The European Journal of Research and Development

journals.orclever.com/ejrnd

OR CLEVER Science & Research Group

Research Article

Prediction of Daily Streamflow Data Using Ensemble Learning Models

Levent Latifoğlu[1*], Ümit Canpolat[2*]

[1] ErciyesUniversity, Türkiye, Orcid ID: 0000-0002-2837-3306, E-mail: latifoglu@erciyes.edu.tr
[2] ErciyesUniversity, Türkiye, Orcid ID: 0000-0002-8486-293X, E-mail: canpolat.umit@hotmail.com
[*] Correspondence: latifoglu@erciyes.edu.tr

(First received August 04, 2022 and in final form December 18, 2022)

Reference: Latifoğlu, L., Canpolat Ü. Prediction of Daily Streamflow Data Using Ensemble Learning Models. The European Journal of Research and Development,2(4), 356-371.

Abstract

Estimating river streamflow is a key task for both flood protection and optimal water resource management. The high degree of uncertainty regarding watershed characteristics, hydrological processes, and climatic factors affecting river flows makes streamflow estimation a challenging problem. These reasons, combined with the increasing prevalence of data on streamflow and precipitation, often lead to data-driven models being preferred over physically-based or conceptual forecasting models. The goal of this study is to predict daily river streamflow data with high accuracy using bagging and boosting approaches, which are ensemble learning methods. In addition, the effect of tributary streamflow on the forecast performance was analyzed in the estimation of the streamflow data. According to the results obtained, it has been shown that ensemble learning models are successful in estimating daily streamflow data, and if the tributary streamflow data is also used as input in the estimation of the streamflow, the determination and correlation performance parameters are improved, and the streamflow data can be estimated using tributary streamflow data.

Keywords: *Ensemble Learning, Bagging, Boosting, Streamflow, Forecasting*

1. Introduction

The sustainability of water is endangered day by day due to many negative factors such as global warming, droughts, and unplanned and unconscious water consumption. Making future estimations of water resources ensures planned, efficient and economical use of existing resources in all aspects in the future. Rivers play an important role in the water cycle. For this reason, precise and accurate streamflow forecasting is critical for water resource management, disaster prevention, and the protection of the aquatic environment [1–4]. The physical formation process of streamflow in basins is influenced by various factors such as precipitation, evaporation, topography and human activities,

so the process is very complex and difficult to understand [4]. Improving the performance of predictive models, especially for rivers with streamflow changes, is a difficult problem and studies are ongoing [5]. This is due to the fact that the physical relationships underlying the streamflow process are still not fully understood, have high complexity, are non-linear and non-stationary [6–9]. The relationship between the influence factors of the different periods in the sub-process and the flow will be able to determine the whole characteristic of the system. In most studies, a single data-driven model is often used to model the entire flow process by generalizing the system understanding, but further generalization can have a bad effect on the performance of the prediction model. Data-driven models, which have the potential to achieve high accuracy with low computational cost, have recently been widely used for streamflow estimation to predict hydrological processes without prior knowledge [9–12]. Instead of physically modeling the formation process of the streamflow, data-driven approaches are adopted using time series analysis, regression analysis and machine learning models to predict future streamflow using long-term streamflow data [13]. For data-driven streamflow estimation, the relationship between input and output characteristics of streamflow data is often non-linear. With the analysis of time series models including Autoregressive (AR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARMA) [14–16] and multiple linear regression (MLR) model [17], it is possible to obtain equations showing linear input-output relations.

However, these models typically analyze the relationship between input and output linearly. The nonlinearity of the processes related to the formation of streamflows cannot be explained by these models and therefore the model prediction performance is negatively affected [18]. With advanced regression models, artificial intelligence techniques and machine learning models, the input-output relationship can be made with non-linear relations and offers an alternative powerful approach for streamflow estimation [5–7, 19]. Support vector regression (SVR) [8], artificial neural network (ANN) [9], Bayesian regression (BR) [10], random forest (RF) [11] and long short-term memory network , LSTM) [12][20] have been increasingly used in recent years and show better prediction performance than linear approaches. [22, 24, 31–33]. Although successful prediction results have been obtained by using machine learning methods, recently it has been tried to increase the success of machine learning algorithms by using ensemble models. Ensemble learning, by its nature, consists of various approaches based on different methodologies. Among these, bagging and boosting approaches are widely used [21, 22].

In this study, it is aimed to estimate the streamflow data. Since the streamflow data obtained as a result of the output of a complex process are non-linear and non-stationary, suitable methods for the estimation of these data and their performance have been

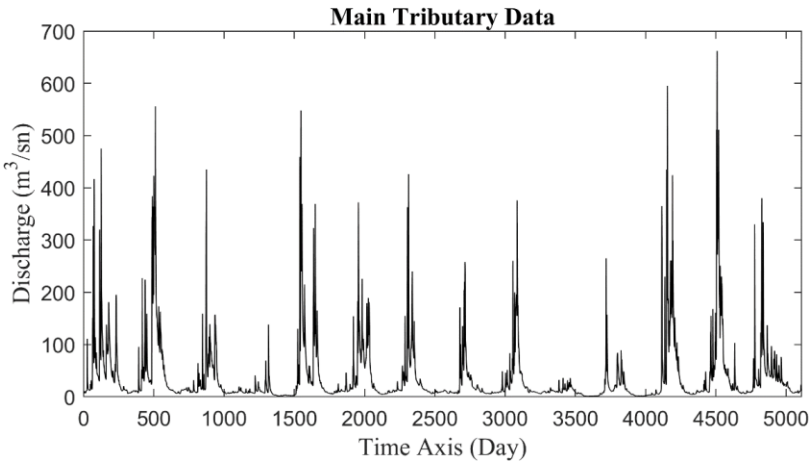
analyzed. Bagging and boosting approaches from Ensemble learning methods have been used in this study.

Daily flow data obtained from the Susurluk basin Simav Stream were used. Streamflow estimation has been made on Simav Stream with tributary streamflow data. In the estimation of streamflow data, the effect of tributary data on estimation performance and the performance of Bagging and Boosting methods were analyzed.

