A comparative study of wavelet and empirical mode decomposition-based GPR models for river discharge relationship modeling at consecutive hydrometric stations
Kiyoumars Roushangar, Masoumeh Chamani, Roghayeh Ghasempour, Hazi Mohammad Azamathulla and Farhad Alizadeh

ABSTRACT
The river stage–discharge relationship has an important impact on modeling, planning, and management of river basins and water resources. In this study, the capability of the Gaussian Process Regressions (GPR) kernel-based approach was assessed in predicting the daily river stage–discharge (RSD) relationship. Three successive hydrometric stations of the Housatonic River were considered, and based on the flow characteristics during the period of 2002–2006 several models were developed and tested via GPR. To enhance the applied model efficiency, two pre-processing techniques, namely Wavelet Transform (WT) and Ensemble Empirical Mode Decomposition (EEMD), were used. Also, two states of the RSD modeling were investigated. In state 1, each station’s own data was used and in state 2, the upstream stations’ datasets were used as input to model the RSD downstream of the river. The single and integrated model results showed that the integrated WT- and EEMD-GPR models resulted in more accurate outcomes. Data processing enhanced the models’ capability between 25% and 40%. The results showed that the RSD modeling in state 1 led to better results; however, when the station’s own data were not available the integrated methods could be applied successfully for the RSD modeling using the previous stations’ data.

Key words | EEMD, GPR, pre-processing, river stage-discharge, WT

HIGHLIGHTS
● The advantages of two WT and EEMD pre-processing techniques and kernel-based model were merged for RSD modeling.
● Two states of modeling based on station’s own data or previous stations’ data were investigated.
● Integrated hybrid techniques outperformed the single GPR method.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).
doi: 10.2166/ws.2021.073
INTRODUCTION

Accurate prediction of river stage–discharge is of great importance for the utilization and management of sustainable water resources. Reliable river discharge prediction is particularly important for hydrological operations and hydro-environmental management (Hasanpour Kashani et al. 2018). According to Petersen-Øverleir (2006), both the stage and the discharge of a river vary depending on the magnitude of rainfall intensity that the river basin receives. To obtain a continuous record of discharge, the stage is recorded and the discharge is computed from a correlation of stage and measured discharge. This correlation, or training (calibration), is known as the stage–discharge relationship (Baiamonte & Ferro 2013). Accurate information about the flow rates of rivers is important for a variety of hydrologic applications such as water and sediment bed material load estimation, water resource planning, operation and development, and hydraulic and hydrologic modeling (Wahlin & Clemmens 2006). However, collecting data for discharge on a continuous basis is costly, especially during large flood events. An alternative approach is to convert records of water stages into discharges using a stage–discharge relationship. So far, various and complex relationships have been proposed to predict river stage–discharge. However, due to the complexity and nonlinearity of the process and validated models, it is difficult to simulate the accurate amount of discharge carried by rivers.

It is said that the stage–discharge relationship is not a simple, unique relation (Habib & Meselhe 2006). Therefore, it is necessary to use other methods with higher efficiency. A number of researchers have developed several prediction models using traditional statistical regression methods and artificial intelligence techniques with this aim. The existing regression methods often do not show the desired accuracy due to the uncertainties of the stage–discharge relationship process. On the other hand, numerical models require a lot of data, which is time-consuming and costly. As the number of variables and their interactions increase, the accuracy of operation predicting will decrease. However, meta-model approaches can easily switch from linear to nonlinear separation. These techniques have been successfully used in a number of diverse fields including water resources (e.g., Anmala et al. 2000; Coulibaly et al. 2000; Lameski et al. 2015; Zeinali et al. 2020) whereas there are many meta-model techniques that are driven by knowledge discovery
Wavelet analysis is one of the most commonly used approaches that is used by researchers for this aim (Adamowski et al. 2009). In the past few decades several artificial intelligence methods (e.g., Artificial Neural Networks (ANNs), Neuro-Fuzzy models (NF), Gaussian Process Regression (GPR), Gene Expression Programming (GEP), Support Vector Machine (SVM), and Kernel Extreme Learning Machine (KELM)) have been developed and applied for assessing complex hydraulic and hydrologic phenomena efficiently. Real-time flood forecasting (Ren et al. 2016), longitudinal dispersion coefficient computing in natural streams (Azamathulla & Wu 2011), daily suspended sediment concentration modeling (Kaveh et al. 2017), predicting bedload in sewer pipes (Roushangar & Ghasempour 2015), side weir discharge coefficient modeling (Azamathulla et al. 2016), monthly streamflow modeling (Zhou et al. 2018), relative energy dissipation prediction in rough and smooth bed channels (Saghebian 2019), evaluation of energy losses of rectangular and circular culverts (Roushangar et al. 2019), and predictions of precipitation and temperature (Ahmed et al. 2020) are some examples.

