Çukurova Tarım Gıda Bil. Der. Çukurova J. Agric. Food Sci.
39(2) : 401-416
doi: 10.36846/CJAFS.2024.160

Research Article

Estimating Daily Streamflow Values Using Artificial Neural Networks, Support Vector Regression and Multiple Linear Regression Models for Ceyhan River Basin

Ahmet İRVEM[1*], Mustafa ÖZBULDU[1]

ABSTRACT
Streamflow data are very important for effective planning and management of water resources in basins. In this study, Artificial Neural Networks (ANN), Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models were developed to estimate the daily streamflow of three different rivers in the Ceyhan River Basin. Daily precipitation and temperature data obtained from The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) re-analysis data were used as predictor variables in the models. The estimation performances of the models were evaluated with different statistical performance measures. According to the evaluation results, the SVR model demonstrated the best performance in daily streamflow estimation for the Ceyhan River, achieving R^2 = 0.95 and RMSE = 28.20 m^3 s^{-1}. Additionally, for Söğütlü Creek, the results were R^2 = 0.82 and RMSE = 6.57 m^3 s^{-1}, while for Keşiş Creek, R^2 = 0.93 and RMSE = 1.45 m^3 s^{-1} were obtained. The findings indicate that the SVR model predicts daily streamflow more successfully than the other models. Furthermore, it was found that the performance of the models developed using machine learning algorithms was superior to that of the linear regression model.
Keywords: Artificial neural networks, support vector regression, MATLAB, streamflow estimation, MERRA-2

Ceyhan Nehir Havzası için Yapay Sinir Ağları, Destek Vektör Regresyonu ve Çoklu Doğrusal Regresyon Modelleri Kullanılarak Günlük Akış Değerlerinin Tahmini

ÖZ
Akış verileri, havzalardaki su kaynaklarının etkin planlanması ve yönetilebilmesi için oldukça önemlidir. Bu çalışmada, Ceyhan Nehri Havzası'nda bulunan üç farklı akarsuyun günlük akışını tahmin etmek için Yapay Sinir Ağları (YSA), Destek Vektör Regresyonu (DVR) ve Çoklu Doğrusal Regresyon (ÇDR) modelleri oluşturulmuştur. Modellerde tahmin edici değişkenler olarak Araştırma ve Uygulamalar için Modern Çağ Retrospektif Analizi, sürüm 2 (MERRA-2) re-analiz verilerinden elde edilen günlük yağış ve sıcaklık verileri kullanılmıştır. Modellerin tahmin performansları farklı istatistiksel performans ölçütleri ile değerlendirilmiştir. Değerlendirme sonuçlarına göre, DVR modelinin Ceyhan Nehri için günlük akış tahmininde R^2=0.95 ve RMSE=28.20 m^3 s^{-1} değerleri ile en başarılı performansı gösterdiği belirlenmiştir. Ayrıca, Söğütlü Çayı için R^2=0.82 ve RMSE=6.57 m^3 s^{-1}, Keşiş Dere için ise R^2=0.93 ve RMSE=1.45 m^3 s^{-1} sonuçları elde edilmiştir. Elde edilen bulgular, DVR modelinin günlük akarsu akışını diğer modellere göre daha başarılı bir şekilde tahmin ettiği belirlenmiştir. Ayrıca makine öğrenmesi algoritmaları kullanılarak oluşturulan modellerin performansının doğrusal regresyon modeline göre daha üstün olduğu tespit edilmiştir.
Anahtar Kelimeler: Yapay sinir ağları, destek vektör regresyonu, MATLAB, akış tahmini, MERRA-2

ORCID ID (Yazar sırasına göre)
0000-0002-3838-1924, 0000-0002-5359-8750

Yayın Kuruluna Geliş Tarihi: 14.05.2024
Kabul Tarihi: 04.11.2024
[1]Department of Biosystems Engineering, Hatay Mustafa Kemal University, Hatay, Türkiye
*E-posta: airvem@mku.edu.tr

Estimating Daily Streamflow Values Using Artificial Neural Networks, Support Vector Regression and Multiple Linear Regression Models for Ceyhan River Basin

Introduction

The phenomenon of climate change is anticipated to increase both the frequency and intensity of natural disasters, including floods and droughts. It is therefore of critical importance to develop and implement accurate hydrological models in order to mitigate the potential impacts of such disasters. (Tongal and Booji, 2018; Sauquet et al., 2021; Wang et al., 2019). The analysis of streamflow data plays a pivotal role in the field of hydrological modelling, particularly in the context of flood planning, the effective management of water resources and the comprehensive planning of river basins (Bhadrecha et al., 2016; Meng et al., 2019). However, many regions suffer from inadequate streamflow observation networks, leading to challenges in obtaining accurate and long-term streamflow data for effective modeling (Kumar and Sen, 2018; Shrestha et al., 2021). As such, it is crucial to develop new techniques to obtain and utilize streamflow data effectively in hydrological modeling, particularly in regions with limited resources.

Hydrologists have long been developing hydrological models to estimate streamflow, but the complexity of the hydrological cycle and the inherent differences in basin characteristics make it challenging to accurately estimate streamflow (Peng et al., 2017). As a consequence of this, there has been a marked increase in interest in the development of data-driven hydrological models for estimating runoff, with a view to utilizing available data. Data-driven models have been shown to be more practical and simpler to use compared to models that extensively examine hydrological processes (Masselot et al., 2016; Noori and Kalın, 2016). Consequently, the application of machine learning (ML) algorithms has become increasingly prevalent in hydrological studies, including the estimation of precipitation, the assessment of water quality, and the modelling of groundwater. (Cheng et al., 2015; Hosseini and Mahjouri, 2016; Young et al., 2017; Khatri et al., 2020). However, it is essential to consider the limitations of these models as well, such as their inability to account for complex physical processes and the need for adequate training data. Hydrological models need to be appropriately applied and validated with field observations to ensure their reliability and applicability in hydrological research and practice.

In recent times, ML algorithms, including artificial neural networks (ANNs), support vector regression (SVR), fuzzy logic and wavelet transform, have become increasingly popular in the field of streamflow estimation. This is due to their ability to provide accurate results that are comparable to those of physical process-based models. Among the algorithms available for this purpose, ANNs have been employed with considerable success for the estimation of streamflow in a variety of hydrological studies. Riad et al. (2004) utilized ANNs to model the precipitation-flow relationship in a semi-arid basin. Their findings reported that the ANN model demonstrated superior performance compared to the classical regression model in streamflow estimation. Similarly, Mehr et al. (2015) evaluated a backpropagation ANN model for streamflow estimation in a poorly observed basin. Their findings indicated that the ANN model has the potential to be an effective alternative for monthly flow estimation. Nevertheless, studies have indicated that SVR can achieve superior simulation accuracy in hydrological applications in comparison to ANNs, as reported by Tongal and Booji (2018) and Parisouj et al. (2020). Furthermore, previous research has indicated that ANNs and SVRs are the most preferred ML models for streamflow estimation, in comparison to other algorithms (Wang et al., 2019). In light of the aforementioned findings, SVR and ANN models were preferred as the most appropriate ML algorithms for the estimation of streamflow in the study area, given their prevalence and reliability.

