Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2021, 14(3), 914-925
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2021, 14(3), 914-925
DOI: 10.18185/erzifbed.949126
Research Article

Daily Flow Modeling With Random Forest and K-Nearest Neighbor Methods

H. Yildirim DALKILIC [1], S. Nur YESILYURT[2*], Pijush SAMUI[3]

[1] Department of Civil Engineering, Erzincan Binali Yıldırım University, Erzincan 24000, Turkey.
[2*]School of Natural and Applied Sciences, Erzincan Binali Yıldırım University, Erzincan 24000, Turkey,
[3] Department of Civil Engineering, National Institute of Technology Patna, 800005.

Geliş / Received: 07/06/2021, Kabul / Accepted: 21/09/2021

Abstract

Water is an indispensable natural resource for life. Therefore, protection and control of water resources are of great importance. Since river flow estimation and modeling are very important in cases such as the management of water resources, irrigation, it is included in the literature as an issue that needs constant research and development. A large number of techniques are being used for estimation and modeling; thus, the estimation results are gradually improving with the development of the studies carried out, the comparison of techniques, and the determination and removal of the shortcomings. In this study, Random Forest and K-Nearest Neighbors nonlinear regression models, which are two of the machine learning methods, were used to evaluating the estimation results, to find the better estimation method, and to determine the advantages and disadvantages of these methods. In addition, Random Search and Grid Search methods were used to make the hyperparameter selection and comparison for the Random Forest model. In this study, in which daily flow data of 1981-2011 of the two stations in the Euphrates were used, and, when compared to other models, it was observed that better results were obtained when Random Search was applied to determine the hyperparameters of the Random Forest model.

Keywords: Random Forest, K-Nearest Neighbors, Random Search, Grid Search.

Random Forest ve K-Nearest Neighbor Yöntemleri ile Günlük Akım Modellemesi

Öz

Su, canlı yaşamı için vazgeçilemez bir doğal kaynaktır. Bu nedenle su kaynaklarının korunması ve kontrolü büyük önem arz etmektedir. Nehir akımı tahmini ve modellemesi; su kaynaklarının yönetimi, sulama faaliyetleri gibi durumlarda önem arz ettiği için sürekli araştırılmaya ve geliştirilmeye ihtiyaç duyulmuş bir konu olarak literatürde yer almaktadır. Tahmin ve modelleme için çok sayıda teknik kullanılmakta, yapılan çalışmaların gelişmesi, tekniklerin kıyaslanması ve eksik yönlerin görülmesi ile tahmin sonuçları giderek iyileşmektedir. Bu çalışmada da tahmin sonuçlarını değerlendirmek ve daha iyi olan tahmin yöntemini bulabilmek, yöntemlerin avantaj ve dezavantajlarını tespit edebilmek için makine öğrenmesi yöntemlerinden Rastgele Orman ve K-En Yakın Komşu doğrusal olmayan regresyon modelleri kullanılmıştır. Ayrıca RF modeli için hiperparametre seçimi Random Search (Rastgele Arama) ve Grid Search (Grid Arama) yöntemleri kullanılarak da oluşturulup kıyaslaması yapılmıştır. Fırat havzasında yer alan iki istasyona ait 1981-2011 yılları için günlük akım verileri kullanılan çalışmada; rank analizi ile nihai sonuca ulaşılmış olup diğer modellere göre Random Forest modelinin hiperparametrelerinin belirlenebilmesi için Random Search uygulandığında daha iyi sonuç alındığı görülmüştür.

Anahtar Kelimeler: Rastgele Orman, K-En Yakın Komşular, Random Search (Rastgele Arama), Grid Search (Grid Arama).