Among others, Gaussian Process Regression (GPR), an effective kernel-based machine learning algorithm, is capable of being applied to prediction of hydrological process (Sun et al. 2014). GPR is a good nonparametric regression tool since it has theoretical simplicity and good distribution capability, as well as providing probabilistic results (Rasmussen & Nickisch 2010). A well-advised application of this tool is to select an appropriate covariance function or kernel and tune related hyper-parameters. On the other hand, hybrid models based on signal decomposition can be effective in increasing the time series prediction method efficiency (Pachori et al. 2013). Since most hydrological time series such as river stage–discharge are non-stationary, trendy, or have seasonal fluctuations the use of other methods for identifying the dominant periodicities of time series are necessary. Wavelet analysis is one of the most commonly used approaches that is used by researchers for this aim (Adamowski et al. 2009). Additionally, the Empirical Mode Decomposition (EMD) method, which is suitable for non-linear and non-stationary time series (Huang et al. 1998), has been used recently. Unlike wavelet decomposition, empirical mode decomposition extracts the data oscillatory-mode components without a priori determining the basis functions or level of decomposition (Labate et al. 2013). A time series in the Discrete Wavelet Transform (DWT) breaks down into a series of linearly independent detail signals and one approximation signal by using a specific wavelet function. DWT can detect sudden signal changes well, because it transforms original time series data into two types of wavelet coefficients: approximation and detail. EMD has been proven to be a quite effective method for extracting signals from data generated in noisy nonlinear and non-stationary processes. Research has proved that proper data pre-processing by using WT and EMD can lead models to adequately illustrate the real specifications of the basic system.

By considering the aforementioned issues, river discharge is very important in the hydrological cycle and hydro-environmental management and plays a basic role in designing and operating irrigation schemes on a watershed scale. One of the most important problems in analysis and management of river discharge is the modeling/simulation. Drawing upon the knowledge of the previous data-mining models, this paper aimed to enhance the quality of estimation of river stage–discharge using a kernel-based model (GER) and two data-processing methods under two scenarios. In the first scenario, the intended parameter of each station was estimated using the station’s own data, and in the second scenario, the previous stations’ data were used as inputs to investigate the issue that the existing sub-basins between the consecutive stations may have noticeable impacts on the flow regime of the downstream station. Discrete Wavelet Transform (DWT) and Ensemble Empirical Mode Decomposition (EEMD) were used for this aim. Kernel-based approaches such as GPR are relatively new, important methods based on different kernel types. Such models are based on statistical learning theory and are capable of adapting themselves to predict any variable of interest via sufficient inputs. The training of the GPR method is fast, has high accuracy, and the probability of the occurrence of data overtraining in this method is less. Also, WT and EEMD are pre-processing methods that give remarkable vision to the physical form of the data by representing information in both time and frequency domains. Therefore, in this study, the input data were decomposed into sub-series by WT and EEMD. Then, these subseries were used as inputs in the GPR...
method. The daily river stage–discharge data of the Housatonic River in the period of 2002–2006 were used to test the developed models.

In general, due to the complex nature of the river stage–discharge process, the present study tried to develop a unique strategy to predict it precisely. The proposed conceptual strategy has some advantages to cover the shortcomings. First, based on quadratic optimization of the convex function in GPR, it can easily switch from linear to nonlinear separation by nonlinear mapping using so-called kernel functions. Therefore, it is capable of being applied to prediction of the river discharge process. On the other hand, EEMD and WT were employed as a hybrid pre-processing approach conjugated to GPR. WT along with EEMD decomposes a non-stationary signal into a given quantity of stationary sub-signals. Then, AI methods can be combined with WT and EEMD to improve the preciseness of the prediction. Also, by applying sensitivity analysis, the most effective parameters in the modeling process can be selected. It is expected that the integrated hybrid techniques outperform the single GPR method and improve the modeling efficiency.

### MATERIALS AND METHODS

#### Study area and used data

The Housatonic River is a river, approximately 149 miles (240 km) long, in western Massachusetts and western Connecticut in the United States. It flows south to southeast, and drains about 1,950 square miles (5,100 km²) of southwestern Connecticut into Long Island Sound. Its watershed is just to the west of the watershed of the lower Connecticut River. In this study, the datasets of three successive stations were used in which the distance between stations was approximately 50 km. Daily discharge and stage data of the river are available for the three selected hydrometric stations on the Housatonic River at the USGS website (https://waterdata.usgs.gov/nwis/sw). Table 1 demonstrates the characteristics of the Housatonic River and the used data. Also, Figure 1 shows the location of the selected stations. The dataset used herein are in daily scale and cover the period from 2002 to 2006. Time series were separated into calibration (70% from 01/01/2002 to 07/01/2005) and verification (30% from 07/02/2005 to 12/31/2006) parts.

