Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Data division effect on machine learning performance for prediction of streamflow

Okan Mert KATİPOĞLU [1]

[1] Erzincan Binali Yıldırım University, Civil Engineering Department, okatipoglu@erzincan.edu.tr, Orcid No: 0000-0001-6421-6087

ARTICLE INFO

Article history:

Received 7 August 2022
Received in revised form 13 October 2022
Accepted 25 November 2022
Available online 31 December 2022

Keywords:

Stream flows, XGBoost, K-Nearest Neighbours, Data division, Euphrates basin

Doi: 10.24012/dumf.1158748

* Corresponding author

ABSTRACT

Accurate estimation of streamflow has an important role in water resources management, disaster preparedness and early warning, reservoir operation, and sizing of water structures. In this study, Extreme gradient boosting (XGBoost) and K-Nearest Neighbours (KNN) algorithms are used for the estimation of streamflow. In order to reveal the appropriate model, the raw model and models with optimized parameters were evaluated while the models were being built. In the setup of the models, various training test rates were also tried, and it was investigated which data division showed more effective results. For this purpose, the data were divided into ratios such as 60-40, 70-30, 80-20, and 90-10, respectively, and the model results were compared. Various statistical indicators such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) were used when comparing the models. As a result of the analysis, it was determined that the most suitable model for monthly streamflow estimation was obtained by using the optimized Xgboost algorithm and 60-40% data division. The obtained outputs constitute a vital resource for decision-makers regarding water resources planning and flood and drought management.

Introduction

Increasing water requirements, climate change and extreme weather events have increased the importance of effective planning and management of water resources [1]. The effective management of water resources depends on the prediction of possible future flows. Accurate and reliable streamflow estimation is crucial regarding irrigation planning, water and electricity supply, flood and drought risks, and reservoir operation. [2]. In particular, estimating maximum and minimum flow values is critical for their use in flood and drought management. The physical formation process of streamflow in a basin is complicated because it depends on several factors such as precipitation, evapotranspiration, infiltration, ground moisture, soil permeability, terrain conditions and vegetation. For this reason, many researchers have modeled past precipitation and streamflow values with various black box models that require less data and can accurately model non-linear relationships [3].

Artificial intelligence (AI) technologies predict streamflows accurately and reliably with the development of AI technologies. However, it has attracted the attention of many researchers about which algorithm can predict streamflows with higher accuracy. In addition, the most important parameter affecting the success of the AI model is the correct selection of the training and test rate. The determination of this ratio is one of the topics that attracts the attention of many researchers. For this reason, the streamflow estimation performances of various AI techniques and data division ratios have been investigated with different statistical and graphical indicators. Humphrey, et al. [4] employed the bayesian artificial neural network approach and successful estimation results were obtained in the monthly streamflow estimation model by separating 80% of the data for training. Tosunoğlu, et al. [5] used Support Vector Machines (SVM), Adaptive Boosting (AdaBoost), K-Nearest Neighbors (KNN) and Random Forest (RF) methods to estimate monthly flows at station 2305, located in the Euphrates basin. As a result, it has been determined that the RF algorithm is the best. Parisouj, et al. [2] estimated daily flows of Support Vector Regression (SVR), backpropagation Artificial Neural Network (ANN) and Extreme Learning Machine (ELM) in 4 rivers in the United States. Adnan, et al. [6] employed an ANN and genetic algorithm (ANN-GA) and ANFIS-genetic algorithm (ANFIS-GA) and M5 Regression Tree to predict the monthly flow data of Pakistan's Neelum and Kunhar Rivers. It has been determined that the model of