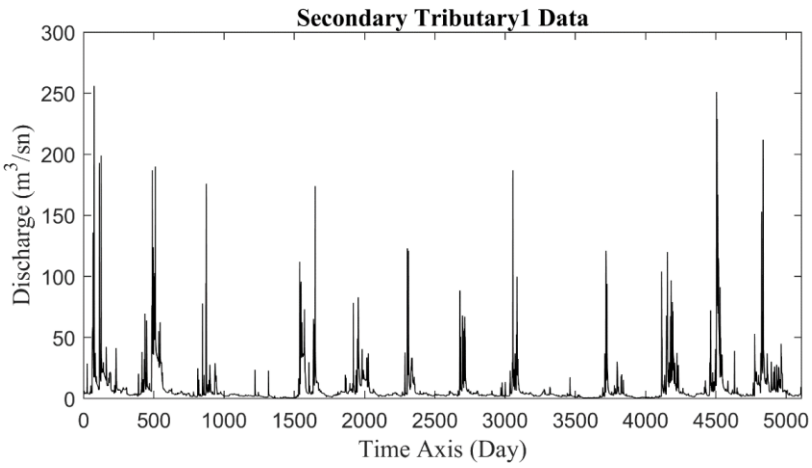
2. Materials and Methods

2.1. Streamflow Data Used in the Forecasting Study

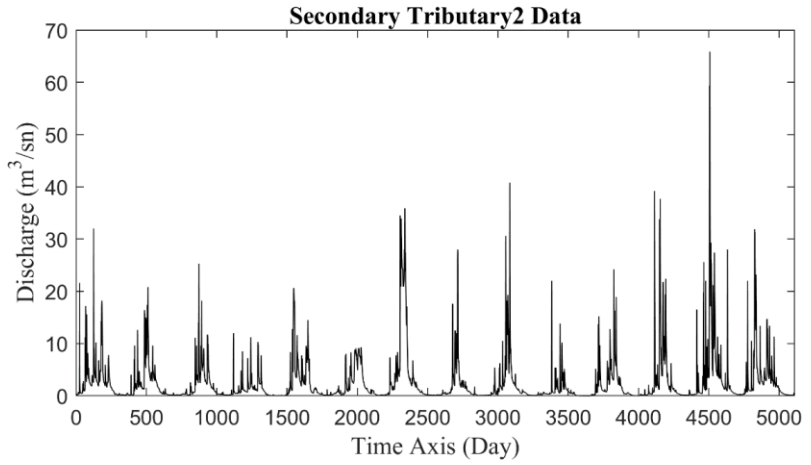
The data recorded from stations 316, 324 and 332 on the Simav Stream in the Susurluk Basin operated by the General Directorate of State Hydraulic Works (Devlet Su İşleri, DSI) in Türkiye, are the 14-year average daily flow data between 1997 and 2011 [23]. The data used in the estimation study is shown in Figure 1. and 3579 of these data were used as training data and 1534 as test data. Statistical values of training and test data are given in Table 1.



a



b



c

Figure 1: Daily streamflow data obtained from hydrological observation stations numbered as a) st 316 b) st 324 c) st 332 used in the estimation study

Table 1: Statistical values of the streamflow data used in the estimation study

TRAINING DATA						
	Maximum (m^3/sn)	Minimum (m^3/sn)	Mean (m^3/sn)	Variance	Skewness	Kurtosis
Main Tributary	556	1.16	35.888	3377.3	3.6885	20.427
Secondary Tributary 1	256	0.083	7.952	234.1	6.9101	72.423
Secondary Tributary 2	40.8	0.006	2.319	17.6	3.8076	20.993

TESTING DATA						
	Maximum (m^3/sn)	Minimum (m^3/sn)	Mean (m^3/sn)	Variance	Skewness	Kurtosis
Main Tributary	662	0.784	39.9601	4959.7	3.9105	22.452
Secondary Tributary 1	251	0.38	9.6310	394.0	6.1378	53.528
Secondary Tributary 2	65.9	0	2.9480	31.1	4.2820	29.678

The time indices of the data used for estimation and the input and output variables used are given in Table 2.

Table 2: Input-output data and Input variables used in the estimation study

Inputs for One Time Lag Model	Input Variables	Output
Main Tributary (Model 1)	Main Tributary(t-1)	Main Tributary(t)
Main Tributary + Secondary Tributary1 (Model 2)	Main Tributary(t-1) + Secondary Tributary1(t-1)	Main Tributary(t)
Main Tributary + Secondary Tributary2 (Model 3)	Main Tributary(t-1) + Secondary Tributary2(t-1)	Main Tributary(t)
Main Tributary + Secondary Tributary1 + Secondary Tributary2 (Model 4)	Main Tributary(t-1)+ Secondary Tributary1(t-1) + Secondary Tributary2(t-1)	Main Tributary(t)
Secondary Tributary1 + Secondary Tributary2 (Model 5)	Secondary Tributary1(t-1) + Secondary Tributary2(t-1)	Main Tributary(t)

2.2. Ensemble Learning Algorithm

The basic logic of ensemble learning methods is based on the principle that decisions taken with many samples give more accurate results than a decision taken from a single sample. The risk of one expert's decision being wrong is greater than the risk that the joint decision of several experts will be wrong. The ensemble learning method is inspired by nature and sociological relationships between humans and has recently been widely used to machine learning methods [24].

In order to obtain better prediction performance than using a single decision tree, ensemble methods have been developed that combine several decision trees. The basic principle in the community model is that a group of weak learners come together to form

a strong learner [25, 26]. Community models are examined in two classes, Bagging (Bootstrap Aggregation) and Boosting.

2.2.1. Bagging Ensemble Learning Method

The bagging ensemble learning method, which is an abbreviation of Bootstrap aggregating, was developed by Breiman [26]. Based on the Bootstrap sampling method, this method is based on training different subsets of the training data set. In this method, different training samples are obtained from the training dataset from the samples that displace each time. Classifiers are trained simultaneously with each sub-training set created. The bagging method uses the majority voting technique to combine the estimates of the classifiers. Bagging process steps are shown in Figure 2.

