

Received August 23, 2019, accepted September 15, 2019, date of publication September 24, 2019, date of current version October 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943515

Non-Linear Input Variable Selection Approach Integrated With Non-Tuned Data Intelligence Model for Streamflow Pattern Simulation

SINAN JASIM HADI¹, S. I. ABBA², SAAD SH. SAMMEN³, SINAN Q. SALIH⁴,
NADHIR AL-ANSARI⁵, AND ZAHER MUNDHER YASEEN⁶

¹Department of the Real Estate Development and Management, Ankara University, 06100 Ankara, Turkey

²Department of Physical Planning Development, Maitama Sule University Kano, Kano 700221, Nigeria

³Department of Civil Engineering, College of Engineering, Diyala University, Diyala Governorate, Iraq

⁴Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

⁵Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden

⁶Sustainable Developments in Civil Engineering Research Group, Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Corresponding author: Zaher Mundher Yaseen (yaseen@tdtu.edu.vn)


ABSTRACT Streamflow modeling is considered as an essential component for water resources planning and management. There are numerous challenges related to streamflow prediction that are facing water resources engineers. These challenges due to the complex processes associated with several natural variables such as non-stationarity, non-linearity, and randomness. In this study, a new model is proposed to predict long-term streamflow. Several lags that cover several years are abstracted using the potential of Extreme Gradient Boosting (XGB) then after the selected inputs variables are imposed into the predictive model (i.e., Extreme Learning Machine (ELM)). The proposed model is compared with the stand-alone schema in which the optimum lags of the variables are supplied into the XGB and ELM models. Hydrological variables including rainfall, temperature and evapotranspiration are used to build the model and predict the streamflow at Goksu-Himmeti basin in Turkey. The results showed that XGB model performed an excellent result in which can be used for predicting the streamflow pattern. Also, it is clear from the attained results that the accuracy of the streamflow prediction using XGB technique could be improved when the high number of lags was used. However, the implementation of the XGB is tree-based technique in which several issues could be raised such as overfitting problem. The proposed schema XGBELM in which XGB approach is selected the correlated inputs and ranking them according to their importance; then after, the selected inputs are supplied into the ELM model for the prediction process. The XGBELM model outperformed the stand-alone schema of both XGB and ELM models and the high-lagged schema of the XGB. It is important to indicate that the XGBELM model found to improve the prediction ability with minimum variables number.

INDEX TERMS Correlated variables, non-linear XGB approach, extreme learning machine, streamflow simulation.

I. INTRODUCTION

Scientific studies have evidenced the complex nature of streamflow pattern due to the association to natural variabilities (such as their randomness, non-stationarity, complex nature, and non-linearity) [1]. In the field of hydrology, several efforts have been dedicated to the enhancement of the accuracy and reliability of predicting hydrological variables [2]–[6]. Until now, several hydro-meteorological

studies have been performed but no single approach has been identified as applicable in modeling hydrological events under different conditions, especially with different catchment features due to certain physical processes such as the periodicity or randomness of the methods, the existing patterns in the model data, as well as the natural stochasticity that often describes streamflow datasets [7], [8]. Based on this perspective, there is currently no widely accepted model which can outperform other models in different hydrological conditions. The generation of a consistent prediction using several models may not be achievable due to the dynamic

The associate editor coordinating the review of this manuscript and approving it for publication was Weiping Ding .

nature non-stationarity of historical data. This has made it necessary that researchers should develop stronger and efficient models using the available historical data [9]. Furthermore, the advantage of the rapidly evolving computational models which can improve the accuracy of modeling methodologies must be considered as well [10]–[12]. Newly developed data intelligence methodologies should be investigated in the development of these robust and efficient prediction models.

To have a perfect understanding of water resource management, it is necessary to understand the hydrological processes that govern streamflow pattern and its phenomenon [13], [14]. The past few decades have witnessed several studies on streamflow phenomenon as a result of the interests in studying both regional and global pattern of hydrologic changes that cause flooding and drought [15]–[17]. According to the literature, the streamflow pattern is modeled using two main approaches: (i) using physically based models such as models that deploy partial differential equations, and (ii) using artificial intelligence (AI) based statistical models such as soft computing methods [14]. However, several studies need to be done and especially on the hydrological variables that will provide the initial and boundary criteria required for the simulation of the elemental processes of a given watershed using the physically based models.

Based on the established streamflow simulation studies, classical regression tools have been widely used but they have been generally associated with low accuracy levels, giving room to the development of the AI methods which are considered as accurate and non-linear hydrologic tools. Based on several comprehensive review researches in the field of hydrology, numerous AI models explored for streamflow simulation such as support vector machine (SVM), artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), complementary wavelet-AI model, and hybrid evolutionary computing models [8], [15]–[18]. However, there is still a notable level of challenges associated with these models in terms of their generalization and implementation as an expert system for river engineering sustainability. Such shortcomings include their time-inefficiency and non-automated modeling process. Therefore, it has become necessary to design a universally applicable automated AI model which can be implemented over different local scales.

The extreme learning machine (ELM) was recently developed as a new learning approach whose major advantage is in its ability to map the internal features without the need to iteratively tune the parameters of the hidden neuron as required in a traditional ANN model [19]. The input and hidden neuron weightings are computed randomly in the ELM from several pre-assigned neurons without having to pass through all the neurons in the model [20]. Furthermore, the generalization capability of the ELM is acceptable, and it requires less computation time [14].

By recalling the recent hydrological studies on the streamflow modeling using the feasibility of the ELM mode

approach have produced optimistic results. Reference [21] made the first attempt on the implementation of using ELM algorithm to model streamflow as a predictive machine learning model (unorganized) for the capturing of the non-linearity of hydrological problems in seasonal streamflow data sourced from the Brazilian hydropower plant. From the results, scholars observed a significant finding with respect to streamflow time series modeling which ushered in the exploration of the suitability of this approach as a hydrological tool. Several other attempts were made by other researchers, such as who deployed a new predictive model on two reservoir monthly inflow in China and reported excellent prediction skills based on the modeling results which were validated using SVR (one of the dominant AI models).

Another study conducted on the integration of ELM model with binary-coded swarm optimization (BCSO) based on implicit base flow separation streamflow prediction [22]. During the modeling process, the digital filter of the input variable was optimized using the BCSO method and the results showed significant base-flow process underestimation due to the inefficiency of the reflected geographical information of the investigated watershed. A conventional ELM model was developed by [23] for hydrological time series prediction in the USA based on the daily time series information. However, the ELM model used in this study showed a slight difference in accuracy compared to an evolutionary computing SVR model. Another version of the ELM called online sequential ELM (OS-ELM) was recently proposed for the simulation of different streamflow pattern scales in Canada by [24]. From the findings, there was a significant level of differences in the forecasting accuracy compared to multi-linear regression (MLR) model which served as a benchmark model. The OS-ELM was also applied for the forecasting of flooding events in Neckar River, Germany, by [25], where the hourly meteorological dataset was used for the construction and evaluation of the OS-ELM. A non-tuned ELM-based predictive model was proposed by for the modeling of the monthly time scale streamflow data in Iraq [26]. Upon the verification of the proposed model against support vector regression and generalized regression neural network models, the proposed ELM model showed an effective predictive model for monthly time-scale over the semi-arid region. Reference [27] studied the potential of ELM model for streamflow prediction in Queensland, Australia using a set of nine predictors to investigate the seasonality of discharge water level. The predictors used in the study were rainfall; Pacific Decadal Oscillation Index; Indian Ocean Dipole Index; Southern Oscillation Index; ENSO Modoki Index; Nino 3.0, Nino 3.4, and Nino 4.0 sea surface temperatures (SSTs). The ELM model did not only have a better accuracy compared to the ANN, but it was also faster than the ANN model by several folds, showing its suitability as a tool for modeling real-time streamflow. Reference [28] performed another study on the ELM application to forecast the daily time scale (of tropical environment) of Johor River located in Malaysia. Research finding evidenced the capacity of ELM

model on this region. Reference [29] developed a hybrid ELM coupled with wavelet denoising method to forecast streamflow based on multi-station information. The performance of the proposed hybrid model compared to that of least square SVM (LSSVM) model and the results were reported to be good. On this premise, an attempt towards the integration of ELM model with the genetic algorithm (GA) (an input selection algorithm) has been made by [30] for the forecasting of the monthly streamflow at the Ajichai Basin. The outcome of the study showed that the integration of ELM with GA enhanced the forecasting accuracy of the ELM model. All the forgoing researches evidenced the capability of the ELM model over the other AI models in simulating streamflow pattern over multiple locations all around the globe.

As a fact, the non-linear relationships between the simulated parameters and the estimators can be established by the conceptually based methods that depend on historical input data without the need for a previous information about the nature of the flow. Such models are beneficial because they require fewer hydrological inputs [31]. The steps in the AI-based models include the issue analysis, data collection and pre-processing, data-driven model selection, optimal model identification from the list of trained models, and final model evaluation. Among these steps, the most vital step is the identification of the data-driven model because it is the stage for the implementation of the learning process and feature extraction; the approximation of the optimal model is achieved by minimizing the training error between the target and forecasted matrix, while the optimal model is selected from the list of trained models (model with the lowest mean square error is selected) based on an independent set of validation processes. Several factors which can affect real hydrological conditions can also affect the approximation accuracy of AI models. Such factors include the model input determination, the time scale or forecast horizon, as well as the model configuration.