#### Pre-processing approaches

One of the most popular approaches in time series processing is the wavelet transform (WT) (Farajzadeh & Alizadeh 2017). The WT uses the flexible window function (mother wavelet) in signal processing. The flexible window function can be changed over time according to the signal shape and compactness (Mehr et al. 2015). After using WT, the signal will decompose into two approximation (large-scale or low-frequency) and detailed (small-scale) components. An illustration of a three-level WT is shown in Figure 2. In the first level, the original signal (x) is decomposed into the two components of approximation (cA1) and detailed (cD1). In the second level, cA1 is again decomposed to approximation (cA2) and detailed (cD2) components. Finally, in the third level, cA2 is decomposed to cA3 approximation and cD3 detailed components. The sum of all detailed subseries and approximation series obtained from the third level will be the original signal (i.e. \( x = cD1 + cD2 + cD3 + cA3 \)). The other approach for time series processing is empirical mode decomposition (EMD). The EMD method is an effective self-adaptive dyadic filter bank which is applied to white noise (a random signal which has equal intensity at different frequencies). By applying this method, each signal can be decomposed into a number of inherent mode functions (IMFs) which can be used to process nonlinear and

| Hydrometric station | Station no. | Area (km²) | Max. discharge (m³/s) | Min. discharge (m³/s) | Max. stage (m) | Min. stage (m) |
|---------------------|-------------|------------|-----------------------|-----------------------|----------------|----------------|
| Station 1           | 01197500    | 730.38     | 193.40                | 1.78                  | 0.33           | 0.06           |
| Station 2           | 01200500    | 1,642      | 492.71                | 1.78                  | 0.31           | 0.03           |
| Station 3           | 01205300    | 2,579      | 928.79                | 2.60                  | 0.46           | 0.02           |
non-stationary signals. One of the advantages of this method is the ability to determine the instantaneous frequency of the signal. At each step of the signal decomposition into its frequency components, the high-frequency components are separated first and this process must continue until the component with the lowest frequency remains. EEMD is developed based on EMD. In fact, using the noise-added data analysis method the EEMD solves the EMD mode mixing issue (Wu & Huang 2009). The EEMD algorithm can be described as: (1) for a given signal $x(t)$, random white noise is added to the signal, (2) the noise-added signal is decomposed using EMD for obtaining IMF series, (3) steps 1 and 2 are repeated until the number of additions of white noise is greater than or equal to the number of trials, (4) for obtaining the ensemble IMF, the average of the sum of all IMFs is computed ($\text{I}_j(t)$), and (5) the original signal is formed as $x(t) = \sum_{j=1}^{n} \text{I}_j(t)$. For selecting the most effective IMFs and using them as inputs in the modeling

Figure 1 | The geographic location of the Housatonic River and location of the hydrometric stations with their series plots.
process, their energy values can be calculated and the IMFs with higher energy can be used as inputs (see Lei et al. (2013) for more details).

 Kernel-based approach

Kernel-based (KB) approaches are new methods which are used for classification and regression purposes. One of the important kernel-based approaches is Gaussian Process Regression (GPR), which works based on the different kernel types such as polynomial, normalized polynomial, radial basis function (RBF) and Pearson functions. Such a model is capable of adapting itself to predict any variable of interest via sufficient inputs. The GPR can model nonlinear decision boundaries, and there are several kernels to choose from. It is also fairly robust against overfitting, especially in high-dimensional space. As an intrinsic extension of the Gaussian distribution, Gaussian Processes (GP) bear their properties with mean and covariance as a vector and matrix respectively (MacKay 1998). A GP is defined as a collection of random variables, any limited number of which has a joint multivariate Gaussian distribution. The appropriate selection of kernel type is the most important step in the GPR due to its direct impact on training and classification precision. In fact, the GPR method is memory-intensive, and trickier to tune due to the importance of picking the right kernel. In kernel-based models we will be able to predict the proper behavior of the system, although we will not be able to characterize its intrinsic structure and behavior (Babovic 2009).

 Performance criteria

In the current study, the proposed model’s efficiency was assessed via correlation coefficient (R), determination coefficient (DC), and root mean square error (RMSE) criteria as follows:

\[
DC = 1 - \frac{\sum_{i=1}^{N} (l_o - l_p)^2}{\sum_{i=1}^{N} (l_o - \bar{l_o})^2}, \quad R = \frac{\sum_{i=1}^{N} (l_o - \bar{l_o}) \times (l_p - \bar{l_p})}{\sqrt{\sum_{i=1}^{N} (l_o - \bar{l_o})^2 \times (l_p - \bar{l_p})^2}},
\]

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (l_o - l_p)^2}{N}}
\]  

where \(l_o\), \(l_p\), \(\bar{l_o}\), \(\bar{l_p}\), \(N\) are the observed values, predicted values, mean observed values, mean predicted values, and number of data samples, respectively. \(R\) provides information for linear dependence between observation and corresponding predicted values. The RMSE is of special interest because it is widely reported in the literature. The RMSE describes the average difference between predicted values and measured values. This measurement is not oversensitive to extreme values (outliers), but is rather sensitive to additive and proportional differences between model predictions and observations. Therefore, correlation-based measures (e.g. DC) can indicate that a model is a good predictor (Legates & McCabe 1999). The DC measures the fraction of the variance in the data explained by the model. The small
quantity for RMSE (down to 0) and high quantity for DC and $R$ (up to 1) represent the higher performance of the model. It should be noted that in this study all input variables were scaled between 0 and 1 in order to eliminate the input and output variable dimensions.

