The Capacity of the Hybridizing Wavelet Transformation Approach With Data-Driven Models for Modeling Monthly-Scale Streamflow

SINAN JASIM HADI1, (Member, IEEE), MUSTAFA TOMBUL2, SINAN Q. SALIH3, NADHIR AL-ANSARI4, AND ZAHER MUNDHER YASEEN5, (Member, IEEE)

1Department of Real Estate Management and Development, Faculty of Applied Sciences, Ankara University, 00026 Ankara, Turkey
2Department of Civil Engineering, Faculty of Engineering, Eskisehir Technical University, 26555 Eskisehir, Turkey
3Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Iraq
4Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden
5Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

Corresponding author: Zaher Mundher Yaseen (zahermundheryaseen@duytan.edu.vn)

ABSTRACT Hybrid models that combine wavelet transformation (WT) as a pre-processing tool with data-driven models (DDMs) as modeling approaches have been widely investigated for forecasting streamflow. The WT approach has been applied to original time series for decomposing processes prior to the application of DDM modeling. This procedure has been applied to eliminate redundant patterns or information that lead to a dramatic increase in the model performance. In this study, three experiments were implemented, including stand-alone data-driven modeling, hindcast decomposing using WT divided and entered into the extreme learning machine (ELM), and the extreme gradient boosting (XGB) model to forecast streamflow data. The WT method was applied in two forms: discrete and continuous (DWT and CWT). In this paper, a new hybrid model is proposed based on an integrative prediction model where XGB is used as an input selection tool for the importance attributes of the prediction matrix that are then supplied to the ELM model as a predictive model. The monthly streamflow, upstream flow, rainfall, temperature, and potential evapotranspiration of a basin named in 1805 and located in the south east of Turkey, are used for development of the model. The modeling results show that applying the WT method improved the performance in the hindcast experiment based on the CWT form with minimum root mean square error (RMSE = 4.910 m$^3$/s).

On the contrary, WT deteriorated the performance of the forecasting and the stand-alone models exhibited a better performance. WT increased the performance of the hindcast experiment due to the inclusion of future information caused by convolution of the time series. However, the forecast experiment experienced deterioration due to the border effect at the end of the time series. Hence, WT was found not to be a useful pre-processing technique in forecasting the streamflow.

INDEX TERMS Streamflow forecasting, gradient boosting, extreme learning machine, wavelet transformation, streamflow monitoring.

I. INTRODUCTION

A. STREAMFLOW MODELING SIGNIFICANCE

Considering hydrological process elements, streamflow is a crucial process on global and regional scales [1], [2]. It is considered the main source of freshwater [3]. Streamflow is highly associated with several hydrological characteristics and thus has a major influence on water resource management in areas susceptible to disasters, where accurate short- and long-term streamflow forecasting is critical [4]–[6]. Short-term forecasting is essential for two main applications: the forecasting of floods and development of a cautionary system [7], [8]. Long-term forecasts (e.g., monthly or annual) are useful for several applications, such as irrigation management decisions, reservoir operations, hydro-power generation, and sediment transportation [9], [10]. The accurate streamflow forecasting can contribute to several watershed catchment sustainability and management and thus its accurate
forecasting is highly beneficial for decision makers and river engineering maintenance [11], [12].

B. LITERATURE REVIEW

In general, hydrological models are divided into data-driven models and physical-based models [13]. Physical-based models are known as white box models that involve the physical process in the hydrological cycle, leading to the need for a large amount of data that is not always available [14]. On the contrary, data-driven models (DDMs) known as black box models map the relation between inputs and outputs through statistical formulation, without involving the physical process. Within engineering applications, the capacity of DDMs has been evidenced remarkably [15]–[19]. Several DDMs have been explored in the literature for streamflow modeling, such as the adaptive neuro-fuzzy inference system (ANFIS), artificial neural network (ANN), support vector machine (SVM), genetic programming (GP), Decision Tree (DT), and ELM [20], [21].

C. RESEARCH MOTIVATION AND ENTHUSIASM

Streamflow forecasting is a complex hydrological approach that features the streamflow trend with non-linearity and non-stationarity [22]. DDMs have the ability to solve the non-linearity and non-stationarity in the mean and variance, but their main disadvantage is the failure they display in handling non-stationary fluctuations [23]. Therefore, pre-processing was found to be important for improving the performance [24]. Discrete wavelet transformation (DWT), singular spectrum analysis (SSA), and empirical mode decomposition (EMD) are pre-processing techniques that decompose the non-stationary and non-linear time series into several components that are easier to model [25]. The capacity of the wavelet transformation method is superior to that of other pre-processing methods due to its capability to abstract non-trivial and significant time series information [2]. Extracting explicit information from historical data time series can resolve the non-linearity and non-stationarity [26].

In recent years, hybrid models have been developed using these pre-processing techniques with DMMs, and the performance of these models in forecasting has increased dramatically. Many researchers have investigated the application of wavelet transformation (WT) (particularly DWT)-DDMs hybrid models, such as [23], [27]–[29]. Continuous wavelet transformation has also been applied with DDMs by several researchers, such as [30]. As an example, Badrzedeh et al. (2017) studied the hybrid model of DWT-ANFIS and found that the use of DWT as a pre-processing tool increased the model’s performance significantly [31]. Kisi and Cimen (2011) conducted a study investigating the improvement in the SVM model obtained by conjugating it with DWT and reported that the conjugated model DWT-SVM had a higher performance in forecasting monthly streamflow [4].

However, most of these studies have applied hybrid models in such a way that future information is sent to the model, but must not be included in the forecasting experiment [32], [33]. In the case of DWT, such studies have been implemented in such a way that all of the data are decomposed and reconstructed to produce the sub time series. Then, the sub time series are divided into calibration/training and validation subsets to be imposed in DDMs. In the use of continuous WT (CWT), the same procedure is followed, except for the fact that choosing the most contributing scale(s) to be imposed in DDM as CWT produces redundant information. Such a procedure of implementation sends future information to the model and therefore, this is named a hindcasting experiment and not a forecasting experiment [34]. A recent study (Zhang, Peng et al. 2015) conducted hindcast and forecast experiments using DWT, SSA, and EMD as pre-processing approaches and concluded that the hybrid models performed worse than the original models Du et al. (2017) investigated the hybrid models using DWT and SSA with ANN and SVM and found that the hybrid models included future information and therefore caused an incorrect increase in the performance of the original models.

