Pol. J. Environ. Stud. Vol. 32, No. 4 (2023), 3161-3173
DOI: *10.15244/pjoes/161515* ONLINE PUBLICATION DATE: 2023-06-15

Original Research

Combining Machine Learning Algorithms with Empirical Mode Decomposition and Discrete Wavelet Transform for Monthly Peak Discharge Prediction

Okan Mert Katipoğlu[1]*, Metin Sarıgöl[2]

[1]Erzincan Binali Yıldırım University, Department of Civil Engineering, Erzincan, Turkey
[2]Erzincan Binali Yıldırım University, Design Department, Erzincan, Turkey

Received: 8 September 2022
Accepted: 21 February 2023

Abstract

Accurate and reliable peak discharge prediction is of great importance in water resource management and flood control studies. The aim of this study was to predict monthly peak discharge by combining various signal decomposition processes and machine learning models. For this purpose, monthly peak discharge data were decomposed into subsignals utilizing Daubechies 3, Coiflet 5, discrete Meyer main wavelets, and empirical mode decomposition methods, which are the ones most commonly used in hydrological studies. The separated signals were subjected to correlation analysis, and the sum of the highly correlated signals was presented as input to the machine learning models. Support vector machines, regression trees, ensemble trees, and adaptive neuro-fuzzy inference system models were used for peak discharge forecasting. The performance of the established models was evaluated with the help of statistical indicators such as mean absolute error, root mean squared error, and determination coefficient, and graphically through Taylor diagrams. At the end of the study, the most effective results were obtained with the hybrid model established by the Daubechies 3 wavelet - 4 decomposition levels and combined with the coarse Gaussian support vector machine.

Keywords: discrete wavelet transform, empirical mode decomposition, machine learning, monthly discharge forecasting

Introduction

Accurate and reliable estimation of peak discharges is essential in planning water resources, design of water structures, and disaster relief. For this reason, meteorology, hydrology, and environmental science are among the critical subjects that attract attention [1, 2]. Natural disasters cause great damage to human life, agriculture, infrastructure, and the socioeconomic system [3, 4]. Direct impacts include loss of life, loss of agricultural production, damage to infrastructure, disruption of trade and education, and indirect social

*e-mail: okatipoglu@erzincan.edu.tr

impacts on communities and human health [5, 6]. In particular, high-intensity precipitation in arid mountainous areas and dry and crusty soil structures can quickly turn rain into runoff, which can cause higher risks. However, early notification of natural disasters and accurate forecasting and prevention can effectively reduce the resulting losses [7-10].

Flood forecasting models are vital in managing extreme weather events and assessing potential flood-related hazards. Accurate estimation allows for developing effective water resources management strategies, conducting comprehensive analyzes and making discharge models more clearly [11]. Therefore, the importance of advanced systems to reduce damage is strongly emphasized when forecasting floods and other disasters [12]. However, because of the variability in climatic conditions, the estimation of where the flood will occur, and the duration of the flood delay, today's widely used flood forecasting models, which include simplified assumptions, are mainly based on data [13]. Therefore, certain techniques are used to model mathematical expressions, such as deterministic, stochastic, continuous, empirical black box, and hybrid models [14]. Furthermore, while flood risk management is conducted to reduce or prevent the adverse effects of floods, early warning systems can contribute to implementing effective emergency strategies such as protecting populations and property and early warnings in severe flood situations [15].

Black box, physically based, and conceptual models are commonly used to predict hydrological variables [16]. Conceptual and physically based models require much knowledge about the fundamental physics of the structure established with differential equations. On the other hand, machine learning (ML) models describe complex relationships between input and output parameters. The input and output data used in the modeling are based on the observed data and no other parameters are needed. In addition, high computational efficiency is another advantage due to the high computational power of ML models [17].

In recent years, ML algorithms have been used frequently in hydrology. In particular, random forest (RF), artificial neural network (ANN), decision tree (DT), wavelet neural network (WNN), support vector machine (SVM), neural wavelet network (NWN), adaptive neuro-fuzzy inference system (ANFIS), and hybrid models have seen wide use in hydrology. Dawson and Wilby [18] applied an ANN for rainfall -runoff modeling and forecasting floods. Talei and Chua [19] used the ANFIS model to investigate the effect of lag time in the rainfall-runoff model. Noury et al. [20] employed SVM and NWN models with wavelet functions to simulate water level fluctuation in Lake Urmia. It was determined that the SVM model showed better results than the NWN. Granata et al. [21] compared the stormwater management model (SWMM) and the SVM-based precipitation runoff model. The SVM model gave outstanding results. Kasiviswanathan et al. [22] combined ANN and WNN models with the block bootstrap (BB). WNN-BB outperformed ANN-BB. Seo et al. [23] used support vector machine and wavelet packet (WPSVM), adaptive neuro-fuzzy inference system and wavelet packet (WPANFIS), and artificial neural network and wavelet packet (WPANN) to predict river stage and evaluated their performance. WPANFIS gave the best results. Yaseen et al. [24] created a heuristic algorithm and fuzzy logic model for monthly flow estimation. The firefly algorithm was used in the training phase of the ANFIS inference system. Compared to the normal ANFIS models, the ANFIS models trained with the firefly algorithm gave more successful results for streamflow estimation. Shafizadeh-Moghadam et al. [25] reported that boosted regression trees (BRT) were the most effective machine learning model. Choi et al. [26] employed ANN, DT, RF, and SVM models to forecast the water level. Choubin et al. [27] combined multivariate discriminant analysis (MDA), classification, and regression trees (CART) with the SVM model to map flood susceptibility, and the MDA model gave the highest prediction accuracy in the analysis of the results. Abedi et al. [28] used BRT, XGBoost, CART, and RF models to create a flash flood map. Yuan et al. [29] determined that LSTM and radial basis function (RBF) gave a better statistical performance than the EEMD-LSTM series. An examination of the existing literature reveals that studies on the performance evaluation of hybrid models established by combining wavelet transform (WT), empirical mode decomposition (EMD), and machine learning methods in the estimation of peak discharge values are lacking.

The most crucial aim in the present study was to predict monthly peak discharge by combining various signal decomposition processes and machine learning models. The first stage of the research involved separating monthly peak discharge data into subcomponents. In the next step, these subcomponent's data were subjected to the machine learning method. Finally, the last stage of the research concerned setting up models after training and testing ML. In this way, it was thought that it would allow peak discharge prediction by adding only input data without the need for any other data.

The aim was to construct a canal network model for peak discharge estimation along a river and to determine the efficiency and effectiveness of these methods. In addition, the most effective signal separation process in peak discharge estimation and the performance of various hybrid ML models were analyzed in the study.