It can be observed from the reported literature, most of the research established using linear statistical approaches (e.g., auto-correlation function (ACF) and partial auto-correlation function (PACF)) to abstract the correlated lags of the historical data time series [26], [32]. Indeed, this is the main drawback as the actual relationship between those correlated lags and the targeted step is characterized by non-linear relationship; hence, this problem needs a serious attention by scholars to be addressed. Using such a method like ACF or PACF leads to choosing the successive lags in predicting any time series, this could not be the case in machine learning models. In these two methods, the correlation used is linear while the relationship between the input and output variables is non-linear in which some of the non-linearity could be explained. Thus, exploring a new approach where the essential correlated lag times are abstracted using non-linear mechanism is highly needed. The objective of current study is to use a hybrid model consists of XGB and ELM for modeling monthly scale streamflow pattern. XGB approach is non-linear tree-based model that has the ability to abstract the significant correlated

attributes for the prediction matrix (i.e., lags time in this case) [33], [34]. According to the significant of the selection of the appropriate inputs “input attributes for the prediction matrix” in sequential manner [35], the current research is devoted on the integration of non-linear input selection with non-tuned learning model for modeling hydrological problem. The proper selection of the relevant candidates has been proved scientifically influencing the prediction capacity for such a complex non-linear and non-stationary problems [36]. The major motivation of this study is to inspect the potential of XGB approach for selecting the most significant related attributes to the targeted variable. For this purpose, the hybrid model is established to improve the prediction performance and skills of the non-tuned data intelligence model for mimicking the streamflow pattern.

II. METHODS AND MODELING DEVELOPMENT

A. EXTREME GRADIENT BOOSTING (XGB)

XGB as a novel non-linear supervised learning approach, was first developed by [37]. It has the computational efficiency, speed learning ability and capable processing of the Gradient tree boosting algorithm developed by [38]. The ensemble learning of decision three was applied in the XGB algorithms and can be used for both regression and classification [39]. However, since the pronouncement of this new algorithm to the best authors’ knowledge there is no published research indicating the application of XGB in streamflow forecasting in general and as essential means of input section with a wavelet. Due to the outstanding distinctive features, this algorithm is employed in this study both in modeling and selection process.

The algorithm is attributed to the set of classification and regression trees (CART) as it differs from decision tree (DT), in that each of the leaves has a real score that helps in richer interpretation that beyond the classification which is the case in DT. In CART, a single tree is used which is incapable and weak to some extent, hence, an ensemble of multiple trees is proposed. For a sample of training data set (with multiple feature) x_i (i.e. lagged downstream, upstream, rainfall, temperature, and potential evapotranspiration in this study) to predict a target variable y_i (i.e., one month ahead downstream), the mathematical expression of the ensemble of k number of trees can be written as [40]:

$$\hat{y} = \sum_{k=1}^K f_k(x_i), \quad f_k \in F \quad (1)$$

where f_k is a function in a set of all likely function in CARTs, F the set of these functions. Optimization is the major objective of this function as [40]:

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (2)$$

where l and Ω part is the training loss and regularization function, respectively with the latter being used to control the complexity and prevent the overfitting.

For the purpose of training a unary function for training is consider, because it is a bit hard to train all the three function simultaneously (f_i). Hence the prediction value at step t as \hat{y}_i^t lead to [40]:

$$\hat{y}_i^t = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i) \quad (3)$$

After the substitution of the new predicted values from Eq. (3), the performance of every single iteration is accomplished by the addition of one tree and the chosen tree is the one that minimizes the objective function in Eq. (4). Keeping in mind that, the mean squared errors (MSE) as the loss function, the objective function becomes [33]:

$$obj^t = \sum_{i=1}^n (y_i - (\hat{y}_i^{t-1} + f_t(x_i)))^2 + \sum_{i=1}^t \Omega(f_i) \quad (4)$$

which can also be in a form of [33]:

$$obj^t = \sum_{i=1}^n [2(\hat{y}_i^{t-1} - y_i)f_t(x_i) + f_t(x_i)^2 + \Omega(f_t) + constant] \quad (5)$$

Generally, for the function in form of first order or quadratic the MSE is approachable while functions like logistic tend more complicated, thus, Taylor expansion up to the second order is employed [33]:

$$obj^t = \sum_{i=1}^n [l(y_i, \hat{y}_i^{t-1}) + g f_t(x_i) + \frac{1}{2} h f_t^2(x_i) + \Omega(f_t) + constant] \quad (6)$$

where $g_i = \partial_{\hat{y}_i^{t-1}} l(y_i - \hat{y}_i^{t-1})$, and $h_i = \partial_{\hat{y}_i^{t-1}}^2 l(y_i - \hat{y}_i^{t-1})$. Note that, the tree can be defined as $f_t(x) = w_{q(x)}$, and the regularization function is considered as in Eq. (7).

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (7)$$

where w , q and T represents vector of scores on leaves, assigning function of each point in the data to the equivalent leaf and number of leaves, respectively. The objective value of the (6) t^{th} can be written as in Eq. (8) this followed by the addition of Eq. (7) and eliminating the constants from Eq. (7).

$$obj^t \approx \sum_{i=1}^n [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 = \sum_{j=1}^T [l(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \quad (8)$$

where $I_j = \{i | q(x_i) = j\}$ signifies the indices of data points given to the j^{th} leaf. By defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, the objective function is written as:

$$obj^t = \sum_{j=1}^T [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T \quad (9)$$

In order to determine the goodness of any given structure tree $q(x)$, the best w_j and objective function can be written as:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (10)$$

$$obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (11)$$

Meanwhile, computing all the tress options and select the best one is not an inflexible process of solving the problem; hence, Eq. (12) is applied for both optimization of the tree and score gain in splitting a leaf into two leaves. Eq. (12) consists of the score on the new left leaf (L), the score of the new tight leaf (R), the score in the original leaf, and the regularization term. In case of the gain is less than it would be better not to add the branch.

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (12)$$

With the addition of the gain information of the parameter which all together measure a certain parameter for the entire trees network, this algorithm is used to obtain all the feature importance. For the purpose of forecasting the stream flow, in this research the characteristic is applied for proposing a new tool of the important scales in the CWT.

B. EXTREME LEARNING MACHINE (ELM)

Learning in the context of ELM was first proposed by [20] as a relatively emerging system of data-driven models (DDMs) comprises of single layer feed forward neural networks (SLFNs) architecture, but quite different from the traditional neural network. The promising manners of the ELM could be attributed to its generalization ability and fast learning speed. With regards to common neural network (i.e., Back-propagation NN), ELM overcomes the problems of slow learning speed, local minima and overfitting [19]. In the last decades the ELM algorithms have been applied in various field of hydrological modeling by several researchers due to it is high performance ability and precisely for streamflow forecasting such as [41].

In this study, an ELM model was developed using training data set $\{(x_1, y_1), \dots, (x_t, y_t)\}$ where x_t and y_t are the explanatory and response variable, respectively. For this purpose, the explanatory variables (input vector) denoted as x_1, x_2, \dots, x_t defined as the lagged streamflow and the output vector y_1, y_2, \dots, y_t represent the observed one step ahead streamflow. For set of N training samples (i.e. $t = 1, 2, \dots, N$) in which $x_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$, a SLFN with H hidden nodes is mathematically expressed as [20]:

$$\sum_{i=1}^H B_i g_i (\alpha_i x_t + \beta_i) = z_t, \quad (13)$$

where $B \in \mathbb{R}^H$, $Z(z_t \in \mathbb{R})$ and $G(\alpha, \beta, x)$ represents the forecasted weights in the output layer, model output

and activation function of the hidden layer, respectively. While α_i , β_i , i and d signifies the weights of the randomized layers, biases of these randomized layers, the index of the specific node in the hidden layer and the number of inputs, respectively.

As stated above, the study employed activation function using trial and error approach to check the best activation function and found that sigmoid is the best as [42]:

$$G(x) = \frac{1}{1 + \exp(-x)}, \quad (14)$$

The output layer contained a linear transfer function, in an ELM model a proper number of hidden neurons, randomized input layer weights (α), and randomized hidden layer biases (β) can lead to a zero error (Eq. (15)). Therefore, produced the weights of the output layer can be obtained analytically for any training dataset [20]:

$$\sum_{t=1}^N \|z_t - y_t\| = 0, \quad (15)$$

The system of linear equation can be used to obtain the value of B for any input-output training samples:

$$Y = GB \quad (16)$$

in which

$$\begin{aligned} G(\alpha, \beta, x) &= \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_N) \end{bmatrix} \\ &= \begin{bmatrix} g_1(\alpha_1 \cdot x_1 + \beta_1) & \cdots & g_L(w_H \cdot x_1 + \beta_H) \\ \vdots & \dots & \vdots \\ g_1(\alpha_N \cdot x_N + \beta_1) & \cdots & g_L(w_H \cdot x_N + \beta_H) \end{bmatrix}_{N \times H} \end{aligned} \quad (17)$$

and

$$B = \begin{bmatrix} B_1^T \\ \vdots \\ B_H^T \end{bmatrix}_{H \times 1} \quad (18)$$

and

$$Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times 1} \quad (19)$$

where G here known as the hidden layer output, T is the transpose of the matrix. The output weights \hat{B} can be estimated by inverting the matrix of the hidden layer using Moore-penrose generalized inverse function (+):

$$\hat{B} = G^+ Y \quad (20)$$

Eventually, the estimated values \hat{y} (i.e. represents the one month ahead streamflow in this study) can be determined by:

$$\hat{y} = \sum_{i=1}^H \hat{B}_i g_i(\alpha_i \cdot x_t + \beta_i) \quad (21)$$

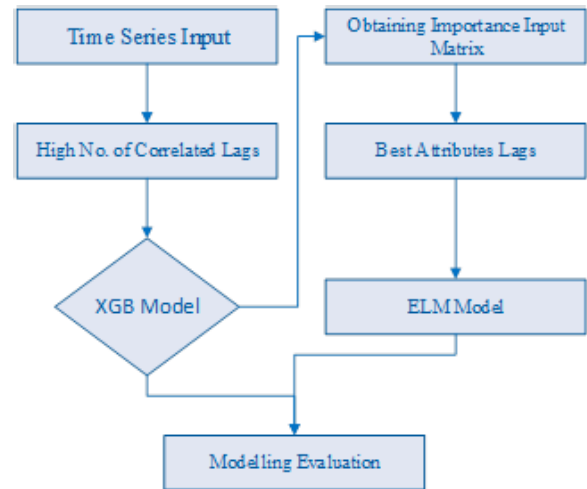


FIGURE 1. The proposed modeling input selection schema for the streamflow according to their importance.