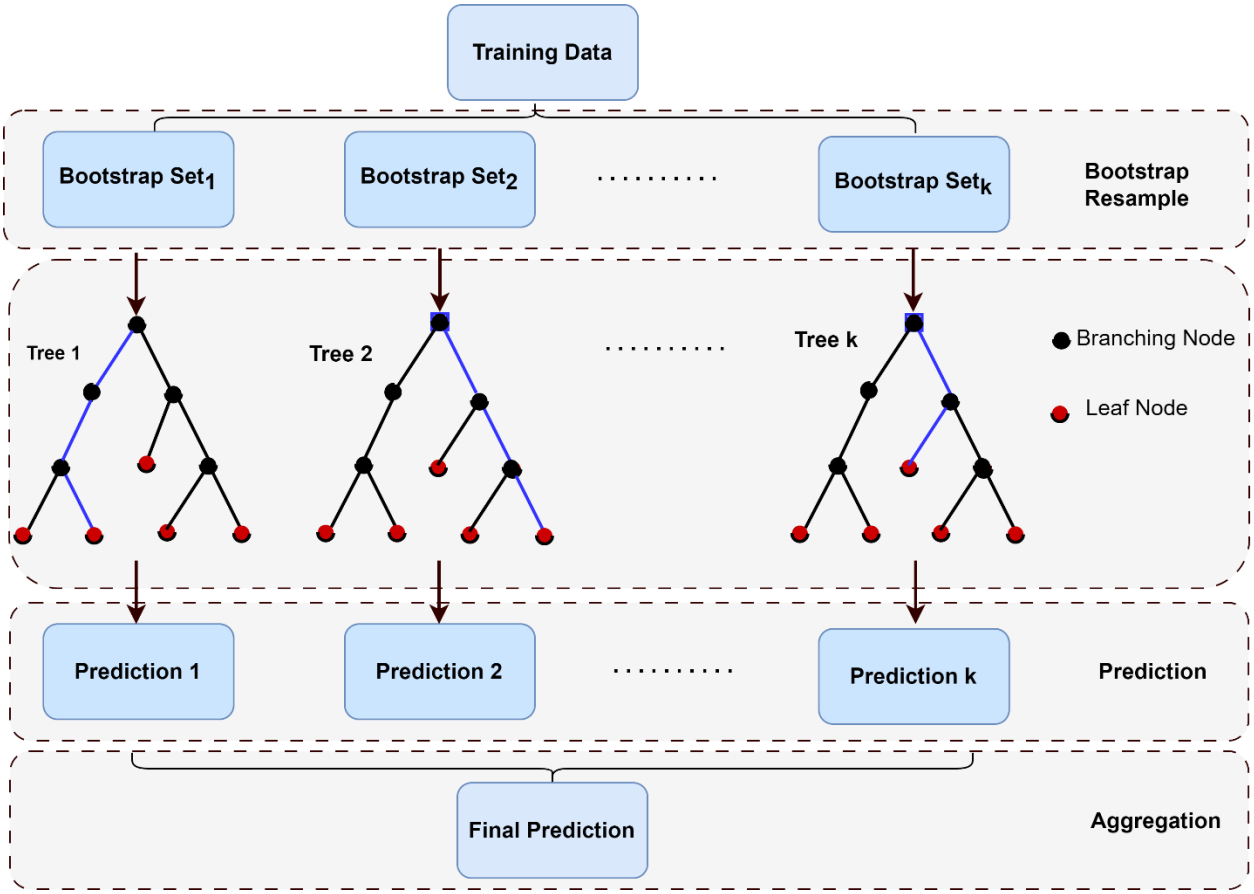


Figure 2: Bagging algorithm steps

In the bagging ensemble learning method, data is trained with a single type of classifier rather than different types of classifiers. The prediction results obtained by simultaneously training different sub-datasets of the same type of classifier as much as the number of Bootstrap sub-sample training datasets are combined. In this study, tree

learner was used in the implementation of the bagging method. In addition, the Number of Learning Cycles and Number of Split parameters have been optimized using the Bayesian Optimization method to obtain the best model with performance in the prediction study.

2.2.2. Boosting Ensemble Learning Method

The Boosting ensemble learning method uses the approach of combining classifiers for weak learners to obtain a stronger classifier than a single classifier. In this method, by combining many weak classifiers, the disadvantages of strong classifiers having a complex structure and being difficult to train are eliminated [25, 27].

While many classifiers are trained simultaneously in the Bagging method, the classifiers are not trained simultaneously by using iterative operations in the Boosting method. A powerful classifier is obtained with the boosting method in order to prevent the same errors from occurring by taking into account the errors in the previous estimation in each iterative process of the boosting method.

Basically, three classifiers are used in the boosting algorithm and each classifier generates predictions. Firstly, the first classifier estimate is obtained, which classifies the randomly selected data set from the training data set. The second classifier is trained on a data set, half of which is misclassified by the first classifier and the other half is correctly classified. The final classifier is trained on data that the two classifiers did not match before.

In the boosting ensemble learning method, the predictions of the classifier that receives the ensemble prediction majority vote are accepted [24, 27]. The process steps of the Boosting ensemble learning method are shown in Figure 3..