Corresponding Author: sefa.yesilyurt@erzincan.edu.tr

1. Introduction

Water is of great importance for the continuation of living life. However, since water is not an infinite resource, existing water resources are gradually decreasing due to the development of economic activities and increased consumption caused by the increasing population, thus, estimated water shortage in the future poses a risk to living life. The control and efficient use of these resources are of great importance for life and therefore the researchers are trying to make efficient planning for water control and consumption by aiming to determine the water potential. River flow estimation is one of the most important issues to be considered in these plans. In many studies, numerous methods were used for river flow estimation and it was aimed to achieve the best model structure and best estimation performance. Although the studies prove that estimations can be made by some mathematical methods, it is also seen that more successful results are obtained with artificial intelligence techniques and fuzzy logic methods, and modeling can be carried out in a shorter time through these. However, it has not become possible to reduce these methods to a single one or to create a universal one that is superior to the others (Yaseen et al., 2019). (Altunkaynak & Basakin, 2018) In their study, used the daily flow data of the Columbia River in the USA to compare the estimation performances of Adaptive Network-Based Fuzzy Logic Inference System (ANFIS), Artificial Neural Networks and Nonlinear Autoregressive Model (NAR) and Autoregressive Moving Average Models (ARIMA), and found out that ARIMA gave better results. (Chenga et al., 2020) used artificial neural network (ANN) and a Long Short-Term Memory (LSTM) method to estimate the flow up to 20 days ago. It was observed that the LSTM model worked better and it was emphasized that daily flow estimates are very important for water resources management.

There are numerous studies conducted in which Random Forest and K-Nearest Neighbor methods have been used. In one of these studies, (Tosunoglu et al., 2020) used Support Vector Machines (SVM), Adaptive Upgrade (AdaBoost), KNN and RF methods for monthly river flow modeling in Coruh basin and they observed that the RF model worked better according to the test results. In another study, river flow was estimated up to seven days ago using RF and Prophet methods and it was concluded that the RF model represented sudden flow fluctuations better than the other method (Papacharalampous & Tyralis, 2018). (Li et al., 2019) analyzed the daily flow data through five different models and compared their estimation capacities. When Extreme Learning Machine (ELM), extreme learning machine with kernels (ELM-kernel), RF, back-propagation neural network (BPNN) and support vector machine (SVM) methods were compared, it was found that the ELM-Kernel model gave the best results while the basic ELM model was found to have the worst predictive results. In addition, low and high flow data were evaluated in the study and it was observed that the Kernel-ELM model was superior in the estimation of these data sets, while it was also emphasized that it did not have an obvious advantage in any river flow modeling compared to other models. In their study, (Modares et al., 2018) used Artificial Neural Network (ANN), Generalized Regression Neural Network (GRNN), Least Square-Support Vector Regression (LS-SVR), and KNN to make the monthly flow estimations. It was shown that model performances differed in nonlinear and nonlinear cases and it was observed that the KNN model performed better in nonlinear cases.

In this study, RF and KNN models were used to perform flow modeling up to three days ago using the daily flow data of 1981-2011 of two stations located in the Euphrates basin, and they were compared in terms of their estimation capacities. In addition to the RF model basic application, hyperparameter determination was conducted through Random Search and Grid Search and the estimation performances of the models were evaluated. Model estimation results were evaluated by applying the square of correlation coefficient (R2), Root Mean Squared Error

(RMSE), and Mean Absolute Error (MAE) performance evaluation indices, and for final evaluation rank analysis was performed for all indices. Hence, it is aimed to find a better performing method for river flow estimation and to determine the advantages and disadvantages of all these methods.

2. Method

2.1. Random Forest (RF)

Developed by Breiman in 2001, RF, which is widely used because of its stability and good generalization, has an algorithm that combines multiple decision trees to predict (Breiman, 2001). The algorithm that largely solves the overfitting problem as it generates random sub-datasets from the datasets adjusts the number parameters of trees (ntree) and the number parameters of estimators tested at each node (mtry) to obtain the minimal correlation and generalization error (Breiman, 2001; Were et al., 2015). Empirical methods applied to find the mtry parameter have previously been used by different researchers as in Equations 1-3 (Huang, 2014; Al-Abadi & Shahid, 2016).

$$m_{try} = \log_2(M + 1) \quad (1)$$

$$m_{try} = \sqrt{M} \quad (2)$$

$$m_{try} = \frac{M}{3} \quad (3)$$

The M used in these equations refers to the number of input variables defined in the dataset. But empirical methods may not always give the best outcomes. Therefore, the model was handled with R-Programming language to select the most reasonable m_{try} value. Moreover, 10 times cross-validation method was also used to make the estimation and performance of the model more stable and safer (Li, Sha, & Wang, 2019).

2.2. Hyperparameter Optimization

When the RF model is created with its basic structure, it comes along with some disadvantages. For example, the trials are conducted with a limited number of parameters and the predictions in terms of the problem are limited in this case. Therefore, the presence of the most appropriate parameters to evaluate the model parameters can be effective in improving model performance. Random Search and Grid Search methods, which are widely used due to their ease of application and good results, were used with Caret package in R-programming language.