C. PROPOSED MODELING SCHEMA

ELM has been proved as robust non-linear machine learning approach that could be used for modeling complex time series but the problem not only in ELM rather in most of the machine learning techniques is the selection of the inputs as having too much inputs deteriorate the model performance and having less input may not unhide all the hidden information in the time series [41], [36]. XGB is tree-based model that have the ability to show the importance of the inputs used in predicting the output [43]. Therefore; in this paper, a modeling schema is proposed in which the inputs used for streamflow prediction is arranged according to their importance before being used for the prediction process. In other worlds, as can be seen in Figure 1, the XGB is used for ranking the inputs according to their importance in predicting the output, and then these inputs are taking into account according to their rank starting with the most important and ending with the less important input. Accordingly, the first model includes only the most important input and the second model includes the most important and the second most important inputs, and the last model includes all the inputs has been ranked in the important matrix. These models are then used as ELM models and eventually all the models are evaluated, and the inputs of the best performed model chosen as the best inputs.

$$\begin{aligned} m_1 &= im_1 \\ m_2 &= im_1 + im_2 \\ m_n &= im_1 + im_2 \dots + im_n \end{aligned} \quad (22)$$

where m is the model, im is the input, and n is the number of the inputs included in the importance matrix.

D. MODELING DEVELOPMENT

According to autoregression theory, most of the time series data including the hydrologic variable such as the streamflow as the case in this study can be predicted using the lags

TABLE 1. The developed input variables models.

No	Model	Variables
M1	S	Streamflow
M2	SR	Streamflow + Rainfall
M3	SRT	Streamflow + Rainfall + Temperature
M4	SRE	Streamflow + Rainfall + Evapotranspiration
M5	SRTE	Streamflow + Rainfall + Temperature + Evapotranspiration

of the same time series or variables that contributes in the output variables. In order to define the optimum number of lags if the same variable considered, Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) can be used and Cross correlation function (CCF) can be used if other variables are used. In this study, streamflow, rainfall, temperature, and potential evapotranspiration was used as variables/inputs to predict the streamflow one month ahead. The optimum number of lags for every variable was obtained using the ACF, PACF, CCF before imposing these lags as inputs into the ELM and XGB techniques.

Several schemas were applied in this study:

- i. Stand-alone schema in which the optimum lags obtained from the ACF, PACF, and CCF were imposed into ELM and XGB by using several combinations shown in Table 1. Through the manuscript these models will be referred as SELM and SXGB.
- ii. High-lagged schema in which 30 lags of every variables obtained before imposing them into XGB to choose from the ranked inputs (i.e. lagged variables) according to the importance matrix. The model of this schema is referred as XGB.
- iii. Combined XGBELM in which the ranked inputs are imposed into the ELM model. The model of this schema is referred as XGBELM. See the proposed model section for the details of this schema.

The XGB model used in this study has several parameters that needs to be tuned in order to achieve the best results. Therefore, hyper tuning techniques implemented in order to achieve the optimum values of these parameters. The best tuned parameters of gamma ($\Gamma\gamma$) and eta ($H\eta$) are shown in Table 2. The used ELM has only the number of neurons that could affect the results of the prediction. Accordingly, 1 to 50 number of neurons used in ELM that 50 models were developed and evaluated and the number of neurons for the best performed are used as the neurons of the best model.

E. EVALUATION OF PERFORMANCE

Three different evaluation metrics namely, Root Mean Square Error (RMSE), Nash-Sutcliffe (NC), and Mean Absolute Error (MAE) were computed for evaluating the predictability performance of the applied predictive models [44]–[46].

TABLE 2. Best tuned parameters of XGB.

	Max depth	$H\eta$	$\Gamma\gamma$	Min child weight
SXGB	1	0.7	0.1	6
	6	0.1	0.1	7
	9	0.1	0.1	2
	4	0.1	0.1	6
	9	0.1	0.1	6
XGB	2	0.1	0.1	3
	6	0.1	0.1	6
	5	0.1	0.1	5
	5	0.1	0.1	5
	2	0.1	0.1	7

The RMSE is calculated using:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\hat{Q}_i - Q_i)^2}{N}} \tag{23}$$

where Q represents the observed values, \hat{Q} represents the predicted values, and N is the number of examined dataset.

The NC is calculated as follows:

$$NC = 1 - \frac{\sum_{i=1}^N (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^N (Q_i - \bar{Q})^2} \tag{24}$$

where \bar{Q} represents the mean of the observed values. The values of NC vary between $-\infty$ and one.

The MAE is calculated as follows:

$$MAE = \frac{\sum_{i=1}^N e_i}{N} \tag{25}$$

where e_i is the difference between the observed and predicted values: $\hat{Q}_i - Q_i$.

III. CASE STUDY AND DATA DESCRIPTION

The studied basin is located in southwest of Turkey between $36^{\circ}1'33''E - 36^{\circ}40'48''E$ and $37^{\circ}40'9''N - 38^{\circ}40'30''N$ namely Goksu-Himmeti (coded 1801) covers an area of 2400 km^2 was chosen as case study (Figure 2). The elevation in the basin varies between 6670-2880 m. Daily streamflow for the period Feb-1973 to Sep-2000 measured at a station named Goksu-Himmeti which is located in the outlet of the basin was collected from the Ministry of forests and water affairs- the general directory of Water affairs. Daily rainfall and temperature data for the same period measured at a meteorological station (namely Sariz) located in the middle of the basin was collected from the Ministry of forests and water affairs- the general directory of meteorology. The potential evapotranspiration data was also collected for the same period but due to the high number of missing data these data was not sufficient to be used. Therefore, evapotranspiration data were obtained from CruTS data which was assessed for same study area by [47]. The evapotranspiration collected from

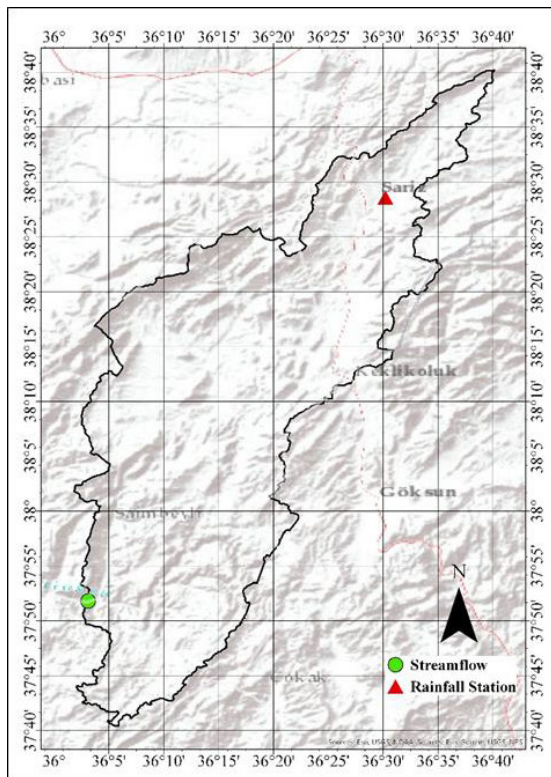


FIGURE 2. The location of the studied catchment.

TABLE 3. Basic descriptive of the statistics for the utilized hydrological information.

Statistics	Stream.	Rainfall	Temp.	Eva.
Min	9.2	0	-8.9	0.6
Max	149.4	187.5	21.9	6.6
Range	140.2	187.5	30.8	6
Median	20	39.5	7.8	2.7
Mean	29.4	44.6	7.4	3
Var	542	1190.4	71.3	3.4
Coef.Var	0.8	0.8	1.1	0.6

CruTS is monthly data while others are daily, therefore all the daily variable were converted into monthly data. Time series of the collected data after the conversion are shown in Figure 3.

The dataset was divided into two subsets 75% for training and 25% for testing to maintain generality. This data division was attained using trial and error procedure for the best predictability performance of the applied methodology. The descriptive statistics of the used data are summarized in Table 3.

IV. APPLICATION RESULTS AND ANALYSIS

This section is reported the application results of the modeling development. A first, the relationship between the recognized lags using ACF and PACF for the streamflow was figured for short-term with 40 month (i.e., about 3 years) and

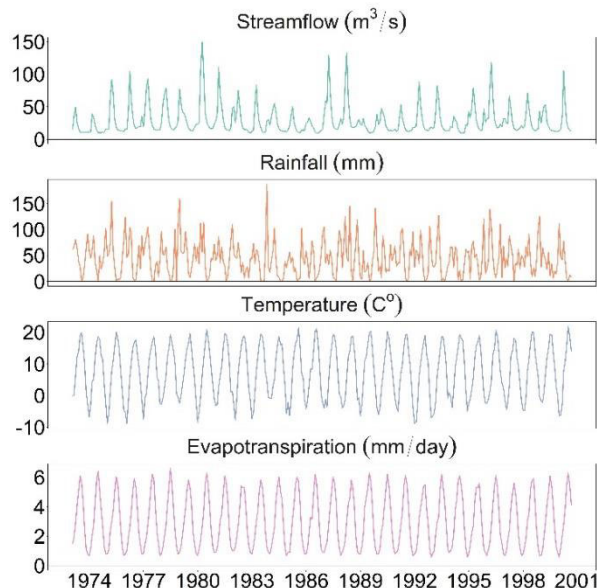


FIGURE 3. The actual time series of the used hydrological data.