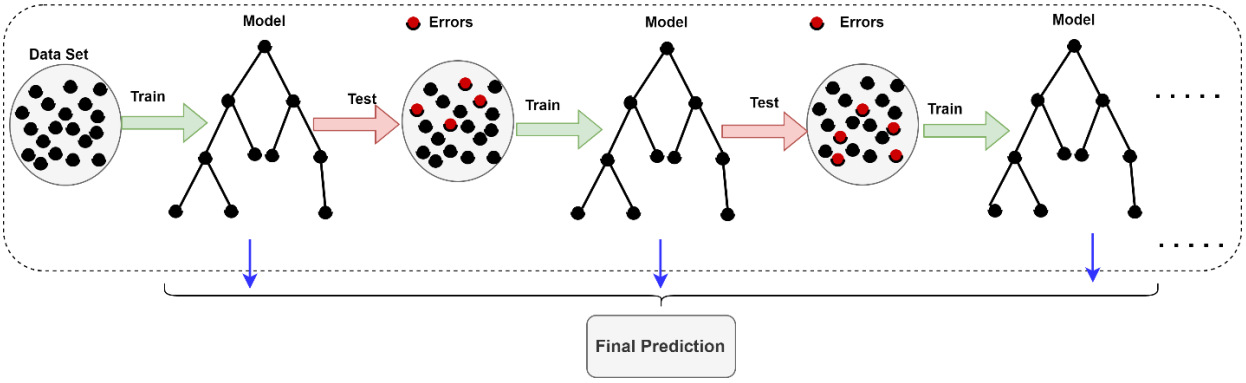


Figure 3: Boosting algorithm steps

In this study, the LSBoost algorithm was used during the implementation of the boosting method. In this algorithm the ensemble adjusts a new learner to the difference

between the observed response and the sum of all previously grown learners at each phase. The mean-squared error is minimized by the ensemble fitting. In addition, as in the bagging method, the number of learning cycles, learning rate and number of split parameters have been optimized using the Bayesian Optimization method in order to obtain the best model in this method..

2.3. Performance Evaluation Parameters

Mean Square Error (MSE), Mean Absolute Error (MAE), Correlation coefficient (R) and Dtermination coefficient (R^2), which are the statistical parameters commonly used in the streamflow estimation study, were used to evaluate the model performance [28].

Mean Squared Error: MSE is a measure of the performance of the prediction model. It is obtained by dividing the sum of the squares of the difference of the observed and predicted values in the data series by the total number of data. Since it is derived from the square of the Euclidean distance, it shows that the estimated value converges strongly to the true as the error approaches zero, and it always takes a positive value [28].

Mean Absolute Error: Mean absolute error is a measure of prediction error commonly used in time series analysis and is defined as the mean of the absolute values of errors between all samples in the observed data set and the predicted data. The fact that the MAE value is close to zero indicates that the result produced by the prediction model is strongly close to the desired value.

Determination coefficient: The certainty coefficient is a measure of the fit of a modeled data with the observed data. It shows that the correlation between the observed and predicted data increases as the R^2 value gets closer to 1 [29].

Correlation Coefficient: Correlation coefficient is the coefficient indicating the direction and magnitude of the relationship between observed and predicted data and takes a value between (-1) and (+1). The positive values of the correlation coefficient indicate the direct linear relationship between the variables; negative values indicate an inverse linear relationship. A correlation coefficient of 0 indicates that there is no linear relationship between these variables [29].

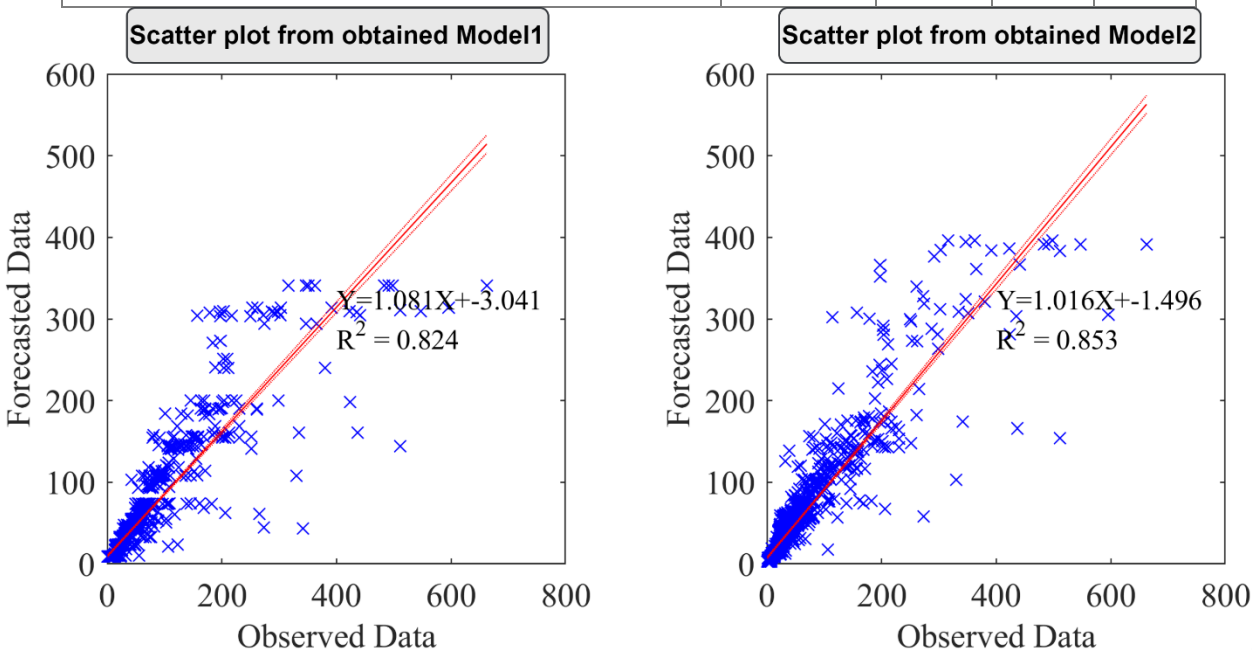
3. Results

In this study, ensemble models were developed on the estimation of the daily streamflow data of the main tributary obtained from station 316 and the secondary tributary streamflow data obtained from stations 324 and 332 on Simav Stream in Susurluk basin. For the estimation of the daily flow data, the main tributary data itself and the secondary tributary data were also included in the forecasting study, and the effect of the secondary tributary data on the forecast performance was analyzed.

The estimation results of the daily streamflow data using the estimation model with the optimized parameters of the Bagging approach are shown in Table 3. The scatter plots between the observed data and the estimated data are shown in Figure 4. In addition, the graph of the forecasted data, in which the main tributary streamflow data is used as input (Model 1), is shown in Figure 5 as an example.