D. RESEARCH OBJECTIVES

The main objective of this study is to conduct three experiments, including stand-alone DDM, hindcast, and forecast experiments, to predict one month-ahead streamflow. In the stand-alone experiment, no WT was applied, whilst for both hindcast and forecast experiments, DWT and CWT were applied as pre-processing techniques with the extreme learning machine (ELM), which, to the knowledge of the authors, has not been applied with WT in the literature. CWT produces redundant information, so has not been applied as extensively as DWT [23]. However, in this study, a recent approach named extreme gradient boosting (XGB) is applied for choosing the contributing scales from the CWT scales to be imposed in the ELM, although XGB itself is a stand-alone model.

II. PRE-PROCESSING METHOD

Wavelet transformation is categorized into CWT and DWT. CWT can be described as the summation through time of the signal multiplied by shifted and scaled versions of the wavelet known as the mother wavelet \( \psi_{a,b}(t) \):

\[
W_c(a, b) = |a|^{-1/2} \int_{-\infty}^{+\infty} f(t) \psi_{a,b}^* \left( \frac{t - b}{a} \right) dt, \tag{1}
\]

where \( W_c(a, b) \) are the CWT coefficients, * indicates the conjugate function complexity, and \( a \) and \( b \) are the scale shifting parameters. According to Eq. (1), every scale is assigned for a number of coefficients equivalent to the length of the original time series.

DWT is thought of as the dyadic sampling form of CWT, with \( a = 2^j \) and \( b = k2^j \), where \( k \) is the location index and \( j \)
is the decomposition level. The discrete wavelet is written as
\[ \psi_{j,k}(t) = 2^{-j/2} \psi \left( 2^{-j} t - k \right), \tag{2} \]
and the DWT is written as
\[ W_\psi (j, k) = 2^{-j/2} \int_{\infty}^{\infty} f(t) \psi^*_j (2^{-j} t - k) \, dt. \tag{3} \]
For a discrete time series \( x_n \), the DWT is
\[ W_\psi (j, k) = 2^{-j/2} \sum_{n=0}^{N} \psi (2^{-j} t - k) x_n, \tag{4} \]
where \( W_\psi (j, k) \) are the DWT coefficients and to reconstruct the original time series, the inverse DWT is implemented, given as
\[ x_n = A + \sum_{j=1}^{J} \sum_{k=0}^{2^{j-1}-1} W_\psi (j, k) 2^{-j/2} \left( 2^{-j} t - k \right) \tag{5} \]
and can be written as
\[ x_n = A(t) + \sum_{j=1}^{J} W_j (t) \), \tag{6} \]
where \( A(t) \) is the approximation at level \( J \) and \( W_j (t) \) are the coefficients of details at levels \( j = 1, 2, \ldots, J \).

III. DATA-DRIVEN MODELS

A. EXTREME LEARNING MACHINE (ELM)

The ELM is an emerging DDM of the single hidden layer feed forward networks (SLFNs) of ANN first proposed by [36]. ELM overcomes the disadvantages of over fitting, local minima, and slow learning of the traditional Back-propagation ANN. Since the development of the algorithm, it has been applied in several applications of hydrological modeling by researchers and for streamflow forecasting in particular [21]. One of the recent studies [37] compared the performance of ELM with SVM and the generalized regression neural network and concluded that the ELM was superior to other approaches, in accordance with the predictability performance.

The ELM model is developed using a set of training samples as \( \{(x_1, y_1), \ldots, (x_t, y_t)\} \), where \( x_t \) is the independent variable and \( y_t \) is the target variable. In this study, the dependent variable is the input vector denoted as \( x_1, x_2, \ldots, x_t \), defined as the lagged streamflow, or any other variable used in this study; the variables used as inputs are explained in the used data section. The targets or output vector \( y_1, y_2, \ldots, y_t \) signify the targeted one month-ahead streamflow. For a given dataset with \( N \)-many observations (i.e., \( t = 1, 2, \ldots, N \)), \( x_t \in \mathbb{R}^d \), and \( y_t \in \mathbb{R} \), an SLFN with \( H \) hidden nodes is mathematically expressed as follows:
\[ \sum_{i=1}^{H} B_i g_i (\alpha_i x_t + \beta_i) = z_t, \tag{7} \]
where \( B \in \mathbb{R}^H \) is the estimated weight of the network connection between the hidden and output layers (i.e., \( Z(z_t \in \mathbb{R}) \)).

\( G(\alpha, \beta, x) \) is the activation function of the hidden layer, \( \alpha_i \) represents the randomized weights, \( \beta_i \) represents the randomized biases, \( i \) is the indicator of the particular node in the hidden layer, and \( d \) is the number of input variables.

In the current research, a sigmoid activation function is implemented, in accordance with the previously established research [38].
\[ G(x) = \frac{1}{1 + \exp(-x)} \tag{8} \]
A linear transfer function is applied for the output layer. According to [39], an optimal learning process for the ELM model can be attained with random weight assignment for the input layer and hidden neurons. In addition, randomized hidden layer biases \( \beta \) can reduce the error as much as possible to zero (Eq. (9)). Therefore, the following equation can be applied for calculating the weights of the output layer:
\[ \sum_{t=1}^{N} \| y_t - z_t \| = 0. \tag{9} \]
The \( B \) values of \( N \) input–output training samples can be estimated using a system of linear equations:
\[ Y = GB, \tag{10} \]
where
\[ G(\alpha, \beta, x) = \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_N) \end{bmatrix} = \begin{bmatrix} g_1 (\alpha_1 x_1 + \beta_1) & \cdots & g_L (\alpha_H x_L + \beta_L) \\ \vdots & \ddots & \vdots \\ g_1 (\alpha_N x_N + \beta_1) & \cdots & g_L (\alpha_H x_L + \beta_L) \end{bmatrix}, \tag{11} \]
and
\[ B = \begin{bmatrix} B_1^T \\ \vdots \\ B_H^T \end{bmatrix}_{H \times 1}, \tag{12} \]
and
\[ Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times 1}, \tag{13} \]
where \( G \) is the hidden layer output and \( T \) is the transpose of the matrix. The Moore–Penrose generalized inverse function (+) can be used for inverting the matrix of the hidden layer in the output weights \( B \):
\[ \hat{B} = G^+ Y. \tag{14} \]
Ultimately, the value of \( \hat{y} \) (i.e., representing streamflow data one month ahead in this study) can be estimated by
\[ \hat{y} = \sum_{i=1}^{H} \hat{B}_i g_i (\alpha_i x_t + \beta_i). \tag{15} \]
**B. EXTREME GRADIENT BOOSTING (XGB)**

XGB was first developed by [40], as an efficient, fast, and scalable implementation of the gradient tree boosting algorithm developed by [41]. XGB is classified as a supervised learning technique that uses an ensemble of decision trees [42]. It can be used for both classification and regression problems. Being a new algorithm, to the best of the authors’ knowledge, it has not been applied before in streamflow forecasting in general, and as an input selection tool with a wavelet in particular. In this study, this algorithm is applied as a model and selection tool for the important scales of CWT by taking advantage of the feature importance characteristic of this algorithm.