Material and Method

Study Area and Data

The data from Değirmenocağı discharge observation station (DOS), numbered E18A027, at an altitude of 740 m at 37°51′18″ North, 35°29′10″ East, and

Ergenuşağı DOS, numbered E18A027, at an altitude of 360 m, at 37°39′55″ North, 35°34′47″ East, were used. DOSs E18A027 and E18A026 were opened in 1987. The monthly peak discharge rates used in the study cover the years 1987-2015. Six hundred seventy-two peak discharge rates were obtained and used in the study. The characteristics of the DOSs are shown in Table 1 and graphics are plotted in Fig. 1.

The most important stream with its source in Kayseri Province is the River Zamantı. It is one of the two large branches of the River Seyhan, and its length in the region is 230 km. The River Zamantı, which has a precipitation area of approximately 8,700 km², has an average annual flow of 65,603 m³. The highest flow rate of the river is 970 m³ per second, and the water depth at this time goes up to 410 cm (Waybackmachinearchive 2021). DOSs on the River Zamantı are located in the Seyhan Basin. The DOSs and their locations are shown in Figs 2-3.

According to the climatic characteristics of the precipitation area, large floods do not occur in the region during the summer months. As can be understood from examining the measurement values of the hydrometry stations, no major floods were observed during the summer months. Therefore, the flood hydrographs observed at the hydrometry stations took place in December-May following the climatic characteristics of the River Zamantı precipitation area. These hydrographs were formed by a combination of rain or rain and snow melt flow [30].

Machine Learning Models

In our study, we aimed to compare the effectiveness of hybrid models by employing various machine learning techniques such as fine, medium, and coarse tree, linear and nonlinear SVM, Gaussian SVR, boosted and bagged trees, and ANFIS, which are commonly used in the literature. Furthermore, we investigated the impact of different signal decomposition methods on the performance of these models.

Table 1. Discharge observation stations used in the study.

Province/ District Name	Station Number	Station Code	Station Name	Basin Name	River Name	Station Elevation
Kayseri / Yahyalı	1	E18A027	Değirmenocağı	Seyhan	Zamanti	740
	2	E18A026	Ergenuşağı	Seyhan	Zamanti	360

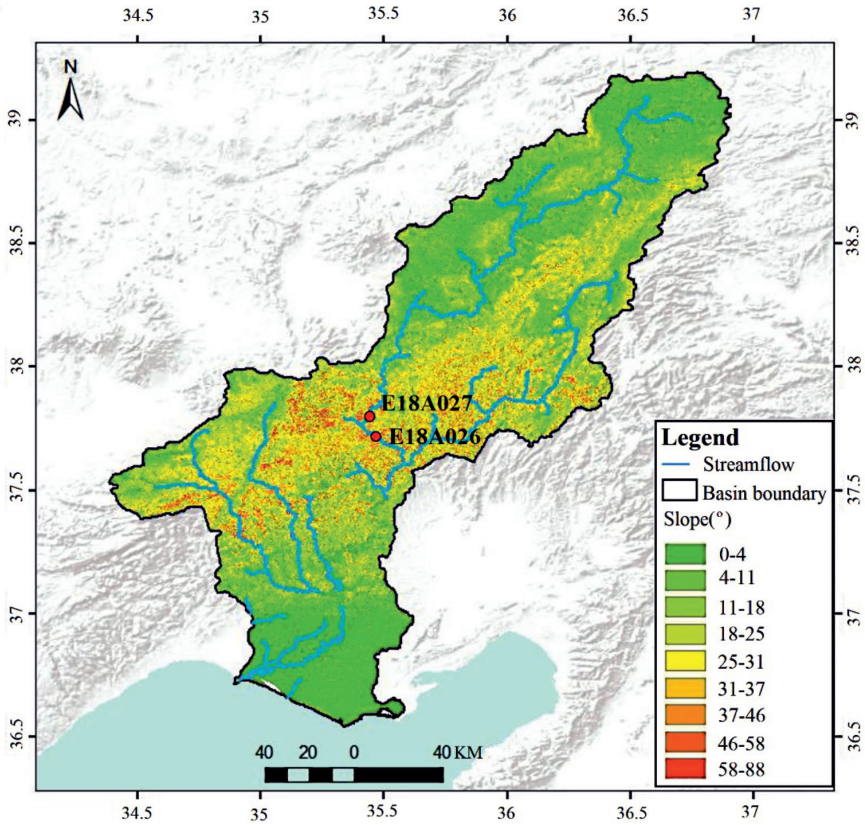

Fig. 1. The Seyhan Basin and river network.

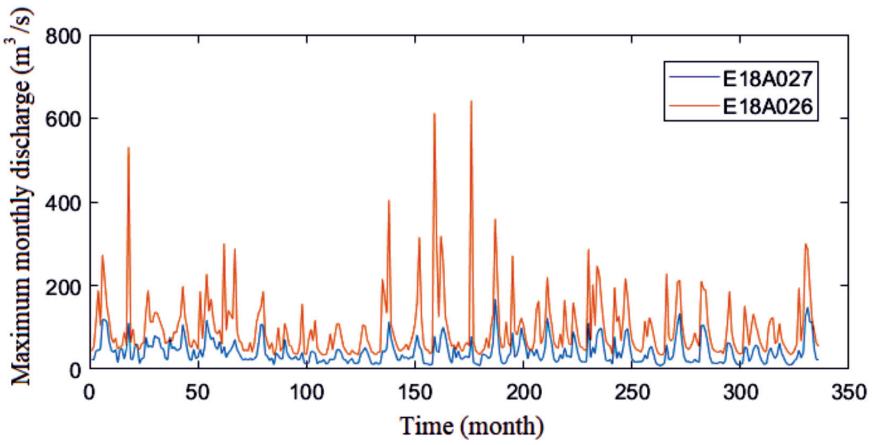

Fig. 2. Maximum monthly discharge at the E18A027 a) and E18A026 b) DOSs.

Fig. 3. E18A027 and E18A026 DOSs.

Adaptive Neuro-Fuzzy Inference System (ANFIS)

Zadeh [31] developed the logic system that an object can be included in more than one set. In classical logic, representation can take two values (1 or 0), expressed as presence and absence. In fuzzy logic (FL), the representation can take infinite values between 1 and 0. ANFIS, on the other hand, can be defined as the adaptation of neural networks to Takagi and Sugeno [32], one of the inference methods in the FL system. Takagi–Sugeno, whose inference can be expressed with exact numbers, does not need clarification. However, the parameters of the result functions need to be optimized. This optimization process was carried out using ANNs. Combining the decision-making ability of FL with the learning ability of an ANN, ANFIS provides fast results for analyzing numerical data [33].

Support Vector Machine (SVM)

This method is an optimization-based algorithm designed by Vapnik [34] that minimizes the error, and was later used for regression with the name SVR algorithm. Since the SVM algorithm depends on kernel functions, it is installed by taking the maximum value into the model as a non-parametric method. While the advantages of this algorithm are ease of implementation and compatibility with nonlinear and linear data, the disadvantages are the difficulty of interpreting the model parameters and the long duration of algorithm training.