**Model developing**

Selection of appropriate variables as inputs is the most important step in modeling via intelligence methods. In this research, previous values of daily river stage–discharge over the period of 2002–2006 were used as inputs to model the river discharge values. In the modeling process two states were considered: in the first state, the river daily discharge of the selected stations was predicted based on each station’s data, and in the second state, modeling was done based on the data from previous stations. Also, the impact of data pre-processing on improving model accuracy was assessed using the WT and EEMD methods. According to Aussem et al. (1998), the minimum decomposition level in the WT method can be obtained as follows:

$$L = \text{int} \log N$$

(2)

where $L$ is the decomposition level and $N$ is the number of time series. Using Equation (2) the value of parameter $L$ was obtained as 3, however, due to the seasonality of the process, $L = 5$ was used. Several combinations of discharge and stage values were used as inputs of the models to predict daily discharge. The developed input combinations for river discharge prediction are shown in Table 2. For all defined combinations, $t$ demonstrates the time step (daily step), $H(t - 1)$ is stage values at time $t - 1$, and $Q(t - 1)$ and $Q(t - 2)$ represent discharge values at time $(t - 1)$ and $(t - 2)$, respectively. The output is discharge values at current time step $Q(t)$. In Figure 3 the considered modeling process is shown.

**RESULTS AND DISCUSSION**

**GPR model development**

It should be noted that each artificial intelligence method has its own parameters, and for achieving the desired results, the optimized amount of these parameters should be determined. For example, in designing the GPR approach the selection of the appropriate type of kernel function is needed. There are various kernel functions which can be used based on the nature of the studied phenomenon. However, most studies indicate that the radial basis function (RBF) kernels result in better prediction for different hydrological issues. In this research, in order to determine the best performance of the GPR model and select the best kernel function, several developed models were tested via the GPR considering various kernels. Also, the optimal value of capacity constant ($C$) and the size of the error-intensive zone ($\epsilon$) in the GPR method are required due to their high impact on the accuracy of the mentioned regression approaches. The coefficient $C$ is a positive constant that influences a trade-off between an approximation error and the regression and must be selected by the user, and $\epsilon$ has an effect on the smoothness of the GPR’s response, so both the complexity and the generalization capability of the network depend on its value. If epsilon is larger than the range of the target values we cannot expect a good result. If epsilon is 0, we can expect overfitting (Smola 1996). Also, the variable parameter used with a kernel function considerably affects the flexibility of the function. These parameters should be selected by the

**Table 2 | Considered models for river discharge modeling**

| State 1 | Model | Input variable | State 2 | Model | Input variable |
|---------|-------|----------------|---------|-------|----------------|
| I       | $Q_1(t), Q_2(t), Q_3(t) = Q(t - 1)$ | I(I) | $Q_2(t) = Q_1(t - 1)$ | H(I) | $Q_2(t) = Q_2(t - 1)$ |
| II      | $Q_1(t), Q_2(t) = H(t - 1)$ | I(II) | $Q_2(t) = H_1(t - 1)$ | H(II) | $Q_3(t) = H_2(t - 1)$ |
| III     | $Q_1(t), Q_2(t) = Q(t - 1)$, $H(t - 1)$ | I(III) | $Q_2(t) = Q_1(t - 1), H_1(t - 1)$ | H(III) | $Q_2(t) = Q_2(t - 1), H_3(t - 1)$ |

$Q(t - 2)$
**Figure 3** Proposed modeling process in the study.

**Figure 4** Performance of GPR method with different kernel functions for model (IV) of station 1.
In the current study, in accordance with Cherkassky & Ma (2002), optimization of these parameters has been done by a systematic grid search of the parameters using cross-validation on the training set. For selecting the optimum parameters and assessing average developed model performance, the RMSE and DC were used based on some well-known studies (Willmott & Matsuura 2005; Roman et al. 2012). Figure 4 indicates the results of statistical parameters of different kernel functions for model (IV) of the river discharge modeling in station 1. According to the results, using the kernel function of RBF led to better prediction accuracy in comparison with the other kernels with values of DC = 0.82 and RMSE = 0.042 for the calibration set and DC = 0.75 and RMSE = 0.05 for the verification set. Therefore, the RBF kernel function was used as the core tool of the GPR, which was applied for the rest of the models.

