Hindawi
Complexity
Volume 2019, Article ID 2782715, 14 pages
https://doi.org/10.1155/2019/2782715

Research Article

Development of Multidecomposition Hybrid Model for Hydrological Time Series Analysis

Hafiza Mamona Nazir[1], **Ijaz Hussain**[1], **Muhammad Faisal**[2,3], **Alaa Mohamd Shoukry**[4,5], **Showkat Gani**[6] and **Ishfaq Ahmad**[7,8]

[1]*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Faculty of Health Studies, University of Bradford, Bradford BD7 1DP, UK*
[3]*Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK*
[4]*Arriyadh Community College, King Saud University, Riyadh, Saudi Arabia*
[5]*KSA Workers University, El-Mansoura, Egypt*
[6]*College of Business Administration, King Saud University, Al-Muzahimiyah, Saudi Arabia*
[7]*Department of Mathematics, College of Science, King Khalid University, Abha 61413, Saudi Arabia*
[8]*Department of Mathematics and Statistics, Faculty of Basic and Applied Sciences, International Islamic University, 44000 Islamabad, Pakistan*

Correspondence should be addressed to Ijaz Hussain; ijaz@qau.edu.pk

Received 1 October 2018; Accepted 13 December 2018; Published 2 January 2019

Guest Editor: Pedro Palos

Accurate prediction of hydrological processes is key for optimal allocation of water resources. In this study, two novel hybrid models are developed to improve the prediction precision of hydrological time series data based on the principal of three stages as denoising, decomposition, and decomposed component prediction and summation. The proposed architecture is applied on daily rivers inflow time series data of Indus Basin System. The performances of the proposed models are compared with traditional single-stage model (without denoised and decomposed), the hybrid two-stage model (with denoised), and existing three-stage hybrid model (with denoised and decomposition). Three evaluation measures are used to assess the prediction accuracy of all models such as Mean Relative Error (MRE), Mean Absolute Error (MAE), and Mean Square Error (MSE). The proposed, three-stage hybrid models have shown improvement in prediction accuracy with minimum MRE, MAE, and MSE for all case studies as compared to other existing one-stage and two-stage models. In summary, the accuracy of prediction is improved by reducing the complexity of hydrological time series data by incorporating the denoising and decomposition.

1. Introduction

Accurate prediction of hydrological processes is key for optimal allocation of water resources. It is challenging because of its nonstationary and multiscale stochastic characteristics of hydrological process which are affected not only by climate change but also by other socioeconomic development projects. The human activities also effected the climate change through contributing in Earth's atmosphere by burning of fossil fuels which release carbon dioxide in atmosphere. Instead of these, greenhouse and aerosols have made effect on Earth's atmosphere by altering in-out coming solar radiations which is the part of Earth's energy balance. This makes the prediction of hydrological time series data challenging. To predict such hydrological processes, two broad types of models are commonly used, one is the process-based models which further included the lumped conceptual models, hydrological model, and one-two-dimensional hydrodynamic models [1], and the second is data driven models which included autoregressive moving averages and artificial neural network (which are also known as black box models). The process-based models considered the physical mechanism of stochastic hydrological processes, which requires a large amount of data for calibration and validation

[2]. Moreover, physical-based models demand the scientific principles of energy and water movements spatiotemporally. Zahidul [3] concluded that unavailability of sufficient amount of data and scientific knowledge of water movement can lead to poor understanding of hydrological system which makes the hydrological modeling a challenging task. In order to overcome these drawbacks, hydrologists used data driven models to efficiently model the hydrological process [4, 5]. The data driven models only take the advantage of inherent the input-output relationship through data manipulation without considering the internal physical process. The data-driven models are efficient over the process-driven models by appreciating the advantage of less demanding the quantitative data, simple formulation with better prediction performance [6]. These data-driven models are further divided into two categories: simple traditional statistical techniques and more complex machine learning methods. In the last few decades, many traditional statistical time series models are developed including Autoregressive (AR), Moving Averages (MA), Autoregressive Moving Averages (ARMA), and Autoregressive Integrated Moving Averages (ARIMA) [7]. Application of ARIMA model to monitoring hydrological processes like river discharge is common and successfully applied [8]. But the problem with all these traditional statistical techniques required that the time series data to be stationary. However, hydrological data was characterized as both nonstationary and nonlinear due to its time varying nature. Therefore, these techniques are not enough to capture the nonlinear characteristics of hydrological series [6]. To rescue the drawbacks of existing traditional models, machine learning (ML) algorithms have been put forward and widely exploited, which provide powerful solution to the instability of hydrological time series data [4]. ML models include Artificial Neural Network (ANN), Support Vector Machine (SVM), and random forest and genetic algorithms [9–14]. Riad et al. [5] developed an ANN to model the nonlinear relation between rainfall and runoff and concluded that ANN model is better to model the complex hydrological system over the traditional statistical models. However, these ML methods have their own drawbacks such as overfitting and being sensitive to parameter selection. In addition, there are two main drawbacks of using ML models: first is that ML models ignore the time varying characteristics of hydrological time series data and secondly the hydrological data contains noises which deprive the researchers to accurately predict the hydrological time series data in an effective way [15]. These time varying and noise corrupted characteristics of hydrological time series data require hybrid approaches to model the complex characteristics of hydrological time series data [16].

To conquer the limitations of existing single models, some hybrid algorithms such as data preprocessing methods are utilized with data-driven models with the hope to enhance the prediction performance of complex hydrological time series data by extracting time varying components with noise reduction. These preprocess based hybrid models have already been applied in hydrology [2]. The framework of hybrid model usually comprised "decomposition," "prediction," and "ensemble" [2, 6, 13]. The most commonly used data preprocessing method is wavelet analysis (WA) which is used to decompose the nonlinear and nonstationary hydrological data into multiscale components [13]. These processed multiscale components are further used as inputs in black box models at prediction stage and finally predicted components are ensemble to get final predictions. Peng et al. [6] proposed hybrid model by using empirical wavelet transform and ANN for reliable stream flow forecasting. They demonstrated their proposed hybrid model efficiency over single models. Later, Wu et al. [11] exploited a two-stage hybrid model by incorporating Wavelet Multi-Resolution Analysis (WMRA), and other data preprocessing methods as MA, singular spectrum analysis with ANN to enhance the estimate of daily flows. They proposed five models including ANN-MA, ANN-SSA1, ANN-SSA2, ANN-WMRA1, and ANN-WMRA2 and suggested that decomposition with MA model performs better than WMRA. An improvement in wavelet decomposition method has been made to get more accurate hybrid results comprising WA [17]. However, the problem which reduces the performance of WA, i.e., selection of mother wavelet basis function, is still an open debate as the selection of mother wavelet is subjectively determined among many wavelet basis functions. The optimality of multiscale characteristics entirely depends on the choice of mother wavelet function as poorly selected mother wavelet function can lead to more uncertainty in time-scale components. To overcome this drawback, Huang et al. [18] proposed a purely data-driven Empirical Mode Decomposition (EMD) technique. The objective of EMD is to decompose the nonlinear and nonstationary data adaptively into number of oscillatory components called Intrinsic Mode Decomposition (IMF). A number of studies have been conducted combining the EMD with data driven models [15, 18–21]. Specifically in hydrology, EMD is used with ANN for wind speed and stream flow prediction [15, 20]. Agana and Homaifar [21] developed the EMD-based predictive deep belief network for accurately predicting and forecasting the Standardized Stream flow Index (SSI). Their study manifested that their proposed model is better than the existing standard methods with the improvement in prediction accuracy of SSI. However, Kang et al. [22] revealed that EMD suffers with mode mixing problem which ultimately affects the efficiency of decomposition. In order to deal with this mode mixing problem, Wu and Hang [23] proposed an improved EMD by successively introducing white Gauss noise in signals, called Ensemble Empirical Mode Decomposition (EEMD) that addresses the issue of frequent apparent of mode mixing in EMD. Later, EEMD was effectively used as data decomposition method to extract the multiscale characteristics [24–26]. Di et al. [2] proposed a four-stage hybrid model (based on EEMD for decomposition) to improve the prediction accuracy by reducing the redundant noises and concluded that coupling the appropriate data decomposition with EEMD method with data driven models could improve the prediction performance compared to existing EMD based hybrid model. Jiang et al. [26] proposed another two-stage hybrid approach coupling EEMD with data-driven models to forecast high speed rail passenger flow to estimate daily ridership. They suggested that their proposed hybrid model is more suitable