Regression Tree (RT)

Node, branch, and leaf are the three basic elements that make up the basic structure of a decision tree. While the node represents each attribute in this tree structure, the leaf, the last part of the tree structure, and branches are the elements of the tree. The root is the upper part of the tree structure. The branches are the parts between the root and the leaves [35]. Nodes represent targets, while links are used for decisions. The rules are written down from root to leaf (IF–THEN rules) [36, 37]. In the decision tree method, action is taken according to the answer to the question in concluding an event.

Regression Tree Ensemble (RTE)

The basis of this algorithm is leaf nodes and decision nodes. First, the standard deviation values between the clusters and target are calculated and then the standard

deviation of the target set. After this stage, the results are subtracted from the standard deviations calculated from the target clusters, and the cluster with the most significant standard deviation is defined as the root. These process steps are continued for each node and finally the tree structure is created [38].

Empirical Mode Decomposition (EMD)

EMD has become a valuable tool widely used and adaptable for analyzing multichannel data flexibly with linear and nonstationary time series, offering intrinsic mode functions (IMFs) based on instantaneous frequency. Hence, it is suitable for linear stationary and nonlinear processes. Finally, EMD results are processed into the energy–frequency–time dispersion [39].

Wavelet Transform (WT)

In this transform, a mathematical structure is used, and analysis of local changes in time series and information from various data sources is performed. WTs improve data quality by providing reliable parsing of an original time series. Prediction accuracy is increased by discrete WT banding of data. This offers better peak discharge forecast lead times. The DWT method separates the first dataset into different resolution levels to extract higher quality data while building the model. It is commonly used in peak discharge time series forecasting due to its valuable properties. In peak discharge modeling, DWTs have been widely applied in areas such as rainfall–runoff, diurnal runoff, and reservoir inflow [40].

To specify the optimal number of decomposition levels in the wavelet transform, the formula based on the signal length in Equation (1) is used.

$$L = \text{int}[\log(N)] \tag{1}$$

Accordingly, since the data length is L = 336, and L = Int[log(336)], approximately 3, the peak discharge data were divided into 3, with 4 decomposition level subcomponents, and hybrid models were established.

Fig. 4 shows the flow chart showing the planning steps of the study. The techniques used in the decomposition and modeling stages of the techniques used in this diagram are presented. In addition, the steps of the model setup and the way of the study about the selection of the best model are expressed.

Testing Routing Success

The most commonly used method to measure the model's success is the root mean square error (RMSE).

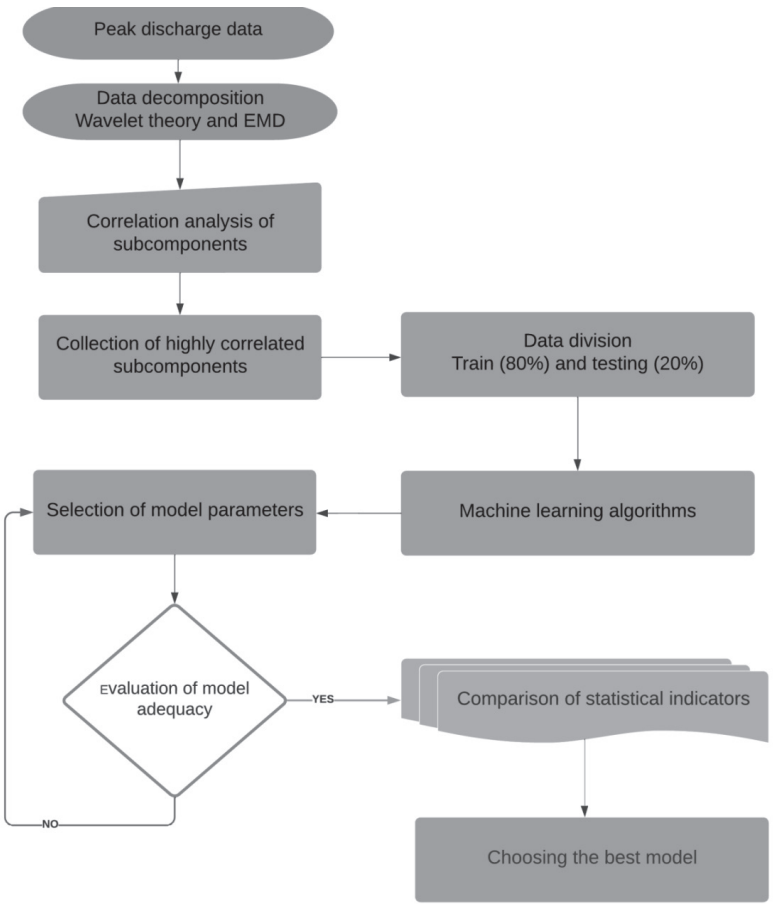

Fig. 4. Flow chart of the study.

The first criterion is standard deviation of the difference between the calculated value and the actual value. This difference measures how far the regression line is from the data points.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Qi_{calculated} - Qi_{measured})^2}{n}} \quad (2)$$

Here $Qi_{Calculated}$: Calculated discharge rate, $Qi_{mesasured}$: Measured discharge, n: Number of data.

In order to assess the effectiveness of the model, a second criterion, namely the determination coefficient (R^2), was utilized. This coefficient represents the linear regression between the predicted and observed values and can be calculated using the formula provided below. R^2 is a value ranging from 0 to 1, and is commonly used in trend analysis. The closer the R^2 value is to 1, the stronger the correlation or relationship between the two variables.

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Qi_{measured} - Q^{mean}{}_{measured})(Qi_{calculated} - Q^{mean}{}_{calculated})}{\sqrt{\sum_{i=1}^{n}(Qi_{measured} - Q^{mean}{}_{measured})^2 \sum_{i=1}^{n}(Qi_{calculated} - Q^{mean}{}_{calculated})^2}} \right)^2 \quad (3)$$

Here $Qi_{Calculated}$: Calculated discharge, $Q^{mean}{}_{Calculated}$: Average of calculated discharge, $Qi_{Measured}$: Measured discharge, $Q^{mean}{}_{Measured}$: Average of measured discharge, n: Number of data. The third criterion used to measure the model's success, the mean absolute error (MAE) shown in Equation (4), is the mean of the absolute difference between the actual value and calculated value expressed by the formula below. It is a linear mean where all individual differences are weighted equally on the mean [41].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Q_{imeasured} - Q_{icalculated}| \quad (4)$$

Here $Q_{icalculated}$: Calculated discharge, $Q_{imeasaured}$: Measured discharge, n: Number of data.

Results and Discussion

The aim of this study was to predict peak discharge by combining various machine learning models such as RT, SVM, ET, and ANFIS with DWT and EMD signal processing methods. For this purpose, monthly peak discharge data of the Değirmenocağı DOS on the River Zamanti in Seyhan Basin were used as input and Ergenuşağı discharges as output. In the establishment of the model, 80% of data between 1987 and 2015 were used as training data and 20% as test data. Five-fold cross validation was used to eliminate the overfitting problem.