THE RESULTS OF STAGE–DISCHARGE MODELING

Modeling based on raw data

For evaluating the river discharge in the selected stations, several models were developed based on the river discharge and stage data. The models were analyzed with the GPR model to carry out the river discharge prediction. Table 3 and Figure 5 show the results of the developed GPR models. From the obtained results of the statistical parameters (RMSE, R, and DC), it can be stated that in the first state the model (IV) with input parameters of \( Q(t - 1) \), \( Q(t - 2) \) performed better than the others. Based on the results of models (I) and (II), it can be seen that in estimation of the RSD, using previous stage data led to more accurate results than using previous discharge data. In the second state, for station 2, the model I(III) with input parameters of \( Q_1(t - 1) \), \( H_1(t - 1) \) was selected as the superior model and for station 3, the model H(IV) with parameters of \( Q_2(t - 1) \), \( Q_2(t - 2) \) led to better predictions. In the stations’ relationship investigation, it seems that using the previous stations’ discharge data yielded better predictions than the previous stations’ stage data. A comparison between the results of the two considered states showed that modeling based on each station’s own data led to more desirable results. However, using the previous stations’ data in the modeling process led to relatively accurate results; therefore, via the GPR approach the previous stations’ data could be applied when the stations’ own data were unavailable.

| Station | Performance criteria | Calibration | Verification | Calibration | Verification |
|---------|----------------------|-------------|--------------|-------------|--------------|
|         | Model                | \( R \)  | \( DC \)  | \( RMSE \) | \( R \)  | \( DC \)  | \( RMSE \) | \( R \)  | \( DC \)  | \( RMSE \) |
| State 1 | 1 I                  | 0.89       | 0.76       | 0.050      | 0.84       | 0.69       | 0.062      | 2          | 0.89       | 0.78       | 0.049      |
|         | II                   | 0.90       | 0.77       | 0.048      | 0.86       | 0.73       | 0.056      | 0.92       | 0.79       | 0.047      |
|         | III                  | 0.92       | 0.08       | 0.044      | 0.87       | 0.74       | 0.053      | 0.94       | 0.83       | 0.039      |
|         | IV                   | 0.93       | 0.82       | 0.042      | 0.90       | 0.75       | 0.050      | 0.95       | 0.86       | 0.033      |
| State 2 | 3-2                  |            |            |            |            |            |            |            |            |            |
|         | 2-I(I)               | 0.90       | 0.75       | 0.043      | 0.86       | 0.72       | 0.060      | H(I)       | 0.87       | 0.66       | 0.051      |
|         | I(II)               | 0.90       | 0.78       | 0.041      | 0.86       | 0.71       | 0.061      | H(II)      | 0.86       | 0.68       | 0.049      |
|         | I(III)              | 0.91       | 0.79       | 0.040      | 0.87       | 0.75       | 0.058      | H(III)     | 0.87       | 0.69       | 0.048      |
|         |                     |            |            |            |            |            |            | H(IV)      | 0.87       | 0.70       | 0.045      |
|         |                     |            |            |            |            |            |            | H(V)       | 0.87       | 0.65       | 0.052      |
|         |                     |            |            |            |            |            |            | H(VI)      | 0.87       | 0.68       | 0.049      |
Modeling based on pre-processing data

In this section, the effect of time series pre-processing on increasing model accuracy was investigated. The input time series were decomposed using the WT and EEMD methods. To decompose the time series by WT, a mother wavelet which is more similar to the signal should be selected. In this study, the Daubechies (db2 and db4) and symlet (sym2 and sym4) mother wavelets were trained and it was found that the db4 mother wavelet led to better outcomes (see Figure 6). Therefore, the db4 mother wavelet was used for time series decomposition. Also, the decomposition level was selected as five ($L = 5$), which means five detail and one approximation subseries would be obtained (see Figure 7(a)). Figure 7 illustrates the decomposed time series of station 1 using db4 and $L = 5$. In the second step, data was decomposed via EEMD. The principle of EEMD is the decomposition of a signal to different IMFs and one residual signal. The sum of these signals will be the same original signal. The formation of IMFs is based on subtracting the basic function from the original signal. This process continues until the residual signal remains almost constant. In the EEMD, two critical parameters directly affect the decomposition performance of the algorithm.
Figure 7: Decomposition of daily discharge time series of station 1 by (a) Daubechies 4 wavelet (level 5), and (b) EEMD pre-processing methods.
The number of the ensemble and the amplitude of white noise need to improve the daily discharge time series forecasting accuracy. The well-established statistical rule between the ensemble number, \( m \), the amplitude of the added white noise, \( \epsilon \), and the standard deviation of error, \( e_m \), is described by Equation (3) (Wu & Huang 2009). Wu & Huang (2009) concluded that the effect of the added white noise should decrease using Equation (3), and pointed out that, if the added noise amplitude is too small, then it may not cause the change of extreme that the EMD depends on; instead, if the added noise is too large, it will result in redundant IMF components. Based on experimental observations in most cases (Wu & Huang 2009), the amplitude of added noise should be about 0.2 times the standard deviation of that of the sample data. In this study, the EEMD method was employed with the ensemble number of 100 and the white noise amplitude of 0.2 times standard deviation. More detailed analysis of noise and ensemble averaging effects on decomposition regarding these parameters can be found in Wu & Huang (2009).