2.2.1. Random search

The Random Search method proposed by Bengio in 2012 is the process of using the preliminary information of the problem and determining the hyperparameter ranges and then finding the optimum values. In the method, in which hyperparameter ranges are determined using preliminary information, performance is observed by training the models with different parameter groups randomly instead of trying each of the values in these ranges. The most suitable hyperparameter group is determined according to the results. Since randomly selected

subsets are engaged instead of all combinations, Random Search structure, which is faster and allows for wide range parameter determination, is as given in Figure 1 (Bergstra & Bengio, 2012).

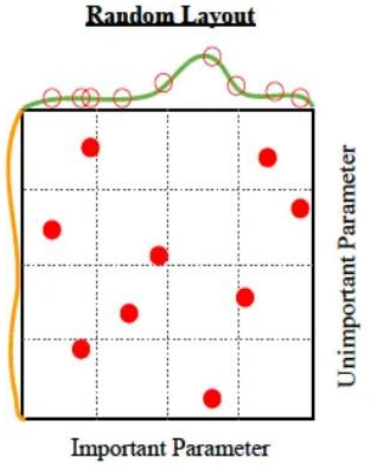


Figure 1. Random Search structure of nine trials to optimize a sample function (Bergstra & Bengio, 2012).

2.2.2. Grid search

Grid Search is a parameter optimization method that is applied as the process of selecting main points after determining the hyperparameter ranges by finding the preliminary information. The method, which creates a value list for parameters by means of the specified main points, trains the network for combinations of all values in the specified range, observes the results, and provides the best value. Although Grid Search structure, shown in Figure 2, takes a long time to be completed as it tries different parameter combinations, it is a widely used one since it can be examined in wide ranges (Bergstra & Bengio, 2012). The study, it is aimed to examine a wide range and improve the model performance by choosing a range such as 100,200,300,500,1000,2000.

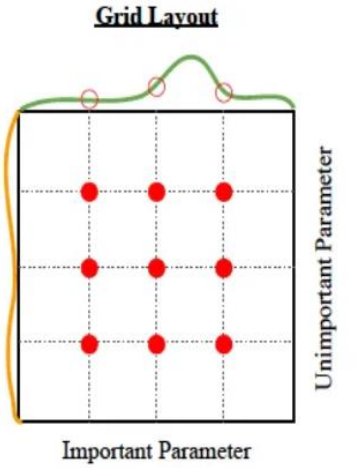


Figure 2. Grid Search structure of nine trials to optimize a sample function (Bergstra & Bengio, 2012).

2.3. K-Nearest Neighbors (KNN)

KNN, with the basic aim to find the points closest to the new point, has a structure that can be used for classification and regression. In both cases, it is of great importance to determine the number of k neighbors to predict the results. In the input feature area, k consists of the closest training samples, the output values for the regression problems applied in this study are the feature values and are calculated by taking the median of the values of their nearest neighbors. That is, the KNN structure is a preferred method due to its pattern-based simplicity and easy interpretation of its outputs. (Peterson, 2009). As the random assignment of k value while creating the KNN model would prevent the best result from obtaining, in this study, the number of model k neighbors was optimized by performing 10-storey cross-validation. In this way, it is aimed to improve the model performance. In addition, the most widely used Euclidean distance criterion for KNN was used in the study (Altunkaynak et al., 2020).

2.4. Model performance evaluation criteria

Three performance evaluation indices called R^2, RMSE, MAE were used to evaluate the performance of the developed models. These indices can be found through the following equations;

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_i - y_j\right| \tag{6}$$

y refers to the measured value, $\bar{y}$ refers to the average of the measured values, N represents the total number of data and the R^2 value can have the best value of 1, while RMSE and MAE can have the best value of 0. Rank analysis, on the other hand, is a method applied to determine the best-performing model taking all the evaluation criteria into account. The method, aiming to determine the performance evaluation score of the models and find the model that gives the best result, is carried out by assigning a rank to the models according to their proximity to the best value for each data set and comparing the scores for all data sets. If R_i is represented as the rank value in the selected model of each data set and n is the number of models, the total rank value will be as follows (Zhang et al., 2020):

$$Modal\ Total\ Rank = \sum_{i=1}^{n} R_i \tag{7}$$

3. Study Area and Data

The Euphrates Basin, which arises from the mountains in the east of Turkey and disembogues into the Persian Gulf and takes its name from the Euphrates River, has a precipitation area of 127,300 square kilometers. The basin, which has the largest drainage area in Turkey and is fed

by the longest river in Western Asia, the Euphrates River, has an average height of 1009.87 m, average annual flow value of 31.61cubic meters, and average precipitation value of 540.1 mm/year. Thus, it is crucial to examine this basin (EIEI, 2000). At the same time, the Euphrates basin is of great importance for Turkey because it has waters that cross the borders of Turkey, and it is also of great importance for coastal countries (Figure 3).