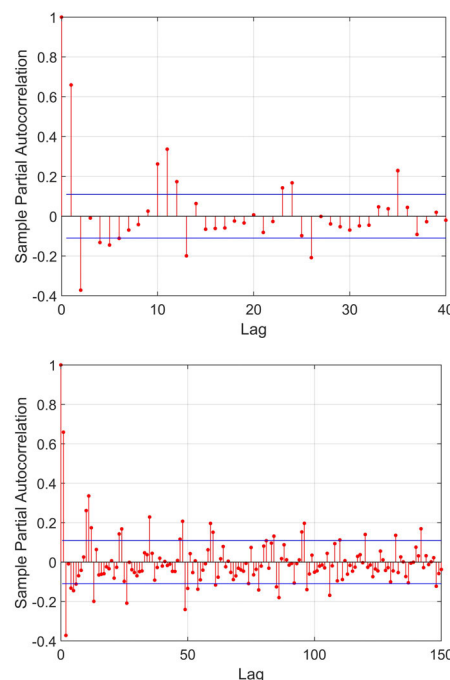


FIGURE 4. Autocorrelation and Partial Autocorrelation function of the streamflow with a) 40 and b) 150 lags.

long-term with 150 months (i.e. about 13 years), as presented in Figure 4. Figure 4 showed that the seasonal correlation is stronger than the successive correlation. According to the ACF statistic of the streamflow, the seasonal correlation was clearly obvious and that is normal due to the nature of the streamflow time series. The prediction of streamflow using the simplest prediction methods such as autoregression or even the seasonal adjusted methods the first lags will be considered as predictands. Figure 4 indicated that 150 lags

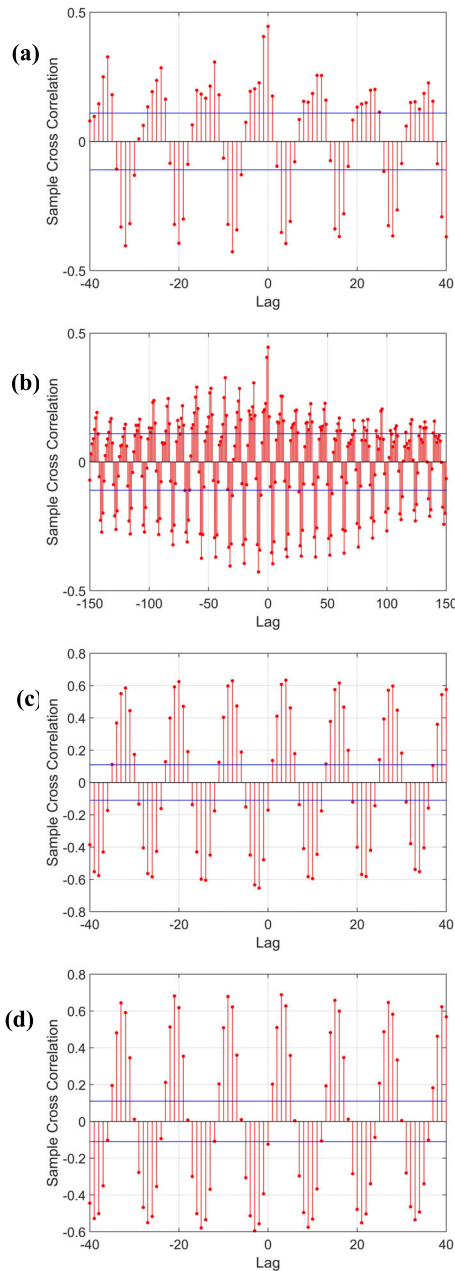


FIGURE 5. The cross-correlation function between streamflow and the other variables: a) cross correlation of streamflow and rainfall with 40 lags, b) cross correlation of streamflow and rainfall with 150 lags, c) cross correlation of streamflow and temperature with 40 lags, d) cross correlation of streamflow and evapotranspiration with 40 lags.

(about 13 years) autocorrelation, lags up to 150 lags should be considered. However, if more lags were used the longer lags would have demonstrated a significant but due to the limitation of the length of the available data only 150 lags used. This is also was observed using the PACF statistic in which the correlation is significant in the annual cycling up to the maximum lags used. Although in the autoregression theory the lags should be considered up to the first cut in the PACF but that does not mean the correlation in the lags after the first cut is neglectable. Hence, the question raised here is

TABLE 4. The statistical performance of the SELM using the optimum normal lags (2, 1, 1, and 1 for streamflow, rainfall, temperature and evapotranspiration respectively). Bold is the best results.

	Training			Testing		
	NC	RMSE (m ³ /s)	MAE (m ³ /s)	NC	RMSE (m ³ /s)	MAE (m ³ /s)
S	0.619	14.858	8.840	0.430	15.801	10.410
SR	0.565	15.888	10.697	0.454	15.475	11.369
SRE	0.800	10.779	6.923	0.658	12.250	8.866
SRT	0.849	9.364	6.468	0.690	11.660	8.724
SRTE	0.838	9.689	6.705	0.704	11.393	8.228

that if all the significantly correlated lags considering, will the predictability performance of the models be improved? The issue of using several inputs deteriorate the performance of all the data intelligence models. Another important issue is the ACF and PACF are linear correlation functions in which the non-linearity could not be explained by using such linear methods. Therefore, in this study a novel approach is implemented to elect the essential lags to impose them as inputs to the DDM.

Considering the cross correlation between the streamflow and the other variables (e.g., rainfall: Figure 5a and b, temperature: Figure 5c, and evapotranspiration: Figure 5d), the correlation was examined for 40 and 150 lags. 150 lags were only examined for the rainfall variable; whereas, the two other variables were almost completely similar in terms of the interpretation. Using the cross correlation led to the choice of the first four significant correlation. On the other, the seasonal cycling showed that the recession of the correlation after these four months. Although this process has gradual decreasing correlation every year; yet, that extends up to the period chosen in this study that 150 lags (i.e., about 13 years). In another words, using the correlation of the CCF in the consideration as a selection criterion, all the lags that have significant correlation should be selected and not only the first four. As an example, using the 150 lags cross correlation function into account, for modeling the streamflow all the significant correlated lags should be used as inputs and this example about 100 lags are significant. Hence, choosing this number of lags as input attributes certainly would deteriorate the model learning process due to the redundant information. This is applicable on the temperature and evapotranspiration as well.

The optimum lags of the variables that used as inputs into the models were obtained using ACF, PACF, and CCF shown in Figure 5. The trial and error by adding one lag every time and evaluate the model. Accordingly, the optimum lags are 2, 1, 1, and 2 for streamflow, rainfall, temperature, and evapotranspiration, respectively. These number of lags were used for the stand-alone schema (i.e. SELM, and SXGB) as inputs based on the combinations listed in Table 1.

The results of the stand-alone schema SELM and SXGB in which the only identified optimum lags of the variables were imposed as inputs predictors and the output is the one month ahead streamflow are shown in Table 4 and Figure 6.

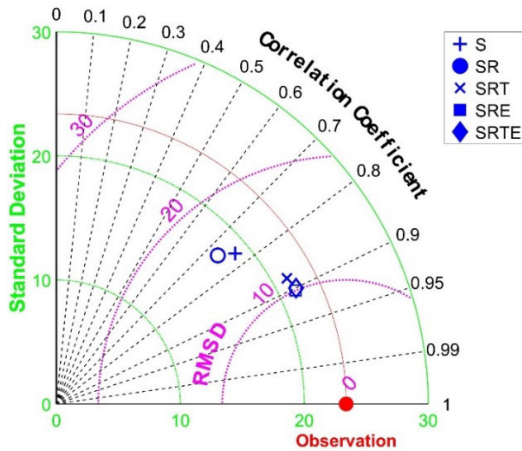


FIGURE 6. Taylor diagram presentation for the SELM model.

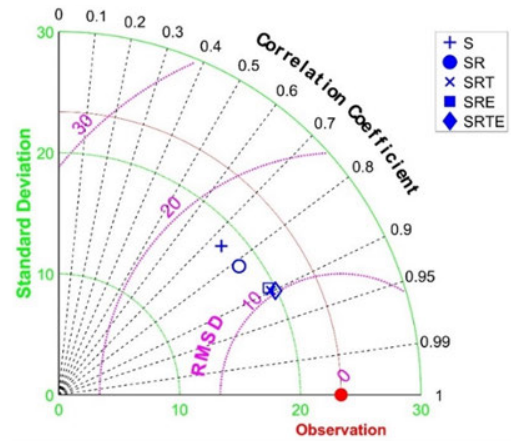


FIGURE 7. Taylor diagram presentation for the SXGB model.

TABLE 5. The statistical performance of the SXGB using the optimum normal lags (2, 1, 1, and 1 for streamflow, rainfall, temperature and evapotranspiration respectively). Bold is the best results.