Fig. 5 shows the peak discharge values, divided into subcomponents with the EMD signal process. EMD produced 7 IMFs and one residual series since it does not have any predetermined fundamental functions. Fig. 6 shows db 3, coif 5, and dmey 3 mother

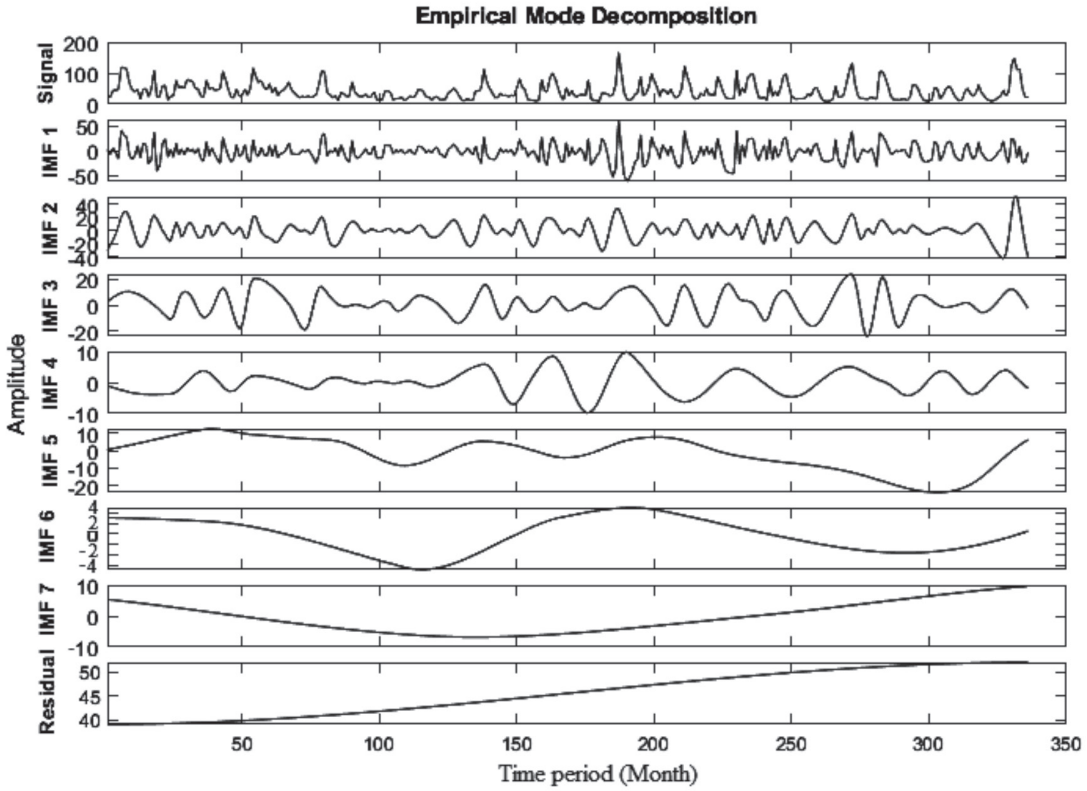

Fig. 5. Peak discharge data of E18A027 Değirmenocağı separated into subsignals by EMD operation.

wavelets, and peak discharge values divided into 3 and 4 decomposition levels. Peak discharge, divided into various details and approximate components, was used to select the input components of the hybrid models by subjecting the time series correlation analysis.

Table 2 shows the correlation coefficients between the input variables divided into subcomponents by the wavelet transform and EMD signal processing methods and the output variable. Hybrid ML models were established by collecting the decomposition variables with a relationship above 0.3 with the output variable. Due to the high predictive success of the model performance, ML models were established with the sum of the highly correlated values. When Table 2 is examined, it is noteworthy that the D2, D2, and D3 detail components and the IMF1, IMF2, and IMF3 decomposition variables have a very high correlation with the outputs. It is also seen that the highest correlation is found in the IMF 1 series. This indicates that sub-signals with high frequency are the most important component in estimating peak discharge.

The parameters with the best results of the hybrid W-ANFIS and EMD-ANFIS models are shown in Table 3. While establishing hybrid ANFIS models, various membership functions such as Trapmf, Gaussmf, Trimf, and 50, 100, 200, and 300 iterations were tried, and the model results giving the lowest error value were selected. Accordingly, 200 iterations and the Trimf membership function generally obtained the most successful results. Here, M2 refers to the ANFIS model formed with input components divided into 2 subsets, ... M5 refers to the ANFIS model created with input components divided into 5 subsets.