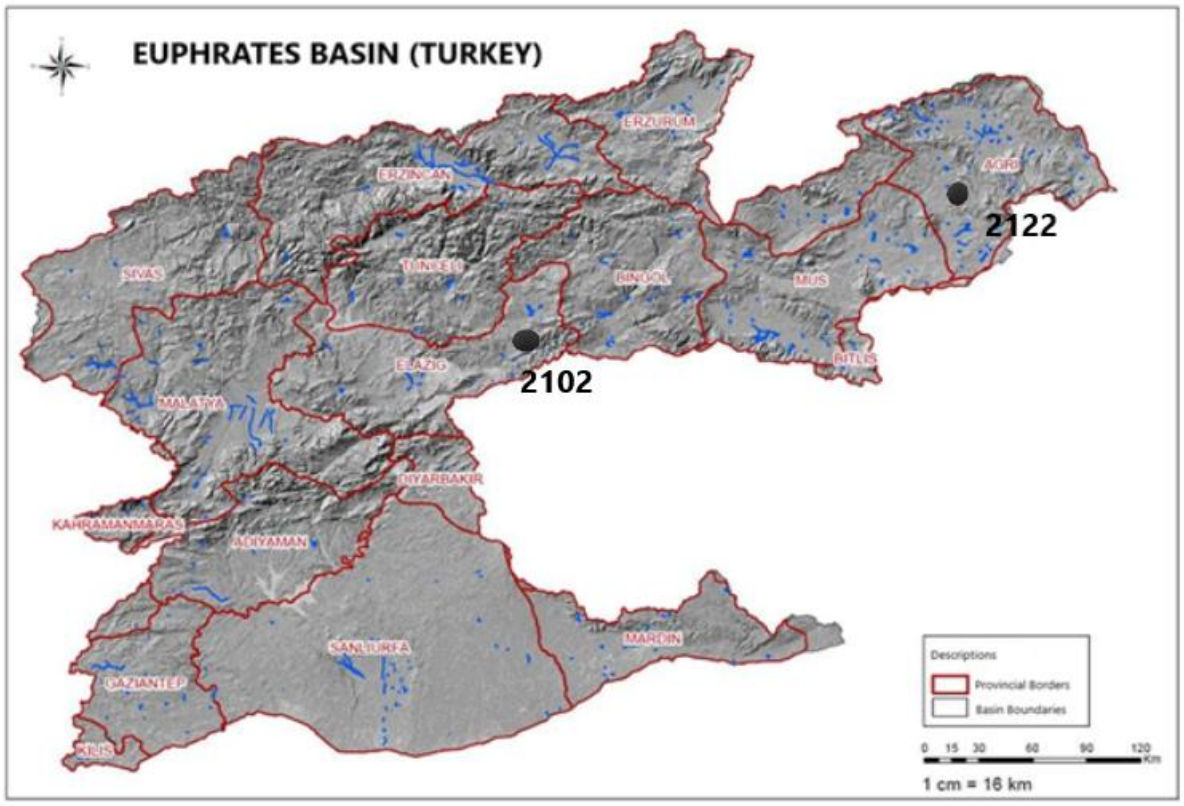


Figure 3. Euphrates Basin (the part within the borders of Turkey) (Yenigün & Gümüs, 2007)