Table 3: Forecasting performance results obtained with the Bagging Method

Inputs for One Time Lag Model	MSE	MAE	R	R^2
Main Tributary	8.4594e+03	50.7477	0.9079	0.8243
Main Tributary + Secondary Tributary1	9.0548e+03	52.6944	0.9236	0.8531
Main Tributary + Secondary Tributary2	8.8197e+03	52.0825	0.9125	0.8326
Main Tributary + Secondary Tributary1 + Secondary Tributary2	9.1235e+03	53.2197	0.9297	0.8643
Secondary Tributary1 + Secondary Tributary2	9.2670e+03	54.8024	0.9006	0.8110



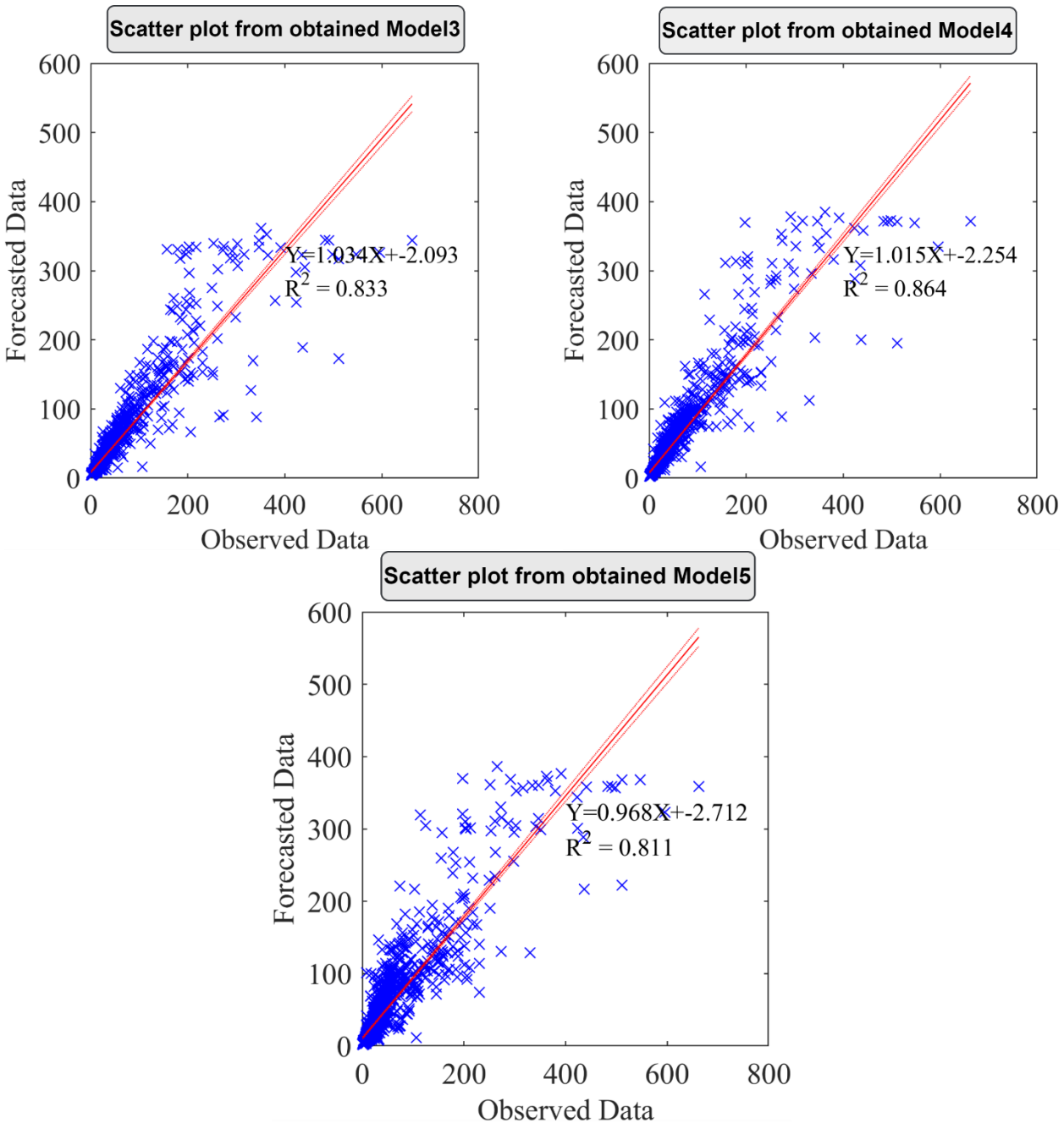


Figure 4: Scatter plots obtained from using Bagging algorithm