The XGB algorithm is part of the Classification And Regression Trees (CART) model of classification and regression trees and it is different from the DT algorithm in that each leaf presents an actual score that helps produce a superior simple classification technique. The CART model has been proposed to be lacking, since it uses a single tree which is not robust enough; therefore, a model which can handle the issue was proposed, consisting of an ensemble of multiple trees. An ensemble of multiple trees is built to perform the learning process based on multiple features (i.e., climate variables in this study) \(x_i\) for predicting one month-ahead streamflow downstream of the catchment. The mathematical formula used to present the ensemble of \(K\) (number of trees) can be expressed as follows:

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

where \(f_k\) is a function in a set of possible functions represented in CARTs and \(F\) is the set of these functions. The main objective of this function is the optimization of

\[
obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{T} \Omega(f_i),\tag{17}
\]

where the first phase \(l\) represents the training loss function and the second phase \(\Omega\) defines the regularization to reduce complications and prevent the overfitting problem. After the inclusion of the tree function and the regularization part, and removal of the constants from Eq. (21), the objective value of the \(t^{th}\) tree can be written as

\[
obj^t = \sum_{i=1}^{n} \left[2 \left(\hat{y}_{i}^{t-1} - y_i\right) f_t (x_i) + f_t (x_i)^2 + \Omega(f_t) + \text{constant}\right].\tag{20}
\]

The MSE metric in both first order or quadratic form is friendly, but functions such as logistics are more complex. Therefore, Taylor expansion was applied up to the second order:

\[
obj^t = \sum_{i=1}^{n} \left[2 \left(\hat{y}_{i}^{t-1} - y_i\right) f_t (x_i) + \frac{1}{2} h_t f_t (x_i)^2 + \Omega(f_t) + \text{constant}\right],\tag{21}
\]

where \(g_t = \frac{\partial y_i}{\partial \hat{y}_i} (y_i - \hat{y}_i)\) and \(h_t = \frac{\partial^2 y_i}{\partial \hat{y}_i^2} (y_i - \hat{y}_i)\). All the steps were conducted in the training loss function part. Considering the regularization part and bearing in mind that the tree is defined as \(f_t (x) = w_q(x)\), the regularization can be written as

\[
\Omega(f) = \gamma T + \frac{1}{2} \sum_{j=1}^{T} w_j^2,\tag{22}
\]

where \(w\) represents the vector of scores on leaves, \(q\) represents the assigning function of each point in the data to the corresponding leaf, and \(T\) represents the number of leaves. After the inclusion of the tree function and the regularization part, and removal of the constants from Eq. (21), the objective value of the \(t^{th}\) tree can be written as

\[
obj^t \approx \sum_{i=1}^{n} \left[\left|g_{i} w_{q(i)} + \frac{1}{2} h_{i} w_{q(i)}^2\right| + \gamma T + \frac{1}{2} \sum_{j=1}^{T} w_j^2\right],
\]

\[
= \sum_{j=1}^{T} \left((\sum_{i \in l_j} g_i) w_j + \frac{1}{2} \left(\sum_{i \in l_j} h_j + \lambda\right) w_j^2 + \gamma T\right),\tag{23}
\]

where \(l_i = \{ i \mid q(x_i) = j\}\) represents the indices of data points assigned to the \(j^{th}\) leaf. Defining \(G_j = \sum_{i \in l_j} g_i\) and \(H_j = \sum_{i \in l_j} h_i\), the objective function can be presented as

\[
obj^t = \sum_{j=1}^{T} \left(G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 + \gamma T\right).\tag{24}
\]

For a given structure of tree \(q(x)\), the best \(w_j\) and objective reduction, which are used for measuring how good the structure is, can be obtained by

\[
w_j^* = -\frac{G_j}{H_j + \lambda},\tag{25}
\]

\[
obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \gamma T.	ag{26}
\]

Enumerating all of the possible trees and picking the best one is not an intractable way of resolving the issue. Therefore, Eq. (27) is used for optimizing one level of the tree at a time and obtaining the score produced by dividing a single leaf into two, creating two new leaves. Eq. (27) consists of the scores of both sides of the new leaf, in which \((L)\) is the score for the
new left leaf, \( R \) is the score for the new right leaf. The score for the original leaf and the regularization is termed \( \gamma \). If the improvement is less than the value of \( \gamma \), it would be better not to consider adding it to the branch.

\[
Gain = \frac{1}{2} \left[ \frac{G^2_L}{H_L + \lambda} + \frac{G^2_R}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + 2\lambda} \right] - \gamma. \quad (27)
\]

By using this gain information in the inclusion of the variable that measures the gain of including a certain variable in the whole tree network, the feature importance is obtained from this algorithm. This characteristic is utilized in this study for suggesting a new selection tool for the main scales in the CWT for forecasting streamflow.

C. PROPOSED MODELING SCHEMA

In this study, two modeling schemas are proposed. In the first schema, the inputs are transformed using DWT (CWT) and all of the reconstructed sub time series (scales) are imposed in the XGB and ELM (only XGB for CWT). In the second schema, the inputs are converted by applying the CWT using the highest possible scales and all of the scales are entered into XGB, while only the important scales obtained from XGB are forced into ELM, according to their order of importance. These two schemas are applied in both hindcast and forecast experiments (see the experiment section). In the following, the proposed schema is explained in detail. The framework of the proposed schema is shown in Figure 1.

CWT transforms the time sequence in the time-frequency domain into several scales for the same length of the original time series. Using a large number of scales produces too many scales that, in the case of imposing them in DDMs, produces bad models due to the inclusion of redundant information that deteriorates the model’s performance [43], [44]. XGB, as mentioned earlier, has the ability to produce the ordered feature importance (i.e., scales in this specific study), which shows the importance of a specific feature in modeling the dependent variables, starting with most important feature and ending with the least significant one. Features not adding any gain to the model are not even shown in the importance matrix. In this study, XGB was proposed as a selection tool, in addition to being a modeling approach itself. In both the hindcast and forecast experiment, the proposed schema was applied as follows:

i. The lags of the climate variables (i.e., \( V = v_1, v_2, \ldots, v_n \)) were converted by applying the CWT with the highest possible scale of 128;