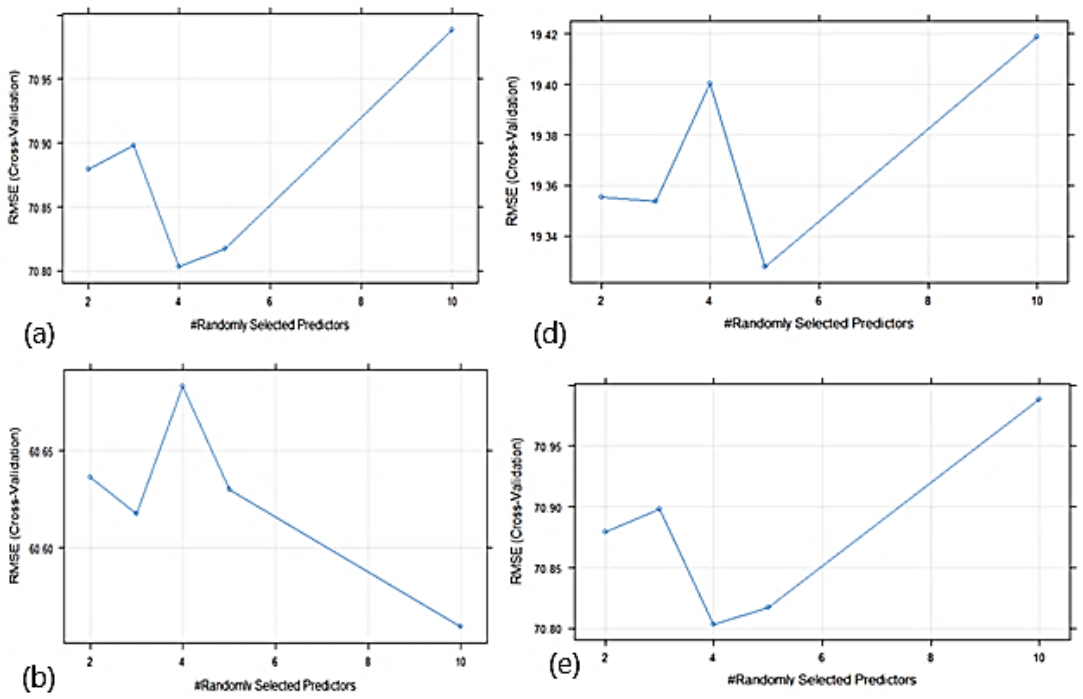
In the study, daily flow estimates were made using daily current data of the two stations (Table 1) in the Euphrates basin for the years 1981-2010 (IPCC projection reference range. The daily flow data of the Euphrates Basin, which has the most data within the borders of Turkey, have been obtained from the flow observation annuals of the relevant years published by the Electricity Administration Survey Works (EIEI) (EIEI, 2000) (DSI , 1981-2010). Various combinations were created using the flow data. As the best results in estimating the current flow data were obtained through the combinations formed with flow data from one, two and three days ago, these combinations were employed in the study. Studies in the literature were examined and the train-test ratio was decided. These studies, it was observed that better results were obtained when 70% or 80% train ratio was selected (Okkan & Inan, 2014; Alexis et al., 2017). In this study, 80% of the data was for training; 20% was used for testing. In addition, it is aimed to get better results by making cross validation.

Table 1. Information about stations

Station Number	Name	Longitute-Latitute	Mean (flow) (m^3/sn)	Max (flow) (m^3/s)	Min (flow) (m^3/s)	Standard deviation (flow)
2102	Murat River - Palu	(39° 56' 22" E - 38° 41' 49" N)	179,23	997	12,1	207,606
2122	Murat River- Tutak	(42° 46' 49" E - 39° 32' 19" N)	47,48	821	1,97	73,041

4. Findings

Flow data for stations 2102 and 2122 in the Euphrates basin were estimated using daily flow data from one, two and three days ago. If Q(t) is taken as flow data available and Q(t-n) is taken as flow data n days ago, model performance results can be shown as in Table 2. The RMSE graph for RF model results is given in Figure 4 and the RMSE graph for KNN model results is given in Figure5.



