1. INTRODUCTION

As long as life exists on Earth, water will continue to be both a solution and a problem for living organisms. One of the tasks of hydrology science is to offer solution options and remove problems as much as possible. Of course, the best use of water skills that human beings have acquired from history has an impact on the development of this science. Although the hydrological cycle causes water to return to the surface, the movements of water can be predicted and optimized. Natural events are affected by environmental conditions. These effects occur due to human interference in river directions or as a result of waste. Due to these effects, predicting irregular river flow becomes more difficult. Especially in countries with high sector diversity due to industrialization and urbanization activities caused by population growth, the human factor affects the hydrological balance of local basins [1,2].

Today, although water resources seem sufficient for living beings, the water crisis is becoming a problem worldwide due to the effects of climate change and increasing population density [3,4,5,6]. There is approximately 1384.12 million km³ of water in the world. Of this water, 1348 million km³, which is over approximately 97%, constitutes the oceans, while 27.82 million km³ of it consists of glaciers and 8.062 million km³ of it consists of underground waters [7]. The accessible surface waters for living beings are around 0.0225 million km³. In addition, 0.013 million km³ of water is found in the atmosphere as water vapor [8].

884 million people in the world cannot access clean water. Moreover, more than 2 million people are deprived of quality drinking water [9].

In Turkey, an average of 574 billion m3 of precipitation falls annually, against an average of 450 billion m3 of water [10,11]. The amount of water that evaporates from soil and water surfaces is around 274 billion m3. In addition, 158 billion m3 of water flows from our rivers. 7 billion m3 of water comes to our country from neighboring countries through rivers. Although the annual usable water potential of our country is 112 billion m3, only 95 billion m3 of it can be used efficiently [12]. The usable surface water is estimated to be a total of 98 billion m3 annually, while the potential for groundwater is estimated to be 14 billion m3 [13]. There are 25 water basins in our country, and their existing water potentials are not distributed equally for each region. Also, each basin has different dynamics and problems. For example, the Büyük Menderes and Ergene basins are facing pollution problems [14].

The average annual water per capita in Turkey was 1652 m³ in 2000, 1544 m³ in 2009, and 1346 m³ in 2020, which is decreasing, and this value clearly shows that Turkey is not a water-rich country in terms of available resources, but rather a country experiencing water stress [15,16].

Based on the data, it is necessary to work towards increasing the efficient use of available water resources in the world and our country. Maximum benefit should be obtained from water resources for living beings. By examining the dynamic structure of water, vital recommendations can be made for urban water management by predicting future water potential.

In studies related to water resources, meaningful results can be obtained by examining water and atmospheric data (ground and surface flow, wind, humidity, temperature, precipitation, etc.) with different comparisons. Researchers have also used different mathematical models to estimate groundwater levels in recent years [17,18]. These models are designed to solve the difficulties of predicting water levels due to factors such as

Additionally, researchers have been using different mathematical models to predict groundwater levels in recent years [17,18]. These models are created to address the difficulties of predicting water levels due to factors such as multiple input and structural parameters, various calibrations, and so on [19,20].

Time series are numerical values ​​of different variable values ​​repeated within a certain period in time. The values ​​in the time series show interrelated changes. It differs from other series data with this feature [21]. From the values ​​in the series, the most meaningful information is extracted and future values ​​are predicted. Analyses can be done in the time-frequency domain [22]. Regression and time series methods, as well as wavelet models, can be used to determine events that occur at certain intervals in past years by combining data from previous years in the same regions. Local and temporal forward prediction results can be obtained. Information can be obtained on where and how water should be used as a result of these predictions. Areas and ways of use that will provide suitable and efficient use of water can be determined [23,24]. There are numerous research studies that use regression models to predict water resources and hydrological data. In one study, flow data was attempted to be estimated using regression models and flow data in the Euphrates basin. As a result, good results were obtained with both methods, but they stated that regression models gave more successful results [25]. Regression analysis is also commonly used in completing missing data [26, 27]. Other statistical or developed models based on optimization algorithms such as SVR and DT used to predict groundwater levels have high success rates in predicting groundwater levels [28, 29,30]. The RF algorithm, a machine learning model, is also frequently used in water resources prediction studies due to its simplicity and speed with high prediction performance [31,32,33]. Data preprocessing techniques can be used to increase the performance of predictions made using machine learning methods. WT is one of these techniques. The wavelet transformation technique was first applied to precipitation data by Kumar and Foufoula-Georgiou (1993), and it is also used in many studies in river-stream flow analysis, yielding more successful results [34,35]. The wavelet transformation simultaneously analyzes time series in both frequency and time domains. A method that is one of the best for non-stationary time series, wavelet transformation allows the data to be divided into different frequency bands, divided into sections in time, and thus enables its own frequency content to be analyzed. In the study presented here, groundwater flow prediction was made by comparing the performance of LR, SVR, DT, RF, and ETR algorithms. The Tunceli Munzur water sub-basin of the Upper Euphrates basin and the Beşdeğirmen rivers in the Kütahya Porsuk River Basin were selected for the study. The daily average flow, daily total precipitation for the region, and daily average air temperature parameters were used to predict flow rates. Method results were compared by creating different combinations of input parameters. In addition, a hybrid method was developed for each algorithm using Wavelet analysis to improve the prediction performance of the methods.