	Training			Testing		
	NC	RMSE (m ³ /s)	MAE (m ³ /s)	NC	RMSE (m ³ /s)	MAE (m ³ /s)
S	0.601	15.190	10.147	0.311	17.363	12.127
SR	0.742	12.235	8.206	0.352	16.855	11.854
SRE	0.845	9.472	5.941	0.631	12.709	8.449
SRT	0.825	10.090	6.546	0.662	12.170	7.647
SRTE	0.839	9.681	6.345	0.702	11.418	6.900

According to the tabulated results, the SRT combination presenting two lags of the streamflow, one of rainfall and one of temperature are considered as inputs, has the highest prediction performance with (NC = 0.849, RMSE = 9.364 m³/s RMSE and MAE = 6.468 m³/s) over the training subset. For the testing subset, SRTE combination which has a one lag evapotranspiration in addition to the lags of the SRT combination has the highest prediction performance with (NC = 0.704, RMSE = 11.393 m³/s and MAE = 8.228 m³/s). Figure 6 exhibits Taylor diagram graphical presentation indicates the SRT and SRTE combinations has the highest predictability performance.

Table 5 reported the statistical results of the SXGB schema. The best input combination performed over the training phase is SRE (i.e. streamflow, rainfall, and evapotranspiration) with (NC = 0.845, RMSE = 9.472 m³/s and MAE = 5.941 m³/s). On the other hand, the best performed model over the testing subset is SRTE by having (NC = 0.704, RMSE = 11.393 m³/s and MAE = 8.228 m³/s). Considering both subset and examining Taylor diagram shown in Figure 7, SRTE is identified as the best performed combination of the SGXB model. Comparing the best performed combinations in both SELM and SXGB, it is obvious the performance is close to each other with small differences based on the statistical the evaluation measurements. In general, the SELM model has no indication of the overfitting that the training subset has close results to the testing subset while in the

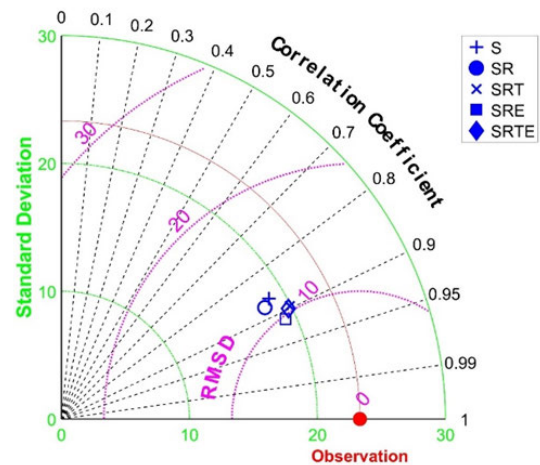


FIGURE 8. Taylor diagram presentation for the XGB model.

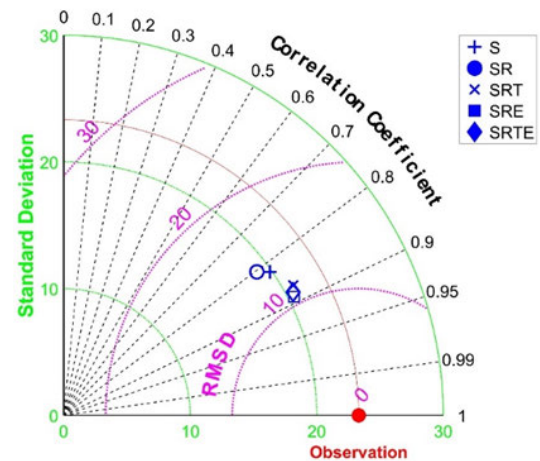


FIGURE 9. Taylor diagram presentation for the XGBELM model.

SXGB model the overfitting is clear and that due to the tree nature of the XGB approach.

After using the stand-alone models, a proposed schema is proposed here by using several lags (i.e. 30 lags in this

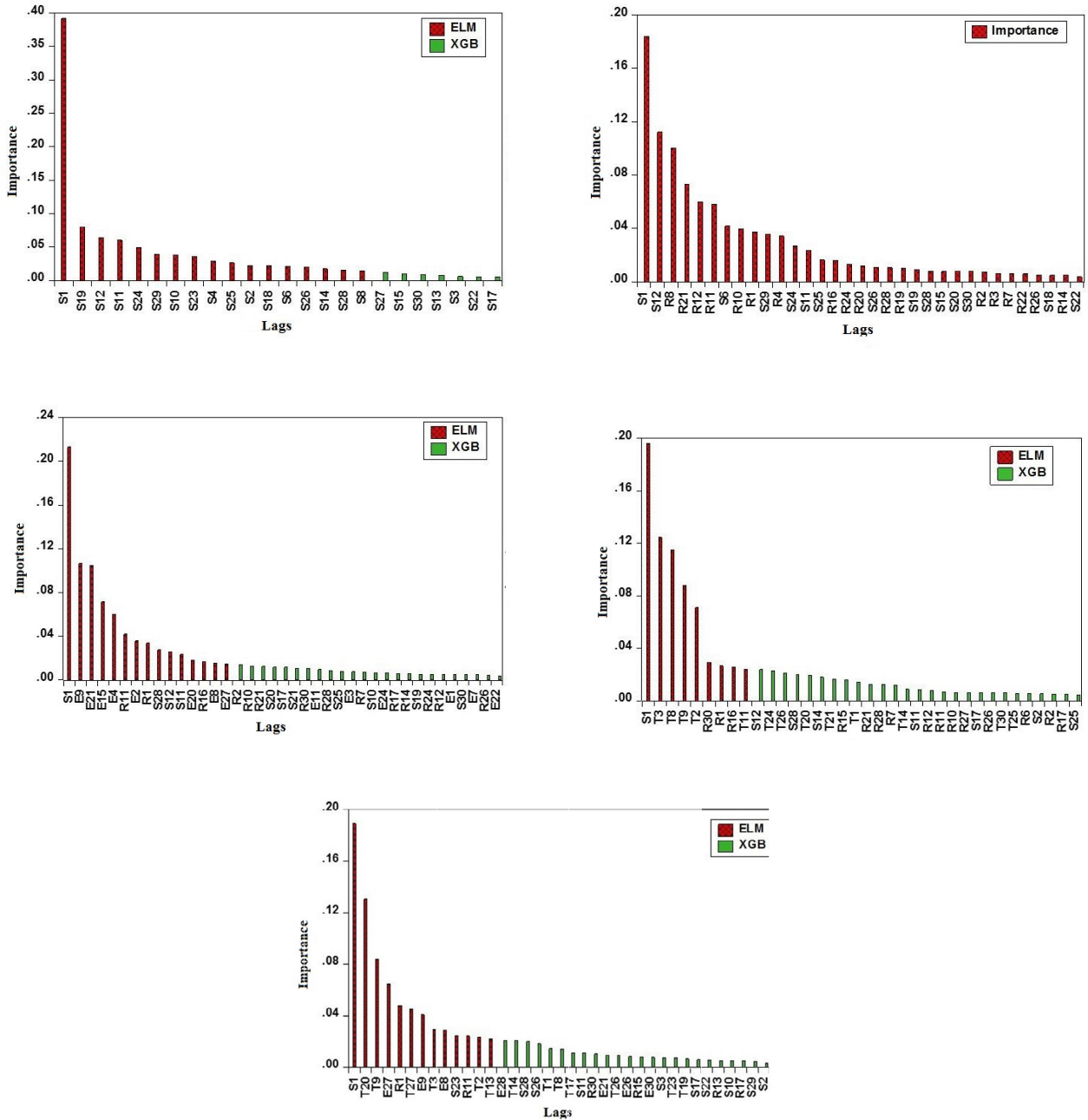


FIGURE 10. The importance of the inputs of both XGB and XGBELM schemas. All lags shown are used in XGB model, red colored are used in ELM.

study) more than the optimum as inputs and utilize the advantage of the ability of the XGB approach in separating the important inputs. The results of this scheme are tabulated in Table 6 and Figure 8. The highest performance is for SRT combination in both training and testing subsets. Considering the highest performed combination, the performance of the XGB model improved in comparison with the stand-alone schema SXGB. NC values of the XGB model over the training and testing are 0.856 and 0.745 for the SRT. Whereas,

SRTE I attained NC values 0.839 and 0.702 for the training and testing subsets. XGB also outperforms SXGB in both subsets considering the RMSE. It is important to mention that the first two combinations S and SR are performed poorly for both models of the stand-alone schema and the SELM and SXGB have dramatically increasing performance. This increasing performance is very important as it indicates that the model is gaining sufficient information to improve the performance without having to add another variable and that

TABLE 6. The statistical performance of the Proposed XGB with 30 lags results. Bold is the best results.

	Training			Testing		
	NC	RMSE (m ³ /s)	MAE (m ³ /s)	NC	RMSE (m ³ /s)	MAE (m ³ /s)
S	0.801	10.722	7.055	0.654	12.731	7.387
SR	0.842	9.572	6.408	0.559	14.370	9.150
SRE	0.851	9.281	5.950	0.694	11.961	7.201
SRT	0.856	9.147	5.988	0.745	10.921	7.418
SRTE	0.832	9.872	6.237	0.731	11.220	7.274

TABLE 7. The statistical performance of the Proposed XGBELM with 30 lags. Bold is the best results.

	Training			Testing		
	NC	RMSE (m ³ /s)	MAE (m ³ /s)	NC	RMSE (m ³ /s)	MAE (m ³ /s)
S	0.709	12.992	8.307	0.647	12.850	9.154
SR	0.665	13.928	9.412	0.644	12.915	9.957
SRE	0.787	11.117	7.311	0.713	11.598	7.973
SRT	0.803	10.682	6.697	0.762	10.556	7.301
SRTE	0.786	11.131	7.329	0.754	10.719	7.869

could reduce the demand of the collected data. The overfitting indication is also existed in this schema although this schema outperforms SXGB schema and that lead to the next proposed schema.

