*(1), 41-57.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). *Extreme learning machine: theory and applications. Neurocomputing, 70*(1-3), 489-501.

Islam, K. I., Elias, E., Carroll, K. C., & Brown, C. (2023). Exploring random forest machine learning and remote sensing data for streamflow prediction: An alternative approach to a process-based hydrologic modeling in a snowmelt-driven watershed. *Remote Sensing, 15*(16), 3999.

Kalra, A., & Ahmad, S. (2009). Using oceanic-atmospheric oscillations for long lead time stream flow forecasting. *Water Resources Research, 45*(3), 233-242.

Khan, M. A., Farooq, F., Javed, M. F., Zafar, A., Ostrowski, K. A., Aslam, F., ... & Maślak, M. (2021). Simulation of depth of wear of eco-friendly concrete using machine learning based computational approaches. *Materials, 15*(1), 58-70.

Kişi, Ö. (2007). Stream flow forecasting using different artificial neural network algorithms. *Journal of Hydrologic Engineering, 12*(5), 532-539.

Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., & Caloiero, T. (2023). Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models. *Water, 15*(14), 2572.

Li, P. H., Kwon, H. H., Sun, L., Lall, U., & Kao, J. J. (2010). A modified support vector machine based prediction model on stream flow at the Shihmen Reservoir, Taiwan. *International Journal of Climatology, 30*(8), 1256-1268.

Luna, I., Soares, S., Magalhaes, M. H., & Ballini, R. (2005, July). *Stream flow forecasting using neural networks and fuzzy clustering techniques*. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. (Vol. 4, pp. 2631-2636). IEEE.

Miche, Y., Sorjamaa, A., & Lendasse, A. (2008, September). *OP-ELM: theory, experiments and a toolbox*. In International conference on Artificial Neural networks (pp. 145-154). Berlin, Heidelberg: Springer Berlin Heidelberg.

Orellana-Alvear, J., Célleri, R., Rollenbeck, R., Muñoz, P., Contreras, P., & Bendix, J. (2020). Assessment of native radar reflectivity and radar rainfall estimates for discharge forecasting in mountain catchments with a random forest model. *Remote Sensing, 12*(12), 1986.

Pustokhina, I., Seraj, A., Hafsan, H., Mostafavi, S. M., & Alizadeh, S. M. (2021). Developing a robust model based on the gaussian process regression approach to predict biodiesel properties. *International Journal of Chemical Engineering, 2021*, 1-12.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2010). *Short lead-time stream flow forecasting by machine learning methods, with climate variability incorporated*. In World Environmental and Water Resources Congress 2010: Challenges of Change (pp. 4608-4619).

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily stream flow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology, 414*, 284-293.

Sahranavard, H., & Naseri, M. (2022). Investigating the effect of observational data length on ARIMA modeling and forecasting accuracy: a case study of Kortian Stream watershed, Iran. *Water Supply, 22*(11), 8253-8265.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology, 85*, 1-16.

She, N., & Basketfield, D. (2005). Long range forecast of stream flow using support vector machine. *In Impacts of global climate change*, 1-9.

Shiri, J., & Kisi, O. (2010). Short-term and long-term stream flow forecasting using a wavelet and neuro-fuzzy conjunction model. *Journal of Hydrology, 394*(3-4), 486-493.

Singh, A. K., Kumar, P., Ali, R., Al-Ansari, N., Vishwakarma, D. K., Kushwaha, K. S., ... & Heddam, S. (2022). An integrated statistical-machine learning approach for runoff prediction. *Sustainability*, *14*(13), 8209.

Sun, A. Y., Wang, D., & Xu, X. (2014). Monthly stream flow forecasting using Gaussian process regression. *Journal of Hydrology, 511*, 72-81.

Tongal, H., & Booij, M. J. (2018). Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of hydrology, 564*, 266-282.

Tosunoğlu, F., HANAY, S., Çintaş, E., & Özyer, B. (2020). Monthly stream flow forecasting using machine learning. *Erzincan University Journal of Science and Technology, 13*(3), 1242-1251.

Tyralis, H., Papacharalampous, G., & Langousis, A. (2021). Super ensemble learning for daily stream flow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Computing and Applications*, *33*(8), 3053-3068.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators, 52*, 394-403.

Zakaria, Z. A., & Shabri, A. (2012). Stream flow forecasting at ungaged sites using support vector machines. *Applied Mathematical Sciences, 6*(60), 3003-3014.