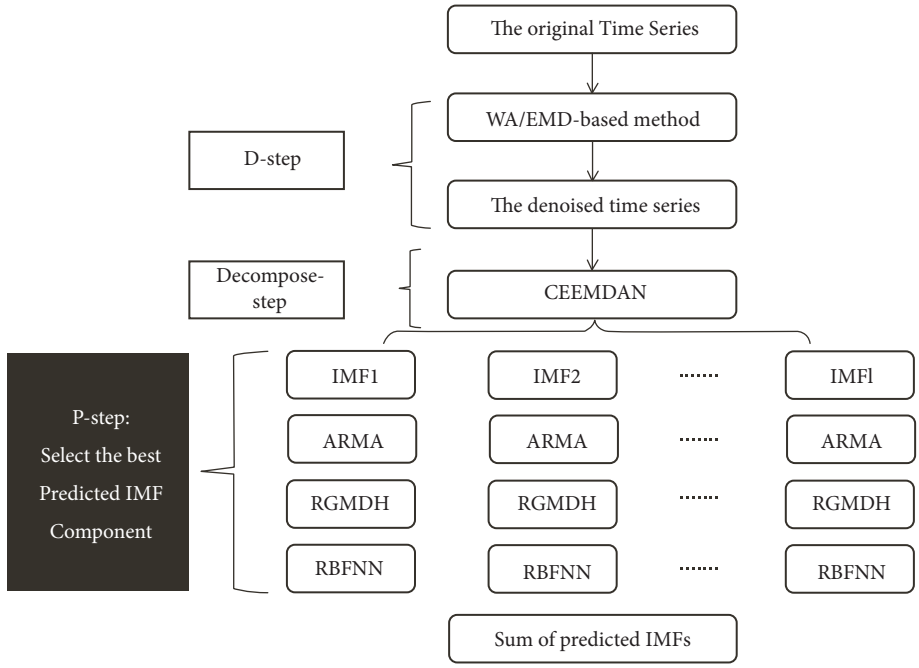

Figure 1: The proposed WA/EMD-CEEMDAN-MM structure to predict hydrological time series data.

for short term prediction by accounting for the day to day variation over other hybrid and single models. However, due to successive addition of independent white Gauss noise, the performance of EEMD is affected which reduces the accuracy of extracted IMFs through EEMD algorithm. Dai et al. [27] reported in their study that EEMD based hybrid models did not perform appropriately due to independent noise addition.

This study aimed to develop a robust hybrid model to decompose the hydrological time varying characteristics using CEEMDAN [28]. The CEEMDAN successively adds white noise, following the steps of EEMD, with interference in each decomposition level to overcome the drawback of EEMD algorithm. Dai et al. [27] developed a model comprising CEEMDAN to daily peak load forecasting which shows robust decomposed ability for reliable forecasting. Therefore, the purpose of using CEEMDAN method for decomposition in this study is to find an effective way to decompose the nonlinear data which enhances the prediction accuracy of the complex hydrological time series data [27].

2. Proposed Methods

In this study, two novel approaches are proposed to enhance the prediction accuracy of the hydrological time series. Both models have the same layout except in stage of denoising, where two different approaches have been used to remove noises from hydrological time series data. In both models, at decomposition stage, an improved version of EEMD, i.e., CEEMDAN, is used to find oscillations, i.e., the high to low frequencies in terms of IMF. At prediction stage, multi-models are used to accurately predict the extracted IMFs by considering the nature of IMFs instead of using only single stochastic model. The purpose of using multimodel is two-way: one is for accurately predicting the IMFs by considering the nature of IMFs and the other is to assess the performance of simple and complex models after reducing the complexity of hydrological time series data through decomposition. Predicted IMFs are added to get the final prediction of hydrological time series. The proposed three stages involve denoising (D-step), decomposition (Decompose-step), and component prediction (P-step), which are briefly described below:

(1) **D-step**: WA and EMD based denoising methods are presented to remove the noises from hydrological time series data.

(2) **Decomposed-step:** Using CEEMDAN, two separately denoised series are decomposed into k IMFs and one residual.

(3) **P-step:** The denoised-decomposed series into k IMFs and one residual are predicted with linear stochastic and nonlinear machine learning models. The model with the lowest error rate of prediction is selected by three performance evaluation measures. Finally the predicted results are added to get the final prediction.

For convenient, two proposed methods as named as EMD (denoising), CEEMDAN (decomposing), MM (multi-models) i.e. EMD-CEEMDAN-MM and WA (denoising), CEEMDAN (denoising) and MM (multi-models) i.e. WA-CEEMDAN-MM. The proposed architecture of WA/EMD-CEEMDAN-MM is given in Figure 1.

2.1. D-Step. In hydrology time series data, noises or stochastic volatiles are inevitable component which ultimately reduced

the performance of prediction. To reduce the noise from data, many algorithms have been proposed in literature such as Fourier analysis, spectral analysis, WA, and EMD [29], as besides decomposition, these techniques have the ability to remove the noises from data. However, the spectral and Fourier analysis only considered the linear and stationary signals, whereas WA and EMD have the ability to address the nonlinear and nonstationary data with better performance. In this study, WA- and EMD-based threshold are adopted to reduce the stochastic volatiles from the hydrological data.

(i) Wavelet analysis based denoising: in order to remove noises, discrete wavelet threshold method is recognized as powerful mathematical functions with hard and soft threshold. With the help of symlet 8 mother wavelet [30], hydrological time series data is decomposed into approximation and details coefficients with the following equations, respectively [31];

$$a_{j,k} = \sum_{k=0}^{2^{N-j}-1} 2^{-j/2} \emptyset \left(2^{-j}t - k\right) \tag{1}$$

and

$$d_{j,k} = \sum_{j=1}^{J} \sum_{k=0}^{2^{N-j}-1} 2^{-j/2} \varphi \left(2^{-j}t - k\right) \tag{2}$$

After estimating the approximation and details coefficients, threshold is calculated for each coefficient to remove noises. The energy of data is distributed only on few wavelet coefficients with high magnitude whereas most of the wavelet coefficients are noisiest with low magnitude. To calculate the noise free coefficient, hard and soft threshold rules are opted, which are listed as follows, respectively [31]:

$$d'_{j,k} = \begin{cases} d_{j,k} & \left|d_{j,k}\right| \geq T_j \\ 0 & \left|d_{j,k}\right| < T_j \end{cases} \tag{3}$$

and

$$d'_{j,k} = \begin{cases} \operatorname{sgn}\left(d_{j,k}\right)\left(\left|d_{j,k}\right| - T_j\right) & \left|d_{j,k}\right| \geq T_j \\ 0 & \left|d_{j,k}\right| < T_j \end{cases} \tag{4}$$

where T_i is the threshold calculated as $T_j = a\sqrt{2E_j \ln(N)}$, $j = 1, 2, \ldots, J$, where a is constant which takes the values between 0.4 and 1.4 with step of 0.1 and $E_j = median(|d_{j,k}|,\ k = 1, 2, \ldots, N)/0.6745$ is median deviation of all details. Then, the decomposed data is reconstructed using the noise free details and approximations using the following equation:

$$\begin{aligned} \widehat{x(t)} = &\sum_{k=0}^{2^{N-j}-1} a'_{j,k} 2^{-j/2} \emptyset \left(2^{-j}t - k\right) \\ &+ \sum_{j=1}^{J} \sum_{k=0}^{2^{N-j}-1} d'_{j,k} 2^{-j/2} \varphi \left(2^{-j}t - k\right) \end{aligned} \tag{5}$$

where $a'_{j,k}$ is threshold approximation coefficient and $d'_{j,k}$ is threshold detailed coefficient.

(ii) Empirical mode decomposition based denoising: an EMD is data-driven algorithm which has been recently proposed to decompose nonlinear and nonstationary data into several oscillatory modes [18]. Due to adaptive nature, EMD directly decomposes data into number of IMFs by satisfying two conditions as follows: **(a)** From complete data set, the number of zero crossings and the number of extremes must be equal or differ at most by one; **(b)** the mean value of the envelope which is smoothed, through cubic spline interpolation, based on the local maxima and minima should be zero at all points.

The EMD structure is defined as follows:

(1) Identify all local maxima and minima from time series $x(t)$, $(t = 1, 2, \ldots, N)$ and make upper envelope of maxima $e_{\max(t)}$ and lower envelope minima $e_{\min(t)}$ through cubic spline interpolation.

(2) Find the mean of upper and lower envelope $m(t) = (e_{\max(t)} + e_{\min(t)})/2$. Find the difference between original series and extracted mean as

$$h(t) = x(t) - m(t) \tag{6}$$

(3) Check the properties defined in **(a)** and **(b)** of $h(t)$; if both conditions are satisfied then mark this $h(t)$ as i^{th} IMF; then the next step will be to replace the original series by $r(t) = x(t) - h(t)$; if $h(t)$ is not IMF just replace $x(t)$ with $h(t)$.