This importance ranking is utilized in choosing the supplied inputs for the ELM according to their ranking. The XGBELM results are reported in Table 7 and graphically presented in Figure 9. In this schema, the best performed combination is SRT with (NC = 0.803 NC, RMSE = 10.682 m³/s and MAE = 6.697 m³/s) for the training subset and (NC = 0.762, RMSE = 10.556 m³/s and MAE = 7.301 m³/s) for the testing subset. Although the training subset has higher performance using XGB schema in all combination, the testing subset of the XGBELM schema has higher performance in all combination except for the S combination considering. The high performance in the testing subset of the DDMs is preferable as training subset is only used in the training, the higher performance in the training subset and less in the testing subset of the XGB schema is considered as an indication of overfitting which is not the case in the XGBELM schema in which training and testing subsets evaluation measurements are very close to each other. Comparing the proposed schema XGBELM with the stand-alone schema SELM and SXGB, the dramatic increase in the performance of the models is very clear in all combinations especially the first two combinations S and SR.

V. DISCUSSION

The selection of the appropriate variables as inputs into the DDMs to predict any variable in general and streamflow in particular is an important issue and must be studied in deep to improve the prediction ability. The stand-alone schema in which only the lagged variables identified using the

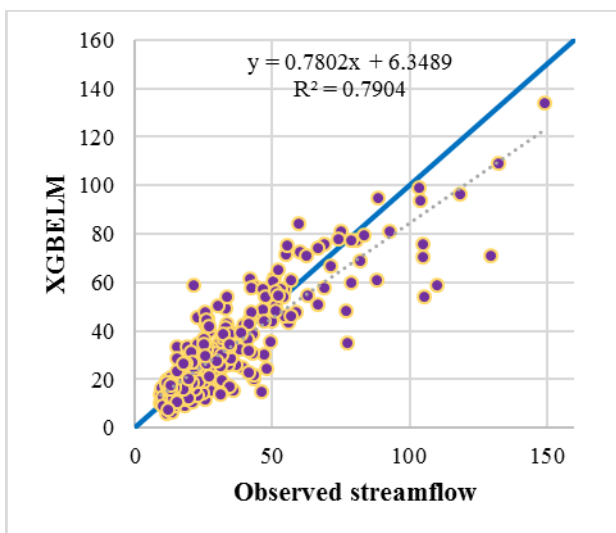
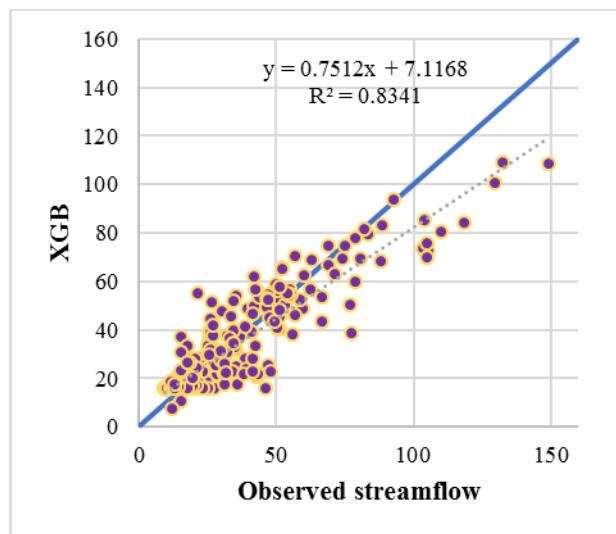


FIGURE 11. Scatter plot of the observed versus the predicted values of ELM, XGB and XGBELM models.

traditional method (i.e., ACF, PACF and CCF) are imposed in the DDMs and showed a poor performance for both models ELM and XGB in the combination that only contains streamflow and the rainfall. Adding the temperature and the evapotranspiration variables improved the performance dramatically. This is clearly evidenced the potential significant of the temperature and evapotranspiration in addition to the rainfall and streamflow as shown in Figure 10. For example, Figure 10a shows the S combination using only the streamflow lags are indicated that the most important input in the prediction is the first lag and the second important input is the 19 lag and that about and an half year, the 3rd, 4th and 7th most important lags are 12, 11, and 10 which are the previous year lags and the 5th most important input is the 24 which represent the previous two years lag.

Accordingly, the seasonal cycling correlation is very important to be included in the prediction process. The evidence is obvious in the results listed in Table 4-6 in which the

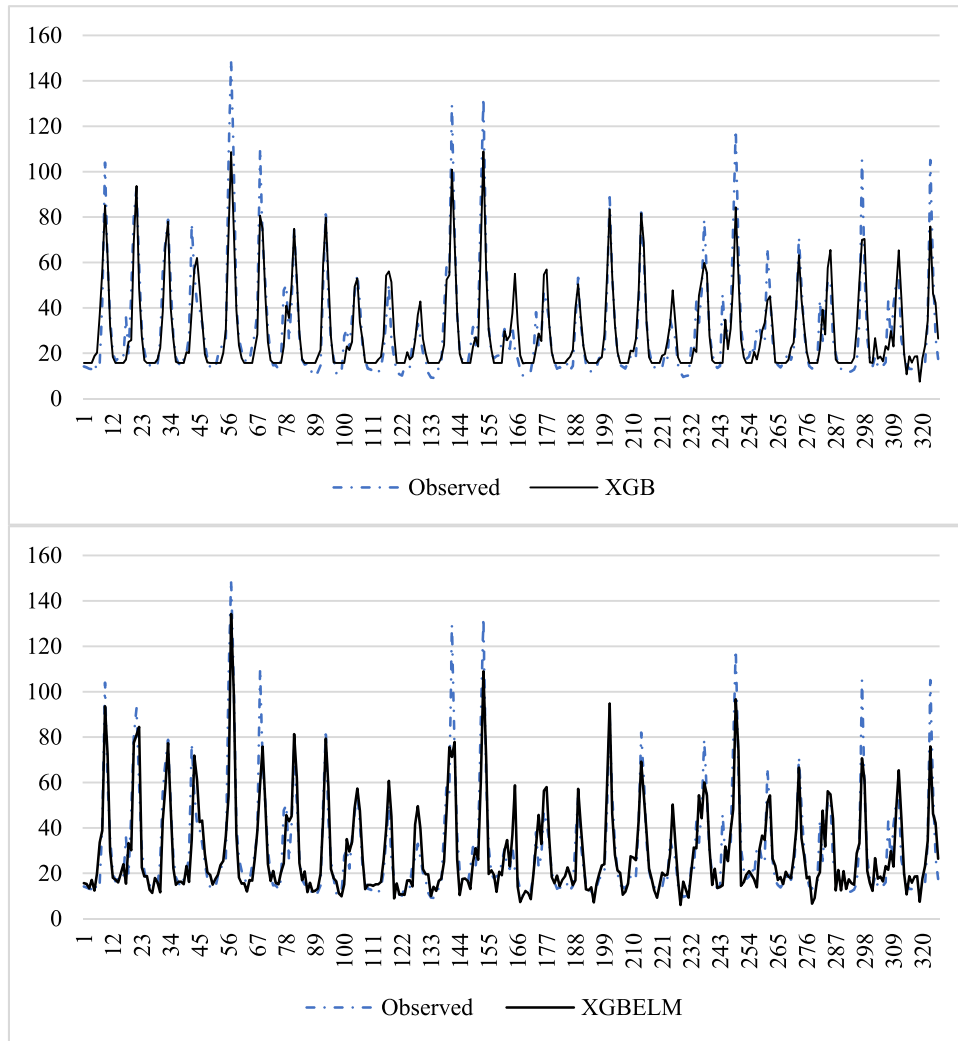


FIGURE 12. Observed and predicted values in the form of time series using XGB and XGBELM models.

performance of the S combination is dramatically increased in the XGB schema in comparison with the stand-alone schema for both SELM and SXGB. Considering the NC metric for the testing subset as an example, $NC = 0.430$ and 0.311 for SELM and SXGB while it is 0.654 for XGB. Examining the other combinations lead to the same results that the seasonality in the lags of the streamflow variable and other variables are very important in the prediction. On the other hand, it is important to remember that the models used here are data-driven in which some time the behavior could not be interpreted and that may explain the existence of some lags that could not be interpreted seasonally but rather explained as additional information to the model. In the Figure 10, the presented lags in every combination are all used as inputs into the XGB while only those red coloreds are used for the ELM. All the combinations have less lags used in the ELM than in XGB except the SR combination and that is also obvious in the evaluation measurements (Table 7) as adding the rainfall deteriorate the model and that not because of no

information added but this might be due to the low number of lags (i.e., 30) used in this study or due to the high fluctuation in the rainfall series.

As mentioned earlier in the results section, the XGB approach in the high-lagged schema has slightly high difference between the training and the testing subsets and lower performance than imposing the lags into the ELM model in the proposed schema XGBELM in the testing subsets. Although XGB has a high prediction ability; however, due to the nature of the approach which is tree-based method that close input variables are predicted as same values and that is obvious in Figure 11 especially in the low streamflow values. This situation is not seen in the XGBELM schema as ELM approach which is not tree-based predicts the output with no similarity unless the inputs are exactly the same. Therefore, the distribution of the points around the 1:1 line in the scatter plot of the XGBELM. Another evaluation aspect is considered that is the time series plot of the observed and predicted values of the two models shown

in Figure 12. The time series plot shows clearly the high agreement between the observed and the predicted using XGBELM. In the case of the XGB, the agreement of the high, and middle values are close to those of the XGBELM while the low values are far from the observed values on the contrast of the XGBELM. Therefore, the proposed XGBELM schema is proved clearly as higher performance than all the other models.