Streamflow estimation models, including SVR and ANN, have been employed in a range of studies across diverse fields, including agriculture, water resources management, and climate change. In agriculture, streamflow estimation models can be used to estimate water availability for irrigation purposes. By accurately estimating streamflow, farmers can make informed decisions about when and how much water to use for irrigation, optimizing crop yield and water use efficiency. Moreover, flow simulation models can assist in the administration of water resources by furnishing data on the availability of water and the potential for drought conditions. In the Awash River Basin in Ethiopia, SVR models have been used to estimate short-term drought conditions, enabling proactive water management strategies (Belayneh et al., 2015). In water resources management, streamflow estimation models are essential for planning and decision-making. These models can provide valuable information on water availability, reservoir management, and flood control. For instance, in the Karkheh Basin in Iran, ML models like ANN and SVR were used to estimate streamflow and analyze flood hazards (Kamali et al., 2022). These models can help in designing and implementing effective water management strategies, ensuring sustainable water supply and minimizing the risk of floods.

Accurately estimating daily streamflow is crucial for successful decisions on water allocation, reservoir management, flood control and irrigation planning. In recent years, the application of ANN and SVR in particular has demonstrated considerable potential for enhancing the accuracy of daily streamflow estimations. (Adamowski and Sun 2010; Kisi et al. 2013; Fotovatikhah et al. 2018; Rauf et al. 2018; Siddiqi et al. 2021). However, studies that compared the performance of different data-driven models revealed that the success rate of the models may vary according to region (Chau and Li, 2009; Tongal and Berndtsson, 2016; Bafitlhile and Li, 2019).

Consequently, it is of paramount importance to ascertain the most efficacious modelling method for the study area in order to achieve more accurate daily streamflow estimation. Furthermore, no previous study has been conducted on the utilization of ML algorithms for the estimation of daily streamflow in the Ceyhan River Basin. In order to achieve this objective, the performance of various models, including ANN, SVR, and MLR models, were evaluated in terms of their ability to estimate daily streamflow using precipitation and temperature data from the Ceyhan River Basin. The findings of this study will address an important research gap in the study area regarding the use of ML methods for streamflow estimation and the selection of the most suitable method to improve accuracy. Additionally, the objective is to contribute to the improvement of water management and decision-making processes in the region by estimating streamflow data.

Material and Method

Study area and data

The models employed for streamflow estimation were applied to three streams within the Ceyhan Basin in Turkey. The Ceyhan Basin is characterized by a climate that is broadly similar to the Mediterranean climate. In accordance with the Köppen classification, the climate is characterized by high temperatures and low humidity during the summer months, with cooler, more humid conditions prevailing during the winter season. The quantity of precipitation varies along the basin as a consequence of altitude differences. The Ministry of Agriculture and Forestry of the Republic of Turkey (TOB) has reported that the average annual total precipitation in the Ceyhan basin is 727.3 mm, with an average annual temperature of 14.8 °C. The total precipitation area of the basin is 21,391 km^2, and the annual total surface water potential is 7,833 hm^3 (TOB, 2019). The streamflow data utilized in the models was sourced from three observation stations of the State Hydraulic Works situated within the basin. The locations of the streamflow observation stations are shown in Figure 1 and Table 1. The stations were selected due to their long-standing records of annual flow measurement. The daily streamflow data for these stations were obtained from State Hydraulic Works for the period 1988-2015. (DSI, 2022).

Table 1. The features of stations used for the observation of streamflow

Stations	Mean annual streamflow ($m^3 s^{-1}$)	Latitude	Longitude
Ceyhan River-Misis	190,176	36°57'28" N	35°38'03" E
Söğütlü Creek- Hanköy	3,741	38°15'20" N	37°32'01" E
Keşiş Creek- Sarıdanışmanlı	10,597	37°28'24" N	36°15'07" E

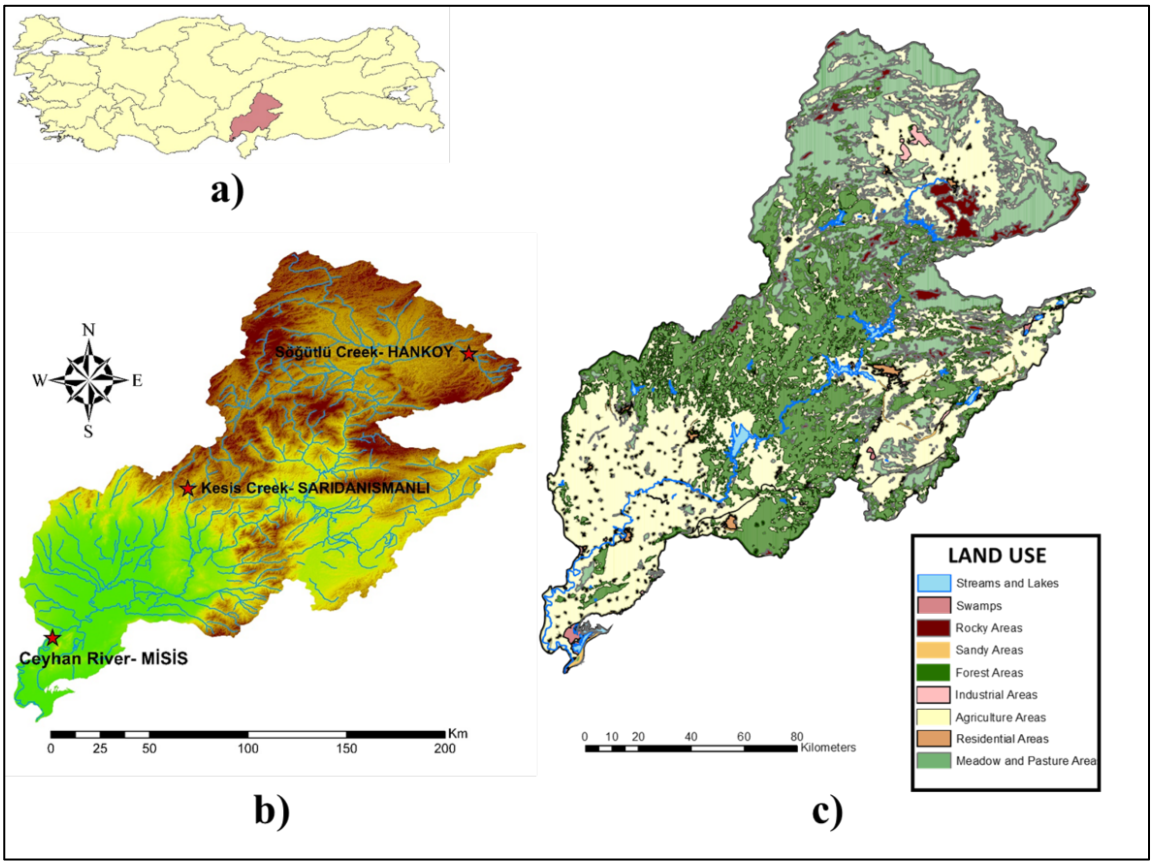

Figure 1. a) location of the study area, b) location of observation stations, c) land use map of the study area (Corine, 2000)

As the literature has shown, precipitation and temperature data are used in many studies to estimate streamflow (Rauf et al., 2018; Parisouj et al., 2020). Therefore, this study used these parameters and their antecedents as predictor variables. However, there are insufficient long-term records in the meteorological observation stations in the study area.