2. Material and Experimental Method

2.1 Study Area

The study utilized flow data from the Melekbahçe stream flow observation station, E21A033, in the Fırat Basin and the Beşdeğirmen stream flow observation station, E12A003, in the Porsuk Creek sub-basin of the Sakarya Basin. The daily average flow data from the Melekbahçe station between 1997 and 2010 and the daily average flow data from the Beşdeğirmen station between 1980 and 2011 were used. The precipitation and temperature data between 1997 and 2010 for the Tunceli station, number 17165, and between 1980 and 2011 for the Kütahya station, number 17155, were also used for the analysis.

Figure 1. Study area (E21A033) location of Melekbahçe stream flow observation station [36].

The Beşdeğirmen station is located near the Kütahya-Eskişehir highway, approximately 17 km from Kütahya.

Figure 2. Study area (E12A033) location of Porsuk Creek Beşdeğirmen stream flow observation station [37].

2.2 Methodology

The flowchart in Figure 3 presents the steps followed in the study.

Figure 3. Flowchart

As shown in Figure 3, not only the flow data from the observation stations but also daily climate data were analyzed. The flow data were analyzed together with precipitation and temperature data to estimate flow velocity. The first step was to apply pre-processing to correct erroneous or missing data. The data were then normalized to prevent outliers and improve the performance of the model. In this study, the data were normalized to the [-1,1] range. Descriptive statistical analyses were used to identify relationships between the data, and these analyses were also used to interpret the results. Next, the data were analyzed using five different machine learning algorithms. To improve the prediction performance, a hybrid method was developed using the Wavelet Transform method, and the data were re-analyzed. Finally, the prediction performances were compared, and the results of the river flow prediction for the future were presented.

2.3 Method

2.3.1 Machine Learning Algorithms

Machine learning is a subcategory of artificial intelligence that involves obtaining information and developing calculation methods from raw or processed data, continually learning, and making predictions based on this data, and then creating special models [38]. Machine learning methods can be used to develop prediction methods and models using past data.

2.3.2.1 Linear Regression

Linear regression is a statistical method that examines the linear relationship between two variables. In machine learning, the linear model is calculated as Equation 1. The independent variable(s) (x,...) are expressed with the dependent variable (y) [41].

y′=b+W1X1 (1)

y′ is the predicted label (desired output). b is the bias (the y-intercept), w1 is the weight of feature 1. Weight is the same concept as "slope." x1 is a feature (known input).

2.3.2.2 Support Vector Regression

SVR is a type of universal feedforward neural network that uses the principle of structural risk minimization to minimize not only the experimental measurement error but also the upper bound of the generalization error when compared to traditional supervised learning methods of neural networks [42].

2.3.2.3 Decision Tree

Decision trees are the most commonly used machine learning method in prediction and classification solutions due to their easy-to-understand structure. The decision output is obtained by

2.2. Methodology

In this study, the flow chart presented in Figure 3 was used. As seen in Figure 3, not only the flow data from the observation station but also the daily climate data were analyzed. In order to estimate flow velocity, precipitation and temperature data were analyzed together with the data obtained from the flow observation station. The first process after obtaining the data was preprocessing to correct incomplete or erroneous data. Then, the data was normalized. Normalization is a process that prevents excessive emissions of data and improves the performance of the model. In this study, the data was normalized to the range [-1,1]. After data preprocessing, descriptive statistical analyzes were used to find relationships between the data. Descriptive statistical analyzes were also used when interpreting the results. In the next step, the data was analyzed using five different machine learning algorithms. A hybrid method was developed using the Wavelet Transform method to improve prediction performance and the data was reanalyzed. In the final step, prediction performances were compared, and river flow prediction results for the future were explained.