ANN-GA and ANFIS-GA is superior to the Regression Tree. Yu, et al. [1] estimated the inputs that were separated into various components with Fourier transform and 10-day streamflow values using SVR and extreme gradient boosting (XGBoost) algorithms. Ni, et al. [7] used XGBoost and the Gaussian mixture model (GMM) to forecast streamflow data. The GMM-XGBoost model showed the best performance. Tao, et al. [8] were evaluated the effects of various machine learning algorithms and three data divisions, (train-test: 70%–30%, 80–20%, and 90%–10%), to predict streamflows in the semi-arid region of Iraq. As a result of the study, the genetic algorithm and support vector regression hybrid model established using 90% training-10% testing rates showed the best monthly river flow estimation performance. Al-Juboori [9] combined KNN and RF algorithms to generate monthly flow data for a river from annual flow data. Dornpunya, et al. [10] used the XGBoost algorithm and three different data divisions to estimate the daily and monthly reservoir inflow values of Sirikit Dam in Thailand. As a result of the study, the results of the appropriate estimation were obtained in the case of the training and testing ratio of 80:20. Tyralis, et al. [11] estimated daily stream flows with various models such as extremely randomized trees, XGBoost, random forests, MARS, lasso, support vector regression, ARIMA. Adnan, et al. [12] used bio-inspired algorithms and various machine learning models for flow prediction in Pakistan. Meshram, et al. [13] applied ANFIS, GP and ANN to predict streamflow in the Shakkar watershed, India. Katipoğlu [14] modeled the monthly average stream flows of the Karasu river, 2154 no, with ANN. In the model setup, the data were separated as 70% train, 15% test, and 15% validation. As a result, successful predictions were obtained.

The main objectives of this study are:

• Comparison of the performances of XGBoost and K-NN algorithms from the estimation of monthly flows

• Evaluation of the effect of data division ratios on the performance of machine learning models.

Extreme gradient boosting (XGBoost)

XGBoost is a widely applied machine learning algorithm for tree boosting proposed by Chen and Guestrin [15]. XGBoost is a faster and better-performing variant of gradient-assisted decision trees. Gradient boosting establishes new models to predict previous model errors. The latest model combines all installed models. In this algorithm, the tree ensemble model has trained additively until the stopping criteria are met. Xgboost is based on the Classification and Regression Tree (CART) and uses the minimization of the loss function to reveal partition attributes. XGBoost minimizes the loss function via Equation (1) to detect the most suitable feature.

$$\text{Obj}\ (\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (1)$$

Where $L(y_i, \hat{y}_i)$ shows the training error [16].

K-Nearest Neighbours (KNN)

It is an algorithm that makes classification based on the distance between the data of a problem. The logic of this classification is expressed by the distance between the samples from the same class or from different classes according to their similarities. In the variable estimation of the KNN algorithm, it applies an input from the training data by comparing it with the value of the nearest neighbors. First, classification is done by the proximity of a selected feature to its closest feature. Then, the distances between the objects are calculated using Equation (2) [17].

$$\mathrm{d(i,j)} = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2} \qquad (2)$$

Material and Methods

Study area and data

The mean annual flow of the Euphrates is roughly 32 000 $m^{3,}$ and 80% of this amount is located in the upper basin to the north of the Keban Dam. The maximum flow in April and May corresponds to 42% of the total annual flow. Flow values in the Euphrates River basin vary between 200 and 2000 m^3/s. The lowest flows are observed in the winter, while the highest flows occur in the spring [18].

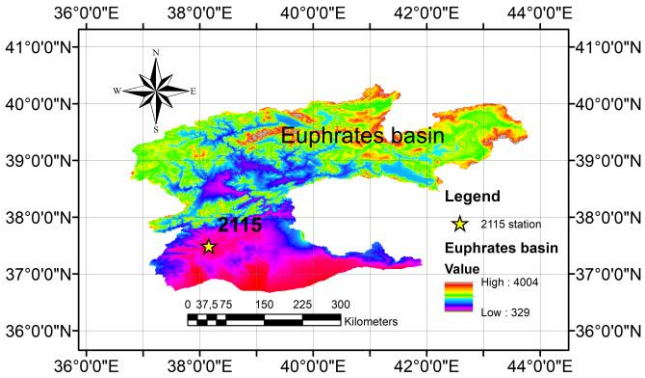

Figure 1. Digital elevation model of Euphrates basin.

The data used in the study were taken from the streamflow observation yearbooks organized by the general directorate of electric power resources survey and development administration (EIEI). These data cover the years 1970 to 2009 (40 years). The location map of the streamflow observation station used in the study is presented in Fig. 1. In Table A1, some statistical properties of the stream flows of the Göksu river numbered 2115 are given. The stream

flow station is located at 38° 9' 26'' East - 37° 29' 36'' North coordinates and 397 m altitude. In addition, the precipitation area of the basin is 185,000 km^2.