Many studies have reported that the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) reanalysis data is a good alternative when daily precipitation and temperature observation data are missing (Gupta et al., 2020; Hamal et al., 2020). Therefore, precipitation and temperature data used as input data in the models for the years 1988-2015 were obtained from the MERRA-2 data at for the locations of the flow observation stations (NASA, 2022). The MERRA-2 data provides worldwide climate data from 1980 to the present. The data provides a spatial resolution of approximately 55 kilometers at latitude (0.5° latitude x 0.625° longitude) (Reichle et al., 2017). The basic statistics of the

Table 2. The basic statistics of the MERRA-2 data

Stations	Temperature (°C)					Precipitation (mm)				
	Min	Max	Mean	Std	Skw	Min	Max	Mean	Std	Skw
Ceyhan River-Misis	-0.1	38.1	19.5	8.5	-0.08	0.0	54.3	1.7	4.7	4.46
Söğütlü Creek-Hanköy	-6.6	35.3	15.6	9.1	-0.04	0.0	51.1	1.5	3.8	4.30
Keşiş Creek-Sarıdanışmanlı	-17.3	29.7	9.3	9.6	-0.13	0.0	32.1	1.1	2.6	4.35

* Min: Minimum value, Max: Maximum value, Std: Standard deviation, Skw: Skewness

MERRA-2 data are presented in Table 2. The input set was constructed using daily precipitation (Pt-1, Pt-2 and Pt) and temperature (Tt-1, Tt-2 and Tt) data and runoff data from the previous two days (Qt-1 and Qt-2). These lagged data are included in the model to capture how the hydrological system is affected by past conditions. In the literature, many studies have reported that precipitation and temperature data from previous days play an important role in hydrological processes and improve model performance (Riad et al., 2004; Tongal and Booji, 2018).

Support vector regression (SVR)

The objective of SVR is to calculate a linear regression function in a higher-dimensional input space (feature space), to which the input data is mapped via a nonlinear function (Raghavendra and Paresh, 2014). Consequently, a linear problem of regression in a complex input space is derived from a nonlinear problem of regression in a relatively simple input space. The solution is then performed in this space. The challenge is to identify a function f(x) hat exhibits the greatest discrepancy ε from the observed lateral displacements for the training data while maintaining a minimal degree of curvature. In order to achieve this, the e-insensitive loss function is employed, whereby errors below a threshold of ε are deemed acceptable, whereas deviations above this threshold are penalized and considered unacceptable (Misra et al., 2009). Mathematically,

$$f(x) = \langle w, x \rangle + b \quad \text{with } w \in X,\ b \in R \tag{1}$$

where $\langle w, x \rangle$ represents the magnitude of the dot product in X. In Equation 1, flatness is indicated by a small value of w. This can be achieved by reducing the Euclidean norm to a minimum, i.e.,. $\|w\|^2$. In this case, the problem can be formulated as given in Equation 2 and 3.

$$\text{minimise } \frac{1}{2}\|w\|^2 \tag{2}$$

$$\text{subject to} \begin{Bmatrix} (\langle w, x_i \rangle + b) - y_i \le \varepsilon \\ y_i - (\langle w, x_i \rangle + b) \le \varepsilon \end{Bmatrix} \tag{3}$$

Nevertheless, there are instances when it is not feasible to derive a flat function f with an error margin of less than ε. In order to solve this problem, Equation 4 and 5 are obtained by adding the parameters called slackness, which are expressed in ξ_i^-, ξ_i^+ notation, to the constraints in the optimization problem (Smola and Schölkopf, 2004; Misra et al., 2009).

$$\text{minimise } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^- + \xi_i^+) \tag{4}$$

$$\text{subject to} \begin{Bmatrix} (\langle w, x_i \rangle + b) - y_i \le \varepsilon + \xi_i^- \\ y_i - (\langle w, x_i \rangle + b) \le \varepsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ \ge 0 \end{Bmatrix} \tag{5}$$

where, C represents the predetermined term that governs the magnitude of the penalty associated with errors that fall outside the negligible margin of error. The slack variables, ξ_i^- and ξ_i^+ represent the lower and upper constraints on the outputs, respectively. It is required that the constant C be greater than zero. The value of C serves to establish a balance between the degree of flatness exhibited by the function and the extent of permissible deviations exceeding the specified tolerance limit of ε. In conclusion, this optimization problem. Finally, this optimization problem is expressed in Equation 6 given below by taking the dual of the problem.

Table 3. Hyperparameter optimization results

Stations	SVR		
	ε	**C**	**y**
Ceyhan River-Misis	1.91	999.1	347.9
Söğütlü Creek- Hanköy	0.03	255.5	133.8
Keşiş Creek- Sarıdanışmanlı	0.04	112.5	32.6

$$f(x) = \sum_{i=1}^{n}(a_i - a_i^*)\, k\langle x_i, x\rangle + b \qquad (6)$$

Here, a_i, a_i^* is the dual variables of the constraints of the problem in Equation 3 and takes values in the range of $[0, C]$.In Equation 6, $k\langle x_i, x\rangle$ is called the kernel function. Consequently, the nonlinear SVR is transformed within the high-dimensional input space in order to facilitate linear analysis. A variety of kernel functions may be employed in support of SVR, including linear, polynomial, sigmoid, and radial basis functions.

In comparison to other kernel functions, there is a notable prevalence of the RBF kernel function in the literature, due to its superior performance in producing more satisfactory results than other functions (Liong and Sivapragasam, 2002; Lin et al., 2006; Tongal and Booji, 2018). Therefore, the radial basis kernel function is used in this study as given Equation 7 (Misra et al., 2009).

$$k\langle x_i, x\rangle = \exp\left(-\frac{\|x_i - x\|}{2y^2}\right) \qquad (7)$$

where, y is kernel scale. To obtain best results of SVR model, three parameters need to be optimized which are ε value, box constraint (C), and kernel scale (γ). These parameters were determined by applying hyperparameter optimization to the training data in the MATLAB R2021 program, and the value ranges were selected according to the default parameters determined by the MATLAB Regression Learner App. The parameters obtained for each river for SVR models are given in Table 3.

Artificial neural network (ANN)

ANNs are systems for information processing that are constructed from nodes or neurons, which are analogous in structure to the human brain. They are employed for the resolution of complex problems. A number of studies have demonstrated the efficacy of ANN as a methodology for modelling nonlinear relationships between inputs and outputs in hydrological studies (Kisi, 2004; Corzo and Solomatine, 2007; Guimaraes Santos and Lima, 2014; Sušanj et al., 2016). The general structure of an ANN is composed of three layers: an input layer, a hidden layer, and an output layer. Each layer contains nodes that represent the variables associated with that layer.

In accordance with Equation 8, each node j receives signals from node i in the preceding layer. A weight, denoted as W_{ij} is assigned to correspond with each incoming signal, represented by X_i. The active incoming signal (S_j) to node j is the weighted sum of all incoming signals, and b_j is the neuron threshold value.

$$S_j = \sum_{i=1}^{n} X_i W_{ij} + b_j \qquad (8)$$

The active incoming signal (S_j) is subjected to a nonlinear activation function, thereby generating the node's outgoing signal as illustrated in Equation 9.

$$f(S_j) = \frac{1}{1 + e^{-S_j}} \qquad (9)$$

The structure of the network, the activation functions and the training algorithms all contribute to the determination of ANN models. Among the various types of ANNs, multilayer perceptron feedforward neural networks are among the most commonly employed (Singh et al., 2009; Lohani et al., 2011). In these networks, each neuron is associated with all neurons in the previous layer. Furthermore, the output of each layer is utilized as input for the subsequent layer. A number of studies have indicated that quasi-Newtonian optimization methods are capable of achieving rapid convergence times and low mean squared errors in hydrological processes. (Aqil et al., 2007; Badrzadeh et al., 2013; Linares Rodriguez et al., 2015). For this reason, the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) and the feedforward neural network algorithm were employed as optimization techniques.