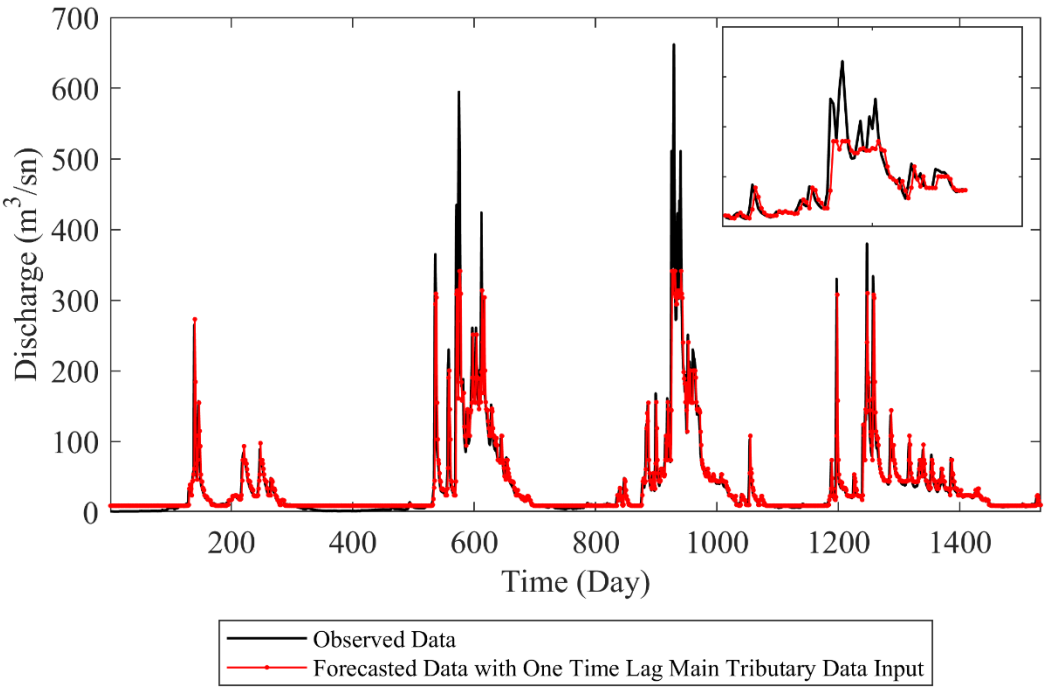
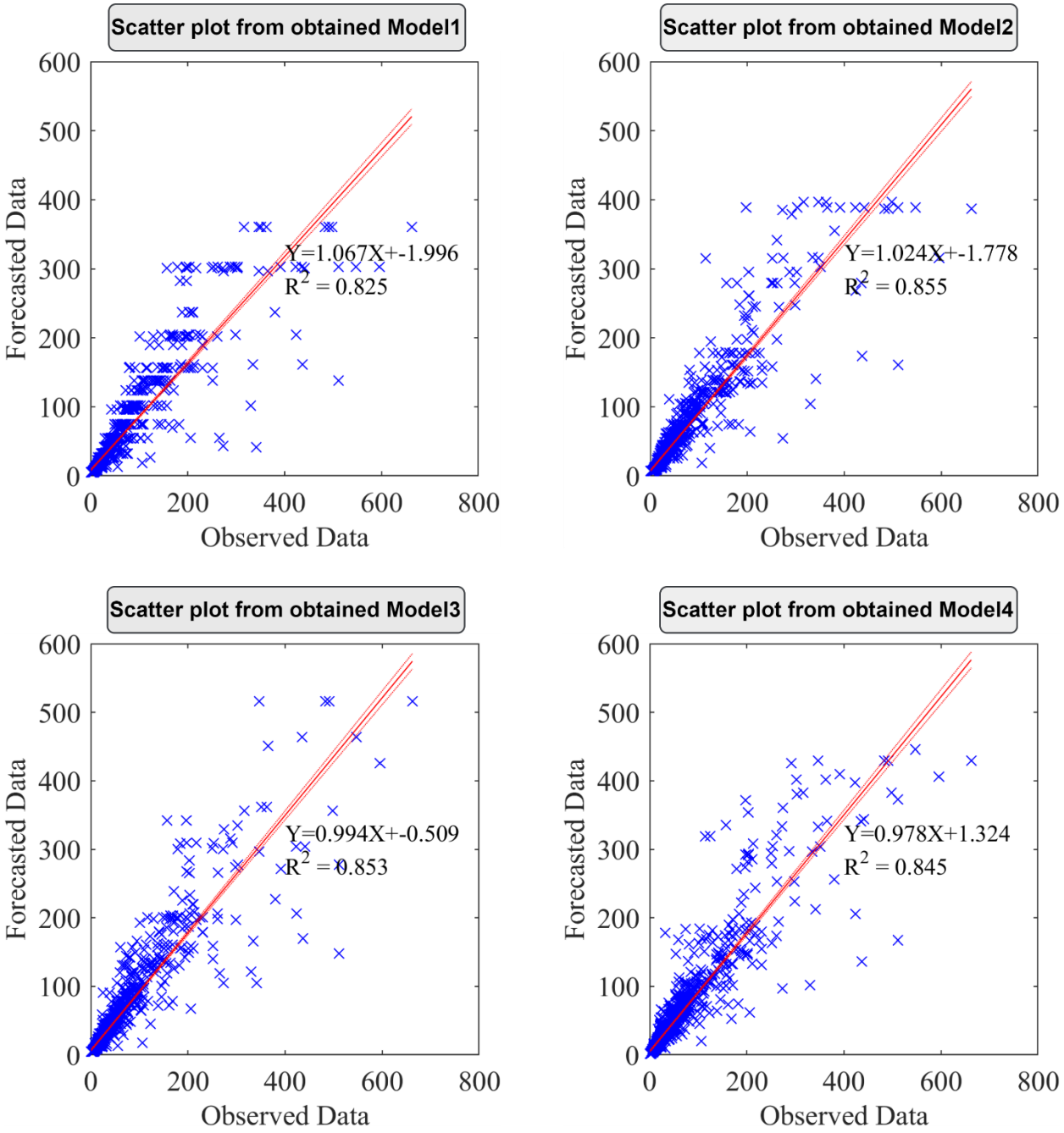


Figure 5: Forecasting graphics obtained from Model1 using Bagging algorithm

The estimation results of the daily flow data by using the estimation model containing the LSBoosting approach with optimized parameters are shown in Table 4. The scatter plots between the observed data and the estimated data are shown in Figure 6. In addition, the graph of the forecasted data, in which the main tributary data is used as input (Model 1), is shown in Figure 7 as an example.