(4) Repeat the process of (1-3), until the residue $r(t)$ becomes a monotone function from which no further IMFs can be extracted.

Finally, original series can be written as the sum of all extracted IMFs and residue as

$$x(t) = \sum_{i}^{m} h_i(t) + r(t) \tag{7}$$

where m is the number of IMFs, as $(i = 1, 2, \ldots, m)$ and $h_i(t)$ is the i^{th} IMF, and $r(t)$ is the trend of the signal. The way of denoised IMF is the same as mentioned in (3)-(5), except the last two IMFs which are used completely without denoising due to low frequencies. The subscript in EMD-based threshold case in (3)-(5) is replaced with i^{th} according to number of IMFs. The denoised signal is reconstructed as follows:

$$\widehat{x(t)} = \sum_{i=1}^{m-2} h_i(t) + \sum_{i=m-2}^{m} h_i(t) + r(t) \tag{8}$$

2.2. Decompose-Step: Decomposition Step. The EEMD method: the EEMD is a technique to stabilize the problem of mode mixing which arises in EMD and decomposes the nonlinear signals into number which contains the information of local time varying characteristics. The procedure of EEMD is as follows:

(a) Add a white Gaussian noise series to the original data set.

(b) Decompose the signals with added white noise into IMFs using conventional EMD method.

(c) Repeat steps (a) and (b) m^{th} time by adding different white noises ($m = 1, 2, \ldots, l$) in original signals.

(d) Obtain the ensemble means of all IMFs m^{th} ensemble time as the final results as $\overline{IMF}_k = \sum_{m=1}^{l} IMF_k^m / l$, where $k = 1, 2, \ldots, K$ is k^{th} IMF.

The CEEMDAN based decomposition: although the EEMD can reduce the mode mixing problem to some extent, due to the successive addition of white Gauss noise in EEMD, the error cannot be completely eliminated from IMFs. To overcome this situation, CEEMDAN function is introduced by Torres et al. [28]. We employed the CEEMDAN to decompose the hydrological time series data. The CEEMDAN is briefly described as follows:

(1) In CEEMDAN, extracted modes are defined as $\widetilde{IMF}_k$; in order to get complete decomposition we need to calculate the first residual by using the first $\widetilde{IMF}_1$, which is calculated by EEMD as $\widetilde{IMF}_1 = \sum_{m=1}^{l} IMF_1^m / l$.

(2) Then replace $x(t)$ by $r_1(t)$ where $r_1(t) = x(t) - \widetilde{IMF}_1$ and add white Gaussian noises, i.e., $w^m(t)$, m^{th} time in $r_1(t)$, and find the IMF by taking the average of first IMF to get the $\widetilde{IMF}_2$. Calculate $r_2(t) = r_1(t) - \widetilde{IMF}_2$ and repeat (2) until stoppage criteria are meet. However, selection of number of ensemble members and amplitude of white noise is still an open challenge but here in this paper the number of ensemble members is fixed as 100 and standard deviation of white noise is settled as 0.2.

(3) The resulting k^{th} decomposed modes, i.e., $\sum_{k=1}^{K} \widetilde{IMF}_k$, and one residual $R(t)$ are used for further prediction of hydrological time series.

More details of EMD, EEMD, and CEEMDAN are given in [20, 28].

2.3. P-Step

Prediction of All IMF's. In prediction stage, denoised IMFs are further used to predict the hydrological time series data as inputs by using simple stochastic and complex machine learning time series algorithms. The reason of using two types of model is that as first few IMFs contain high frequencies which are accurately predicted through complex ML models whenever, last IMFs contain low frequencies which are accurately predictable through simple stochastic models. The selected models are briefly described as follows.

The IMF prediction with ARIMA model: to predict the IMFs, autoregressive moving average model is used as follows:

$$IMF_t{}^i = \alpha_1 IMF_{t-1}{}^i + \cdots + \alpha_p IMF_{t-p}{}^i + \varepsilon_t{}^i + \beta_1 \varepsilon_{t-1}{}^i + \cdots + \beta_q \varepsilon_{t-q}{}^i \quad (9)$$

Table 1: Transfer functions of GMDH-NN algorithms.

Transfer Functions	
Sigmoid function	$z = \frac{1}{(1+e^{-y})}$
Tangent function	$z = \tan(y)$
Polynomial function	$z = y$
Radial basis function	$z = e^{-y^2}$

Here, $IMF_t{}^i$ is the i^{th} IMF and $\varepsilon_t{}^i$ is the i^{th} residual of CEEMDAN where p is autoregressive lag and q is moving average lag value. Often the case, time series is not stationary; [7] made a proposal that differencing to an appropriate degree can make the time series stationary; if this is the case then the model is said to be ARIMA (p, d, q) where d is the difference value which is used to make the series stationary.

The IMF prediction with group method of data handling type neural network: ANN has been proved to be a powerful tool to model complex nonlinear system. One of the submodels of NN, which is constructed to improve explicit polynomial model by self-organizing, is Group Method of Data Handling-type Neural Network (GMDH-NN) [32]. The GMDH-NN has a successful application in a diverse range of area; however, in hydrological modeling it is still scarce. The algorithm of GMDH-NN worked by considering the pairwise relationship between all selected lagged variables. Each selected combination of pairs entered in a neuron and output is constructed for each neuron. The structure of GMDH-NN is illustrated in Figure 2 with four variables, two hidden and one output layer. According to evaluation criteria, some neurons are selected as shown in Figure 2, four neurons are selected, and the output of these neurons becomes the input for next layer. A prediction mean square criterion is used for neuron output selection. The process is continued till the last layer. In the final layer, only single best predicted neuron is selected. However, the GMDH-NN only considers the two variable relations by ignoring the individual effect of each variable. The Architecture Group Method of Data Handling type Neural Network (RGMDH-NN), an improved form of GMDH-NN, is used which simulates not only the two-variable relation but also their individuals. The model for RGMDH-NN is described in the following equation:

$$y_t = a + \sum_{i=1}^{r} b_i x_i + \sum_{i=1}^{r} \sum_{j=1}^{r} c_{ij} x_i x_j \quad (10)$$

The rest of the procedure of RGMDH-NN is the same as GMDH-NN. The selected neuron with minimum Mean Square Error is transferred in the next layer by using the transfer functions listed in Table 1. The coefficients of each neuron are estimated with regularized least square estimation method as this method of estimation has the ability to solve the multicollinearity problem which is usually the inherited part of time series data with multiple lagged variables.

Radial basis function neural network: to predict the denoised IMFs, nonlinear neural network, i.e., Radial Basis

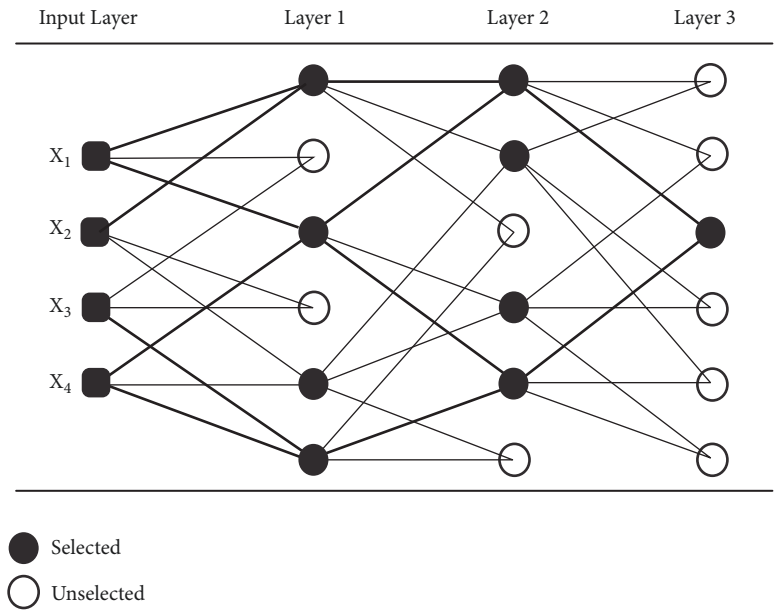

Figure 2: Architecture of GMDH-type neural network (NN) algorithms.