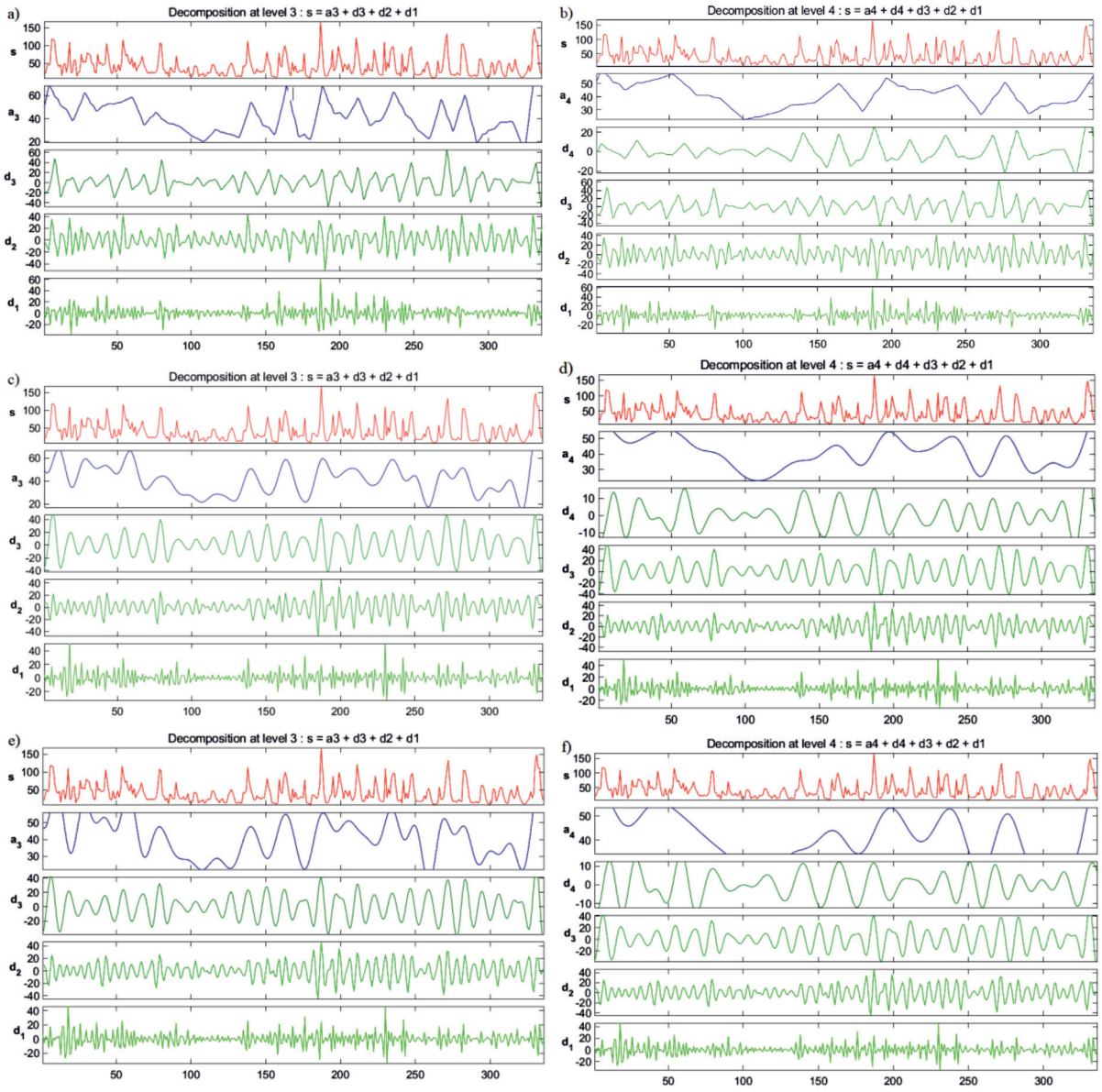

Fig. 6. Separation of E18A027 Değirmenocağı discharge data into subcomponents with a wavelet transform a) db 3 wavelet -3 level, b) db 3 wavelet -4 level, c) coif 5 wavelet - 3 level, d) coif 5 wavelet - 4 level, e) dmey 3 wavelet - 3 level, f) dmey 3 wavelet - 4 level

Table 2. Correlation coefficients of the sub-signals.

Wavelet transform								
DB3-3 Level					DB3- 4 Level			
D1	D2	D3	A3	D1	D2	D3	D4	A4
0.40	**0.40**	**0.39**	0.29	**0.40**	**0.40**	**0.39**	0.22	0.22
Coif 5-3 Level					Coif 5-4 Level			
D1	D2	D3	A3	D1	D2	D3	D4	A4
0.42	**0.31**	**0.45**	0.28	**0.42**	**0.31**	**0.45**	0.18	0.22
Dmey-3 Level					Dmey-4 Level			
D1	D2	D3	A3	D1	D2	D3	D4	A4
0.42	**0.33**	**0.45**	0.26	**0.42**	**0.33**	**0.45**	0.14	0.23
EMD								
	IMF 1	IMF 2	IMF 3	IMF 4	IMF 5	IMF 6	IMF 7	
	0.52	**0.46**	**0.32**	0.13	0.10	0.19	0.01	

Note: Bold characters indicate selected sub-signals.

The automatic selection option of the tree and SVM model parameters used in this study was used by the Matlab regression learner toolbox. The parameters of tree algorithms are generally chosen as minimum leaf size: 8 and 12, Number of learners: 30, Learning rate: 0.1. The parameters of the SVM model are generally kernel scale: 1 and 4, and various kernel types are tested, and the most appropriate kernel function is selected.

Table 4 shows the statistical performance criteria of hybrid models combined with machine learning and various signal processing methods applied to estimate peak discharge values. Statistical parameters such as RMSE, MAE, and R^2 belonging to training, testing, and validation data were evaluated to determine the most successful machine learning model. These parameters, MGSVM and CGSVR algorithms show the highest prediction success. However, it can be deduced that the CGSVR algorithm is somewhat superior. In addition, the FGSVM algorithm provided the worst results in peak discharge prediction.

Taylor diagrams used for the graphical evaluation of the test performance of hybrid ML models constructed are shown in Fig. 7. According to the models established by db 3 wavelet -3 level decomposition in Fig. 7a), it can be deduced that CGSVM and ANFIS M2 models are the most successful models since they have the highest R^2 and lowest error rates and are close to the reference line. According to the models established with db 3 wavelet- 4 level decomposition in Fig. 7b),

Table 3. Membership function and iteration numbers of ANFIS models.

	M2	**M3**	**M4**	**M5**
Db 3-3	Trapmf 200 iterations	Gaussmf 200 iterations	Trimf 300 iterations	Trimf 300 iterations
Db 3-4	Trimf 300 iterations	Trimf 200 iterations	Trimf 200 iterations	Gaussmf 200 iterations
Coif 5-3	Gbellmf 200 iterations	Trimf 200 iterations	Trimf 200 iterations	Gaussmf 300 iterations
Coif 5-4	Gaussmf 300 iterations	Gaussmf 300 iterations	Trimf 200 iterations	Trapmf 200 iterations
Dmey 3	Gaussmf 200 iterations	Trimf 200 iterations	Trimf-200 iterations	Gaussmf 200 iterations
Dmey 4	Trimf 200 iterations	Trimf 200 iterations	Trimf 200 iterations	Trimf 200 iterations
EMD	Gauss2mf 200	Trimf 200 iterations	Gauss2mf 300 iterations	Gaussmf 200 iterations

Table 4. Statistical performance of established hybrid models.