Table 4: Forecast performance results obtained with the Boosting Method

Inputs for One Time Lag Model	MSE	MAE	R	R^2
Main Tributary	8.5549e+03	51.1694	0.9085	0.8254
Main Tributary + Secondary Tributary1	9.0069e+03	51.8523	0.9247	0.8551
Main Tributary + Secondary Tributary2	9.2399e+03	52.5929	0.9235	0.8529
Main Tributary + Secondary Tributary1 + Secondary Tributary2	9.3348e+03	52.7173	0.9191	0.8447
Secondary Tributary1 + Secondary Tributary2	8.9361e+03	53.1081	0.9056	0.8201


Scatter plot from obtained Model1
Y=1.067X+-1.996
R2 = 0.825
Scatter plot from obtained Model2
Y=1.024X+-1.778
R2 = 0.855
Scatter plot from obtained Model3
Y=0.994X+-0.509
R2 = 0.853
Scatter plot from obtained Model4
Y=0.978X+1.324
R2 = 0.845
Forecasted Data
Observed Data
0
100
200
300
400
500
600
800

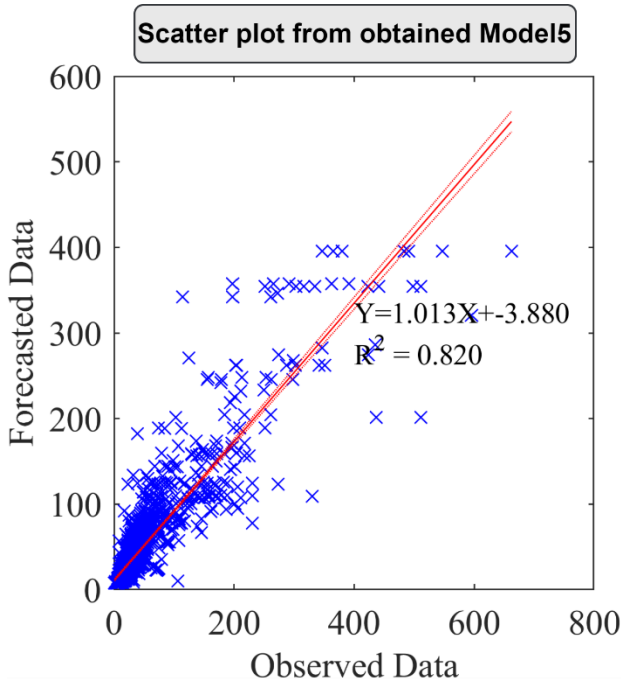


Figure 6: Scatter plots obtained from using Boosting algorithm

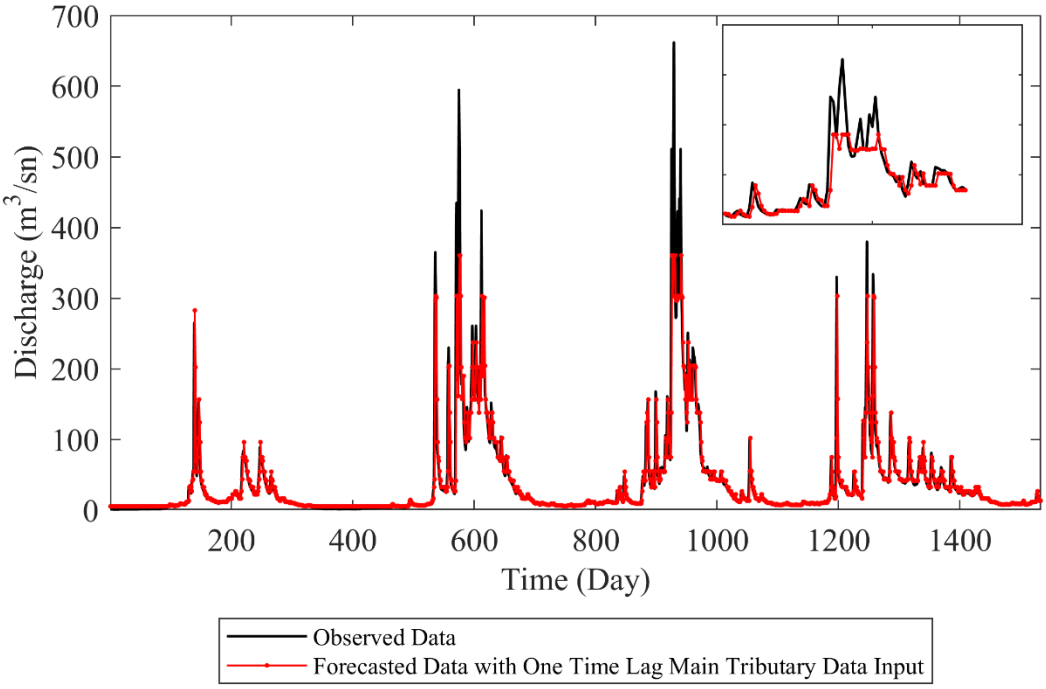


Figure 7: Forecasting graphics obtained from Model1 using Boosting algorithm

4. Discussion and Conclusion

The characteristic model of streamflow cannot be easily predicted due to its high complexity, non-stationary, dynamic and non-linear properties. Modeling and forecasting of streamflow are essential for water resource planning and management, understanding and estimation of suspended sediment load, hydropower generation, design of an irrigation system suitable for plants, and optimum drainage policy from the reservoir. Streamflow forecasting is basically two-fold, real-time forecasting which is crucial for the reliable operation of flood and containment systems, and long-term forecasting important in reservoir operation and planning, hydropower generation, sediment transport, irrigation management decisions and many other applications performed in this category.

When the literature studies are examined, it is stated that standard learning techniques are used in hydrological and river streamflow estimation of ML methods [30, 31]. However, due to their higher efficiency in modelling, the applications of ensemble ML models in hydrological modeling have increased significantly in recent years [24].

In this study, a forward-time estimation of river flow data was performed using the ensemble models Bagging and LSBoosting. Estimation study was carried out by using daily streamflow data obtained from Simav Stream in Susurluk basin. During the establishment of the model, main and secondary tributary (Tributary1 and Tributary2) streamflow data were used separately and together as inputs in the model.

In this study, the model with one-time lag index the highest R^2 value is obtained as 0.8643 using bagging method with Model4. However, the smallest MSE value was 8.4594e+03, which was obtained when only the main tributary data were used.

Also the highest R^2 value is obtained as 0.8551 using boosting method with Model2. However, the smallest MSE value was 8.5549e+03, which was obtained when only the main tributary data were used.