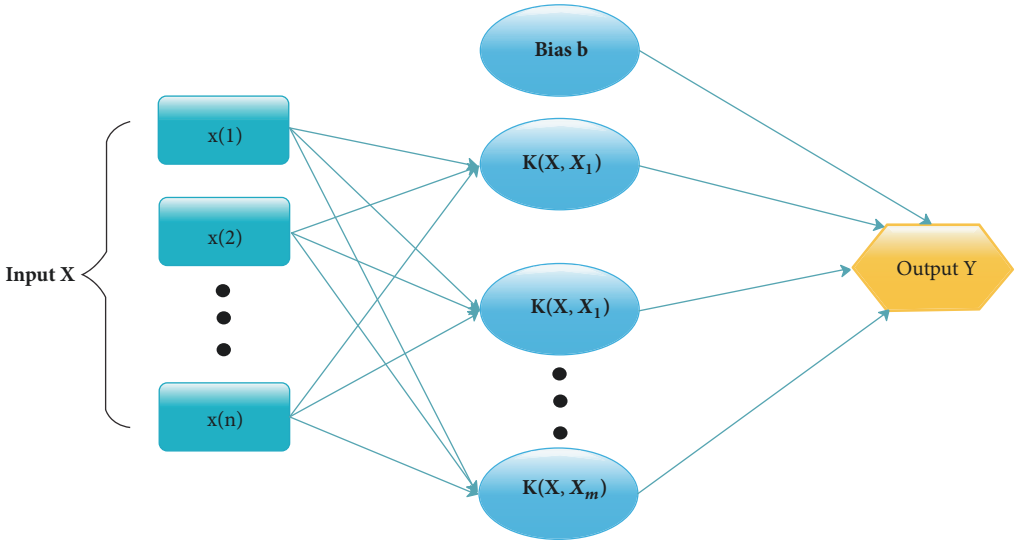

Figure 3: Topological structure of radial basis function.

Function (RBFNN), is also adopted. The reason for selecting RBFNN is that it has a nonlinear structure to find relation between lagged variables. The RBFNN is a three-layer feed-forward neural network which consists of an input layer, a hidden layer, and an output layer. The whole algorithm is illustrated in Figure 3. Unlike GMDH-NN, RBFNN takes all inputs in each neuron with corresponding weights and then hidden layer transfers the output by using radial basis function with weights to output. The sigmoid basis function is used to transfer the complex relation between lagged variables as follows:

$$\emptyset_i(x) = \frac{1}{1 + \exp\left(b^T x - b_0\right)} \tag{11}$$

3. Case Study and Experimental Design

Selection of Study Area. In this study, the Indus Basin System (IBS), known to be the largest river system in Pakistan, is considered which plays a vital role in the power generation and irrigation system. The major tributaries of this river are River Jhelum, River Chenab, and River Kabul. These rivers get their inflows mostly from rainfall, snow, and glacier melt. As in Pakistan, glaciers covered 13,680 km^2 area in which estimated 13% of the areas are covered by Upper Indus Basin (UIB) [33]. About 50% melted water from these 13% areas adds the significant contribution of water in these major rivers. The Indus river and its tributaries cause flooding due to glacier and snow melting and rainfall [34]. The major events

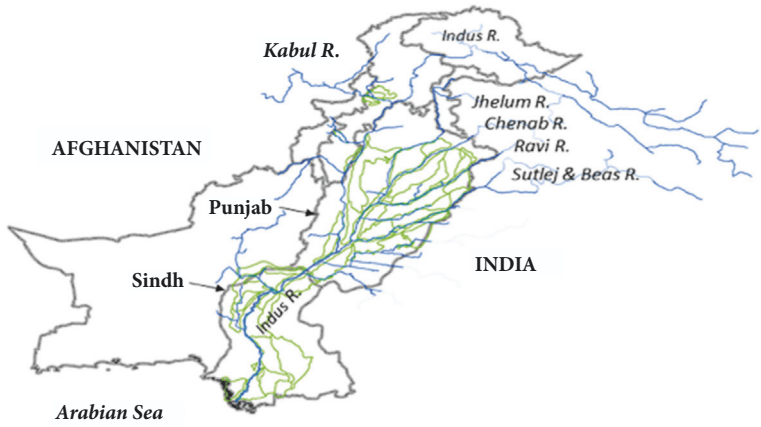

Figure 4: Rivers and irrigation network of Pakistan.

of flood usually occur in summer due to heavy monsoon rainfall which starts from July and end in September. It was reported [35] that, due to excessive monsoon rainfall in 2010, floods have been generated in IBS which affected 14 million people and around 20,000,000 inhabitants were displaced. Moreover, surface water system of Pakistan is also based on flows of IBS and its tributaries [36]. Pappas [37] mentioned that around 65% of agriculture land is irrigated with the Indus water system. Therefore, for effective water resources management and improving sustainable economic and social development and for proactive and integrated flood management, there is a need to appropriately analyze and predict the rivers inflow data of IBS and its tributaries.

Data. To thoroughly investigate the proposed models, four rivers' inflow data is used in this study which is comprised of daily rivers inflow (1^{st}-January to 19^{th}-June) for the period of 2015-2018. We consider the main river inflow of Indus at Tarbela with its two principal, one left and one right, bank tributaries [38]: Jhelum inflow at Mangla, Chenab at Marala, and Kabul at Nowshera, respectively (see Figure 4). Data is measured in 1000 cusecs. The rivers inflow data was acquired from the site of Pakistan Water and Power Development Authority (WAPDA).

Comparison of Proposed Study with Other Methods. The proposed models are compared with other prediction approaches by considering with and without principals of denoising and decomposition. For that purpose, the following types of models are selected:

(I) Without denoising and decomposing, only single statistical model is selected, i.e., ARIMA (for convenience, we call one-stage model 1-S) as used in [8].

(II) Only denoised based models: in this stage, the noise removal capabilities of WA and EMD are assessed. The wavelet based models are WA-ARIMA, WA-RBFNN, and WA-RGMDH whereas the empirical mode decomposition based models are EMD-ARIMA, EMD-RBFNN, and EMD-RGMDH. The different prediction models are chosen for the comparison of traditional statistical models with artificial intelligence based models as RBFN and RGMDH (for convenience, we call two-stage model 2-S). The 2-S selected models for comparison are used from [15, 17] for the comparison with the proposed model.

(III) With denoising and decomposition (existing method): for that purpose, three-stage EMD-EEMD-MM model is used from [2] for the comparison with proposed models. Under this, the multiple models are selected by keeping the prediction characteristics similar to proposed model for comparison purpose (for convenience, we call three-stage model 3-S).

Evaluation Criteria. The prediction accuracy of models is assessed using three evaluation measures such as Mean Relative Error (MRE), Mean Absolute Error (MAE), and Mean Square Error (MSE). The following are their equations, respectively:

$$MRE = \frac{1}{n}\frac{\sum_{t=1}^{n}\left|f(t) - \widehat{f(t)}\right|}{f(t)} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|f(t) - \widehat{f(t)}\right| \tag{13}$$

and

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(f(t) - \widehat{f(t)}\right)^2 \tag{14}$$

All proposed and selected models are evaluated using these criteria. Moreover, in GMDH-NN and RGMDH-NN models, neurons are selected according to MSE.

4. Results

D-stage results: the results of two noise removal filters, i.e., WA and EMD, are described below.