Performance criteria

Root mean square error (RMSE) gives the standard deviation of the best fit, which shows the closeness of the data to the best fit line. The mean of the absolute errors in a set of predictions is expressed as the mean absolute error (MAE). Having RMSE and MAE values close to zero will show the model's success. The coefficient of determination (R^2) expresses how well the estimated and actual data represent the regression line. R^2 value close to 1 indicates the model's success [19]. The R^2 value measures the linear relationship between actual and predicted values. Henseler, et al. [20] stated that 0.75, 0.50 and 0.25 values for R^2, respectively, indicate significant, moderate and poor fit. The statistical indicators are calculated with Equations (3, 4, 5).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Q_i - \hat{Q}_i\right)^2} \quad (3)$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}\left|\hat{Q}_i - Q_i\right|}{N} \quad (4)$$

$$R^2 = 1 - \frac{\left(Q_i - \hat{Q}_i\right)^2}{(Q_i - \bar{Q})^2} \quad (5)$$

where N, Q_i, $\hat{Q}_i$, and $\bar{Q}$ show the sample's length, real value, predicted value, and the mean of the real values, respectively.

Results and Discussion

This study used KNN and Xgboost algorithms to estimate monthly flows. In addition, the effects of data division ratios on model success were investigated in flow estimation. For machine learning models' setup, autocorrelation (ACF) and partial autocorrelation (PACF) graphs of flow data were obtained. Lagged streamflow values with an autocorrelation above 95% limits in the graphs were chosen as inputs to the ML models (Fig. 2).

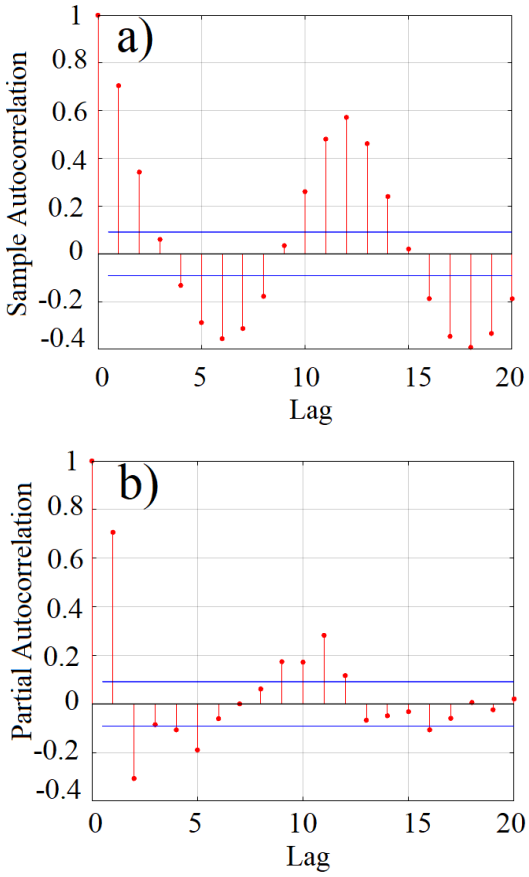

Figure 2. a) ACF, b) PACF graphs.

Fig. 2 shows the ACF and PACF graphs of the 2115 flow observation station. According to the ACF and PACF graphs, the flow values showing the highest intrinsic dependency were used as inputs to the machine learning models for flow estimation. According to the graphs, it was decided to estimate the flow values using the data of the past eight months.

Established model: f(Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5), Q(t-6), Q(t-7), Q(t-8)) = Q(t)

The data division rates tried are shown in Table 2. Here, the most commonly used data rates for establishing ML models in the literature are tested and how the performance of ML models changes is analyzed.