ii. For every variable, 128 time series \( S_{vi} = s_{vi1}, s_{vi2}, \ldots, s_{vi128}, i = 1, 2, \ldots, n \) were acquired. All of the scales were used as inputs and thus (i.e., \( TS = [S_{v1}, S_{v2}, \ldots, S_{vn}] \)) were then entered into the XGB model, and the result showed the one month-ahead downstream flow \( Q_{t+1} \);

iii. The assessment criteria performance was attained for the XGB model;

iv. The importance of the scales was attained from XGB in order that the most important scale was first and the least important scale was last. \( M = \left( s^m_{vij} \right) \), where \( i \) represents the variable, \( j \) signifies the scale of that variable, and \( m \) represents the order of the importance of the variable and takes values according to the number of features/scales chosen in the importance matrix \( M \);

v. The most significant scales were then chosen and imposed in ELM in a sequential manner. The first model only considered the most important scales \( Q_{t+1} = f(s^1_{vij}) \), whilst the second model included the first and second significant scales \( Q_{t+1} = f(s^1_{vij}, s^2_{vij}) \), and this procedure was continuous until the last model containing all of the scales in the importance matrix \( Q_{t+1} = f(s^1_{vij}, s^2_{vij}, \ldots, s^m_{vij}) \).

---

**FIGURE 1.** Framework of the proposed model schema.
vi. The highest performing model was chosen with their scales to compare it with the two other schemas: standalone and DWT preprocessed models.

IV. MODELING EXPERIMENT

A. THE STUDY AREA LOCATION AND DESCRIPTION

In order to examine the three experiments conducted in this study, Goksu-Gokdere basin located in the south east of Turkey was chosen (Figure 2), which was named in 1805. The area of the basin covers about 1790 km\(^2\), with a steep average slope of 23\%, a highly varying elevation of 319-2967 m, and a log water path of 192 km. The streamflow measured at Goksu-Gokdere and that upstream measured at Goksu-Himmetli stations were collected from the General Directory of Water Affairs (Ministry of Forests and Water Affairs) for the period February 1973 to September 1994. The rainfall and temperature observations were interpolated by the Inverse Distance Weight method using 17 stations around the basin due to the non-existence of any meteorological station inside the basin. The potential evapotranspiration was obtained from the CruTS3.23 (i.e., locally assessed by [45]) as the observations collected from the General Directory of Water Affairs had more missing values than existing values. The time series of the studied hydrological variables are shown in Figure 3.

B. MODEL DEVELOPMENT

One of the main issues in time series modeling is recognizing the number of delays to be used in the model, which increases the model performance. One of the most common methods used includes the Autocorrelation Function (ACF) and Cross-Correlation Function (CCF) [46]. This method has been criticized by several researchers, such as [47], [48], because the relation between the variables or lags of one variable can be non-linear and cannot be captured by ACF and CCF, which are linear. Another method is a sequential approach in which one lag is added every iteration, until the model performance does not increase or starts decreasing; this lag is identified as the optimum lag [49]. In this study, both of these methods were applied, and the optimum lags found were 2, 2, 1, 1, and 1 for downstream flow, upstream flow, rainfall, temperature, and evapotranspiration, respectively.

After determination of the optimal lags, a combination of several models was developed. The preliminary model only contained the downstream variables and the other variables were added to the following models one by one, to investigate their effect on the modeling performance. Finally, a model containing all of the variables was developed (Table 1). It is worth mentioning that the variable in the developed model does not refer to the variable itself, but to its optimum lag(s).

Normalization was applied to the data set for every model combination using

\[
y = (b - a) \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} + a,
\]

where \(x\), \(x_{\text{min}}\), and \(x_{\text{max}}\) represent the least and the maximum values of the variable \(x\), respectively, and \(a\) and \(b\) represent the lowest and highest values of the normalized data, respectively. Normalization was implemented between 0.1 and 1 in this study.

### TABLE 1. The constructed input combinations for river flow modeling.

<table>
<thead>
<tr>
<th>Combination No.</th>
<th>Abbreviation</th>
<th>Model Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DS</td>
<td>DS only</td>
</tr>
<tr>
<td>2</td>
<td>DSUS</td>
<td>DS + US</td>
</tr>
<tr>
<td>3</td>
<td>DSR</td>
<td>DS + R</td>
</tr>
<tr>
<td>4</td>
<td>DST</td>
<td>DS + T</td>
</tr>
<tr>
<td>5</td>
<td>DSPET</td>
<td>DS + PET</td>
</tr>
<tr>
<td>6</td>
<td>DSUSRPTET</td>
<td>DS + US + R + T + PET</td>
</tr>
</tbody>
</table>

DWT and CWT use a mother wavelet function to transform the time series into the time-frequency domain. There are a number of functions that can be used and identification of the function that gives the highest performance is another task in WT-based hybrid models. In this study, db7 and Haar (i.e., db1) were determined as the best functions for DWT and CWT, respectively. Another important issue in DWT is the determination of the optimum decomposition level. Two levels were found to be the best in this study. For CWT, 128 scales were used as the highest possible number of scales based on the number of data points utilized. Two schemas were applied: schema I: DWT-ELM, DWT-XGB, and CWT-XGB, and schema II: CWT-XGB-ELM.

**C. STAND-ALONE EXPERIMENT**

The first experiment was conducted to estimate the one month-ahead downstream flow using the bare models without WT. In this experiment, the lagged variables were imposed in the XGB and ELM directly as inputs and the one month-ahead flow was employed as the output. In terms of dividing the time series, the normalized data were divided into 75% for training and 25% for testing. This is obviously forecasting because of the inclusion of $Q_{t+1}$ as the output and taking the first model combination as an example, the inputs were the current $Q_t$ and one month-earlier $Q_{t-1}$ downstream flow.

**D. HINDCAST EXPERIMENT**

A large number of studies which have applied WT-based hybrid models used a hindcast experiment and incorrectly named it forecasting [35]. In the hindcast experiment, the data set was transformed with WT and then divided into training and testing subsets to be imposed in DDMs. The outline of the hindcast experiment is shown in Figure 4 and the steps are as follows:

i. After obtaining the optimum lags and normalizing them, a set of variables was obtained for every model combination $V = v_1, v_2, \ldots, v_n$.

ii. The normalized lags of the variables were decomposed using a certain mother wavelet, a specific decomposition level for DWT, and the highest possible scale for CWT. For DWT, the decomposed coefficients were reconstructed to obtain the sub time series, namely $A_1, D_1, D_2, \ldots, D_j$, a set of sub time series, was obtained for each variable $v_i^{\text{dwt}} = \{A_j, D_1, D_2, \ldots, D_j\}$, where $j$ is the decomposition level. For CWT, $v_i^{\text{cwt}} = s_{v1}, s_{v2}, \ldots, s_{vij}, i = 1, 2, \ldots, n$, where $j$ is the highest possible scale and $s_{v1}$ is a vector of the coefficients of the first scale of the $i$ variable;

iii. The sub time series/scales for all variables were gathered as $TS_{\text{dwt}}^{\text{calib}} = \{S_v^{S1} v_2^{\ldots} v_n^{\text{dwt}}\}$ and $TS_{\text{cwt}}^{\text{calib}} = \{S_v^{S1} S_{v2}^{\ldots} S^{\text{cwt}}\}$ for DWT and CWT, respectively, and divided into two subsets, in which 75% of them were considered for calibration $TS_{\text{dwt}}^{\text{calib}} (TS_{\text{cwt}}^{\text{calib}})$ and 25% for were considered testing purposes $TS_{\text{dwt}}^{\text{test}} (TS_{\text{cwt}}^{\text{test}})$ in DWT (CWT).