In the ANN model, the system's nonlinearity is captured through the use of activation functions. In the field of hydrological modelling, the tangent sigmoid function is the activation function that is most frequently employed. (Dawson and Wilby, 2001; Zadeh et al., 2010). Consequently, the tangent sigmoid function was employed in this study. The function has an output range of 0 to 1. However, since the input data are outside this range, it is necessary to standardize the data. Consequently, the input data is subjected to standardization in accordance with Equation 10 (Rajurkar et al., 2002; Lohani et al., 2011).

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (10)$$

where X is the original data, $\overline{X}$ is the standardized value of the input, X_{min} is minimum and X_{max} is maximum of the measured values in all observations.

Once the data has been normalized, the next step is to determine the optimal number of neurons in the hidden layer. In the event that the number of neurons is insufficient, the network will be unable to structure the complex data, resulting in suboptimal outcomes. A high number of neurons is typically beneficial in enabling the network to address complex systems, although it can also lead to overfitting issues. Consequently, it is of paramount importance to ascertain the optimal number of nodes in the hidden layer, as this has a significant impact on the performance of the trained network. Consequently, the optimal number of neurons in the hidden layer was identified through a trial-and-error approach, wherein the number of neurons in the hidden layer was varied from 2 to 10 (Lohani et al., 2011; Tongal and Booji, 2018). The results of the experiments indicated that the optimal number of neurons in the hidden layer was 6 for Ceyhan River-Misis, 4 for Söğütlü Creek-Hanköy and 7 for Keşiş Creek- Sarıdanısmanlı.

Multiple linear regression (MLR)
Regression analysis is one of the oldest and most widely used methods in long-term hydrological forecasting (Rezaeianzadeh et al., 2014). The MLR Model is a linear model that explains the relationship of a y variable with two or more x variables. The general structure of the model is as given in Equation 11;

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \qquad (11)$$

Here, y is the dependent (or response) variable, x is the independent (or predictive) variable. ε is random error. β is the regression coefficient (constant). Each β coefficient is simply multiplied by a variable x.

Evaluation of model performance
In order to assess the performance of the models against the station data, the four statistical methods of coefficient of determination (R^2), root mean square error to standard deviation ratio (RSR) and Nash-Sutcliffe model efficiency (NSE) were employed.
R^2 is a statistic metric that indicates what proportion of variation in observations a model is able to explain. In addition, R^2 is a statistic that gives some information about whether a model fits well. A high R^2 value is desirable because the higher it is, the less unexplained variation there is. These coefficients are statistical measures of how closely a regression line approximates real data (Moriasi et al., 2007). It is computed by Equation 12.

$$R^2 = \left(\frac{\sum_{i=1}^{n} (O_i - \overline{O}) \times (M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n} (O_i - \overline{O})^2} \times \sqrt{\sum_{i=1}^{n} (M_i - \overline{M})^2}} \right)^2 \qquad (12)$$

One of the commonly used statistical measures of error is the Root Mean Square Error (RMSE) (Singh et al., 2005; Moriasi et al., 2007). It is widely acknowledged that a lower RMSE value indicates superior model performance. However, the actual value of this parameter varies according to the size of the data. In light of this, RSR has been developed as a model evaluation statistic that facilitates the interpretation of these values (Singh et al., 2005). RSR converts the RMSE values to a standard value using the standard deviation values of the observations. The RSR is calculated by dividing the RMSE and the observation standard deviations, as given in Equation 13.

$$RSR = \frac{\mathrm{RMSE}}{\mathrm{Std.Dev.}} = \frac{\sqrt{\sum_{i=1}^{n} |O_i - M_i|^2}}{\sqrt{\sum_{i=1}^{n} |O_i - \bar{O}|^2}} \qquad (13)$$

The Nash-Sutcliffe efficiency coefficient offers a quantitative measure of the predictive efficacy of a given model. Its values range from negative infinity to one. It is desirable that the value it receives is between zero and one, and as it approaches one, it indicates that the model provides a satisfactory estimation result (Moriasi et al., 2007). The NSE is calculated using the equation provided in Equation 14.

$$NSE = 1 - \frac{\sum_{i=1}^{n} (O_i - M_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \qquad (14)$$

In this equations; O_i represents the observation station measurement, $\bar{O}$ denotes the average of the observation station measurements, M_i signifies the model output, $\bar{M}$ signifies the model output. Finally, n denotes the number of data points. The results of the RSR and NSE evaluation are presented in Table 4.

Results and Discussion

In this study, ANN, SVR, and MLR models were applied for three rivers in the Ceyhan River basin to estimate daily streamflow values. Model performances are evaluated using three statistical measures. The input set of the model was constructed using precipitation (Pt-1, Pt-2 and Pt) and temperature (Tt-1, Tt-2 and Tt) data from the previous two days, as well as flow data from the previous days (Qt-1 and Qt-2). This choice was made in order to account for the lagged effects of variables such as precipitation, temperature, and runoff in order to obtain more accurate predictions. The use of previous flow data (Qt-1 and Qt-2) can contribute to a better performance of the model, especially during periods of rise and fall of the hydrograph. In addition, lagged temperature and precipitation data allow the process of infiltration of precipitation into the soil and its contribution to streamflow to be incorporated into the model (Riad et al., 2004; Tongal and Booji, 2018; Rauf et al., 2018). In the modeling, 70% of the data was used as a training data and 30% as a test data. Test results were used in the evaluation of the models.

Streamflow estimation results for Ceyhan River-Misis

The performance results for daily flow measurements between 1988 and 2015 of the SVR, ANN, and MLR models are presented in Table 5, respectively. Based on the results from the performance assessment criteria, the model estimates were found to be similar. However, it has been determined that the model obtained by using the SVR algorithm makes estimations with a lower error amount than other models.

The models' performances in terms of RMSE and R^2 for the test phases are shown in Table 5. RMSE values for MLR are lower than SVM and ANN. Due to the maximum R^2 (0.95) and minimum RMSE values (28.20), the SVR model was determined to be the best model for the daily streamflow estimation of the Ceyhan river. Similar results were observed in the ANN model (Table 5). Among the models, the MLR model had the worst performance compared to the other models (R^2 = 0.93 and RMSE = 28.94).

Table 4. Performance evaluation table for RSR and NSE (Moriasi et al. 2007)

Rank	NSE	RSR
Very Good	0.75 - 1.00	0.00 - 0.50
Good	0.65 - 0.75	0.50 - 0.60
Satisfactory	0.50 - 0.65	0.60 - 0.70
Unsatisfactory	< 0.50	> 0.70

* NSE: Nash-Sutcliffe model efficiency, RSR: Root mean square error to standard deviation ratio

However, when NSE and RSR were evaluated as model performance success, it was determined that the models made very good estimations according to the threshold values given in the literature (RSR<0.50, NSE>0.75). In general, ML regression analysis (SVR and ANN) has been found to give more accurate results than linear regression analysis (MLR) when statistical performance criteria are taken into account. A comparison of the observed and estimated daily flow for all models on a timeline is given in Figure 2. When the figure is examined, it is seen that the estimations obtained from the models make similar estimations to the station observation values and the estimation errors are quite low. As a result of the evaluations, it has been determined that the daily flow of the Ceyhan River, which has a high annual average flow, can be successfully estimated using the model created with a small number of parameters.