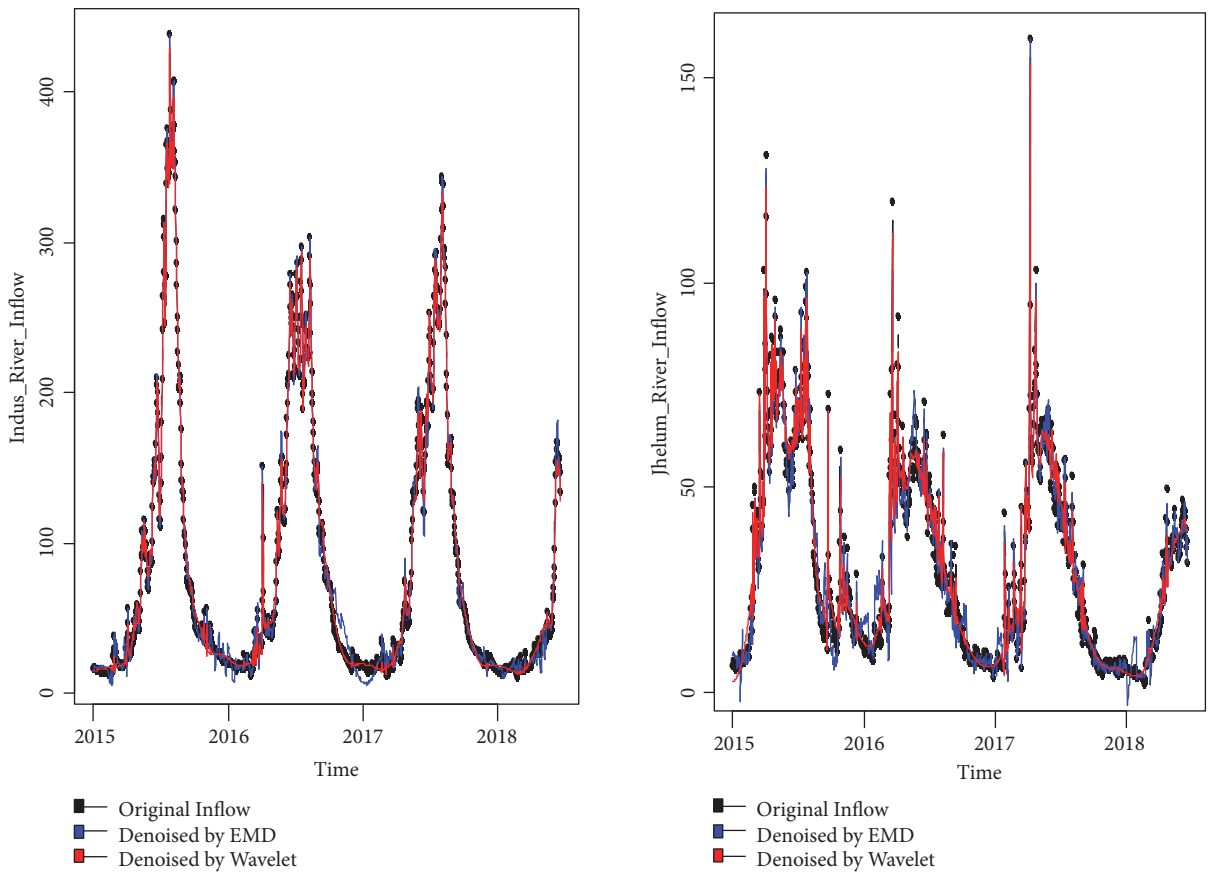

FIGURE 5: The denoised series of the two hydrological time series of Indus and Jhelum rivers inflow. The figure shows the denoised results obtained through the EMD-based threshold method (in red color) and the wavelet analysis-based threshold method (in blue color).

Wavelet based denoised: after calculating the approximations from (1) and details from (2), the hard and soft rule of thresholding are used to remove noises from hydrological time series coefficients. Both hard and soft rules are calculated from (3) and (4) respectively. On behalf of lower MSE, hard threshold based denoised series are reconstructed through (5) for WA.

EMD-based threshold: to remove noises through EMD, intrinsic mode functions are calculated from (7), and then hard and soft thresholds are used to denoise the calculated IMFs except the last two IMFs as, due to smooth and low frequency characteristics, there is no need to denoise the last two IMFs. Hard threshold based denoised IMFs are further used to reconstruct the noise free hydrological time series data from (8).

The WA and EMD based denoised Indus and Jhelum rivers inflow are shown in Figure 5. The statistical measures including mean ($\overline{x}$), standard deviation (σ), MRE, MAE, and MSE of original and denoised series for all case studies of both noise removal methods are presented in Table 2. The results show that the statistical measures are almost the same for both denoising methods except MSE, as for Indus and Jhelum inflow, WA-based denoised series have lower MSE than EMD; however, for Kabul and Chenab inflow, EMD-based denoised series have lower MSE than WA-based denoised series. Overall, it was concluded that both methods have equal performance in denoising the hydrological time series data. In decomposing stage, both of WA and EMD based denoised series are separately used as input to derive the time varying characteristics in terms of high and low frequencies.

Decompose-stage results: to extract the local time varying features from denoised hydrological data, the WA/EMD-based denoised hydrological time series data are further decomposed into nine IMFs and one residual. The CEEMDAN decomposition method is used to extract the IMFs from all four rivers. EMD-based denoised-CEEMDAN-based decomposed results of Indus and Jhelum rivers inflow are shown in Figure 6 whenever WA-CEEMDAN-based noise free decomposed results of Indus and Jhelum rivers inflow are shown in Figure 7. All four rivers are decomposed into nine IMFs and one residual showing similar characteristics for both methods. The extracted IMFs show the characteristics of hydrological time series data where the starting IMFs represent the higher frequency whereas last half IMFs show the low frequencies and residual are shown as trends as shown in Figures 6 and 7. The amplitude of white noise is set as 0.2 as in [2] and numbers of ensemble members are selected as maximum which is 1000.

Table 2: Statistical measures of WA- and EMD-based denoised rivers inflow of four hydrological time series data sets.

River Inflow	Mode	$\overline{x}$	σ	MRE	MAE	MSE
Indus Inflow	Original series	80.2194	87.5044			
	EMD	80.5931	87.3925	3.9118	0.1275	36.7636
	Wavelet	80.2267	86.1632	3.8188	0.0987	22.9626
Jhelum Inflow	Original series	30.2001	23.6743			
	EMD	30.1412	23.1641	2.7118	0.1666	16.4864
	Wavelet	30.2023	22.7799	2.5579	0.1418	10.8837
Kabul Inflow	Original series	29.1746	25.2352			
	EMD	25.23524	25.1181	2.5474	0.2036	12.5216
	Wavelet	29.18118	24.29148	2.7386	0.1615	12.2447
Chenab Inflow	Original series	31.9557	29.4916			
	EMD	32.0024	29.2734	2.271784	0.1470	10.6797
	Wavelet	31.9585	28.2591	3.1958	0.17228	17.8353

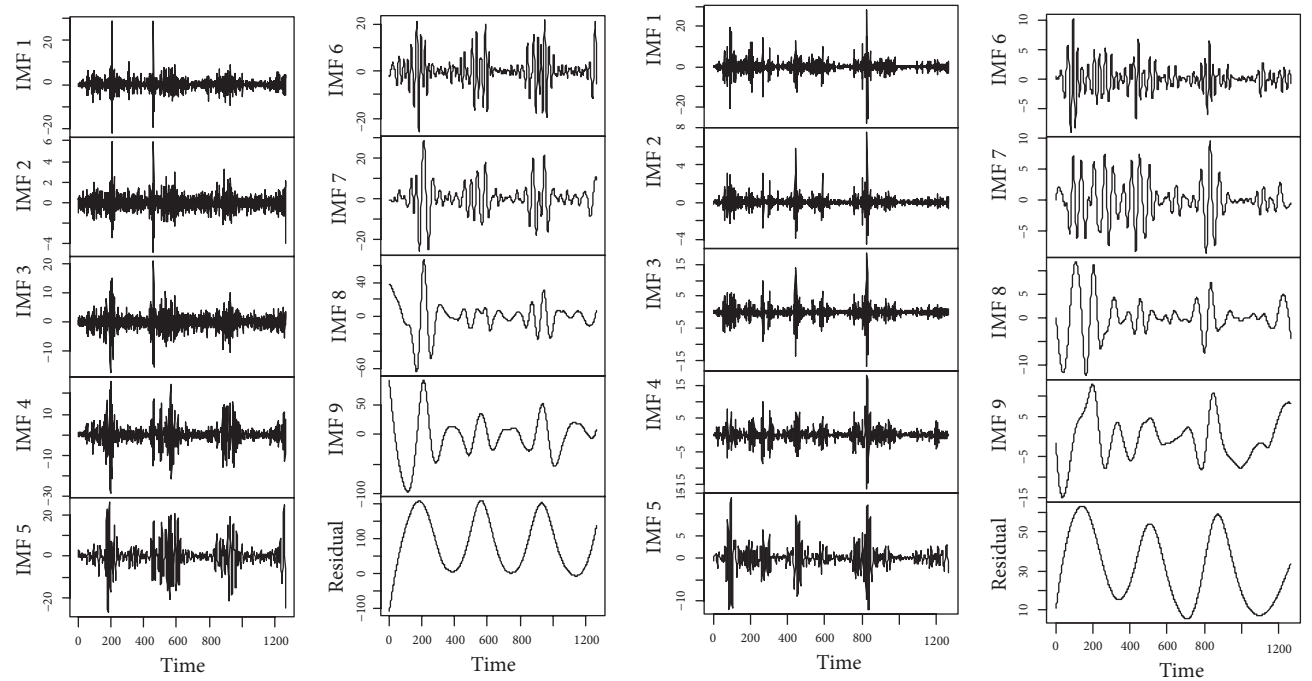

Figure 6: The EMD-CEEMDAN decomposition of Indus (left) and Jhelum rivers inflow (right). The two series are decomposed into nine IMFs and one residue.