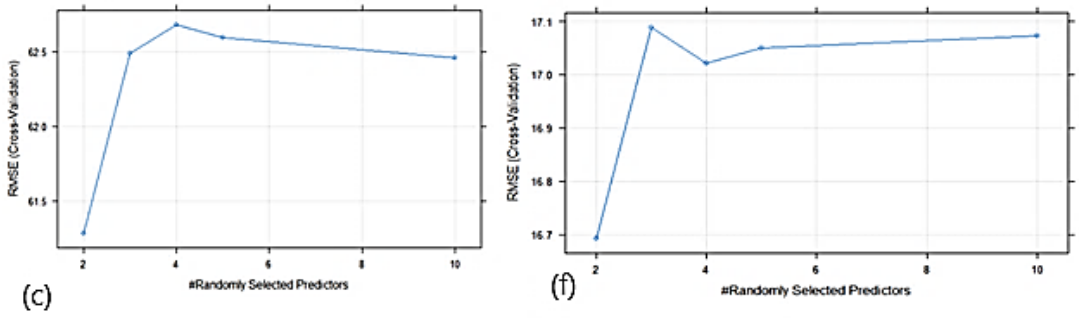


Figure 4. For the RF model structure; model performance of station 2102 (a) Q(t-1)-Q(t), (b) Q(t-1)&Q(t-2)- Q(t) model performance, (c) Q(t-1)&Q(t-2)&Q(T-3)-Q(t) model performance. Model performance of station 2122 (d) Q(t-1)-Q(t) model performance, (e) Q(t-1)&Q(t-2)- Q(t) model performance, (f) Q(t-1)&Q(t-2)&Q(T-3)-Q(t) model performance.

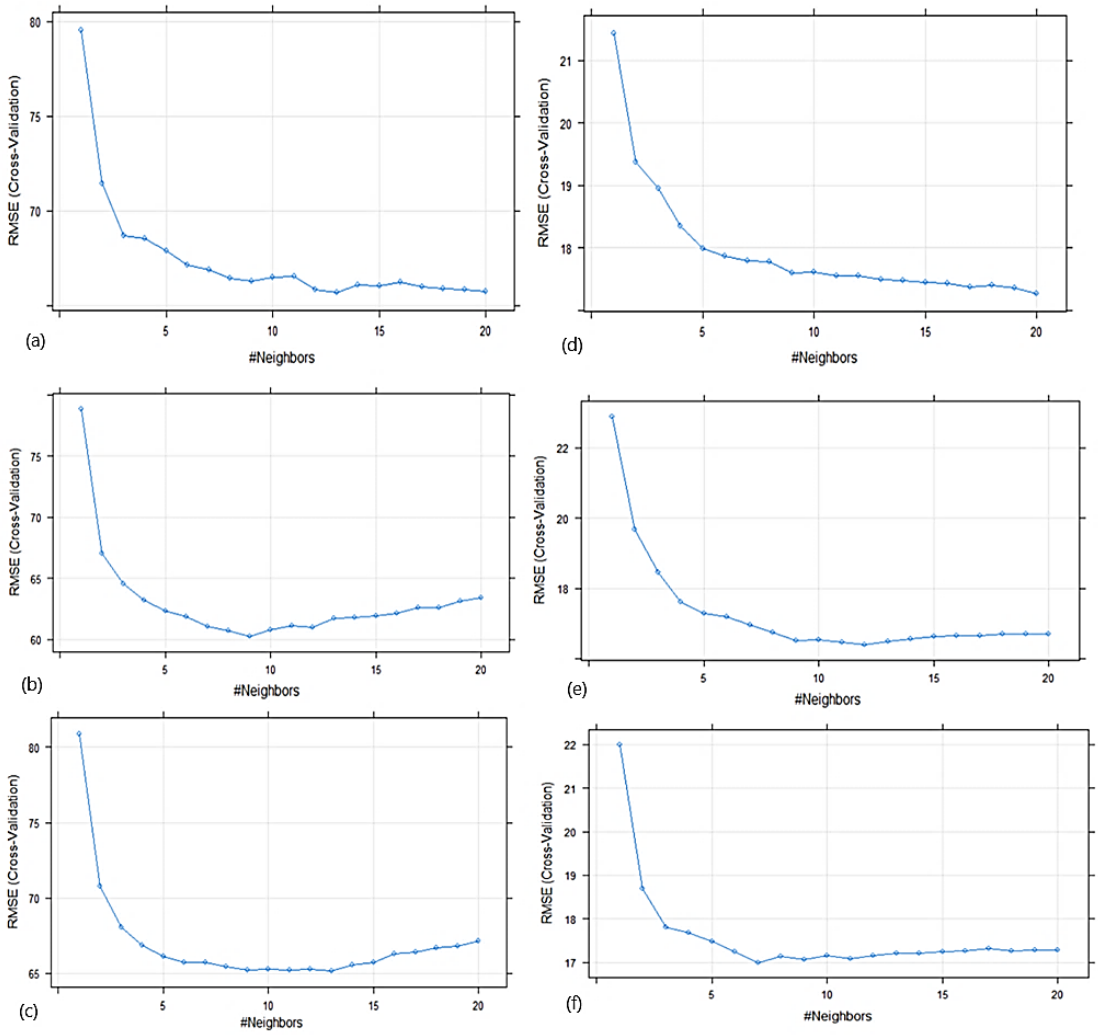


Figure 5. For the RF model structure; model performance of station 2102 (a) Q(t-1)-Q(t) model performance, (b) Q(t-1)&Q(t-2)- Q(t) model performance, (c) Q(t-1)&Q(t-2)&Q(T-3)-Q(t) model performance. Model performance of station 2122 (d) Q(t-1)-Q(t) model performance, (e) Q(t-1)&Q(t-2)- Q(t) model performance, (f) Q(t-1)&Q(t-2)&Q(T-3)-Q(t) model performance.