Streamflow estimation results for Söğütlü Creek- Hanköy

The statistical evaluation results of the estimates made for the Hanköy measurement station located on the Söğütlü Stream in the northern part of the Ceyhan basin are given in Table 6. It was seen that ANN, MLR, and SVM models gave similar results to the results in the Ceyhan River. It has been determined that the models created using ANN and SVM ML algorithms are more successful than the flow estimations of the traditional MLR model.

As shown in Table 6, the best result was for the SVR, as demonstrated by the minimum error rate, RMSE = 6.57, and the best correlation rate, R^2 = 0.82. Additionally, the ANN model also

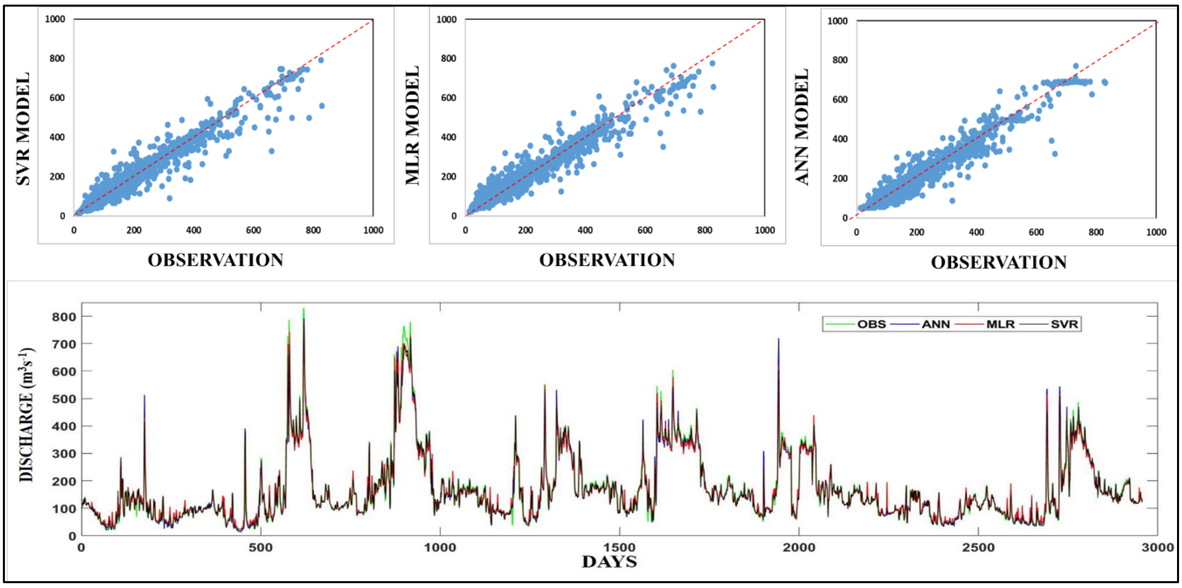

Figure 2. Time series and scatter plot of models for Ceyhan River-Misis

Table 5. Performance results of testing phase for Ceyhan River

Station	Method	Testing Period			
		R^2	RMSE	NSE	RSR
Ceyhan River-Misis	MLR	0.93	28.94	0.92	0.25
	ANN	0.95	28.25	0.95	0.23
	SVR	0.95	28.20	0.95	0.22

*R^2: Coefficient of determination, RMSE: Root mean square error, NSE: Nash-Sutcliffe model efficiency, RSR: Root mean square error to standard deviation ratio

performed well like the SVR model, with the minimum error rate, RMSE = 6.61, and the best correlation rate, R^2 = 0.81, refers to the most precise correlation value. On the other hand, the MLR model showed worse estimation performance than the others, with the minimum error rate RMSE = 6.93 and the best correlation rate R^2 = 0.79. However, when the model performance success was evaluated according to the threshold values (RSR<0.50, NSE>0.75), it was determined that it made very good estimations. The correlation and similarity of the estimations of all models with the observation data can be observed in Figure 3. Similar to the Ceyhan River; It is seen that all three models underestimate the daily flow for Söğütlü Creek-Hanköy, especially the peak flow values, then the observed values. The use of delayed flow data as input to ANN and SVR models can create difficulties in determining and estimating the effects of sudden extreme events such as floods and droughts on stream flow (Chau and Li, 2009; Sazib et al., 2020).

In addition, extreme events in climatic factors such as temperature and precipitation used as input in the models can affect hydrological processes and river flows (Zamanisabzi et al., 2018). These changes may cause over- or underestimation of the model in certain time periods.

Streamflow estimation results for Keşiş Creek- Sarıdanışmanlı

The statistical evaluation of different types of model performances in daily streamflow estimations of the Sarıdanışmanlı measurement station located on the Keşiş Creek for the testing phase is presented in Table 7. In general, it has been observed that ML models (ANN and SVR) give better results than MLR.

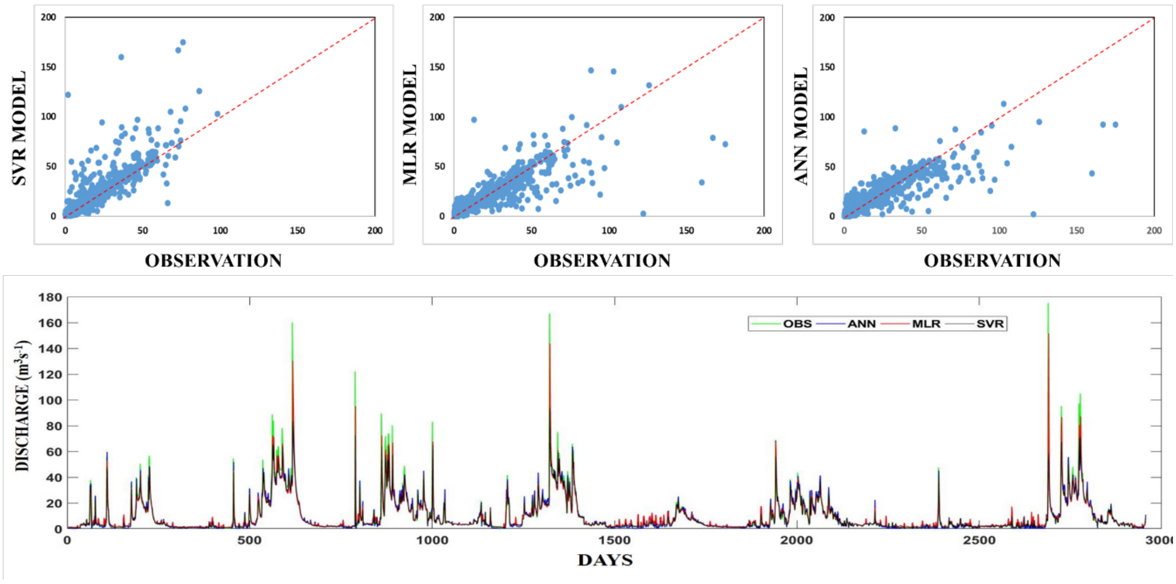

Figure 3. Time series and scatter plot of models for Söğütlü Creek- Hanköy

Table 6. Performance results of testing phase for Söğütlü Creek- Hanköy

Station	Method	Testing Period			
		R^2	RMSE	NSE	RSR
Söğütlü Creek- Hanköy	MLR	0.79	6.93	0.79	0.46
	ANN	0.81	6.61	0.81	0.44
	SVR	0.82	6.57	0.81	0.44