Table 2. Data division combinations

Model	Train	Test
M1	%60	%40
M2	%70	%30
M3	%80	%20
M4	%90	%10

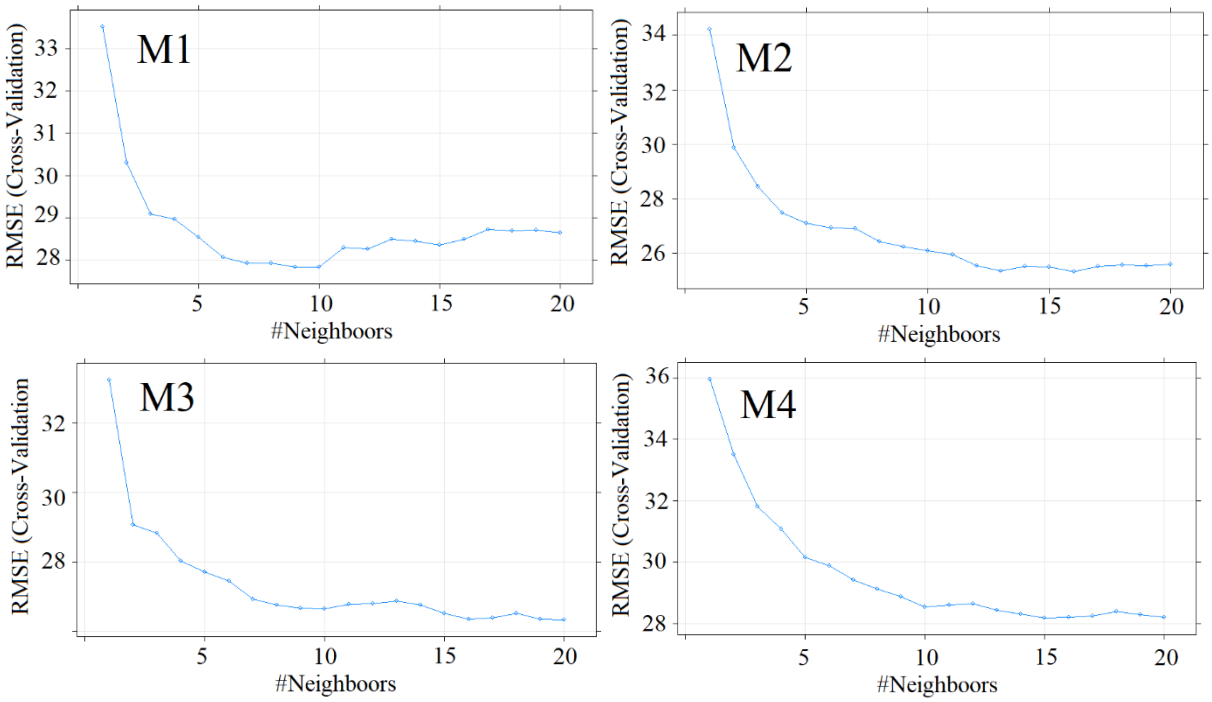

Figure 3 Selection of the optimum k parameter of the KNN model.

In Fig. 3, the change of k parameters of the KNN model is shown. To optimize the KNN model, the model with the lowest error rate was selected as the best model by being tested between k = 1:20. Accordingly, a regression model was established with 10, 16, 20 and 15 nearest neighbor values in M1, M2, M3 and M4 models, respectively.

Table 3. Test results of KNN and Xgboost models in various data divisions

	KNN				**Xgboost**			
	RMSE	R^2	MAE	Total Rank	RMSE	R^2	MAE	Total Rank
M1	**28.240**	**0.559**	**18.733**	**9**	**25.602**	**0.639**	**16.060**	**10**
Rank	**3**	**3**	**3**		**3**	**4**	**3**	
M2	33.374	0.582	19.276	8	30.296	0.627	16.754	7
Rank	2	4	2		2	3	2	
M3	34.348	0.517	20.141	4	32.743	0.539	18.618	3
Rank	1	2	1		1	1	1	
M4	24.584	0.482	16.158	9	23.049	0.571	13.252	10
Rank	4	1	4		4	2	4	