According to the obtained results, it has been seen that the performance of the Bagging method and the performance of the LSBoosting method are close to each other in the estimation of daily streamflow data, but the Bagging method is slightly better. In addition, it has been observed that the R^2 parameter in the models where the main tributary and the secondary tributary streamflow are used together as inputs is better than the models in which other inputs are used. However, the lowest MSE values were obtained only in models that used main strand data as input.

In this study, it has been shown that the main tributary streamflow data can be estimated with high performance using the secondary tributary data. In fact, when the results obtained are examined from the tables, it is seen that the estimation obtained using only the secondary tributary streamflow data is quite close to the estimation performance made with only the main tributary streamflow data.

According to the obtained results, it has been shown that daily streamflow can be predicted with high performance with Bagging and Boosting methods and the main

tributary streamflow estimation can be performed using the secondary tributary streamflow data.

5. Acknowledge

The Erciyes University Scientific Research Projects Unit funded this work under the FBA-2022-11291 project number.

References

[1] Khazaee Poul, A., Shourian, M., & Ebrahimi, H., (2019). A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly streamflow prediction. Water Resources Management, 33(8), 2907–2923.

[2] de Santana Moreira, R. M., & Celeste, A. B., (2017). Performance evaluation of implicit stochastic reservoir operation optimization supported by long-term mean inflow forecast. Stochastic Environmental Research and Risk Assessment, 31(9), 2357–2364.

[3] Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., et al.,(2017). Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the US Southwest. Geophysical Research Letters, 44(24), 12–208.

[4] Hamlet, A. F., Huppert, D., & Lettenmaier, D. P., (2002). Economic value of long-lead streamflow forecasts for Columbia River hydropower. Journal of Water Resources Planning and Management, 128(2), 91–101.

[5] 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 applications in water resource engineering area and future research direction. Journal of Hydrology, 569, 387–408.

[6] Hadi, S. J., & Tombul, M., (2018). Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. Journal of Hydrology, 561, 674–687.

[7] Hadi, S. J., & Tombul, M., (2018). Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study. Water Resources Management, 32(14): 4661–4679.

[8] Toth, E., (2009). Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. Hydrology and Earth System Sciences, 13(9), 1555–1566.

[9] Boucher, M., Quilty, J., & Adamowski, J., (2020). Data assimilation for streamflow forecasting using extreme learning machines and multilayer perceptrons. Water Resources Research, 56(6), e2019WR026226.

[10] Wu, C. L., Chau, K. W., & Li, Y. S., (2009). Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resources Research, 45(8).

[11] Quilty, J., & Adamowski, J., (2020). A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. Environmental Modelling & Software, 130, 104718.

[12] Wu, C. L., & Chau, K.-W., (2010). Data-driven models for monthly streamflow time series prediction. Engineering Applications of Artificial Intelligence, 23(8), 1350–1367.

[13] Wang, W., Van Gelder, P. H., Vrijling, J. K., & Ma, J., (2006). Forecasting daily streamflow using hybrid ANN models. Journal of Hydrology, 324(1–4), 383–399.

[14] Makridakis, S., & Hibon, M., (1997). ARMA models and the Box–Jenkins methodology. Journal of forecasting, 16(3), 147–163.

[15] Kurunç, A., Yürekli, K., & Cevik, O., (2005). Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey. Environmental Modelling & Software, 20(9), 1195–1200.

[16] Musa, J. J., (2013). Stochastic Modeling of Shiroro River Streamflow Process. .

[17] Darlington, R. B., & Hayes, A. F., (2016). Regression Analysis and Linear Models: Concepts, Applications, and Implementation. Guilford Publications.

[18] Tian, P., Lu, H., Feng, W., Guan, Y., et al., (2020). Large decrease in streamflow and sediment load of Qinghai–Tibetan Plateau driven by future climate change: A case study in Lhasa River Basin. Catena, 187, 104340.

[19] Chu, H., Wei, J., Wu, W., Jiang, Y., et al., (2021). A classification-based deep belief networks model framework for daily streamflow forecasting. Journal of Hydrology, 595 (January), 125967. (https://doi.org/10.1016/j.jhydrol.2021.125967)

[20] Jothiprakash, V., & Magar, R. B., (2012). Multi-time-step ahead daily and hourly intermittent reservoir inflow prediction by artificial intelligent techniques using lumped and distributed data. Journal of hydrology, 450, 293–307.

[21] Dietterich, T. G., (2002). Ensemble learning. The handbook of brain theory and neural networks, 2(1): 110–125.

[22] Zhou, Z.-H., (2009). When semi-supervised learning meets ensemble learning. International Workshop on Multiple Classifier Systems, 529–538, Springer.

[23] DSİ Akım Gözlem Yıllıkları. https://www.dsi.gov.tr/Sayfa/Detay/744.

[24] Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R., (2021). Ensemble machine learning paradigms in hydrology: A review. Journal of Hydrology, 598, 126266. (https://doi.org/10.1016/j.jhydrol.2021.126266)

[25] Opitz, D., & Maclin, R., (1999). Popular ensemble methods: An empirical study. Journal of artificial intelligence research, 11, 169–198.

[26] Collins, R., (2018). Machine Learning with Bagging and Boosting. Amazon Digital Services LLC - Kdp Print Us. Retrieved from https://books.google.com.tr/books?id=Ch23vAEACAAJ

[27] Bühlmann, P., (2012). Bagging, boosting and ensemble methods. In Handbook of computational statistics, 985–1022, Springer.

[28] Willmott, C. J., (1982). Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society, 63(11), 1309–1313.

[29] Chicco, D., Warrens, M. J., & Jurman, G., (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7: e623.

[30] Kim, S., Alizamir, M., Kim, N. W., & Kisi, O., (2020). Bayesian model averaging: A unique model enhancing forecasting accuracy for daily streamflow based on different antecedent time series. Sustainability (Switzerland), 12(22), 1–22. (https://doi.org/10.3390/su12229720)

[31] Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., et al., (2019). Daily streamflow prediction using optimally pruned extreme learning machine. Journal of Hydrology, 577, 123981.