Table 2. Model Performances

Stations	Data Combination	Model	RMSE	RANK	R^2	RANK	MAE	RANK	TOTAL RANK
1	Q(t-1)-Q(t)	RF	70.804	1	0.960	3	26.144	1	5
		RF- Random Search	70.268	3	0.961	4	25.854	3	10
		RF - Grid Search	70.773	2	0.960	2	26.035	2	6
		KNN	75.096	4	0.955	1	25.089	4	9
	Q(t-1)&Q(t-2) - Q(t)	RF	60.560	4	0.970	3	21.447	4	11
		RF- Random Search	61.675	3	0.970	4	21.461	3	10
		RF - Grid Search	70.777	2	0.960	2	26.035	1	5
		KNN	73.154	1	0.958	1	22.367	2	4
	Q(t-1)&Q(t-2) & Q(t-3)-Q(t)	RF	69.035	4	0.964	2	23.831	2	8
		RF- Random Search	61.457	2	0.969	3	21.098	4	9
		RF - Grid Search	60.704	3	0.970	4	21.128	3	10
		KNN	71.866	1	0.961	1	23.982	1	3
2.1	Q(t-1)-Q(t)	RF	19.328	4	0.926	3	6.361	2	9
		RF- Random Search	19.482	3	0.925	2	6.356	3	8
		RF - Grid Search	70.773	1	0.960	4	26.035	1	6
		KNN	20.543	2	0.921	1	5.849	4	7
	Q(t-1)&Q(t-2) - Q(t)	RF	21.741	2	0.916	2	7.354	1	5
		RF- Random Search	16.646	3	0.943	3	5.455	4	10
		RF - Grid Search	16.540	4	0.943	4	5.571	3	11
		KNN	24.964	1	0.889	1	7.115	2	4
	Q(t-1)&Q(t-2) & Q(t-3)-Q(t)	RF	16.693	2	0.944	2	5.451	2	6
		RF- Random Search	16.553	4	0.945	3	5.435	3	10
		RF - Grid Search	16.694	3	0.948	4	5.263	4	11
		KNN	20.534	1	0.927	1	5.838	1	3

As shown in the table, rank analysis was applied for each data combination and the final evaluation of the model performances, the high-rank values were checked.

Total rank values are given in table 3.

Table 3. Final performance evaluation with total rank values

MODEL	TOTAL RANK
RF	44
RF- RANDOM SEARCH	57
RF -GRID SEARCH	49
KNN	30

Considering the total rank values, the RF model worked better than the KNN model. When the parameters in the RF model are optimized with Random Search and Grid Search methods, it is seen that there is an improvement in model performance. It is also achieved by rank analysis that the Random Search method further improves model performance compared to the Grid Search method. In addition, when the results are examined in detail, it is observed that the combination that estimates Q(t) data by using Q(t-1), Q(t-2) and Q(t-3) flow data as inputs works much better.

5. Result and Discussion

In the study, in which the current flow data of two stations in the Euphrates basin are estimated with the flow data values of one, two, three days ago, it is seen that the RF algorithm gives much better results than the KNN algorithm according to the model results. The fact that the KNN algorithm works with the mean value calculation method and is a simple estimation model, and that the RF model increases the estimation performance through its random processing capability, has led the performance of the RF model to be better. Since model tuning is performed in both model structures (basic RF and KNN), it is seen that the model performances are generally good, and even in the worst-performing cases, it is seen that the R^2 value is 0.889328, the RMSE value is 75.09553 and the MAE value is 26.1493. In addition, it was observed that better results were obtained when the parameters of the RF algorithm were optimized. Considering the two types of parameter optimization methods used, it is seen that the Radom Search method works better than the Grid Search method as it performs random optimization. Moreover, Random Search method can provide the result more quickly as there is no need to make calculations with thousands of trees as in the Grid Search method (Bergstra & Bengio, 2012).