*R^2: Coefficient of determination, RMSE: Root mean square error, NSE: Nash-Sutcliffe model efficiency, RSR: Root mean square error to standard deviation ratio

The R^2 and RMSE values obtained showed agreement with the observed data. The lowest RMSE value was calculated for the SVR model (1.45) and the highest for the MLR model (1.62). The R^2 value that indicates the agreement with the observation data was higher than approximately 0.90 for all models. When NSE and RSR were evaluated as model performance success, it was determined that the models made very good estimations according to the threshold values given in the literature (RSR<0.50, NSE>0.75). Figure 4 shows the daily streamflow estimation graphs obtained by ANN, SVR, and MLR methods for the testing phase of Keşiş Creek- Sarıdanışmanlı. SVR and ANN model estimations were found to be very similar when compared to MLR. Also, it was determined that the model estimates and the observed streamflows are similar except for the peak values. When the comparison of flow observation measurements and daily flow forecasts obtained from the models on the timeline is examined, it is seen that it makes estimations similar to the observation measurements and trend, as in the Ceyhan river forecasts.

As a result of the evaluations, it was seen that the estimations obtained for Keşiş Stream, which has an annual average flow of 10.60 $m^3 s^{-1}$, are better than the Söğütlü Stream measurements but less unsuccessful than the Ceyhan River, which has a higher annual average flow.

When the results were evaluated in general for the three measurement stations, it was seen that the estimations were generally successful (Figure 5). Similar to the findings of this study, Vatanchi et al. (2023) reported that the performance of the ANN prediction model was successful (NSE=0.92) in their study to predict the daily

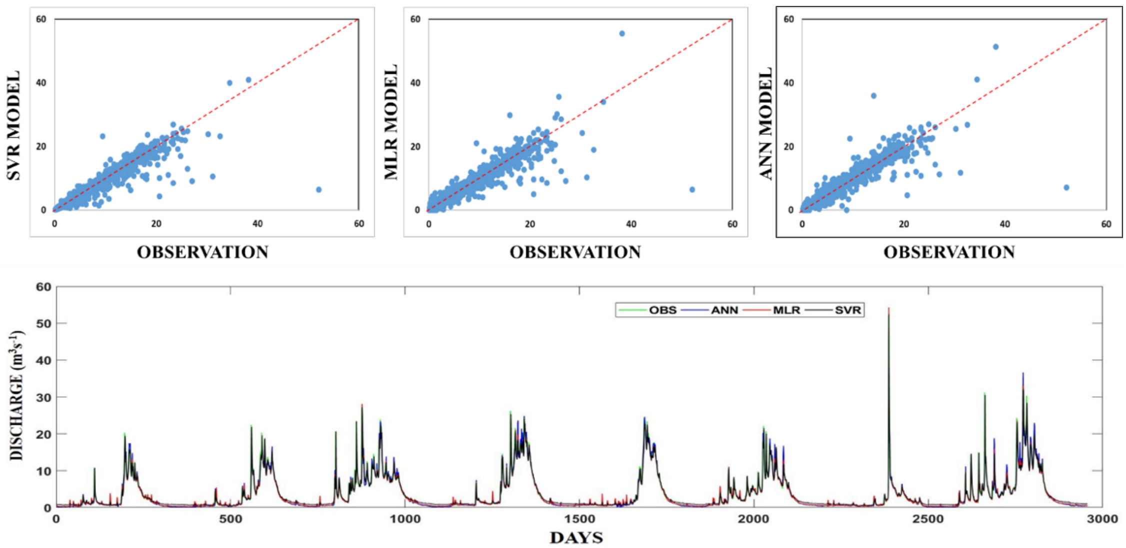

Figure 4. Time series and scatter plot of models for Keşiş Creek- Sarıdanışmanlı

Table 7. Performance results of testing phase for Keşiş Creek- Sarıdanışmanlı

Station	Method	Testing Period			
		R^2	RMSE	NSE	RSR
Keşiş Creek- Sarıdanışmanlı	MLR	0.90	1.62	0.90	0.32
	ANN	0.91	1.50	0.91	0.29
	SVR	0.93	1.45	0.92	0.29

*R^2: Coefficient of determination, RMSE: Root mean square error, NSE: Nash-Sutcliffe model efficiency, RSR: Root mean square error to standard deviation ratio