Note: Bold characters indicate best model and data division

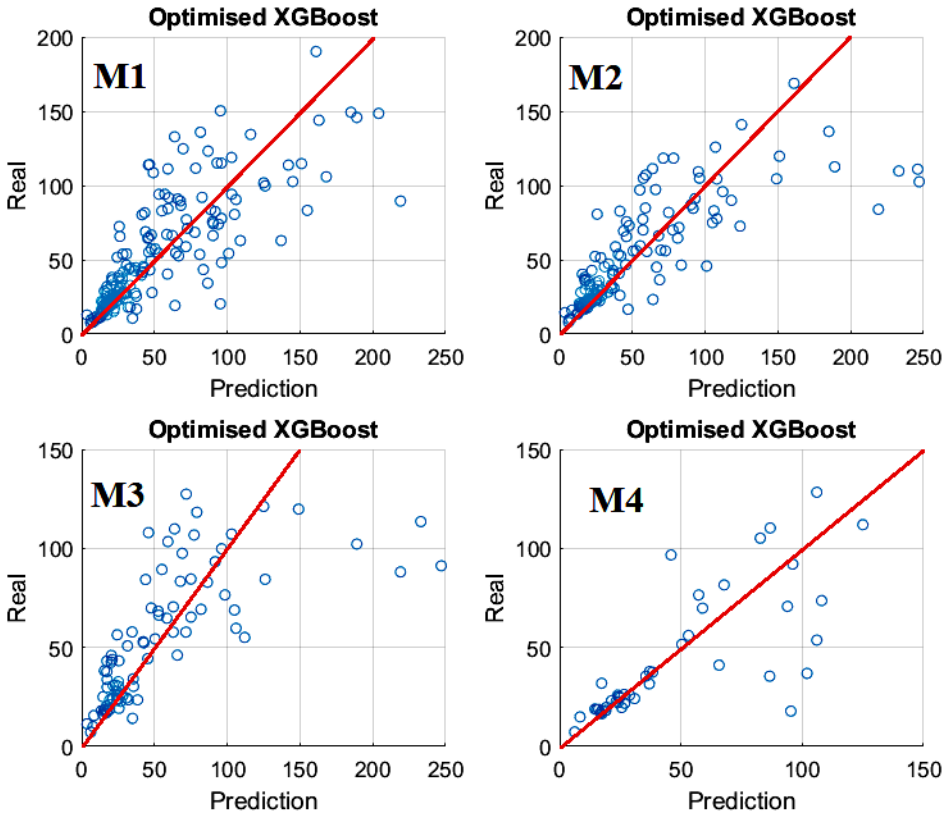

Figure 4. Scatter diagrams of test data of Xgboost models.

Table 3 shows the performances of Xgboost and KNN algorithms in estimating monthly flows using various data division ratios. According to the total rank values, the most successful data divisions are 60-40% and 90-10%. However, when the estimation results obtained with the 90-10% data rates of the KNN model are examined, it has been determined that the rate of 60-40% is better in the flow estimation since the R^2 value is below the acceptable value of 0.5. In addition, it has been determined that the most successful algorithm in monthly flow estimation is Xgboost. In the established XGBoost model, parameters nrounds = 1000, max_depth = 6, eta = 0.01, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 0.5 are used.

Fig. 4 shows the scatter diagrams obtained for various data rates of the Xgboost models. Scattering diagrams show the relationship between actual and predicted current values. It has been determined that the Xgboost model, established using the M1 data division ratio (60-40%), performs the best because the actual and predicted data are distributed around a 45° line.

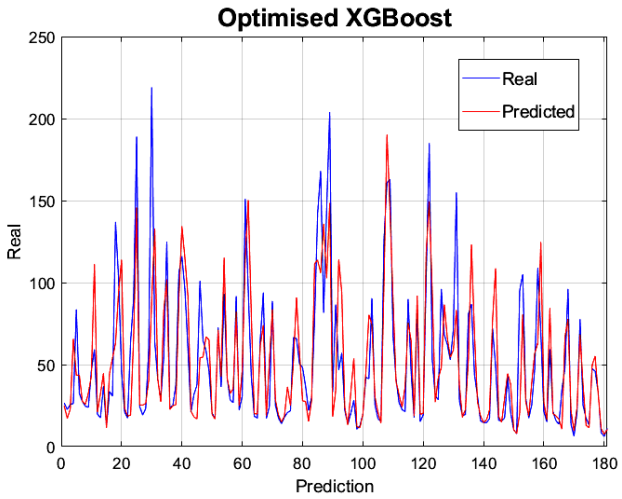

Figure 5. Scatter diagrams of test data of best model.