	Stati.	RT			SVM						Ensemble tree		ANFIS			
		FT	MT	CT	LSVM	QSVM	CSVM	FGSVM	MGSVM	CGSVM*	BT	BAT	M2	M3	M4	M5
EMD	MAE	35.52	30.98	27.77	22.36	20.57	21.80	26.09	19.80	**19.56**	26.40	30.71	28.95	29.41	31.59	28.21
	RMSE	53.15	42.74	41.46	31.06	29.88	33.92	39.08	29.67	**29.02**	36.89	42.15	39.19	38.05	41.83	38.75
	R^2	0.64	0.71	0.66	0.74	0.77	0.77	0.59	0.76	**0.77**	0.74	0.72	0.75	0.76	0.73	0.76
Db 3-3	MAE	38.68	30.80	33.21	28.41	25.03	23.68	27.68	**22.41**	23.27	31.81	32.20	27.46	30.81	29.70	27.95
	RMSE	63.13	42.62	46.14	36.23	34.31	32.19	39.74	**31.11**	31.42	46.47	43.15	35.12	39.48	39.05	36.52
	R^2	0.48	0.61	0.58	0.66	0.69	0.72	0.57	**0.75**	0.74	0.59	0.63	0.75	0.68	0.70	0.74
Db 3-4*	MAE	27.48	28.78	26.29	22.08	20.98	21.64	26.45	21.13	**19.54***	19.54	24.49	26.97	27.14	27.11	26.93
	RMSE	42.62	38.95	38.76	31.11	32.28	32.37	44.86	33.34	**28.90***	33.56	35.51	37.38	34.92	34.82	35.13
	R^2	0.65	0.67	0.63	0.76	0.79	0.74	0.49	0.72	**0.78***	0.71	0.68	0.67	0.77	0.76	0.76
Coif 5-3	MAE	37.67	34.53	31.34	28.63	24.82	23.25	26.21	**22.43**	23.51	29.55	31.62	29.20	31.04	30.24	92.29
	RMSE	56.26	47.31	44.17	37.29	34.77	33.61	35.34	**32.16**	32.94	40.56	42.88	38.27	39.95	39.04	101.37
	R^2	0.50	0.61	0.62	0.63	0.68	0.71	0.68	**0.73**	0.71	0.63	0.63	0.71	0.67	0.70	0.63
Coif 5-4	MAE	30.32	30.52	33.05	24.02	22.61	169.82	26.65	21.36	**21.30**	28.03	30.18	28.25	28.61	30.06	28.43
	RMSE	44.00	44.01	44.01	33.98	32.09	216.41	41.73	32.11	**30.81**	40.07	41.07	34.32	36.27	36.02	37.44
	R^2	0.54	0.57	0.53	0.70	0.73	0.60	0.56	0.73	**0.75**	0.59	0.62	0.74	0.70	0.71	0.68
Dmey 3	MAE	38.45	29.09	27.25	26.36	21.45	67.47	22.93	**20.52**	21.05	27.44	30.28	27.34	28.37	27.88	27.77
	RMSE	56.23	40.82	38.98	33.69	30.20	77.30	32.58	**28.64**	29.43	40.87	43.74	35.11	35.90	35.81	36.11
	R^2	0.68	0.75	0.66	0.70	0.76	0.70	0.71	**0.79**	0.78	0.75	0.75	0.76	0.75	0.76	0.77
Dmey 4	MAE	28.33	27.95	33.14	24.41	22.55	23.05	24.46	21.74	**21.04**	23.32	26.65	29.41	29.86	29.13	28.90
	RMSE	43.39	42.33	45.09	35.01	32.61	35.56	39.96	33.23	**31.53**	36.91	38.29	35.61	36.33	35.46	37.98
	R^2	0.56	0.58	0.50	0.69	0.72	0.71	0.63	0.71	**0.74**	0.64	0.64	0.70	0.68	0.71	0.68

Note: The *mark indicates the best model.

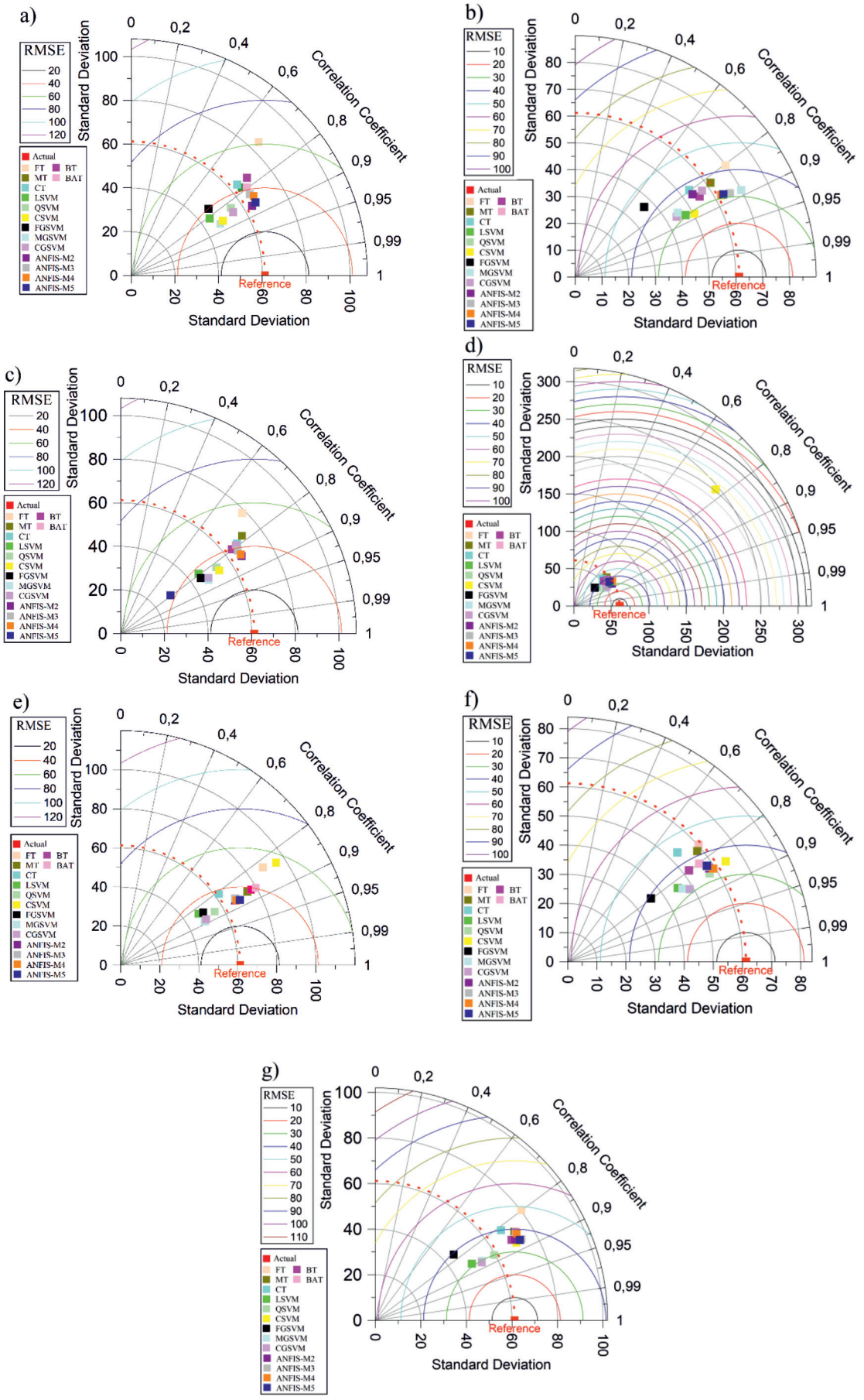

Fig. 7. Taylor diagrams of established hybrid ML models, a) db 3 wavelet -3 level, b) db 3 wavelet -4 level, c) coif 5 wavelet - 3 level, d) coif 5 wavelet - 4 level, e) dmey 3 wavelet - 3 level, f) dmey 3 wavelet - 4 level, g) EMD-ML model.

CSVM is the most effective model. The most successful models in Figs 7(c-e) are CGSVM and MGSVM. Fig. 6f) shows that the most successful models are ANFIS-M3 and MGSVM. Fig. 7g) is CGSVM and QSVM.

The Taylor diagrams of the test data set used to graphically express the predictive power of the models used for peak values are presented in Fig. 7. According to the Taylor diagrams of the test data, CGSVM and MGSVM have the lowest RMSE and highest R^2 values according to the validation chart. In addition, the standard deviation value of the ANFIS-M 5 model is the most successful model since it is closer to the reference data than other models (Fig. 7).