P-step results: for all extracted IMFs and residual, three methods of predictions are adopted to get the more precise and near-to-reality results. For that reason, one traditional statistical method, i.e., ARIMA (p, d, q), with two other nonlinear ML methods, i.e., GMDH-NN and RBFNN, are used to predict the IMFs and residuals of all four river inflows. The rivers inflow data of all four rivers are split: 70% for training set and 30% for testing set. The parameters and structure of models are estimated using 886 observations of rivers inflow. The validity of proposed and selected models is tested using 30% data of rivers inflow. After successful estimation of multimodels on each IMF and residual, the best method with minimum MRE, MAE, and MSE is selected for each IMF prediction. The testing results of proposed models with comparison to all other models for all four rivers' inflow, i.e., Indus inflow, Jhelum inflow, Chenab inflow, and Kabul inflow, are presented in Table 3. The proposed EMD-CEEMDAN-MM and WA-CEEMDAN-MM model prediction results fully demonstrate the effectiveness for all 4 cases with minimum MRE, MAE, and MSE compared to all 1-S [8], 2-S [15, 17], and 3-S [2] evaluation models. However, overall, the proposed WA-CEEMDAN-MM model attains the lowest MSE as compared to other EMD-CEEMDAN-MM proposed models. The worst predicted model is 1-S, i.e., ARIMA, without denoising and without decomposing the hydrological time series data with highest MSE. The predicted graphs of proposed model, i.e., EMD-CEEMDAN-MM, with comparison to 2-S models, i.e., with EMD based denoised for

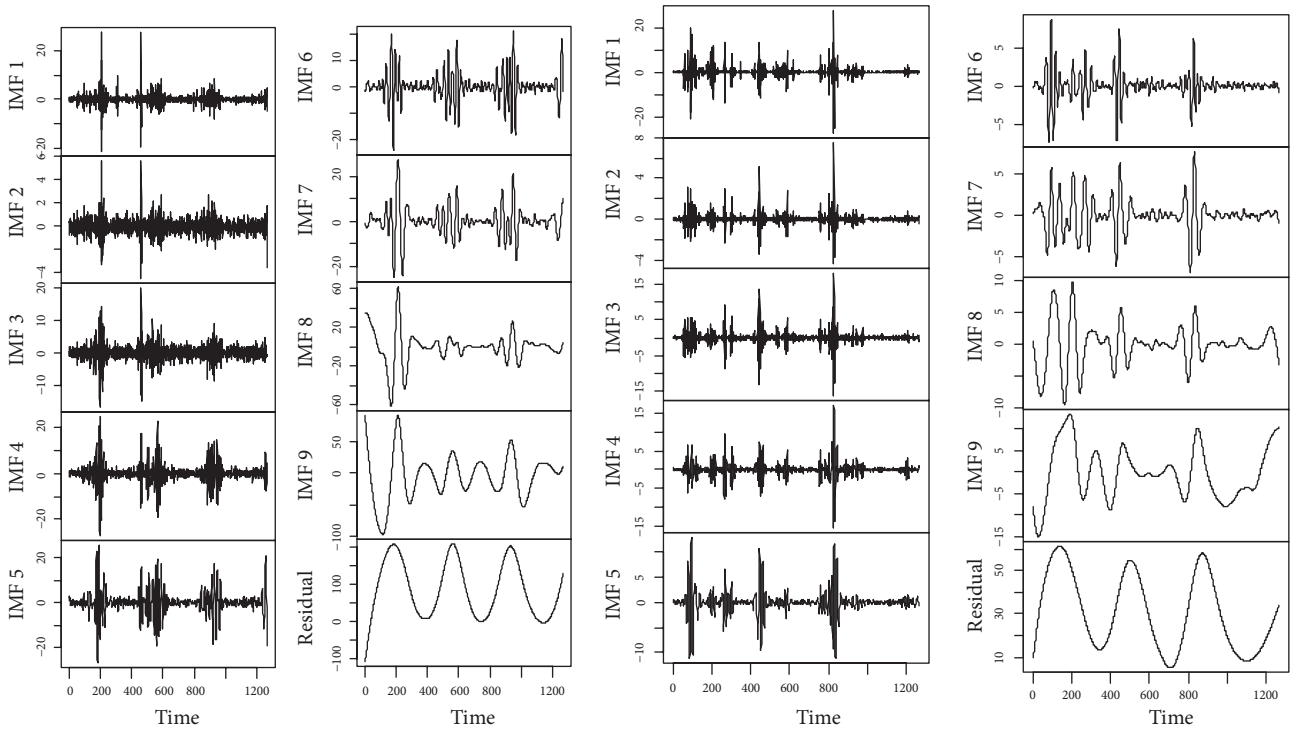

Figure 7: The WA-CEEMDAN decomposition of Indus (left) and Jhelum rivers inflow (right). The two series are decomposed into nine IMFs and one residue.

Indus and Jhelum river inflow, are shown in Figure 8 and WA-CEEMDAN-MM with comparison to 2-S models, i.e., with WA based denoised, are shown in Figure 9.

To improve the prediction accuracy of complex hydrological time series data from simple time series models one can take the advantage from three principals of "denoising," "decomposition," and "ensembling the predicted results." The 2-S model, with simple ARIMA and GMDH, can perform well as compared to 2-S models with complex models and 1-S models by optimal decomposition methods. Moreover, with addition to extracting time varying frequencies from denoised series, one can get the more precise results over 2-S models. However, from Table 3, it can be concluded that the proposed WA-CEEMDAN-MM and EMD-CEEMDAN-MM models perform more efficiently to predict the hydrological time series data by decreasing the complexity of hydrological time series data and enhancing the prediction performance over 1-S, 2-S, and 3-S existing models.

The following conclusions are drawn based on the testing error presented in Table 3.

Overall comparison: the overall performances of proposed models WA-CEEMDAN-MM and WA-CEEMDAN-MM are better than all other evaluation models selected from the study [2, 8, 15, 17] with the lowest MAE, MRE, and MSE values for all case studies. However, among two proposed models, WA-CEEMDAN-MM performs well by attaining on average 8.49%, 24.19%, and 5.43% lowest MAE, MRE, and MSE values, respectively, for all four rivers' inflow prediction as compared to EMD-CEEMDAN-MM as listed in Table 3. It is shown that both proposed models perform well with comparison to 1-S, 2-S, and existing 3-S. Moreover, it is also noticed that most of IMFs are precisely predicted with simple traditional statistical ARIMA model except the first two IMFs as the first IMFs presented high frequencies showing more volatile time varying characteristics with the rest of IMFs. However, overall WA-CEEMDAN-MM is more accurate in predicting the rivers inflow.

Comparison of proposed models with other only denoised series models: removing the noise through WA- and EMD-based threshold filters before statistical analysis improved the prediction accuracy of complex hydrological time series. It can be observed from Table 3 that the MAE, MRE, and MSE values of four cases perform well for 2-S model as compared to 1-S model in both WA and EMD based denoised inputs. However, like overall performance of WA-CEEMDAN-MM, WA-based denoised models perform well compared to EMD-based denoised. Moreover, with denoised series, the several statistical (simple) and machine learning (complex) methods are adopted to further explore the performances between simple and complex methods to predict inflows. This can be seen from Table 3, where WA-RBFN and EMD-RBFN perform the worst compared to WA-ARIMA, WA-RGMDH, EMD-ARIMA, and EMD-RGMDH. This means that with denoising the hydrological series one can move towards simple models as compared to complex models like radial basis function neural network. WA-RGMDH and EMD-RGMDH attain the highest accuracy among all 2-S models.

Comparison of proposed models with other denoised and decomposed models: in addition to denoising, the decomposition of hydrological time series strategy effectively enhances

Table 3: Evaluation index of testing prediction error of proposed models (EMD-CEEMDAN-MM and WA-CEEMDAN-MM) with all selected models for all four case studies.