For the second schema, the $TS_{\text{dwt}}^{\text{calib}} (TS_{\text{cwt}}^{\text{calib}})$ were imposed as inputs in ELM and XGB for training and $Q_{t+1, \text{calib}}$ was the output, and the models with the best parameters were then chosen. The $TS_{\text{dwt}}^{\text{test}} (TS_{\text{cwt}}^{\text{test}})$ and $Q_{t+1, \text{test}}$ were imposed in the best models for testing and the evaluation criteria were obtained. In the third schema, the proposed model (the details are given in the proposed modeling schema section) was applied. The feature importance obtained from XGB was utilized, showing the importance of the scales in the order $s_{v1}, s_{v2}, \ldots, s_{vij}$, where $1$ is the most important scale and $m$ is the least important scale. The training of ELM used the matrix consisting of only the important scale of the calibration subset $TS_{\text{cwt}}^{\text{calib}} = \{S_v^{S1}, S_{v2}^{\ldots}, S_{vij}^{m}\}$ and the testing was implemented using $TS_{\text{cwt}}^{\text{test}}$.

**E. FORECAST EXPERIMENT**

In this experiment, a real forecast experiment was conducted without any future information for building the predictive model. The framework of the forecast experiment is shown in Figure 5 and the steps are as follows:

i. After obtaining the optimum lags and normalizing them, a set of variables was obtained for every model combination $V = v_1, v_2, \ldots, v_n$.

ii. The data were divided into two subsets: 75% for calibration $V_{\text{CALIB}}$ and 25% for testing $V_{\text{TEST}}$. The targets $Q_{t+1}$ were also divided into 75% for calibration $Q_{t+1, \text{CALIB}}$ and 25% to get $Q_{t+1, \text{TEST}}$.
iii. The \( V_{\text{CALIB}} \) and \( Q_{t+1, \text{CALIB}} \) series were denoted by \( 1, 2, \ldots, k \);

iv. The \( V_{\text{CALIB}} \) was decomposed using a certain mother wavelet, a specific decomposition level for DWT, and the highest possible scale for CWT. For DWT, the decomposed coefficients were reconstructed to obtain the sub time series, namely \( A_1, D_1, D_2, \ldots, D_j \). A set of sub time series was obtained for each variable \( S_{dwt}^{vi} = \{A_1, D_1, D_2, \ldots, D_j\} \), where \( j \) is the decomposition level. For CWT, \( \{S_{cwt}^{vi} = s_{vij}, s_{vij+1}, \ldots, s_{vij+i}, i = 1, 2, \ldots, n\} \), where \( j \) is the highest possible scale and \( s_{vij} \) is a vector of the coefficients of the first scale of the \( i \)th variable;

v. The sub time series/scales for all variables were gathered as \( T_{dwt}^{vi} = \{S_{dwt}^{vi} = s_{dwi1}, s_{dwi2}, \ldots, s_{dwiij}\} \) and \( T_{cwt}^{vi} = \{S_{cwt}^{vi} = s_{cwi1}, s_{cwi2}, \ldots, s_{cwiij}\} \) for DWT and CWT, respectively, and divided into two parts: 75% for calibration \( T_{dwt}^{vij, \text{CALIB}} \) (and \( T_{cwt}^{vij, \text{CALIB}} \)) and 25% for testing \( T_{dwt}^{vij, \text{TEST}} \) (and \( T_{cwt}^{vij, \text{TEST}} \)) for DWT(CWT). The target \( Q_{t+1, \text{CALIB}} \) was also divided into \( Q_{t+1, \text{CALIB}} \) and \( Q_{t+1, \text{TEST}} \) with the same percentages;

vi. For the second schema, the \( T_{dwt}^{vij, \text{CALIB} \text{CALIB}} \) and \( T_{cwt}^{vij, \text{CALIB}} \) were imposed as inputs in ELM and XGB for training and \( Q_{t+1, \text{CALIB}} \) was the output, and the models with the best parameters were then chosen and tested using \( T_{dwt}^{vij, \text{TEST}} \) (and \( T_{cwt}^{vij, \text{TEST}} \)) and \( Q_{t+1, \text{TEST}} \);

vii. In the third scheme, the proposed model was applied. The features importance obtained from XGB implemented in step 6 was utilized. The training of ELM used the matrix consisting of only the important scale of the calibration subset \( T_{cwt}^{vij, \text{CALIB} \text{CALIB}} = \{s_{cwi1}, s_{cwi2}, \ldots, s_{cwiij}\} \) and the testing was implemented using \( T_{cwt}^{vij, \text{TEST}} \);

viii. In both schemas, the first value in the \( V_{\text{TEST}} \) was appended into \( V_{\text{CALIB}} \) to obtain a series \( V_{\text{CALIB}} \) with a length of \( 1 - k + 1 \);

ix. The \( V_{\text{CALIB}} \) was decomposed using the same mother wavelet, decomposition level, and highest possible scale. The last value of the decomposed series \( T_{dwt}^{vij, \text{CALIB}} \) (and \( T_{cwt}^{vij, \text{CALIB}} \)) was imposed in the models obtained from 6 and 7 to predict the \( Q_{t+1, \text{CALIB}} \) value and save the predicted value;

x. The first value was appended from \( V_{\text{TEST}} \) into \( V_{\text{CALIB}} \) and from \( Q_{t+1, \text{TEST}} \) into \( Q_{t+1, \text{CALIB}} \) to obtain series with lengths \( 1, 2, \ldots, k + 1 \);

xi. Steps 1–9 were repeated until all of the \( V_{\text{TEST}} \) was appended into \( V_{\text{CALIB}} \);
The evaluation criteria were determined based on the root mean square error (RMSE) and the coefficient of efficiency (CE) [50].