\[
e_m = \epsilon / \sqrt{m}
\]

(3)

In this study, according to Figure 7(b), time series were decomposed into ten IMFs and one residual signal. Then, the obtained subseries were used as inputs of the GPR model to predict the river discharge. The results of the integrated pre-processing models are listed in Table 4 and shown in Figure 8. According to the results presented in Tables 3 and 4, it can be deduced that data pre-processing improved significantly the results accuracy and the applied integrated models were more accurate than the single GPR model. In fact, the use of WT and EEMD led to an improvement in the outcomes. The integrated pre-processing methods increased the model accuracy approximately between 25% and 40%. According to the results, it can be seen that between the two pre-processing methods the EEMD had higher RMSE error criteria in comparison with the WT. Therefore, it can be stated that the WT method performed more successfully than the EEMD method in enhancing the prediction accuracy. Also, it was found that model (IV) with input parameters of \( Q(t - 1) \), \( Q(t - 2) \) in modeling based on each station's own data performed better than the others. In this state adding the \( H(t - 1) \) variable to the input combinations enhanced the model accuracy. In the second state, for station 2, model I(III) with input parameters of \( Q(t - 1) \), \( H(t - 1) \) and for station 5, model H(IV) with parameters of \( Q(t - 1) \), \( Q(t - 2) \) were superior models. From the results, it can be seen that discharge modeling based on the station's own data resulted in more desirable outcomes.

The RMSE error criterion was used to graphically compare the performance of the single and integrated GPR models. The results are shown in Figure 9(a). As can be seen, in the river discharge modeling the distribution ranges of the RMSE were smaller for the integrated models and the WT-GPR model led to more accurate results.

In the next step, for evaluating the impact of each independent parameter, sensitivity analysis was performed for the best model of each state. In this regard, the input variables of the best models were omitted one by one and the WT-GPR model, which had higher efficiency compared with the GPR and EEMD-GPR models, was rerun. Based on the results from Figure 9(b), it can be deduced that parameter \( Q(t - 1) \) had the most significant impact on daily river discharge prediction.

Validation of proposed best models using Arkansas River data

For verifying the applied method's efficiency, datasets of Arkansas River were used. The Arkansas River is a major tributary of the Mississippi River. It generally flows to the east and southeast as it traverses the US states of Colorado, Kansas, Oklahoma, and Arkansas. At 2,364 km, it is the sixth-longest river in the United States, the second-longest tributary in the Mississippi–Missouri system, and the 45th longest river in the world. Its drainage basin covers nearly 440,000 km². Its volume is much smaller than the Missouri and Ohio rivers, with a mean discharge of about 40,000 ft³/s. In this regard, for each state (i.e. modeling based on the station's own data and the previous stations' data) the superior model was run using the single and integrated models and the results were compared with each other. The obtained results are listed in Table 5 and shown in Figure 10. As can be seen from Table 5, the integrated models led to the desired accuracy and the efficiency of WT-GPR was better than that of EEMD-GPR and GPR. Based on the results, it can be deduced that discharge modeling based on the station's own data led to more desirable outcomes.
Table 4 | Statistical parameters of the WT- or EEMD-GPR models for the RSD modeling; state 2 with data processing