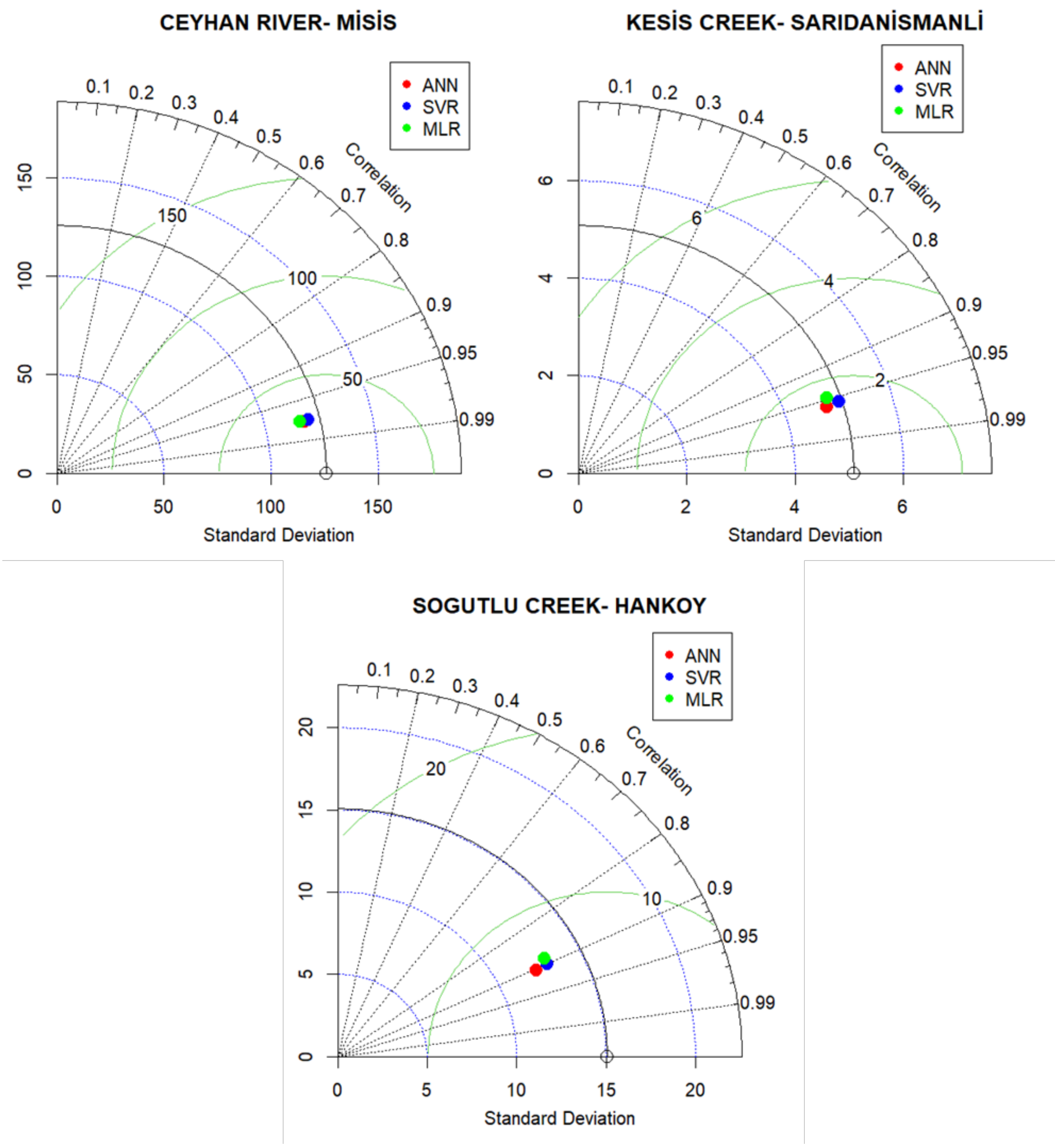

Figure 5. Overall evaluation results obtained for the test phase of the models

flow of the Colorado River in the USA. Kamali et al. (2023) evaluated different machine learning models in a similar study in the Karkheh basin of Iran. It was reported that the SVR model showed high performance with an R^2 and RMSE of 0.85 and 36.49 $m^3 s^{-1}$, respectively, during the test periods. Oad et al. (2023) evaluated the usability of ANN prediction models for stream forecasting in their study for the Goulburn River in Australia and found that the models were acceptable with correlation coefficient values ranging between 0.61-0.95.

In addition, it has been determined that the low amount of flow of the measured river affects the estimation success of the models. Accordingly, it was seen that the estimations for the Ceyhan River, which has high flow, were the most successful, while the estimations made for the Söğütlü Creek, which had the lowest flow, were more unsuccessful. It has been determined that model estimations created using ML algorithms from the models evaluated in the study are more successful than traditional regression approaches. When the results obtained were compared with the information given in the literature, it was seen that similar results and inferences were obtained (Rauf et al., 2018; Tongal and Booji, 2018; Parisouj et al., 2020; Siddiqi et al., 2021;)

Conclusion

In this study, ANN, SVR, and MLR models were developed to estimate daily streamflow values in the Ceyhan River Basin. The performance of the models was determined using daily precipitation and temperature data, which were obtained from MERRA-2 reanalysis data.

The analysis results revealed that the SVR model exhibited the best predictive performance for the Ceyhan River, with $R^2 = 0.95$ and RMSE = 28.20 $m^3 s^{-1}$. Furthermore, high accuracy was also observed in the other rivers, with the Söğütlü Creek showing $R^2 = 0.82$ and RMSE = 6.57 $m^3 s^{-1}$, and the Keşiş Creek showing $R^2 = 0.93$ and RMSE = 1.45 $m^3 s^{-1}$. These results indicate that machine learning algorithms (particularly SVR and ANN) performed significantly better than the traditional MLR model.

The key findings of this study highlight the effectiveness of data-driven approaches in hydrological modeling and emphasize the benefits of using alternative data sources (such as MERRA-2 reanalysis data) for streamflow estimation in areas like the Ceyhan River Basin, which may suffer from data scarcity. However, the limitation of the model to only predict one day ahead streamflow constrains its potential applications. Future research should focus on developing multi-step prediction methods to enhance the accuracy of streamflow forecasts.

In conclusion, this study contributes to the improvement of hydrological modeling in the Ceyhan River Basin and provides valuable insights for water management and decision-making processes. These findings can assist in the more effective management of water resources in the region.

Acknowledgements

We appreciate the State Hydraulic Works for providing the streamflow data needed for the study.

References

Adamowski, J., Sun, K. (2010) Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *J Hydrol* 390(1–2):85–91.

Aqil, M., Kita, I., Yano, A., Nishiyama, S. (2007) Neural networks for real time catchment flow modeling and prediction. *Water Resour Manage* 21:1781–1796.

Badrzadeh, H., Sarukkalige, R., Jayawardena, A.W. (2013) Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting. *J Hydrol* 507:75–85.

Bafitlhile, T. M., Li, Z. (2019) Applicability of ε-support vector machine and artificial neural network for flood forecasting in humid, semi-humid and semi-arid basins in China. *Water* 1(11):85.

Belayneh, A., Adamowski, J., Khalil, B. (2015) Short-term SPI drought forecasting ın the Awash River Basin in Ethiopia using wavelet transforms and machine learning methods. *Sustain Water Resour Manage* 1(2):87-101.

Bhadrecha, M. H., Khatri, N., Tyagi, S. (2016) Rapid integrated water quality evaluation of Mahisagar river using benthic macroinvertebrates. *Environ Monit Assess* 188:254.

Chau, K. W., and Li, Y. (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour Res* 8(45).

Cheng, C. T., Feng, Z. K., Niu, W. J., Liao, S. L. (2015) Heuristic methods for reservoir monthly inflow forecasting: a case study of Xinfengjiang Reservoir in Pearl River, China. *Water* 7:4477–4495.

Corine (2020) Land use and land cover map. https://land.copernicus.eu/en/products/corine-land-cover Accessed 18 December 2022.

Corzo, G., Solomatine, D. (2007) Baseflow separation techniques for modular artificial neural network modelling in flow forecasting. *Hydrological Sciences Journal* 52:491–507.

Dawson, C. W., Wilby, R. L. (2001) Hydrological modelling using artificial neural networks. *Prog Phys Geogr Earth Environ* 25(1):80–108.

DSI (2022) Daily streamflow data. https://www.dsi.gov.tr/ . Accessed 13 December 2022.

Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K., Ardabili, S. F., Piran, M. J. (2018) Survey of computational intelligence as basis to big flood management: challenges, research directions and future work. *Eng Appl Comput Fluid Mech* 12(1):411–437.

Guimaraes Santos, C. A., Silva, G. B. L. D. (2014) Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models. *Hydrol Sci J* 59(2):312–324.

Gupta, P., Verma, S., Bhatla, R., Chandel, A. S., Singh, J., Payra, S. (2020) Validation of surface temperature derived from MERRA-2 reanalysis against IMD gridded data set over India. *Earth Space Sci* 7(1):e2019EA000910.

Hamal, K., Sharma, S., Khadka, N., Baniya, B., Ali, M., Shrestha, M. S., Xu, T., Shrestha, D., Dawadi, B. (2020) Evaluation of MERRA-2 precipitation products using gauge observation in Nepal. *Hydrology* 7(3):40.

Hosseini, S. M., Mahjouri, N. (2016) Integrating support vector regression and a geomorphologic artificial neural network for daily rainfall-runoff modeling. *Appl Soft Comput* 38:329-345.

Kamali, S., Saedi, F., Asghari, K. (2023) Prediction of flood in karkheh basin using data-driven methods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W1-2022: 349-354.

Kamali, S., Saedi, F., Asghari, S. (2022) Prediction of flood in Karkheh Basin using data-driven methods. *ISPRS Ann Photogramm Remote Sens Spatial Inf Sci X-4:W1.*

Khatri, N., Tyagi, S., Rawtani, D., Tharmavaram, M. (2020) Assessment of river water quality through application of indices: a case study River Sabarmati, Gujarat, India. *Sustain Water Resour Manag* 6:101.

Kisi, Ö. (2004) River flow modeling using artificial neural networks. *J Hydrol Eng* 9(1):60–63.

Kisi, O., Shiri, J., Tombul, M. (2013) Modeling rainfall-runoff process using soft computing techniques. *Comput Geosci* 51:108-117.