River Inflow	Model Name	Models	MRE	MAE	MSE
Indus Inflow	1-S	ARIMA	4.2347	0.0685	64.7141
	2-S	WA-ARMA	3.2862	0.0430	53.4782
		WA-RGMDH	3.2548	0.0393	46.7382
		WA-RBFN	20.1949	0.2598	2301.772
		EMD-ARMA	4.9898	0.0960	76.1440
		EMD-RGMDH	4.9653	0.0915	76.0884
		EMD-RBFN	34.3741	0.7762	3931.601
	3-S	EMD-EEMD-MM	5.2710	0.1721	44.0115
		WA-CEEMDAN-MM	**1.5410**	**0.0349**	**5.5734**
		EMD-CEEMDAN-MM	**1.8009**	**0.0462**	**6.6983**
Jhelum Inflow	1-S	ARMA	3.5224	0.1201	47.5529
	2-S	WA-ARMA	2.6129	0.0748	37.1441
		WA-RGMDH	2.6208	0.0773	37.7954
		WA-RBFN	9.8608	0.7714	180.7443
		EMD-ARMA	3.7354	0.1551	48.3164
		EMD-RGMDH	3.7357	0.1620	48.3606
		EMD-RBFN	2.8822	0.2506	51.9916
	3-S	EMD-EEMD-MM	2.0096	0.1269	7.3565
		WA-CEEMDAN-MM	**1.1805**	**0.0457**	**6.8225**
		EMD-CEEMDAN-MM	**1.4480**	**0.0642**	**7.7709**
Kabul Inflow	1-S	ARMA	2.4910	0.0883	25.0136
	2-S	WA-ARMA	1.9999	0.0592	20.6874
		WA-RGMDH	2.0794	0.0729	21.0612
		WA-RBFN	1.6565	0.0997	13.3554
		EMD-ARMA	2.9538	0.1484	28.5767
		EMD-RGMDH	3.0114	0.2280	28.9351
		EMD-RBFN	4.9355	0.7613	69.9346
	3-S	EMD-EEMD-MM	1.8758	0.3166	5.8020
		WA-CEEMDAN-MM	**0.7664**	**0.0363**	**2.1072**
		EMD-CEEMDAN-MM	**0.9599**	**0.0861**	**2.7636**
Chenab Inflow	1-S	ARMA	5.4157	0.4646	108.185
	2-S	WA-ARMA	3.9652	0.1087	84.2359
		WA-RGMDH	3.6147	0.0943	81.6493
		WA-RBFN	4.1424	0.2757	47.6184
		EMD-ARMA	4.7971	0.2721	100.7013
		EMD-RGMDHA	4.4812	0.1865	95.6680
		EMD-RBFN	10.8228	2.1666	284.5627
	3-S	EMD-EEMD-MM	2.7172	0.2298	14.5191
		WA-CEEMDAN-MM	**1.6940**	**0.0705**	**13.5702**
		EMD-CEEMDAN-MM	**1.9345**	**0.1105**	**14.067**

the prediction accuracy by reducing the complexity of hydrological data in multiple dimensions. It is shown from Table 3 that the 3-S (existing) performs better as on average for all four rivers MAE, MRE, and MSE values are 13.76%, -6.55%, and 54.79%, respectively, lower than 1-S model and 63.40, 64.76%, and 96.78% lower than 2-S model (EMD-RBFNN). Further research work can be done to explore the ways to reduce the mathematical complexity of separate denoising and decomposition like only single filter which not only denoises but also decomposes the hydrological time series data with the same filter to effectively predict or simulate the data.

5. Conclusion

The accurate prediction of hydrological time series data is essential for water supply and water resources purposes. Considering the instability and complexity of hydrological

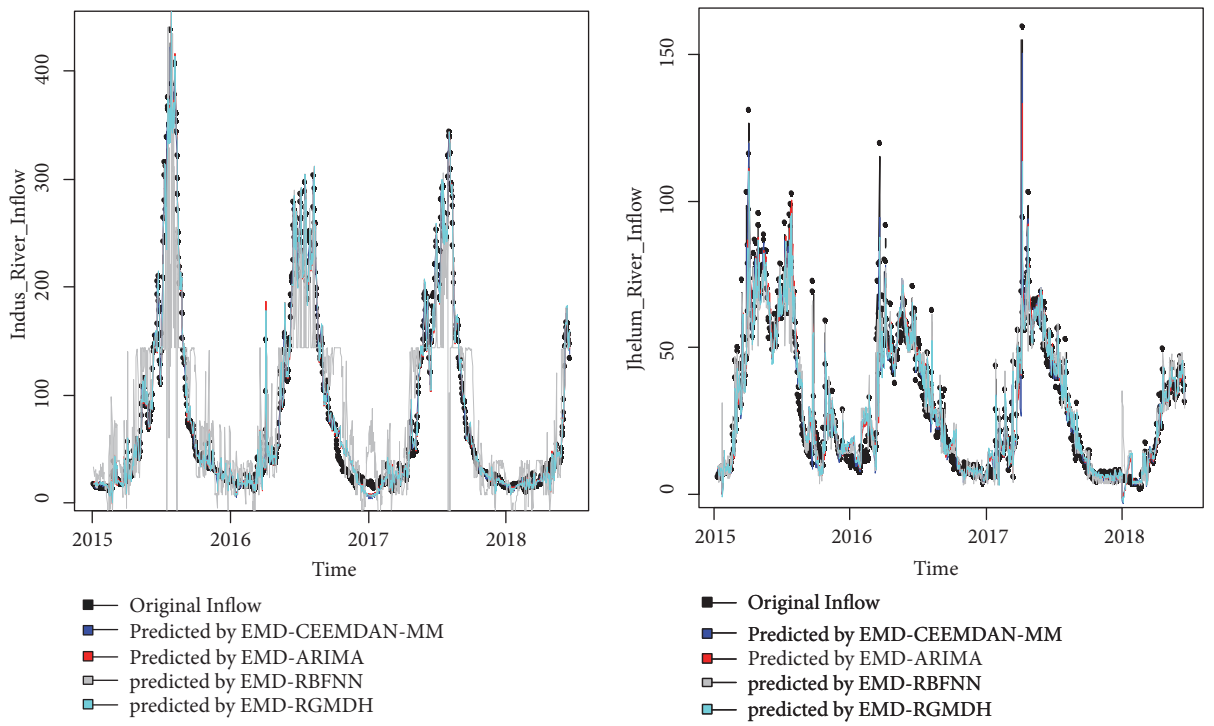

Figure 8: Prediction results of Indus and Jhelum rivers inflow using proposed EMD-CEEMDAN-MM with comparison to other EMD based denoised and predicted models.

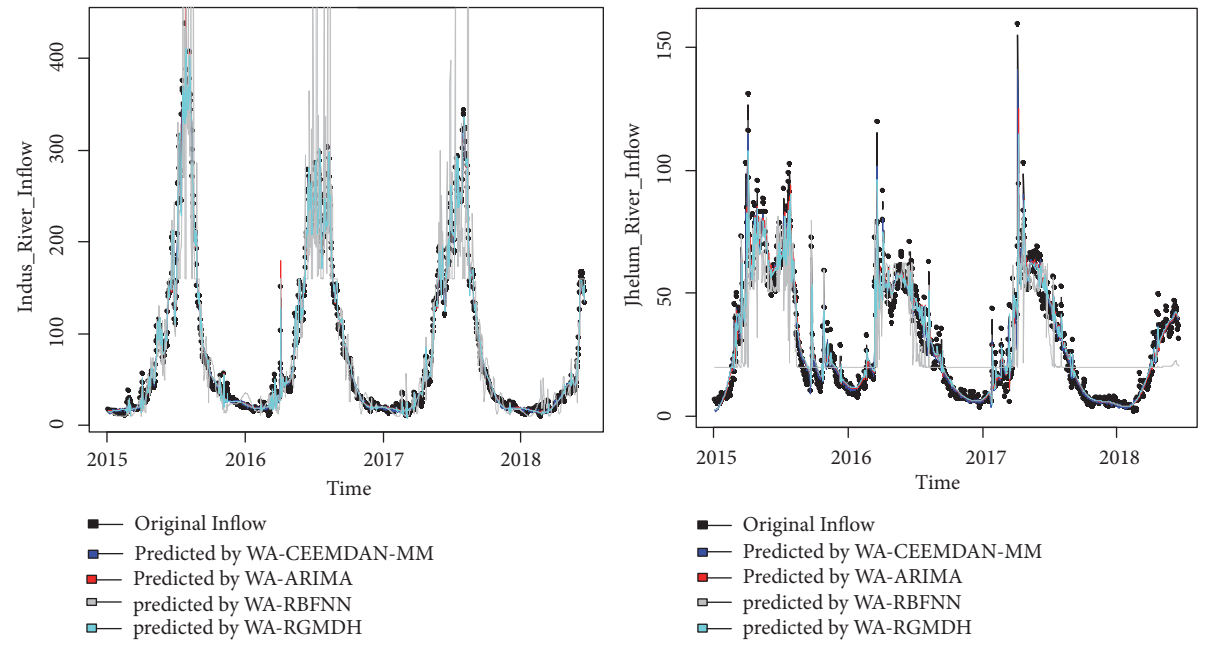

Figure 9: Prediction results of Indus and Jhelum river inflow using proposed WA-CEEMDAN-MM with comparison to WA based denoised predicted models.

time series, some data preprocessing methods are adopted with the aim to enhance the prediction of such stochastic data by decomposing the complexity of hydrological time series data in an effective way. This research proposed two new methods with three stages as "denoised," decomposition, and prediction and summation, named as WA-CEEMDAN-MM and EMD-CEEMDAN-MM, for efficiently predicting the hydrological time series. For the verification of proposed methods, four cases of rivers inflow data from Indus Basin System are utilized. The overall results show that the proposed

hybrid prediction model improves the prediction performance significantly and outperforms some other popular prediction methods. Our two proposed, three-stage hybrid models show improvement in prediction accuracy with minimum MRE, MAE, and MSE for all four rivers as compared to other existing one-stage [8] and two-stage [15, 17] and three-stage [2] models. In summary, the accuracy of prediction is improved by reducing the complexity of hydrological time series data by incorporating the denoising and decomposition. In addition, these new prediction models are also capable of solving other nonlinear prediction problems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no. RG-1437-027.