References

Kagoda, P. A., Ndiritu, J., Ntuli, C., & Mwaka, B. (2010). Application of radial basis function neural networks to short-term streamflow forecasting. *Physics and Chemistry of the Earth, Parts A/B/C, 35*(13-14), 571-581.

Al-Abadi, A. M., & Shahid, S. (2016). Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model. *Modeling Earth Systems and Environment, 2*(2), 1-17.

Alexis, B. L., Lazare, K. K., Séraphin, K. K., Alex, K. Z., Félix, K. K., & Bamory, K. (2017). Rain-Flow Modeling Using a Multi-Layer Artificial Neural Network on the Watershed of the Cavally River (Côte d'Ivoire). *Journal of Water Resource and Protection, 9*(12), 1403-1413.

Altunkaynak, A., & Başakın, E. (2018). Zaman Serileri Kullanılarak Nehir Akım Tahmini ve Farklı Yöntemlerle Karşılaştırılması. *Erzincan University Journal of Science and Technology, 11*(1), 92-101. doi:10.18185/erzifbed.339781

Altunkaynak, A, Başakın, E , Kartal, E . (2020). Dalgacık K-En Yakın Komşuluk Yöntemi ile Hava Kirliliği Tahmini. *Uludağ University Journal of The Faculty of Engineering*, 25 (3), 1547-1556 . doı: 10.17482/uumfd.809938

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research, 13*, 281-305.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

DSI . (1981-2010). Akım Gözlem Yıllıkları : https://www.dsi.gov.tr/Sayfa/Detay/744

EIEI. (2000). Akım Gözlem Yıllığı. Ankara: T.C. Elektrik İşleri Etüt İdaresi .

G.H., H. (2014). An insight into extreme learning machines: random neurons, random features and kernels. *Cognitive Computation, 6*(3), 376-390. https://doi.org/10.1007/s12559-014-9255-2

Li, X., Sha, J., & Wang, Z. L. (2019). Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal, 64*(15), 1857-1866. https://doi.org/10.1080/02626667.2019.1680846

M.Chenga, F.Fanga, T.KinouchibI, M.Navonc, & C.C.Paina. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology, 590*, 125376. https://doi.org/10.1016/j.jhydrol.2020.125376

Modares, F., Araghinejad, S., & Ebrahimi, K. (2018). A Comparative Assessment of Artificial Neural Network, Generalized Regression Neural Network, Least-Square Support Vector Regression, and K-Nearest Neighbor Regression for Monthly Streamflow Forecasting in Linear and Nonlinear Conditions. *Water Resources Management volume* , 243-258.

Papacharalampous, G. A., & Tyralis, H. (2018). Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences*, 201-208. https://doi.org/10.5194/adgeo-45-201-2018

Peterson, L. (2009). K-nearest neighbor. *Scholarpedia*, 1883. doi:10.4249/scholarpedia.1883

Okkan, U., & Inan, G. (2015). Statistical downscaling of monthly reservoir inflows for Kemer watershed in Turkey: use of machine learning methods, multiple GCMs and emission scenarios. *International Journal of Climatology*, 35(11), 3274-3295.

Tosunoglu, F., Hanay, Y. S., Cintas, E., & Özyer, B. (2020). Monthly Streamflow Forecasting Using Machine Learning. *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 13*(3), 1242-1251. doi:10.18185/erzifbed.780477

Were, K., TienBui, D., B.Dick, Ø., & RamSingh, B. (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*, 394-403.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K.-W. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical application in water resource engineering area and future research direction. *Journal of Hydrology, 120*, 387-408.

Yenigün, K., & Gümüş, V. (2007). Fırat Havzası Akımlarında Görülen Trendlerin Nedenlerinin Araştırılması. *V. Ulusal Hidroloji Kongresi*, (s. 239-248). Ankara.

Z.M., Y., S.O., S., R.C., D., & K., C. (2019). An enhanced extreme learning machine model for river flow forecasting; State-of-art, practical application in water resources engineering area and future research direction. *Journal of Hydrology, 120*, 387-408.

Zhang, H., Zhou, J., Jahed Armaghani, D., Tahir, M., Pham, B., & Huynh, V. (2020). A Combination of Feature Selection and Random Forest Techniques to Solve a Problem Related to Blast-Induced Ground Vibration. *Appl. Sci., 10 (3)*, 869. doi:https://doi.org/10.3390/app10030869