V. APPLICATION RESULTS AND DISCUSSION

The stand-alone experiment consisted of imposing the normalized lags of the different combinations in the model as inputs and the one month-ahead downstream flow $Q_{t+1}$ as the output, after the division of the data set into calibration and test subsets. In this type of modeling, the future information is not included and considered as forecasting as no decomposition is applied to the original time series. The results of the two methods of ELM and XGB are listed in Table 2 and 3. The ELM results indicate that including any of the variables lags beside the lags of the downstream flow (i.e., DS combination) increases the performance of the model in general. Although including the lags of all variables (i.e., RTPETUSDS combination) increases the performance in comparison to the DS combination, the model containing the PET and DS (i.e., PETDS combination) exhibits the highest performance with minimum (RMSE = 18.635 m$^3$/s) and maximum (CE = 0.68) over the testing phase. Using XGB model also shows that including other variables with the DS increases the performance of the models, but the highest performance is achieved for the combination that consists of all the variables (i.e., RTPETUSDS combination) and not the PETDS combination, as in ELM. XGB mostly outperforms the ELM in the calibration subset, indicating overtraining that is normal in tree-based models and needs careful parameter tuning. However, the best prediction results attained for the TDS input combination with minimum (RMSE = 19.946 m$^3$/s) and maximum (CE = 0.633) over the testing phase.

### TABLE 2. The CE and RMSE of the stand-alone schema of the extreme learning machine (ELM) method.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Calibration</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CE RMSE (m$^3$/s)</td>
<td>CE RMSE (m$^3$/s)</td>
</tr>
<tr>
<td>DS</td>
<td>0.539 37.135</td>
<td>0.504 23.195</td>
</tr>
<tr>
<td>USDS</td>
<td>0.648 32.427</td>
<td>0.603 20.740</td>
</tr>
<tr>
<td>RDS</td>
<td>0.626 33.436</td>
<td>0.579 21.365</td>
</tr>
<tr>
<td>TDS</td>
<td>0.710 29.429</td>
<td>0.662 19.131</td>
</tr>
<tr>
<td>PETDS</td>
<td>0.789 25.141</td>
<td>0.680 18.635</td>
</tr>
<tr>
<td>RTPETUSDS</td>
<td>0.751 27.282</td>
<td>0.676 18.739</td>
</tr>
</tbody>
</table>

### TABLE 3. The CE and RMSE of the stand-alone schema of the extreme gradient boosting (XGB) method.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Calibration</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CE RMSE (m$^3$/s)</td>
<td>CE RMSE (m$^3$/s)</td>
</tr>
<tr>
<td>DS</td>
<td>0.661 31.844</td>
<td>0.369 26.152</td>
</tr>
<tr>
<td>USDS</td>
<td>0.768 26.309</td>
<td>0.535 22.454</td>
</tr>
<tr>
<td>RDS</td>
<td>0.711 29.393</td>
<td>0.414 25.198</td>
</tr>
<tr>
<td>TDS</td>
<td>0.772 26.082</td>
<td>0.633 19.946</td>
</tr>
<tr>
<td>PETDS</td>
<td>0.831 22.493</td>
<td>0.594 20.986</td>
</tr>
<tr>
<td>RTPETUSDS</td>
<td>0.861 20.394</td>
<td>0.630 20.021</td>
</tr>
</tbody>
</table>

In both the hindcast and forecast experiments, the two schemas were implemented. In schema I, the data were decomposed using DWT (CWT). In this schema, CWT-ELM was not applied as the number of inputs was huge, which would deteriorate the performance dramatically. In schema II, only the important scales of the CWT obtained by XGB were included in the model and the hybrid model produced was CWT-XGB-ELM. The hindcast experiment was implemented in such a way that the inputs were decomposed and divided into calibration and testing subsets, which were then imposed in ELM or XGB. The results of the four hybrid models of this experiment are shown in Table 4. The use of DWT as a pre-processing tool increased the performance of the models dramatically in comparison with the stand-alone models, especially with ELM, which performed better than XGB. However, the CWT-XGB hybrid method behaved better than DWT with XGB in DWT-XGB and better than the stand-alone models. The proposed hybrid model CWT-XGB-ELM resulted in the highest prediction performance, with highest accuracy in the PETDS combination being 0.987 and 0.973 CE and 6.182 and 5.413 m$^3$/s for the calibration and test, respectively.

The reason for this dramatic increase in the hybrid models is that the decomposition and reconstruction of the times series resulted in some future information being included in the model. This information was included in the decomposed time series due to the convolution process of the original time series with the filters. For DWT, level 2 was used, which produced three reconstructed subseries: one approximation (A) and two details (D1, D1). The sub time series of the 200 months and 265 months are displayed in Figure 6. The difference between the 200 sub time series and the first 200 steps of the 256 sub time series of the downstream flow shown in Figure 7 indicates that the decomposition of the 256 time series includes some of the future values. These
TABLE 4. The CE and RMSE of the hindcast experiment.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Calibration CE (m³/s)</th>
<th>Calibration RMSE (m³/s)</th>
<th>Test CE (m³/s)</th>
<th>Test RMSE (m³/s)</th>
<th>DWT-ELM</th>
<th>DWT-XGB</th>
<th>CWT-XGB</th>
<th>CWT-XGB-ELM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>0.965</td>
<td>10.176</td>
<td>0.944</td>
<td>7.783</td>
<td>0.931</td>
<td>14.410</td>
<td>0.841</td>
<td>13.113</td>
</tr>
<tr>
<td>USDS</td>
<td>0.962</td>
<td>10.784</td>
<td>0.948</td>
<td>7.444</td>
<td>0.929</td>
<td>14.594</td>
<td>0.818</td>
<td>14.007</td>
</tr>
<tr>
<td>RDS</td>
<td>0.963</td>
<td>10.489</td>
<td>0.925</td>
<td>8.865</td>
<td>0.931</td>
<td>14.370</td>
<td>0.797</td>
<td>14.842</td>
</tr>
<tr>
<td>TDS</td>
<td>0.967</td>
<td>9.969</td>
<td>0.946</td>
<td>7.624</td>
<td>0.931</td>
<td>14.370</td>
<td>0.797</td>
<td>14.842</td>
</tr>
<tr>
<td>PETDS</td>
<td>0.959</td>
<td>11.101</td>
<td>0.942</td>
<td>7.859</td>
<td>0.929</td>
<td>14.520</td>
<td>0.797</td>
<td>14.841</td>
</tr>
<tr>
<td>RTPETUSDS</td>
<td>0.970</td>
<td>9.552</td>
<td>0.943</td>
<td>7.795</td>
<td>0.933</td>
<td>14.121</td>
<td>0.830</td>
<td>13.569</td>
</tr>
</tbody>
</table>

TABLE 5. The importance matrix of the CWT scales obtained by XGB. The value in the brackets represents the scale.