This study aims to predict peak discharge by combining various signal decomposition processes and machine learning models. For this purpose, upstream data as input and downstream data as output are presented to the model in Zamanti River. As a result of the study, it was concluded that combining various signal decomposition processes with machine learning models are black box models that allow effective and reliable use of peak discharge prediction. The study's results overlap to a large extent in terms of producing effective results in peak discharge forecasting with the hybrid models [2, 42]. Zhu et al. [2] used SVR, EMD-SVR, and DWT-SVR models to estimate streamflow over the Jinsha River in China's upper reaches of the Yangtze River. As a result, it has been determined that both EMD and DWT increase the accuracy of streamflow estimation. DWT is a more efficient signal processing method than the EMD model. The study is compatible with Zhu et al. [2] study in obtaining the most effective model with wavelet transform. Gaussian Process Regression (GPR), Ensemble EMD, and WT method predict the daily river stage-discharge relationship. As a result, it has been determined that the prediction performance of the signal processing techniques is the best.

Alizadeh [43] found that estimating river stage discharge from various systems learning systems and EMD, WT signal decomposition, and mutual information methods should be used together. The result of the study coincides with the study of Alizadeh [43]. Rezaie-Balf [44] in Iran and South Korea used daily river flow forecasting machine learning and EMD techniques. It has been found that EMD-ML methods produce successful outputs. The results of the study are in agreement with Rezaie-Balf [44]. Ghalkhani et al. [45] divided the flow series into a trend and several stationary components, and each sub-sequence prediction model was created with the LSTM model. The estimation results of subseries from LSTM and Radial Basis Function (RBF) showed better statistical performance than the EEMD-LSTM series. Nikoo et al. [46] used Support vector machine and wavelet packet (WPSVM), adaptive neuro-fuzzy inference system and wavelet packet (WPANFIS), and artificial neural network and wavelet packet (WPANN) as three hybrid models which were analyzed to predict river stage and the evaluation of their performance. It was determined that WPANFIS gave the best results. The results of Yuan [29] and Seo et al. [23] in the literature are mainly in line with the presented research. When all the outputs are evaluated together, it was concluded that the hybrid models established using the EMD and WT methods offer the highest performance level in peak discharge estimation. The EMD generally produces slightly more successful assessments than the WT method. However, the most effective estimate was obtained with the db 3 wavelet.

Conclusion

This study aims to model peak flow data, which is divided into sub-components by two different signal processes, DWT and EMD, with various machine learning approaches in the Seyhan basin, which has a mountainous structure and a semi-arid climate. The main results of the study are listed as follows:

- In general, the success order of hybrid machine learning models was found as SVM>ANFIS>ET>RT.
- Although EMD-ML models generally produced better prediction results than W-ML models, the most successful peak discharge prediction was obtained with the Db 3 wavelet-ML model.
- Db 3 wavelet and 4 levels CGSVM model with RMSE (28.90 m^3/s), MAE (19.54 m^3/s), and R^2 (0.78) values was chosen as the best model for peak discharge estimation.
- Among dmey, db 3 coif 5 wavelets, the most effective signal separation was done with db 3.
- Among the subcomponents separated by EMD, high-frequency sub-signals were the most effective in estimating peak discharge.
- The findings of this study are expected to be useful for governments and other stakeholders involved in water infrastructure development, as well as in minimizing the loss of life and property in the region, reducing environmental damage, and managing flood risks. Furthermore, the flood forecasting model developed in this study could enable effective early warning, awareness raising, and preparation efforts, as well as successful implementation of flood risk management strategies.

It was determined that hybrid models built with the combination of artificial intelligence and signal processing techniques will produce satisfactory results in peak discharge forecasting. A major limitation of this work is that only two different DOS's are used. Moreover, given the interconnectedness of various areas that can be integrated and monitored from a central location, it is suggested that hybrid models developed for other regions could potentially yield positive results and help predict the effects of floods. For this reason, it is foreseen that many lives can be saved with proactive

measures by giving advance notice and informing the relevant units.

It is suggested that deep learning methods can be used together with signal operations to increase the success of peak discharge prediction in future studies.

Acknowledgments

The data used in the study were obtained from the General Directorate of State Hydraulic Works Rasatlar Branch Office and DSI Regional Directorates. Thanks to DSI for providing the data. Its online address is https://www.dsi.gov.tr/Sayfa/Detay/744. Figures 2 and 3 was obtained from tarimorman.gov.tr, Seyhan Havzası Taşkın Yönetim Planı. Its online address is https://www.tarimorman.gov.tr/SYGM/Belgeler/Ta%C5%9Fk%C4%B1n%20Y%C3%B6netim%20Planlar%C4%B1/Seyhan_Yonetici_Ozeti_v3.pdf and date of access is 15.06.2022

Conflict of Interest

The authors declare no conflict of interest.