The flow line of the XGBoost model installed with the M1 combination is shown in Figure 5. The fact that the actual and estimated current values are compatible with each other confirms the high performance of the established model. Furthermore, when the figure is examined, the established model predicts the monthly average streamflow data satisfactorily, except for some maximum values.

Tao, et al. [8] investigated the effect of data division in estimating stream flows, and as a result, the most successful results were obtained using 90% training-10% testing rates. Katipoğlu [14] estimated monthly flows with the ANN model. In designing the model, the data was divided into 70% train, 15% test, and 15% validation. As a result of the research, realistic estimation results were produced. When the literature is evaluated, the data rates used in different locations can produce successful output in streamflow estimation. For this reason, it is necessary to choose the appropriate model by trying various data division ratios.

Ghorbani, et al. [21] revealed that the cascade correlation neural network and the random forest algorithm indicated high streamflow prediction achievement. Ni, et al. [7] determined that XGBoost is better than SVM in streamflow prediction. Elkurdy, et al. [22] combine the prediction Variational Mode Decomposition (VMD) and XGBoost models in daily streamflow in the Bow River (Alberta, Canada). As a result, it has been determined that the hybrid VMD-XGBoost model exhibits very high prediction success. The outputs of the study are largely in line with previous studies. In this direction, it can be deduced that the XGBoost algorithm has significantly superior features in estimating monthly flow data. In addition, it is thought that it can be used in other basins and stations.

Conclusion

This study used machine learning models such as KNN and Xgboost to predict monthly flows at the streamflow observation station 2115 in the Euphrates basin. In addition, the effect of data division ratios on model performance in flow estimation was investigated. The main outputs of the study are listed as follows:

• Xgboost algorithm presented the most successful result in estimating monthly flows.

• It has been determined that flows can be predicted effectively using data from the past eight months. R^2: 0.639 value indicates that currents can be modeled satisfactorily.

• The best data division is obtained by using training: 60% and testing 40%.

• It has been understood that ACF and PACF graphics have a remarkable place in the selection of the input combination.

• As a result of the study, streamflow data can be predicted effectively with AI methods. When Figure 5 is examined, while the minimum streamflows can be estimated effectively, the maximum streamflows deviate significantly from the actual values. This proves that the XGBoost algorithm can effectively predict low flows and droughts.

• The study's main limitation is the analysis in a single station. Therefore, to develop the results, it is necessary to test the models in different climatic regions and streams with varying flow regimes.

• To increase the performance of the flow estimation model, it is recommended to try hybrid ML models with various preprocessing methods such as empirical mode decomposition and Wavelet transform in future studies.

Acknowledgement

Thanks to the General Directorate of Electric Power Resources Survey and Development Administration for providing monthly stream flows

References

[1] X. Yu, Y. Wang, L. Wu, G. Chen, L. Wang, and H. Qin, "Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting," Journal of Hydrology, vol. 582, p. 124293, 2020.

[2] P. Parisouj, H. Mohebzadeh, and T. Lee, "Employing machine learning algorithms for streamflow prediction: a case study of four river basins with different climatic zones in the United States," Water Resources Management, vol. 34, no. 13, pp. 4113-4131, 2020.

[3] W. Wang, Stochasticity, nonlinearity and forecasting of streamflow processes. Ios Press, 2006.

[4] G. B. Humphrey, M. S. Gibbs, G. C. Dandy, and H. R. Maier, "A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network," Journal of Hydrology, vol. 540, pp. 623-640, 2016.

[5] F. Tosunoğlu, S. HANAY, E. Çintaş, and B. Özyer, "Monthly streamflow forecasting using machine learning," Erzincan University Journal of Science and Technology, vol. 13, no. 3, pp. 1242-1251, 2020.

[6] R. M. Adnan, Z. Liang, A. Kuriqi, O. Kisi, A. Malik, and B. Li, "Streamflow forecasting using heuristic machine learning methods," in 2020 2nd International Conference on Computer and Information Sciences (ICCIS), 2020: IEEE, pp. 1-6.

[7] L. Ni et al., "Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model," Journal of Hydrology, vol. 586, p. 124901, 2020.