<table>
<thead>
<tr>
<th>Im. Rank</th>
<th>DS</th>
<th>USDS</th>
<th>RDS</th>
<th>TDS</th>
<th>PETDS</th>
<th>RTPETUSDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DS₁₅(5)</td>
<td>DS₁₅(5)</td>
<td>DS₁₅(5)</td>
<td>DS₁₅(5)</td>
<td>DS₁₅(5)</td>
<td>US₁₅(4)</td>
</tr>
<tr>
<td>2</td>
<td>DS₁₈(8)</td>
<td>US₁₈(4)</td>
<td>DS₁₈(3)</td>
<td>DS₁₈(8)</td>
<td>DS₁₈(7)</td>
<td>DS₁₈(5)</td>
</tr>
<tr>
<td>3</td>
<td>DS₁₃(3)</td>
<td>US₁₃(10)</td>
<td>DS₁₃(8)</td>
<td>DS₁₃(6)</td>
<td>DS₁₃(8)</td>
<td>DS₁₃(3)</td>
</tr>
<tr>
<td>4</td>
<td>DS₁₇(7)</td>
<td>DS₁₇(1)</td>
<td>DS₁₇(6)</td>
<td>DS₁₇(7)</td>
<td>DS₁₇(1)</td>
<td>US₁₇(6)</td>
</tr>
<tr>
<td>5</td>
<td>DS₁₆(6)</td>
<td>US₁₆(7)</td>
<td>DS₁₆(11)</td>
<td>DS₁₆(1)</td>
<td>DS₁₆(3)</td>
<td>DS₁₆(1)</td>
</tr>
<tr>
<td>6</td>
<td>DS₁₂(1)</td>
<td>DS₁₂(3)</td>
<td>DS₁₂(11)</td>
<td>DS₁₂(4)</td>
<td>DS₁₂(6)</td>
<td>US₁₂(7)</td>
</tr>
<tr>
<td>7</td>
<td>DS₁₃(4)</td>
<td>US₁₃(8)</td>
<td>DS₁₃(3)</td>
<td>PET₁₃(11)</td>
<td>US₁₃(12)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>DS₁₂(9)</td>
<td>US₁₂(8)</td>
<td>US₁₂(8)</td>
<td>US₁₂(10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>US₁₂(5)</td>
<td>US₁₂(5)</td>
<td>US₁₂(5)</td>
<td>US₁₂(5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>DS₁₄(6)</td>
<td>DS₁₄(6)</td>
<td>DS₁₄(2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>US₁₄(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 7. The difference between the first 200 steps of the sub time series of the downstream flow using continuous wave transformation (DWT) with level 2 for 200 and 256 months.

According to that, the number of different values in D1 is 12 and in D2 and A2 is 36. The calculation of the number of different values can be found in detail in [35].

For CWT, the time series were transformed using 128 scales for every variable. Using the proposed method, the importance matrix of the most important scales involved in the modeling process was obtained using XGB. According to the results of this importance matrix, the 5th scale of the first lag of downstream flow Q₁ was found to be the most important scale for all model combinations (Table 5). Therefore, a comparison of the 200 and 256 time series was conducted on this scale only. Haar (i.e., db1), which has a filter length of 2, was found to be the best function of the CWT analysis. Transformation of the original time series (scale 5 is shown in Figure 8) leads to the same issue of different values at the end of the decomposed finite time series. As the filter length is only 2, there are only three different values between the 200 and 256 decomposed coefficients and the differences vary, based on the border effect method used, which arises in finite time series [25]. As the best wavelet function found is db7 for this study, the filter length is 14.

FIGURE 8. The coefficients of the 5th scale of the downstream flow using CWT.

FIGURE 9. The difference between the first 200 months of coefficients of the 5th scale of the downstream flow using CWT for 200 and 256 months.
TABLE 6. The CE and RMSE of the forecast experiment.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>DWT-ELM CE</th>
<th>DWT-ELM RMSE (m³/s)</th>
<th>DWT-XGB CE</th>
<th>DWT-XGB RMSE (m³/s)</th>
<th>CWT-XGB CE</th>
<th>CWT-XGB RMSE (m³/s)</th>
<th>CWT-XGB-ELM CE</th>
<th>CWT-XGB-ELM RMSE (m³/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>0.40</td>
<td>26.11</td>
<td>0.36</td>
<td>27.05</td>
<td>-0.09</td>
<td>35.21</td>
<td>-0.24</td>
<td>37.62</td>
</tr>
<tr>
<td>USDS</td>
<td>0.31</td>
<td>27.93</td>
<td>0.34</td>
<td>27.50</td>
<td>-0.15</td>
<td>36.14</td>
<td>-1.24</td>
<td>50.57</td>
</tr>
<tr>
<td>RDS</td>
<td>0.41</td>
<td>25.91</td>
<td>-0.35</td>
<td>39.21</td>
<td>-0.06</td>
<td>34.70</td>
<td>-0.33</td>
<td>38.91</td>
</tr>
<tr>
<td>TDS</td>
<td>0.46</td>
<td>24.82</td>
<td>0.52</td>
<td>23.28</td>
<td>-0.15</td>
<td>36.18</td>
<td>-2.39</td>
<td>62.17</td>
</tr>
<tr>
<td>PETDS</td>
<td>0.49</td>
<td>24.21</td>
<td>0.47</td>
<td>24.49</td>
<td>-0.20</td>
<td>37.05</td>
<td>-2.11</td>
<td>59.57</td>
</tr>
<tr>
<td>RTPETUSDS</td>
<td>0.44</td>
<td>25.27</td>
<td>-0.11</td>
<td>35.49</td>
<td>-0.92</td>
<td>46.80</td>
<td>-0.92</td>
<td>46.80</td>
</tr>
</tbody>
</table>


have very high magnitudes (Figure 9). Although the different values are not as high as those obtained when applying db7 in DWT, some future information is still passed to the model, causing a dramatic intensification in the performance of the models.

According to the previous discussion, the most important issue in WT-based models is the issue of the border effect on the finite time series, which is the case for all hydrological time series and not only those relating to streamflow. Therefore, decomposition of the time series before the division is an inaccurate practice of forecasting, as some of the future information is included in the modeling due to convolution. In a real example, the streamflow of the current month to be forecasted contains no future information as the data is recorded up to that month. Therefore, we have to decompose the time series and include all of the previous records in this decomposition, in order to then impose them in the DDM for forecasting. In this case, the ends are treated differently, according to the method of the border effect used. These distorted ends are not included in the hindcast experiment. In fact, they are only included at the end of the whole original time series, which is not the real case. Therefore, a real forecast experiment is conducted.