| Station | Model/Method | Performance criteria | Calibration | Verification | Calibration | Verification |
|---------|--------------|----------------------|-------------|--------------|-------------|--------------|
|         |              |                      | R  | DC | RMSE | R  | DC | RMSE | R  | DC | RMSE | R  | DC | RMSE |
| State 1 |              |                      |    |    |      |    |    |      |    |    |      |    |    |      |
| 1       | I            | EEMD                 | 0.91| 0.86| 0.040| 0.89| 0.81| 0.047| 0.91| 0.87| 0.038| 0.91| 0.83| 0.042|
|         |              | WT                   | 0.92| 0.86| 0.038| 0.90| 0.83| 0.044| 0.94| 0.89| 0.036| 0.91| 0.84| 0.040|
|         | II           | EEMD                 | 0.91| 0.88| 0.039| 0.89| 0.84| 0.043| 0.94| 0.89| 0.033| 0.92| 0.85| 0.038|
|         |              | WT                   | 0.94| 0.87| 0.037| 0.91| 0.85| 0.041| 0.95| 0.90| 0.031| 0.94| 0.87| 0.035|
|         | III          | EEMD                 | 0.93| 0.90| 0.033| 0.93| 0.87| 0.040| 0.94| 0.92| 0.030| 0.93| 0.90| 0.033|
|         |              | WT                   | 0.95| 0.91| 0.031| 0.94| 0.88| 0.038| 0.96| 0.93| 0.029| 0.94| 0.91| 0.032|
|         | IV           | EEMD                 | 0.96| 0.92| 0.030| 0.94| 0.89| 0.035| 0.96| 0.93| 0.028| 0.95| 0.93| 0.031|
|         |              | WT                   | 0.96| 0.93| 0.029| 0.95| 0.90| 0.033| 0.98| 0.95| 0.026| 0.96| 0.94| 0.029|
| 2       |              |                      |    |    |      |    |    |      |    |    |      |    |    |      |
| State 2 |              |                      |    |    |      |    |    |      |    |    |      |    |    |      |
| 2-1     | I(I)         | EEMD                 | 0.84| 0.79| 0.049| 0.77| 0.74| 0.057| 0.90| 0.74| 0.049| 0.82| 0.68| 0.052|
|         |              | WT                   | 0.87| 0.80| 0.047| 0.79| 0.77| 0.051| 0.91| 0.79| 0.045| 0.83| 0.77| 0.046|
|         | I(II)        | EEMD                 | 0.89| 0.83| 0.042| 0.80| 0.79| 0.051| 0.88| 0.76| 0.043| 0.81| 0.66| 0.056|
|         |              | WT                   | 0.90| 0.86| 0.040| 0.82| 0.80| 0.049| 0.89| 0.81| 0.041| 0.83| 0.74| 0.050|
|         | I(III)       | EEMD                 | 0.91| 0.89| 0.036| 0.88| 0.84| 0.044| 0.89| 0.77| 0.041| 0.81| 0.69| 0.051|
|         |              | WT                   | 0.92| 0.90| 0.034| 0.88| 0.86| 0.040| 0.90| 0.82| 0.039| 0.82| 0.76| 0.048|
|         | II           | EEMD                 | 0.95| 0.92| 0.031| 0.94| 0.87| 0.038| 0.95| 0.93| 0.028| 0.95| 0.93| 0.031|
|         |              | WT                   | 0.95| 0.93| 0.050| 0.94| 0.88| 0.036| 0.95| 0.93| 0.050| 0.94| 0.88| 0.036|
Figure 8 | Scatterplot of modeling via (a) WT-GPR and (b) EEMD-GPR superior models for both states in verification series.
Figure 9 | (a) Comparison of the distribution range of the RMSE criterion for superior models of the used methods, and (b) relative significance of each of the input parameters of the best models.
| Station | Performance criteria | Calibration | Verification | Calibration | Verification |
|---------|----------------------|-------------|--------------|-------------|--------------|
|         |                      | $R$ | DC | RMSE | $R$ | DC | RMSE | $R$ | DC | RMSE | $R$ | DC | RMSE |
| State 1 |                      |     |    |      |     |    |      |     |    |      |     |    |      |
| 1       | IV                   | GPR | 0.91 | 0.81 | 0.057 | 0.91 | 0.76 | 0.059 | 2 IV | GPR | 0.93 | 0.85 | 0.053 | 0.89 | 0.81 | 0.056 |
|         |                      | EEMD | 0.97 | 0.91 | 0.044 | 0.95 | 0.89 | 0.048 |      | EEMD | 0.97 | 0.94 | 0.059 | 0.94 | 0.91 | 0.043 |
|         |                      | WT   | 0.99 | 0.94 | 0.038 | 0.96 | 0.91 | 0.042 |      | WT   | 0.97 | 0.96 | 0.036 | 0.95 | 0.93 | 0.041 |
|         |                      | GPR  | 0.86 | 0.81 | 0.056 | 0.84 | 0.71 | 0.066 |      |       |      |      |      |      |      |      |
| 3       | IV                   | EEMD | 0.94 | 0.91 | 0.043 | 0.93 | 0.88 | 0.049 |      |       |      |      |      |      |      |      |
|         |                      | WT   | 0.95 | 0.94 | 0.040 | 0.92 | 0.91 | 0.044 |      |       |      |      |      |      |      |      |
| State 2 |                      |     |    |      |     |    |      |     |    |      |     |    |      |
| 2-1     | I(III)               | GPR  | 0.90 | 0.78 | 0.062 | 0.85 | 0.73 | 0.066 | 3-2 | H(IV) | GPR  | 0.85 | 0.71 | 0.067 | 0.77 | 0.65 | 0.069 |
|         |                      | EEMD | 0.91 | 0.83 | 0.055 | 0.87 | 0.80 | 0.058 |      | EEMD | 0.88 | 0.80 | 0.057 | 0.86 | 0.78 | 0.063 |
|         |                      | WT   | 0.94 | 0.87 | 0.050 | 0.90 | 0.85 | 0.052 |      | WT   | 0.91 | 0.87 | 0.048 | 0.89 | 0.83 | 0.054 |
CONCLUSION