Kumar, V., Sen, S. (2018) Evaluation of spring discharge dynamics using recession curve analysis: a case study in data-scarce region, Lesser Himalayas, India. *Sustain. Water Resour. Manag.* 4:539–557.

Lin, J. Y., Cheng, C. T., Chau, K. W. (2006) Using support vector machines for long-term discharge prediction. *Hydrol Sci J* 51(4):599–612.

Linares Rodriguez, A., Lara-Fanego, V., Pozo-Vazquez, D., Tovar-Pescador, J. (2015) One-day-ahead streamflow forecasting using artificial neural networks and a meteorological mesoscale model. *J Hydrol Eng* 20(9):05015001.

Liong, S. Y., Sivapragasam, C. (2002) Flood stage forecasting with support vector machines. *J Am Water Resour Assoc* 38(1):173–186.

Lohani, A. K., Goel, N., Bhatia, K. (2011) Comparative study of neural network, fuzzy logic and linear transfer function techniques in daily rainfall-runoff modelling under different input domains. *Hydrol Process* 25(2):175–193.

Masselot, P., Dabo-Niang, S., Chebana, F., Ouarda, T. B. M. J. (2016) Streamflow forecasting using functional regression. *J Hydrol* 538:754–766.

Mehr, A. D., Kahya, E., Şahin, A., Nazemosadat, M. J. (2015) Successive-station monthly streamflow prediction using different artificial neural network algorithms. *Int J Environ Sci Technol* 12:2191-2200.

Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., Wang, L. (2019) A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. *J Hydrol* 568:462–478.

Misra, D., Oommen, T., Agarwal, A., Mishra, S. K., Thompson, A. M. (2009) Application and analysis of support vector machine based simulation for runoff and sediment yield. *Biosyst Eng* 103(4):527–535.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L. (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans Am Soc Agric Biol Eng* 50(3):885–900.

NASA (2022) https://gmao.gsfc.nasa.gov/reanalysis/merra-2/data_access/ . Accessed 18 December 2022.

Noori, N., Kalın, L. (2016) Coupling SWAT and ANN models for enhanced daily streamflow prediction. *J Hydrol* 533:141–151.

Oad, S., Imteaz, M. A., Mekanik, F. (2023) Artificial neural network (ann)-based long-term streamflow forecasting models using climate indices for three tributaries of goulburn river, australia. *Climate*, 11 (7): 152.

Parisouj, P., Mohebzadeh, H., Lee, T. (2020) Employing machine learning algorithms for streamflow prediction: a case study of four river basins with different climatic zones in the United States. *Water Resour Manag* 34:4113-4131.

Peng, T., Zhou, J., Zhang, C., Fu, W. (2017) Streamflow forecasting using empirical wavelet transform and artificial neural networks. *Water* 9(6):406.

Raghavendra, N. S., Paresh, C. D. (2014) Support vector machine applications in the field of hydrology: a review. *Appl Soft Comput* 19:372–386.

Rajurkar, M., Kothyari, U., Chaube, U. (2002) Artificial neural networks for daily rainfall—runoff modeling. *Hydrol Sci J* 47(6):865–877.

Rauf, A., Ghumman, A. R., Ahmad, S., Hashmi, H. N. (2018) Performance assessment of artificial neural networks and support vector regression models for stream flow predictions. *Environ Monit Assess* 190:704.

Rezaeianzadeh, M., Tabari, H., Yazdi, A. A., Isik, S., Kalin, L. (2014) Flood flow forecasting using ANN, ANFIS and regression models. *Neur Comput Appl* 25:25–37.

Riad, S., Mania, J., Bouchaou, L., Najjar, Y. (2004) Rainfall-runoff model using an artificial neural network approach. *Math Comput Model* 40(7–8):839–846.

Reichle, R. H., Draper, C. S., Liu, Q., Girotto, M., Mahanama, S. P. P., Koster, R. D., De Lannoy, G. J. M. (2017) Assessment of MERRA-2 land surface hydrology estimates. *J. Climate* 30(8):2937–2960.

Sauquet, E., Beaufort, A., Sarremejane, R., Thirel, G. (2021) Predicting flow intermittence in France under climate change. *Hydrol Sci J* 66:2046-2059.

Sazib, N., Bolten, J. D., Mladenova, I. E. (2020) Exploring spatiotemporal relations between soil moisture, precipitation, and streamflow for a large set of watersheds using google earth engine. *Water*, 5(12):1371.

Shrestha, S., Bae, D. H., Hok, P., Ghimire, S., Pokhrel, Y. (2021) Future hydrology and hydrological extremes under climate change in Asian river basins. *Sci Rep* 11:17089.

Siddiqi, T. A., Ashraf, S., Khan, S. A., Iqbal, M. J. (2021) Estimation of data-driven streamflow predicting models using machine learning methods. *Arab J Geosc* 14:1058.

Singh, J., Knapp, H. V., Arnold, J. G., Demissieet, M. (2005) Hydrologic modeling of the Iroquois River watershed using HSPF and SWAT. *J Am Water Resour Assoc* 41(2):361-375.

Singh, K. P., Basant, A., Malik, A., Jainet, G (2009) Artificial neural network modeling of the river water quality—a case study. *Ecol Model* 220(6):888–895.

Smola, A. J., Schölkopf, B. (2004) A tutorial on support vector regression. *Stat Comput* 14(3):199-222.

Sušanj, I., Ožanić, N., Marovic, I. (2016) Methodology for developing hydrological models based on an artificial neural network to establish an early warning system in small catchments. *Adv Meteorol* 14:9125219.

TOB (2019) Drought management plan of Ceyhan River Basin. Republic of Turkey Ministry of Agriculture and Forestry, General Directorate of Water Management. https://www.tarimorman.gov.tr/SYGM/Belgeler/Kuraklık%20Yönetim%20Planları/Ceyhan%20Havzası%20Kuraklık%20Yönetim%20Planı%20Yönetici%20Özeti.pdf. Accessed 22 August 2022.

Tongal, H., Berndtsson, R. (2016) Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. *Stoch Environ Res Risk Assess* 3(31):661-682.

Tongal, H., Booji, M. J. (2018) Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *J Hydrol* 564:266–282.

Vatanchi, S.M., Etemadfard, H., Maghrebi, M.F. et al. (2023) A Comparative Study on Forecasting of Long-term Daily Streamflow using ANN, ANFIS, BiLSTM and CNN-GRU-LSTM. *Water Resour Manage* 37: 4769–4785.

Wang, L., Li, X., Ma, C., Bai, Y. (2019) Improving the prediction accuracy of monthly streamflow using a data-driven model based on a double-processing strategy. *J Hydrol* 573:733-745.

Young, C. C., Liu, W. C., Wu, M. C. (2017) A physically based and machine learning hybrid approach for accurate rainfall-runoff modeling during extreme typhoon events. *Appl Soft Comput* 53:205-216.

Zadeh, M.R., Amin, S., Khalili, D., Singh, V. P. (2010) Daily outflow prediction by multi layer perceptron with logistic sigmoid and tangent sigmoid activation functions. *Water Resour Manag* 24(11):2673–2688.

Zamanisabzi, H., King, J. P., Dilekli, N., Shoghli, B., Abudu, S. (2018) Developing an ann based streamflow forecast model utilizing data-mining techniques to improve reservoir streamflow prediction accuracy: a case study. *CivileJournal* 5(4):1135.