References

1. DEHGHANI M., SAGHAFIAN B., RIVAZ F., KHODADADI A. Monthly stream flow forecasting via dynamic spatio-temporal models. Stochastic environmental research and risk assessment, **29** (3), 861, **2015**.
2. ZHU S., ZHOU J., YE L., MENG C. Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. Environmental Earth Sciences, **75** (6), 1, **2016**.
3. DANSO-AMOAKO E., SCHOLZ M., KALIMERIS N., YANG Q., SHAO J. Predicting dam failure risk for sustainable flood retention basins: A generic case study for the wider Greater Manchester area. Computers, Environment and Urban Systems, **36** (5), 423, **2012**.
4. HASSANVAND M.R., KARAMI H., MOUSAVI S.-F. Investigation of neural network and fuzzy inference neural network and their optimization using meta-algorithms in river flood routing. Natural Hazards, **94** (3), 1057, **2018**.
5. PANT R., THACKER S., HALL J.W., ALDERSON D., BARR S. Critical infrastructure impact assessment due to flood exposure. Journal of Flood Risk Management, **11** (1), 22, **2018**.
6. JONKMAN S., VRIJLING J. Loss of life due to floods. Journal of Flood Risk Management, **1** (1), 43, **2008**.
7. YUAN X., ZHANG X., TIAN F. Research and application of an intelligent networking model for flood forecasting in the arid mountainous basins. Journal of Flood Risk Management, **13** (3), e12638, **2020**.
8. YASEEN Z.M., EL-SHAFIE A., JAAFAR O., AFAN H.A., SAYL K.N. Artificial intelligence based models for stream-flow forecasting: 2000-2015. Journal of Hydrology, **530**, 829, **2015**.
9. KUNDZEWICZ Z.W., KANAE S., SENEVIRATNE S.I., HANDMER J., NICHOLLS N., PEDUZZI P., SHERSTYUKOV B. Flood risk and climate change: global and regional perspectives. Hydrological Sciences Journal, **59** (1), 1, **2014**.
10. ŞEN Z., KHIYAMI H.A., AL-HARTHY S.G., AL-AMMAWI F.A., AL-BALKHI A.B., AL-ZAHRANI M.I., AL-HAWSAWY H.M. Flash flood inundation map preparation for wadis in arid regions. Arabian Journal of Geosciences, **6** (9), 3563, **2013**.
11. XIE K., OZBAY K., ZHU Y., YANG H. Evacuation zone modeling under climate change: A data-driven method. **2017**.
12. PITT M. Learning lessons from the 2007 floods. Pitt Review, **2008**.
13. LOHANI A.K., GOEL N.K., BHATIA K. Improving real time flood forecasting using fuzzy inference system. Journal of hydrology, **509**, 25, **2014**.
14. MOSAVI A., OZTURK P., CHAU K.W. Flood prediction using machine learning models: Literature review. Water, **10** (11), 1536, **2018**.
15. LIU C., GUO L., YE L., ZHANG S., ZHAO Y., SONG T. A review of advances in China's flash flood early-warning system. Natural hazards, **92** (2), 619, **2018**.
16. ZOUNEMAT-KERMANI M., MATTA E., COMINOLA A., XIA X., ZHANG Q., LIANG Q., HINKELMANN R. Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects. Journal of Hydrology, **588**, 125085, **2020**.
17. DAZZI S., VACONDIOR., MIGNOSA P. Flood stage forecasting using machine-learning methods: a case study on the Parma River (Italy). Water, **13** (12), 1612, **2021**.
18. DAWSON C., WILBY R. Hydrological modelling using artificial neural networks. Progress in physical Geography, **25** (1), 80, **2001**.
19. TALEI A., CHUA L.H. Influence of lag time on event-based rainfall-runoff modeling using the data driven approach. Journal of hydrology, **438**, 223, **2012**.
20. NOURY M., SEDGHI H., BABAZEDEH H., FAHMI H. Urmia lake water level fluctuation hydro informatics modeling using support vector machine and conjunction of wavelet and neural network. Water Resources, **41** (3), 261, **2014**.
21. GRANATA F., GARGANO R., DE MARINIS G. Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model. Water, **8** (3), 69, **2016**.
22. KASIVISWANATHAN K.S., HE J., SUDHEER K.P., TAY J.H. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. Journal of hydrology, **536**, 161, **2016**.
23. SEO Y., KIM S., KISI O., SINGH V.P., PARASURAMAN K. River stage forecasting using wavelet packet decomposition and machine learning models. Water Resources Management, **30** (11), 4011, **2016**.
24. YASEEN Z.M., EBTEHAJ I., BONAKDARI H., DEO R.C., MEHR A.D., MOHTAR W.H.M.W., SINGH V.P. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. Journal of Hydrology, **554**, 263, **2017**.
25. SHAFIZADEH-MOGHADAM H., VALAVI R., SHAHABI H., CHAPI K., SHIRZADI A. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. Journal of environmental management, **217**, 1, **2018**.

26. CHOI C., KIM J., HAN H., HAN D., KIM H.S. Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea. Water, **12** (1), 93, **2019**.
27. CHOUBIN B., MORADI E., GOLSHAN M., ADAMOWSKI J., SAJEDI-HOSSEINI F., MOSAVI A An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Science of the Total Environment, **651**, 2087, **2019**.
28. ABEDI R., COSTACHE R., SHAFIZADEH-MOGHADAM H., PHAM Q.B. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto International, 1, **2021**.
29. YUAN R., CAI S., LIAO W., LEI X., ZHANG Y., YIN Z., XU Y. Daily runoff forecasting using ensemble empirical mode decomposition and long short-term memory. Frontiers in Earth Science, **9**, 621780, **2021**.
30. FORESTRY M.O.A.A. Seyhan basin flood management plan." https://www.tarimorman.gov.tr/SYGM/Belgeler/Ta%C5%9Fk%C4%B1n%20Y%C3%B6netim%20Planlar%C4%B1/Seyhan_Yonetici_Ozeti_v3.pdf. (Recived: 4.7.2022). **2020**.
31. ZADEH L., Fuzzy sets''. Information and Control, **8**, 338, **1965**.
32. TAKAGI T., SUGENO M. Fuzzy identification of systems and its applications to modeling and control. IEEE transactions on systems, man, and cybernetics, **1985** (1), 116, **1985**.
33. JANG J.-S. ANFIS: adaptive-network-based fuzzy inference system. IEEE transactions on systems, man, and cybernetics, **23** (3), 665, **1993**.
34. VAPNIK V.N. Statistical learning theory. Adaptive and learning systems for signal processing communications and control, **1998**.
35. QUINLAN J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 302, **1993**.
36. SHAMIM A., HUSSAIN H., SHAIKH M.U. A framework for generation of rules from decision tree and decision table. in 2010 International Conference on Information and Emerging Technologies. IEEE. **2010**.
37. BREIMAN L., Classification and regression trees. **2017**: Routledge.
38. QUINLAN J.R. Induction of decision trees. Machine learning, **1** (1), 81, **1986**.
39. HUANG N.E., SHEN Z., LONG S.R., TUNG C.C., LIU H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. A, **454**, 903, **1998**.
40. MOSAVI A., BATHLA Y., VARKONYI-KOCZY A. Predicting the future using web knowledge: state of the art survey. in International conference on global research and education. Springer, **2017**.
41. CHAI T., DRAXLER R.R. Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific Model Development Discussions, **7** (1), 1525, **2014**.
42. ROUSHANGAR K., CHAMANI M., GHASEMPOUR R., AZAMATHULLA H.M., ALIZADEH F. A comparative study of wavelet and empirical mode decomposition-based GPR models for river discharge relationship modeling at consecutive hydrometric stations. Water Supply, **21** (6), 3080, **2021**.
43. ALIZADEH F., FAREGH GHARAMALEKI A., JALILZADEH R. A two-stage multiple-point conceptual model to predict river stage-discharge process using machine learning approaches. Journal of Water and Climate Change, **12** (1), 278, **2021**.
44. REZAIE-BALF M., KIM S., FALLAH H., ALAGHMAND S. Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: Application on the perennial rivers in Iran and South Korea. Journal of Hydrology, **572**, 470, **2019**.
45. GHALKHANI H., GOLIAN S., SAGHAFIAN B., FAROKHNIA A., SHAMSELDIN A. Application of surrogate artificial intelligent models for real-time flood routing. Water and Environment Journal, **27** (4), 535, **2013**.
46. NIKOO M., RAMEZANI F., HADZIMA-NYARKO M., NYARKO E.K., NIKOO M. Flood-routing modeling with neural network optimized by social-based algorithm. Natural Hazards, **82** (1), 1, **2016**.
47. YUAN R., CAI S., LIAO W., LEI X., ZHANG Y., YIN Z., XU Y. Daily runoff forecasting using ensemble empirical mode decomposition and long short-term memory. Frontiers in Earth Science, **9**, 621780, **2021**.