In the forecast experiment, the modeling is done in such a way that to forecast the coming month, all of the inputs up until the previous month are used for calibrating the model after the decomposition and these inputs until the current month are decomposed and imposed in the calibrated model for forecasting, as in reality, only data up to the current month exists. This procedure greatly depends on the ends of the decomposed time series. The results of this experiment are shown in Table 6. According to these results, the performance of all of the combinations deteriorated in comparison to the stand-alone experiment. The DWT-ELM showed the highest performance in this forecast experiment. The proposed model (CWT-XGB-ELM) had the worst performance, while in the hindcast experiment, it showed a dramatic increase in the performance. This deterioration in the proposed model caused by the high distortion at the ends of the CWT-based transformed coefficients resulted from the boundary effect.

The fitted values of the best performing models of the three experiments are plotted versus the observed values in Figure 10. The figure shows the almost perfect agreement of points with the best fit line for the hindcast experiment, but this has been proved to be an incorrect application of WT-based hybrid models. For forecasting, only 56 points are shown, as only these values were forecasted, while precedent was used for calibration. The agreement between these points and the best fit line is much worse than that of the stand-alone experiment. In brief, the use of WT in both DWT and CWT deteriorates the model’s performance and stand-alone models outperform hybrid models.

Finally, it is worth to validate the current research results with the reported literature review studies on the streamflow
modeling using hybrid models between the WT and data-driven models. The reported modeling results conducted by [29] stated that the minimum RMSE were attained (24.79 and 16.72 m³/s) using WT-ANN model at Yingluoxia and Zhamaheke stations over the testing phase. In another research [51], authors developed WT-linear genetic programming (WT-LGP) model for streamflow forecasting at Pataveh and Shahmokhtar stations on Beshar River, Iran. The developed predictive model attained minimum (RMSE = 19.664 and 17.96 m³/s) at the Pataveh and Shahmokhtar stations over the testing phase, respectively. Apparently, the proposed hybrid predictive model in the current research demonstrated a superior predictability capacity for streamflow modeling over the established literature studies. Further, the developed hybrid CWT-XGB-ELM model revealed a reliable predictive model for streamflow forecasting and river engineering sustainability.

VI. CONCLUSION
In this study, three modeling experiments were applied, including a stand-alone approach in which no WT was employed; a hindcast experiment where DWT (CWT) was utilized to the time series, which were then divided into calibration and testing before entering them into ELM or XGB; and a forecast experiment in which the data was divided and after the model was trained with the first decomposed subset, a value was added from the second subset to the first subset to be decomposed and entered into the ELM or XGB to implement a real forecast. In both the hindcast and forecast experiment, a novel hybrid model was proposed, in which the importance matrix showing the importance of the features (i.e., scales in this study) was utilized so that only the important scales were imposed in the ELM. According to the results obtained, several points can be made:

i. The use of WT-based hybrid models increases the performance of the models in hindcast experiments due to the inclusion of future information, as a consequence of time series convolution;

ii. WT-based hybrid models in the forecast experiment deteriorate the performance of models and stand-alone models perform better, mostly due to the border effect, which distorts the ends of the decomposed time series;

iii. In hindcast experiments, the distortion at the ends of the decomposed finite time series caused by the border effect is only involved at the end of the original time series, while in the forecast experiment, which is considered as real forecasting, it is involved in forecasting every value in the testing subset;

iv. The proposed hybrid model using XGB as a selection tool, in addition of being a modeling approach itself, has dramatically improved the performance of the hindcast experiment, but not the forecast experiment. Therefore, it has the potential to be applied in other applications for reducing the number of features imposed in the modeling approach;

v. The use of metrological and hydrological variables, such as temperature and potential evapotranspiration, beside the lagged streamflow, which is essential for auto regression, improves the performance of the models.

vi. The proposed hybrid model CWT-XGB-ELM was attained the best prediction accuracy using the input combination of PETDS. The reported performance metrics were (CE = 0.987 and 0.973) and (RMSE = 6.182 and 5.413 m³/s) for the calibration and test phases, respectively.

Based on the reported modeling results, there is still space for further modeling enhancement using the feasibility of nature-inspired optimization algorithms for hyperparameter tuning of the ELM model [52].

ACKNOWLEDGMENT
The authors acknowledge their appreciation for the hydrological data provider (the General Directory of Water Affairs (Ministry of Forests and Water Affairs), Turkey).

CONFLICT OF INTEREST
The authors have no conflicts of interest to declare.

REFERENCES


MUSTafa Tombul received the master’s and Ph.D. degrees from Anadolu University, Turkey, in 1991 and 1996, respectively. He became an Associate Professor, in 2009, and a Full Professor, in 2015. He is currently working as a Full Professor at Eskisehir Technical University. He has published many articles in international journals and conferences. His research interest includes water resources engineering, and in particular, he is interested in flood modeling, climate change, and machine learning in water resources.

Sinan Q. Salih received the M.Sc. degree in computer sciences from Universiti Tenaga Nasional (UNITEN), in 2012, and the Ph.D. degree from Universiti Malaysia Pahang (UMP), Malaysia, in 2018. He is major in machine learning, optimization, and data analytic. His current research interests include optimization algorithms, nature-inspired metaheuristics, machine learning, and feature selection 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 at the Department of Civil, Environmental and Natural Resources Engineering, Lulea Technical University, Sweden. Previously, he worked at Baghdad University, from 1976 to 1995, and then at Al-Bayt University, Jordan, from 1995 to 2007. His research interests include geology, water resources, and environment. He has served several academic administrative posts (Dean and Head of the Department). He has published more than 424 articles in international/national journals, chapters in books, and 13 books. He has 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 has supervised more than 66 postgraduate students at universities in Iraq, Jordan, U.K., and Australia. 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 the Founder and President of the Iraqi Scientific Society for Water Resources. He is also a member of the editorial board of ten international journals. He has been awarded several scientific and educational awards. Among them, the British Council named him as one of the top five scientists in Cultural Relations on its 70th Anniversary.

Zaher Mundher Yaseen (Member, IEEE) received the master’s and Ph.D. degrees from the National University of Malaysia (UKM), Malaysia, in 2012 and 2017, respectively. He is currently a Senior Lecturer and a Senior Researcher in the field of civil engineering. He has published over 130 articles in international journals, with a Google Scholar H-index of 26, and a total of 2113 citations. His major interests include hydrology, water resources engineering, hydrological process modeling, environmental engineering, and climate. In addition, he has interested in machine learning and advanced data analytics.