2.3. Method

2.3.1. Machine Learning Algorithms

Machine Learning is a subcategory of Artificial Intelligence that obtains information from raw or processed data, develops calculation methods, continuously learns, and makes predictions based on this data, and then creates special models [38]. With machine learning methods, appropriate models can be created and prediction methods can be developed by using past data [39,40].

2.3.2.1. Linear Regression

The statistical method that examines the linear relationship between two variables is called linear correlation and simple linear regression. The linear model in machine learning is calculated as Equation 1. The independent variable(s) (x,..) is expressed with the dependent variable (y) [41].

y′=b+W1X1 (1)

y' is the predicted label (desired output). b is the bias (the y-intercept), w1 is the weight of feature 1. Weight is the same concept as "slope". x1 is a feature (known input).

2.3.2.2. Support Vector Regression

SVR, a type of universal feedforward neural network, uses the principle of structural risk minimization, which aims to minimize not only the experimental measurement error (empirical risk minimization) but also the upper bound of the generalization error, compared to traditional supervised learning methods of neural networks [42].

2.3.2.3. Decision Tree

Decision Trees are the most commonly used machine learning method in prediction and classification solutions due to their ease of understanding. The decision output is obtained by dividing the input data into groups with clustering algorithms and repeating them [43].

2.3.2.3. Random Forest

Random Forest is a supervised learning algorithm consisting of trained decision trees. A random forest consisting of multiple decision trees can be used to solve classification and regression problems [43].

2.3.2.4. Extra Trees

Extra Trees have a similar structure to Random Forest. The classification decision is made based on the majority in the tree group. Random cut points are selected to separate the nodes. Extra trees use the entire original sample as input data [44].

2.3.2.5. Wavelet Transform

It is a frequently used model to observe and identify signals in the time-frequency dimension. By examining the change of non-consecutive and stationary signals, wavelet transformation decomposes the time series into its subcomponents by breaking it down into parts and analyzing the prominent repetitions of the time series and how these repetitions change over time [44].

In the study, the original data for average flow velocity, total precipitation, and average temperature were separated into their sub-components using wavelet transformation analysis. A hybrid method was developed

This is a research article on machine learning applications for predicting streamflow rates using daily average flow, daily average temperature, and daily total rainfall data from two streamflow observation stations in Turkey. The study utilized five different machine learning algorithms and compared their results. The Random Forest and Extra Trees models performed the best in terms of predicting streamflow rates. The study also employed the Wavelet transform to improve the accuracy of the machine learning models and achieved better results. The research suggests that machine learning and Wavelet transform techniques can be useful in predicting streamflow rates and improving the accuracy of the predictions.

In this study, daily mean flow, daily total precipitation, and daily mean air temperature data for Porsuk Çayı Beşdeğirmen and Melekbahçe flow observation stations belonging to the Porsuk Creek sub-basin of the Sakarya Basin and the Munzur sub-basin of the Upper Euphrates Basin, respectively, were used between October 1, 1979 and September 30, 2011. A total of 11,688 daily data values were used, and the statistical values are presented in Table 1 for Beşdeğirmen and Table 2 for Melekbahçe. Among the five different machine learning approaches used to analyze the data, the ETR and RF methods were found to have the highest r2\_test values for Beşdeğirmen, with 88% and 85.8%, respectively. For Melekbahçe, the RF, ETR, and SVR algorithms were the most successful, with success rates of 70.9%, and 68.6%, respectively, based on r2\_test results. Hybrid models were developed by applying wavelet transformation to improve the results of the machine learning algorithms used in the analysis. The comparative results for Beşdeğirmen are shown in Table 5 and Figure 6, and the three best-performing methods in the DT analysis were LR, ETR, and RF, with r2\_test values of 97.4, 97.7, and 95.3, respectively. The results for Melekbahçe are shown in Table 6 and Figure 7, and the two most successful methods based on r2\_test results were the RF and ETR algorithms, with success rates of 85.2% and 84.8%, respectively.

In conclusion, a model was attempted to be created to estimate the flow rate of rivers using data obtained from observation stations on two rivers in Turkey. Daily flow, temperature, and precipitation data were tested using various machine learning algorithms, and more successful methods were obtained by creating hybrid models using the wavelet model. The results of five different machine learning algorithms were also compared. The random forest and extra tree models yielded higher rates, and it was observed that the machine learning and wavelet transformation method had an effective impact on the results and increased the model's success.

Wavelet transformation was found to be effective in explaining and visualizing trends on time series and irregular data. The percentage success rates of D-machine learning models were at least 10% closer to the success rates of the W-machine learning models. The results obtained in this study can be used in hydrological modeling and can also be used for flood control and water management purposes.