As mentioned earlier in the results section, the XGB approach in the high-lagged schema has a bit high difference between the training and the testing subsets and lower performance than imposing the lags into the ELM model in the proposed schema XGBELM in the testing subsets. Although XGB has a high prediction ability but due to the nature of the approach which is tree-based method that close input variables are predicted as same values and that is obvious in Figure 11 especially in the low streamflow values. This situation is not seen in the XGBELM schema as ELM approach which is not tree-based predicts the output with no similarity unless the inputs are exactly the same. Therefore, the distribution of the points around the 1:1 line in the scatter plot of the XGBELM. Another evaluation aspect is considered that is the time series plot of the observed and predicted values of the two models shown in Figure 12. The time series plot shows clearly the high agreement between the observed and the predicted using XGBELM. In the case of the XGB, the agreement of the high, and middle values are close to those of the XGBELM while the low values are far from the observed values on the contrast of the XGBELM. Therefore, the proposed XGBELM schema is proved clearly as higher performance than all the other models.

In summary, the stand-alone schema with optimum lags performs good in case of the including of several variables that could contribute in the prediction as the required information can be obtained from these variables. Using less variables such as only the streamflow itself could not be sufficient in the prediction and that might be due to the neglect of the correlation of the streamflow with many lags that exceed the number of lags obtained by traditional methods. Including lags of several years could be very useful in improving the ability of the DDMs in the prediction as several seasonal lags could add information to the model and the need to other variables vanished which is important in the areas that have scarce availability of the data. Adding many lags deteriorate the DDMs performance, therefore entering a selection model such as XGB in this study and then use the selected lags as inputs in the DDMs such as ELM in this study could really improve the predictability of the models.

VI. CONCLUSION

In this study, the main enthusiasm was to improve the predictability performance of non-tuned data intelligence model (i.e., ELM) by integrating a non-linear variable abstraction approach for monthly streamflow prediction. The research was established based on three modeling schemas:

Stand-alone where only the optimum lags were considered, high-lagged where arbitrary high number of lags implemented, and the proposed XGBELM conjunction schema in which the XGB used as selection tool. The prediction matrix of the streamflow was performed based on several variables including rainfall, temperature, and evapotranspiration process.

According to the attained results, several conclusions are remarked as followed:

- i. ELM and XGB has a very similar performance in the use of the optimum lags identified by the traditional method including ACF, PACF, CCF, and sequential additional manner.
- ii. The use of the optimum lags with the stand-alone DDMs is useful in case of using several variables but not sufficient in case of using only the lags of the variable to be predicted itself.
- iii. The use of traditional optimum lags as inputs in the prediction process might not extract all the information required especially the seasonality to obtain high performance of DDMs.
- iv. Using Arbitrary number of lags more than the optimum lags and covers several years helps in extracting the seasonality hidden information and improve the prediction ability of the DDMs but a selection tool/approach should be implemented.
- v. XGB is proved to be a very strong modeling approach and can be used in predicting the monthly streamflow even if a high number of inputs are used as it has the ability in filtering only the important inputs.
- vi. XGB is proved as a robust selection tool that can be applied on selecting the inputs and ranking them according to their importance and after that the selected inputs can be utilized in the predicting.
- vii. The proposed schema XGBELM in which the XGB is used as selection approach and ELM is the predicting approach is proved to be improving the prediction ability of the monthly streamflow and that is especially when low number of variables used and that is very important in the regions where no much data available.
- viii. The use of meteorological and hydrological variables such as rainfall, temperature, and evapotranspiration improve the performance of the DDMs in general and the proposed model in particular.

LIMITATION OF THE STUDY

In this study, only one arbitrary number of lags was used owing to the length of the used data is limited and using higher number of lags could affect the training process of the proposed model. The proposed model uses an arbitrary number of lags and the optimum number of lags higher than those identified by the ACF or CCF cannot be identified exactly. Using several arbitrary numbers of lags could be one of the solutions. In addition, using nature inspired algorithms for selecting the appropriate input variables might

be produced more informative attributes for the prediction matrix [48].

CONFLICT OF INTEREST

The authors have no conflict of interest to any party.

ACKNOWLEDGMENT

The authors would like to appreciate the hydrological data provider: Ministry of forests and water affairs- the general directory of Water affairs, Turkey.

ABBREVIATIONS

Extreme Gradient Boosting: (XBG)
 Extreme Learning Machine: (ELM)
 Online Sequential Extreme Learning Machine: (OS-ELM)
 Artificial Intelligence: (AI)
 Support Vector Machine: (SVM)
 Least Square Support Vector Machine: (LSSVM)
 Artificial Neural Network: (ANN)
 Feed Forward Neural Networks: (SLFNs)
 Adaptive Neuro-Fuzzy Inference System: (ANFIS)
 Multi-Linear Regression: (MLR)
 Genetic Algorithm: (GA)
 Auto-Correlation Function: (ACF)
 Partial Auto-Correlation Function: (PACF)
 Cross Correlation Function: (CCF)
 Classification and Regression Trees: (CART)
 Data-Driven Models: (DDMs)
 Mean Squared Errors: (MSE)
 Root Mean Square Error: (RMSE)
 Nash-Sutcliffe: (NC)
 Mean Absolute Error: (MAE)
 Stand-Alone Schema of Extreme Gradient Boosting: (SXGB)
 Stand-Alone Schema of Extreme Learning Machine: (SELM)
 Binary-coded swarm optimization: (BCSO)
 Sea surface temperatures: (SSTs)
 Training Loss: (l)
 Regularization Function: (Ω)
 Vector of Scores on Leaves: (w)
 Function of each point to the equivalent leaf : (q)
 Number of Leaves: (T)
 Score on the new left leaf: (L)
 The score of the new tight leaf: (R)
 Signifies the weights of the randomized layers: (α_i)
 Biases of these randomized layers: (β_i)
 The index of the specific node in the hidden layer: (i)
 The number of inputs: (d)
 Gamma: ($\Gamma\gamma$)
 Eta: ($H\eta$)

REFERENCES

- [1] Z. M. Yaseen, A. El-shafie, O. Jaafar, H. A. Afan, and K. N. Sayl, "Artificial intelligence based models for stream-flow forecasting: 2000–2015," *J. Hydrol.*, vol. 530, pp. 829–844, Nov. 2015.
- [2] J. Guo, J. Zhou, H. Qin, Q. Zou, and Q. Li, "Monthly streamflow forecasting based on improved support vector machine model," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13073–13081, 2011.
- [3] B. Bou-Fakhreddine, I. Mougharbel, A. Faye, S. A. Chakra, and Y. Pollet, "Daily river flow prediction based on two-phase constructive fuzzy systems modeling: A case of hydrological–meteorological measurements asymmetry," *J. Hydrol.*, vol. 558, pp. 255–265, Mar. 2018.
- [4] E. Shamaei and M. Kaedi, "Suspended sediment concentration estimation by stacking the genetic programming and neuro-fuzzy predictions," *Appl. Soft Comput.*, vol. 45, pp. 187–196, Aug. 2016.
- [5] A. D. Mehr, E. Kahya, and E. Olyaie, "Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique," *J. Hydrol.*, vol. 505, pp. 240–249, Nov. 2013.
- [6] Z. Liu, P. Zhou, G. Chen, and L. Guo, "Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting," *J. Hydrol.*, vol. 519, pp. 2822–2831, Nov. 2014.
- [7] M. Rezaie-Balf, S. R. Naganna, A. Ghaemi, and P. C. Deka, "Wavelet coupled MARS and M5 model tree approaches for groundwater level forecasting," *J. Hydrol.*, vol. 553, pp. 356–373, Oct. 2017.
- [8] N. S. Raghavendra and P. C. Deka, "Support vector machine applications in the field of hydrology: A review," *Appl. Soft Comput.*, vol. 19, pp. 372–386, Jun. 2014.
- [9] E. Meng, S. Huang, Q. Huang, W. Fang, L. Wu, and L. Wang, "A robust method for non-stationary streamflow prediction based on improved EMD-SVM model," *J. Hydrol.*, vol. 568, pp. 462–478, Jan. 2019.
- [10] H. R. Maier, Z. Kapelan, J. Kasprzyk, and L. S. Matott, "Thematic issue on evolutionary algorithms in water resources," *Environ. Model. Softw.*, vol. 69, pp. 222–225, Jul. 2015.
- [11] J. M. Szemis, H. R. Maier, and G. C. Dandy, "A framework for using ant colony optimization to schedule environmental flow management alternatives for rivers, wetlands, and floodplains," *Water Resour. Res.*, vol. 48, no. 8, pp. 1–21, Aug. 2012.
- [12] G. B. Humphrey, S. Galelli, A. Castelletti, H. R. Maier, G. C. Dandy, and M. S. Gibbs, "A new evaluation framework for input variable selection algorithms used in environmental modelling," in *Proc. 7th Int. Congr. Environ. Modelling Softw.*, vol. 62, 2014, pp. 33–51.
- [13] R. S. V. Teegavarapu, "Exploring geometrical patterns in streamflow time series: Utility for forecasting?" *Hydrol. Res.*, vol. 49, no. 6, pp. 1724–1739, 2018.
- [14] M. Sivapalan, "From engineering hydrology to earth system science: Milestones in the transformation of hydrologic science," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 3, pp. 1665–1694, Mar. 2018.
- [15] F. Fahimi, Z. M. Yaseen, and A. El-shafie, "Application of soft computing based hybrid models in hydrological variables modeling: A comprehensive review," *Theor. Appl. Climatol.*, vol. 128, pp. 875–903, May 2017.
- [16] V. Nourani, A. H. Baghanam, J. Adamowski, and O. Kisi, "Applications of hybrid wavelet–artificial intelligence models in hydrology: A review," *J. Hydrol.*, vol. 514, pp. 358–377, Jun. 2014.
- [17] A. D. Mehr, V. Nourani, E. Kahya, B. Hrnjica, A. M. A. Sattar, and Z. M. Yaseen, "Genetic programming in water resources engineering: A state-of-the-art review," *J. Hydrol.*, vol. 566, pp. 643–667, Nov. 2018.
- [18] G. J. Bowden, G. C. Dandy, and H. R. Maier, "Input determination for neural network models in water resources applications. Part 1—Background and methodology," *J. Hydrol.*, vol. 301, nos. 1–4, pp. 75–92, 2005.
- [19] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Netw.*, vol. 61, pp. 32–48, Jan. 2015.
- [20] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
- [21] H. Siqueira, L. Boccatto, R. Attux, and C. Lyra, "Unorganized machines for seasonal streamflow series forecasting," *Int. J. Neural Syst.*, vol. 24, no. 3, 2014, Art. no. 1430009.
- [22] R. Taormina and K.-W. Chau, "Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and extreme learning machines," *J. Hydrol.*, vol. 529, pp. 1617–1632, Oct. 2015.
- [23] M. Atiquzzaman and J. Kandasamy, "Prediction of hydrological time-series using extreme learning machine," *J. Hydroinform.*, vol. 18, no. 2, pp. 345–353, 2015.
- [24] A. R. Lima, A. J. Cannon, and W. W. Hsieh, "Forecasting daily streamflow using online sequential extreme learning machines," *J. Hydrol.*, vol. 537, pp. 431–443, Jun. 2016.