[8] H. Tao et al., "Training and testing data division influence on hybrid machine learning model process:

application of river flow forecasting," Complexity, vol. 2020, 2020.
[9] A. M. Al-Juboori, "A hybrid model to predict monthly streamflow using neighboring rivers annual flows," Water Resources Management, vol. 35, no. 2, pp. 729-743, 2021.
[10] P. Dornpunya et al., "The reservoir inflow prediction of Sirikit dam using artificial intelligence with machine learning: extreme gradient boosting techique," artificial intelligence, vol. 25, p. 26, 2021.
[11] H. Tyralis, G. Papacharalampous, and A. Langousis, "Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms," Neural Computing and Applications, vol. 33, no. 8, pp. 3053-3068, 2021.
[12] R. M. Adnan, R. R. Mostafa, A. Elbeltagi, Z. M. Yaseen, S. Shahid, and O. Kisi, "Development of new machine learning model for streamflow prediction: Case studies in Pakistan," Stochastic Environmental Research and Risk Assessment, vol. 36, no. 4, pp. 999-1033, 2022.
[13] S. G. Meshram, C. Meshram, C. A. G. Santos, B. Benzougagh, and K. M. Khedher, "Streamflow prediction based on artificial intelligence techniques," Iranian Journal of Science and Technology, Transactions of Civil Engineering, vol. 46, no. 3, pp. 2393-2403, 2022.
[14] O. M. Katipoğlu, "Monthly stream flows estimation in the Karasu river of euphrates basin with artificial neural networks approach," Mühendislik Bilimleri ve Tasarım Dergisi, vol. 10, no. 3, pp. 917-928, 2022.
[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.
[16] W. Yucong and W. Bo, "Research on EA-xgboost hybrid model for building energy prediction," in Journal of Physics: Conference Series, 2020, vol. 1518, no. 1: IOP Publishing, p. 012082.
[17] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunalı, M. Şimşek, and A. Özçift, "KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi," Marmara Fen Bilimleri Dergisi, vol. 28, no. 3, pp. 89-94, 2016.
[18] A. Yıldırım, "Karakaya barajı ve doğal çevre etkileri," DÜ Ziya Gökalp Eğitim Fakültesi Dergisi, vol. 6, pp. 32-39, 2006.
[19] M. Rose and N. Chithra, "Tree-based ensemble model prediction for hydrological drought in a tropical river basin of India," International Journal of Environmental Science and Technology, pp. 1-18, 2022.
[20] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The use of partial least squares path modeling in international marketing," in New challenges to international marketing: Emerald Group Publishing Limited, 2009.
[21] M. A. Ghorbani, R. C. Deo, S. Kim, M. Hasanpour Kashani, V. Karimi, and M. Izadkhah, "Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia," Soft Computing, vol. 24, no. 16, pp. 12079-12090, 2020.
[22] M. Elkurdy, A. D. Binns, and B. Gharabaghi, "Improved Streamflow Forecasting Using Variational Mode Decomposition and Extreme Gradient Boosting," in AGU Fall Meeting Abstracts, 2020, vol. 2020, pp. H165-0003.

Appendix

Table A1. Summary statistics of stream flows at station 2115

Statistic	October	November	December	January	February	March	April	May	June	July	August	September
Max.	51.10	89.00	106.00	125.00	148.00	247.00	279.00	163.00	67.80	42.00	28.80	25.60
Min.	7.81	2.70	17.10	16.60	18.60	31.70	25.80	15.60	7.83	3.65	3.60	6.18
Mean	20.89	28.18	44.80	54.40	67.35	101.37	112.78	71.67	38.40	24.15	17.64	16.03
Total	818	1091	1758	2143	2662	3978	4432	2831	1537	968	708	641
Skewness	1.56	2.21	0.84	0.74	0.33	1.77	0.90	0.67	0.11	0.06	-0.08	-0.09
Kurtosis	3.49	6.85	-0.54	-0.31	-0.01	3.47	0.37	-0.15	-0.92	-0.68	-0.70	-0.76
Median	19.15	24.95	34.35	51.10	68.00	87.85	92.05	63.45	37.45	24.30	17.20	15.55
Standart deviation	8.41	15.15	26.29	30.78	29.31	49.26	60.43	34.94	15.25	9.58	6.91	5.59
Variance	70.65	229.59	691.26	947.42	859.06	2427	3651	1220	232	91.84	47.80	31.22