References

[1] D. P. Solomatine and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of Hydroinformatics*, vol. 10, no. 1, pp. 3–22, 2008.

[2] C. Di, X. Yang, and X. Wang, "A four-stage hybrid model for hydrological time series forecasting," *PLoS ONE*, vol. 9, no. 8, Article ID e104663, 2014.

[3] Z. Islam, *Literature Review on Physically Based Hydrological Modeling [Ph. D. thesis]*, pp. 1–45, 2011.

[4] A. R. Ghumman, Y. M. Ghazaw, A. R. Sohail, and K. Watanabe, "Runoff forecasting by artificial neural network and conventional model," *Alexandria Engineering Journal*, vol. 50, no. 4, pp. 345–350, 2011.

[5] S. Riad, J. Mania, L. Bouchaou, and Y. Najjar, "Rainfall-runoff model usingan artificial neural network approach," *Mathematical and Computer Modelling*, vol. 40, no. 7-8, pp. 839–846, 2004.

[6] T. Peng, J. Zhou, C. Zhang, and W. Fu, "Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks," *Water*, vol. 9, no. 6, p. 406, 2017.

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, Calif, USA, 1970.

[8] B. N. S. Ghimire, "Application of ARIMA Model for River Discharges Analysis," *Journal of Nepal Physical Society*, vol. 4, no. 1, pp. 27–32, 2017.

[9] Ö. Kişi, "Streamflow forecasting using different artificial neural network algorithms," *Journal of Hydrologic Engineering*, vol. 12, no. 5, pp. 532–539, 2007.

[10] C. A. G. Santos and G. B. L. D. Silva, "Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models," *Hydrological Sciences Journal*, vol. 59, no. 2, pp. 312–324, 2014.

[11] C. Wu, K. Chau, and Y. Li, "Methods to improve neural network performance in daily flows prediction," *Journal of Hydrology*, vol. 372, no. 1-4, pp. 80–93, 2009.

[12] T. Partal, "Wavelet regression and wavelet neural network models for forecasting monthly streamflow," *Journal of Water and Climate Change*, vol. 8, no. 1, pp. 48–61, 2017.

[13] Z. M. Yaseen, M. Fu, C. Wang, W. H. Mohtar, R. C. Deo, and A. El-shafie, "Application of the Hybrid Artificial Neural Network Coupled with Rolling Mechanism and Grey Model Algorithms for Streamflow Forecasting Over Multiple Time Horizons," *Water Resources Management*, vol. 32, no. 5, pp. 1883–1899, 2018.

[14] M. Rezaie-Balf and O. Kisi, "New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine," *Hydrology Research*, vol. 49, no. 3, pp. 939–953, 2018.

[15] H. Liu, C. Chen, H.-Q. Tian, and Y.-F. Li, "A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks," *Journal of Renewable Energy*, vol. 48, pp. 545–556, 2012.

[16] Z. Qu, K. Zhang, J. Wang, W. Zhang, and W. Leng, "A Hybrid model based on ensemble empirical mode decomposition and fruit fly optimization algorithm for wind speed forecasting," *Advances in Meteorology*, 2016.

[17] Y. Sang, "A Practical Guide to Discrete Wavelet Decomposition of Hydrologic Time Series," *Water Resources Management*, vol. 26, no. 11, pp. 3345–3365, 2012.

[18] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, Article ID 1971, pp. 903–995, 1998.

[19] Z. H. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[20] Z. Wang, J. Qiu, and F. Li, "Hybrid Models Combining EMD/EEMD and ARIMA for Long-Term Streamflow Forecasting," *Water*, vol. 10, no. 7, p. 853, 2018.

[21] N. A. Agana and A. Homaifar, "EMD-based predictive deep belief network for time series prediction: an application to drought forecasting," *Hydrology*, vol. 5, no. 1, p. 18, 2018.

[22] A. Kang, Q. Tan, X. Yuan, X. Lei, and Y. Yuan, "Short-term wind speed prediction using EEMD-LSSVM model," *Advances in Meteorology*, 2017.

[23] Z. H. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis (AADA)*, vol. 1, no. 1, pp. 1–41, 2009.

[24] W.-C. Wang, K.-W. Chau, D.-M. Xu, and X.-Y. Chen, "Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition," *Water Resources Management*, vol. 29, no. 8, pp. 2655–2675, 2015.

[25] H. Su, H. Li, Z. Chen, and Z. Wen, "An approach using ensemble empirical mode decomposition to remove noise from prototypical observations on dam safety," *SpringerPlus*, vol. 5, no. 1, 2016.

[26] X.-S. Jiang, L. Zhang, and M. X. Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support

vector machine with real-world applications in China," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 110–127, 2014.

[27] S. Dai, D. Niu, and Y. Li, "Daily Peak Load Forecasting Based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Support Vector Machine Optimized by Modified Grey Wolf Optimization Algorithm," *Energies*, vol. 11, no. 1, p. 163, 2018.

[28] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4144–4147, Prague, Czech Republic, May 2011.

[29] A. Jayawardena and A. Gurung, "Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach," *Journal of Hydrology*, vol. 228, no. 3-4, pp. 242–264, 2000.

[30] M. Yang, Y. Sang, C. Liu, and Z. Wang, "Discussion on the Choice of Decomposition Level for Wavelet Based Hydrological Time Series Modeling," *Water*, vol. 8, no. 5, p. 197, 2016.

[31] J. Kim, C. Chun, and B. H. Cho, "Comparative analysis of the DWT-based denoising technique selection in noise-riding DCV of the Li-Ion battery pack," in *Proceedings of the 2015 9th International Conference on Power Electronics and ECCE Asia (ICPE 2015-ECCE Asia)*, pp. 2893–2897, Seoul, South Korea, June 2015.

[32] H. Ahmadi, M. Mottaghitalab, and N. Nariman-Zadeh, "Group Method of Data Handling-Type Neural Network Prediction of Broiler Performance Based on Dietary Metabolizable Energy, Methionine, and Lysine," *The Journal of Applied Poultry Research*, vol. 16, no. 4, pp. 494–501, 2007.

[33] A. S. Shakir and S. Ehsan, "Climate Change Impact on River Flows in Chitral Watershed," *Pakistan Journal of Engineering and Applied Sciences*, 2016.

[34] B. Khan, M. J. Iqbal, and M. A. Yosufzai, "Flood risk assessment of River Indus of Pakistan," *Arabian Journal of Geosciences*, vol. 4, no. 1-2, pp. 115–122, 2011.

[35] K. Gaurav, R. Sinha, and P. K. Panda, "The Indus flood of 2010 in Pakistan: a perspective analysis using remote sensing data," *Natural Hazards*, vol. 59, no. 3, pp. 1815–1826, 2011.

[36] A. Sarwar and A. S. Qureshi, "Water management in the indus basin in Pakistan: challenges and opportunities," *Mountain Research and Development*, vol. 31, no. 3, pp. 252–260, 2011.

[37] G. Pappas, "Pakistan and water: new pressures on global security and human health," *American Journal of Public Health*, vol. 101, no. 5, pp. 786–788, 2011.

[38] J. L. Wescoat, A. Siddiqi, and A. Muhammad, "Socio-Hydrology of Channel Flows in Complex River Basins: Rivers, Canals, and Distributaries in Punjab, Pakistan," *Water Resources Research*, vol. 54, no. 1, pp. 464–479, 2018.