The accurate prediction of river discharge is an important factor in improving water resource management. This study assessed the capability of time series pre-processing methods for river daily discharge modeling. In this regard, in the first step, the raw time series (without any data processing) were imposed on the GPR model. Then, the time series were decomposed to several subseries using WT and EEMD and these subseries were used as inputs. According to the results, it was found that the use of both WT and EEMD pre-processing methods increased the models’ accuracy. The applied pre-processing methods enhanced the GPR model performance approximately between 25% and 40%. It was observed that in estimation of the river discharge, using the previous data of flow discharge led to more accurate results. It showed that modeling based on the each stations’ own data led to more desirable results. However, using the integrated GPR approaches, the previous stations’ data could be used in the river daily discharge modeling when the station’s own data was unavailable. Based on the obtained results, it was observed that in state 1, the model with input parameters of \( Q(t - 1), Q(t - 2) \) and in state 2, the model with parameters of \( Q_1(t - 1), H_1(t - 1) \) for station 2 and the model with parameters of \( Q_2(t - 1), Q_2(t - 2) \) for station 3 performed more successfully. The efficiency of the applied methodology was verified using Arkansas River datasets and results showed that the applied methods were successful in the modeling process. In general, the results showed that the integration of the GPR model with pre-processing models could be a suitable solution for more accurate prediction of hydrological variables such as river daily discharge.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Adamowski, K., Prokop, A. & Adamowski, J. 2009 Development of a new method of wavelet aided trend detection and estimation. Hydrology Processes 23 (18), 2686–2696.
Ahmed, K., Sachindra, D. A., Shahid, S., Iqbal, Z., Nawaz, N. & Khan, N. 2020 Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmospheric Research 236, 104806.
Anmala, J., Zhang, B. & Govindaraju, R. S. 2000 Comparison of ANNs and empirical approaches for predicting watershed runoff. Journal of Water Resources Planning and Management 126 (3), 156–166.
Aussem, A., Campbell, J. & Murtagh, F. 1998 Wavelet-based feature extraction and decomposition strategies for financial forecasting. Journal of Computational Intelligence in Finance 6 (2), 5–12.
Azamathulla, H. M. & Wu, F. C. 2011 Support vector machine approach for longitudinal dispersion coefficients in natural streams. Applied Soft Computing 11 (2), 2902–2905.
Azamathulla, H. M., Haggiabi, A. H. & Parsaei, A. 2016 Prediction of side weir discharge coefficient by support vector machine technique. Water Supply 16 (4), 1002–1016.

Babovic, V. 2009 Introducing knowledge into learning based on genetic programming. Journal of Hydroinformatics 11 (3–4), 181–193.

Baiamonte, G. & Ferro, V. 2007 Simple flume for flow measurement in sloping open channel. Journal of Irrigation and Drainage Engineering 133 (1), 71–78.

Cherkassky, V. & Ma, Y. Q. 2002 Selection of meta-parameters for support vector regression. In: Artificial Neural Networks – ICANN 2002 (J. R. Dorronsoro, ed.), Springer-Verlag, Berlin, Germany, pp. 687–693.

Coulibaly, P., Anctil, F. & Bobée, B. 2000 Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. Journal of Hydrology 230 (3–4), 244–257.

Farajzadeh, J. & Alizadeh, F. 2017 A hybrid linear–nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using Wavelet-SARIMAX-LSKELM conjugated model. Journal of Hydroinformatics 20 (1), 246–262.

Habib, E. H. & Meselhe, A. E. 2006 Stage–discharge relations for low-gradient tidal streams using data-driven models. Journal of Hydraulic Engineering 132 (5), 482–492.

Hasanpour Kashani, M., Daneshfaraz, R., Ghorbani, M. A., Najafi, M. R. & Kisi, O. 2015 Comparison of different methods for developing a stage–discharge curve of the Kizilirmak River. Journal of Flood Risk Management 8 (1), 71–86.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. & Liu, H. H. 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences 454, 903–995.

Kaveh, K., Bui, M. D. & Rutschmann, P. 2017 A comparative study of three different learning algorithms applied to ANFIS for predicting daily suspended sediment concentration. International Journal of Sediment Research 32 (3), 340–350.

Labate, D., La Foresta, F., Occhiuto, G., Morabito, F. C., Lay-Ekuakille, A. & Vergallo, P. 2013 Empirical mode decomposition vs wavelet decomposition for the extraction of respiratory signal from single-channel ECG: a comparison. IEEE Sensors Journal 13 (7), 2666–2674.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of ‘goodness of fit’ measures in hydrologic and hydroclimatic model validation. Water Resources Research 35 (1), 233–241.

Lei, Y., He, Z. & Zì, Y. 2009 Application of the EEMD method to rotor fault diagnosis of rotating machinery. Mechanical Systems and Signal Processing 23 (4), 1327–1338.
Wu, Z. H. & Huang, N. E. 2009 Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis* 1, 1–41.

Zeinali, M., Zmanzad-Ghavidel, S., Mehri, Y. & Azamathulla, H. M. 2020 Interaction of hydro-socio-technology-knowledge indicators in integrated water resources management using soft-computing techniques. *Water Supply* 21 (1), 470–491.

Zhou, J., Peng, T., Zhang, C. & Sun, N. 2018 Data pre-analysis and ensemble of various artificial neural networks for monthly streamflow forecasting. *Water* 10 (5), 628.

First received 21 December 2020; accepted in revised form 4 March 2021. Available online 15 March 2021