- [25] B. Yadav, S. Ch, S. Mathur, and J. Adamowski, "Discharge forecasting using an online sequential extreme learning machine (OS-ELM) model: A case study in Neckar River, Germany," *Measurement*, vol. 92, pp. 433–445, Oct. 2016.
- [26] Z. M. Yaseen, O. Jaafar, R. C. Deo, O. Kisi, J. Adamowski, J. Quilty, and A. El-Shafie, "Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq," *J. Hydrol.*, vol. 542, pp. 603–614, Nov. 2016.
- [27] R. C. Deo and M. Şahin, "An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland," *Environ. Monit. Assessment*, vol. 188, p. 90, Feb. 2016.
- [28] Z. M. Yaseen, M. F. Allawi, A. A. Yousif, O. Jaafar, F. M. Hamzah, and A. El-Shafie, "Non-tuned machine learning approach for hydrological time series forecasting," *Neural Comput. Appl.*, vol. 30, no. 5, pp. 1479–1491, 2016.
- [29] V. Nourani, G. Andalib, and F. Sadikoglu, "Multi-station streamflow forecasting using wavelet denoising and artificial intelligence models," *Procedia Comput. Sci.*, vol. 120, pp. 617–624, Jan. 2017.
- [30] A. B. Dariane and S. Azimi, "Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection," *J. Hydroinform.*, vol. 20, no. 2, pp. 520–532, 2017.
- [31] H. Z. Sabzi, J. P. King, and S. Abudu, "Developing an intelligent expert system for streamflow prediction, integrated in a dynamic decision support system for managing multiple reservoirs: A case study," *Expert Syst. Appl.*, vol. 83, pp. 145–163, Oct. 2017.
- [32] V. P. Singh and H. Cui, "Entropy theory for streamflow forecasting," *Environ. Processes*, vol. 2, no. 3, pp. 449–460, 2015.
- [33] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, *XGBoost: Extreme Gradient Boosting. R Package Version 0.6.4.1*. Accessed: 2018. [Online]. Available: <https://CRAN.R-project.org/package=xgboost>
- [34] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme gradient boosting as a method for quantitative structure–activity relationships," *J. Chem. Inf. Model.*, vol. 56, no. 12, pp. 2353–2360, 2016.
- [35] Y. Ji, J. Hao, N. Reyhani, and A. Lendasse, "Direct and recursive prediction of time series using mutual information selection," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, 2005, pp. 1010–1017.
- [36] R. Noori, A. R. Karbassi, A. Moghaddamia, D. Han, M. H. Zokaei-Ashtiani, A. Farokhnia, and M. G. Gousheh, "Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction," *J. Hydrol.*, vol. 401, nos. 3–4, pp. 177–189, 2011.
- [37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [38] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [39] B. Pradhan and M. I. Sameen, "Predicting injury severity of road traffic accidents using a hybrid extreme gradient boosting and deep neural network approach," in *Laser Scanning Systems in Highway and Safety Assessment*. Cham, Switzerland: Springer, 2019, pp. 119–127.
- [40] Z.-Y. Chen, T.-H. Zhang, R. Zhang, Z.-M. Zhu, J. Yang, P.-Y. Chen, C.-Q. Ou, and Y. Guo, "Extreme gradient boosting model to estimate PM_{2.5} concentrations with missing-filled satellite data in China," *Atmos. Environ.*, vol. 202, pp. 180–189, Apr. 2019.
- [41] Z. M. Yaseen, S. O. Sulaiman, R. C. Deo, and K.-W. Chau, "An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction," *J. Hydrol.*, vol. 569, pp. 387–408, Feb. 2019.
- [42] E. Fijani, R. Barzegar, R. Deo, E. Tziritis, and K. Skordas, "Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters," *Sci. Total Environ.*, vol. 648, pp. 839–853, Jan. 2019.
- [43] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, and Y. Xiang, "Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China," *Energy Convers. Manag.*, vol. 164, pp. 102–111, May 2018.
- [44] D. R. Legates and G. J. McCabe, Jr., "Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation," *Water Resour. Res.*, vol. 35, no. 1, pp. 233–241, 1999.
- [45] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—A discussion of principles," *J. Hydrol.*, vol. 10, no. 3, pp. 282–290, 1970.
- [46] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.
- [47] S. J. Hadi and M. Tombul, "Conversion of CruTS 3.23 data and evaluation of precipitation and temperature variables in a local scale," in *Proc. MATEC Web Conf.*, vol. 120, Aug. 2017, Art. no. 05007.
- [48] P. Melin, L. Astudillo, O. Castillo, F. Valdez, and M. Garcia, "Optimal design of type-2 and type-1 fuzzy tracking controllers for autonomous mobile robots under perturbed torques using a new chemical optimization paradigm," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3185–3195, 2013.



SINAN JASIM HADI received the bachelor's degree in civil engineering from the University of Tikrit, Iraq, the master's degree from the University Putra Malaysia, Malaysia, in 2012, and the Ph.D. degree from Anadolu University, Turkey, in 2018. He is currently an Assistant Professor in civil engineering. He has published many articles in several international journals and conferences. His current research interests include water resources engineering, flood modeling, and climate change, in addition to the application of remote sensing and GIS in all these fields.



S. I. ABBA received the bachelor's degree from Bayero University, Kano (BUK) and the master's degree from Sharda University, India. He is currently a Researcher with Near East University (NEU), Cyprus, TRNC. He is also a certified and registered Engineer by the Council for Regulation of Engineering in Nigeria (COREN) and a Researcher in civil and environmental engineering. His current research interests include water resources, environmental modeling and simulation, water quality, and water and wastewater treatment plant. He also has additional skills and expertise in data-driven algorithms, artificial intelligence, and data processing software.



SAAD SH. SAMMEN received the master's degree from the University of Technology, Iraq, in 2009, and the Ph.D. degree from the University Putra Malaysia (UPM), Malaysia, in 2018. He is currently a Senior Lecturer and a Senior Researcher in civil engineering. He has published many articles in international journals. His current research interests include hydraulic engineering, water resources engineering, hydrological processes modeling, flood modeling, and climate change. In addition, he has an expertise in machine learning.



SINAN Q. SALIH received the B.Sc. degree in information systems from the University of Anbar, Anbar, Iraq, in 2010, the M.Sc. degree in computer sciences from the Universiti Tenaga National (UNITEN), Malaysia, in 2012, and the Ph.D. degree in soft modeling and intelligent systems from the Universiti Malaysia Pahang (UMP). His current research interests include optimization algorithms, nature-inspired metaheuristics, machine learning, and feature selection problem for real world problems.



NADHIR AL-ANSARI received the B.Sc. and M.Sc. degrees from the University of Baghdad, in 1968 and 1972, respectively, and the Ph.D. degree in water resources engineering from Dundee University, in 1976. He is currently a Professor with the Department of Civil, Environmental and Natural Resources Engineering, Lulea Technical University, Sweden. He was with Baghdad University, from 1976 to 1995, then with Al-Bayt University, Jordan, from 1995 to 2007.

His current research interests include geology, water resources, and environment. He served several academic administrative posts (Dean and Head of Department). He published more than 424 articles in international/national journals, chapters in books, and 13 books. He executed more than 60 major research projects in Iraq, Jordan, and U.K. He has one patent on physical methods for the separation of iron oxides. He supervised more than 66 post-graduate students at Iraq, Jordan, U.K., and Australia universities. He was a member of the editorial board of ten international journals. He is a member of several scientific societies e.g. International Association of Hydrological Sciences, Chartered Institution of Water and Environment Management, Network of Iraqi Scientists Abroad, and a Founder and the President of the Iraqi Scientific Society for Water Resources. He received several scientific and educational awards, among them is the British Council on its 70th Anniversary awarded him top five scientists in Cultural Relations.



ZAHER MUNDHER YASEEN received the master's and Ph.D. degrees from the National University of Malaysia (UKM), Malaysia, from 2012 to 2017. He is currently a Senior Lecturer and a Senior Researcher in civil engineering. He has published over 80 articles in international journals with a Google Scholar H-Index of 18, and a total of 979 citations. His current research interests include hydrology engineering, water resources engineering, hydrological processes modeling,

and environmental engineering and climate. In addition, he has an excellent expertise in machine learning, advanced data analytics, and environmental sciences.

...