Modeling of contaminant concentration using the classification-based model integrated with data preprocessing algorithms

Saeed Pipelzadeh and Reza Mastouri

ABSTRACT

Water quality is one of the most important factors contributing to a healthy life; meanwhile, total dissolved solids (TDS) and electrical conductivity (EC) are the most important parameters in water quality, and many water developing plans have been implemented for the recognition of these factors. The accurate prediction of water quality parameters (WQPs) is an essential requisite for water quality management, human health, public consumption, and domestic uses. Using three novel data preprocessing algorithms (DPAs), including empirical mode decomposition (EMD), ensemble EMD (EEMD), and variational mode decomposition (VMD) to estimate two important WQPs, TDS and EC, differentiates this study from the existing literature. The acceptability and reliability of the proposed models (e.g., model tree (MT), EMD-MT, EEMD-MT, and VMD-MT) were evaluated using five performance metrics and visual plots. A comparison of the performances of standalone and hybrid models indicated that DPAs can enhance the performance of standalone MT model for both TDS and EC estimations. For instance, the VMD-MT model (root-mean-square error (RMSE) = 24.41 mg/l, ratio of RMSE to SD (RSD) = 0.231, and Nash–Sutcliffe efficiency ($E_{ns}$) = 0.94 (Garmrood) and RMSE = 31.85 mg/l, RSD = 0.133, and $E_{ns}$ = 0.98 (Varand)) outperformed other hybrid models and original MT models for TDS estimations. Regarding the EC estimation results, as for $R^2$, VMD could enhance the accuracy of prediction for the MT model for Garmrood and Varand stations by 10.2 and 7.6%, respectively.

Key words | data preprocessing algorithms, electrical conductivity, model tree, total dissolved solids, water quality parameters

HIGHLIGHTS

- Two important water quality parameters, TDS and EC, were modeled in this study.
- Three data preprocessing algorithms were used to address the nonstationary of the dataset.
- To validate proposed models, a classification-based MT was used as the benchmark model.
- The VMD-MT proves to be an effective tool to provide strong technical support for WQPs.

INTRODUCTION

Conductivity or electrical conductivity (EC) and total dissolved solids (TDS) are some of the water quality parameters (WQPs; dependent variables) that should be controlled in maintaining human health and welfare. Conductivity measurements are used routinely in many industrial and environmental applications as a fast, inexpensive, and reliable way of measuring the ionic content in a solution (Olson & Hawkins 2017; Asadi et al. 2020).
In many cases, conductivity is linked directly to the TDS. TDS is a measure of the solid fraction of a sample which is able to pass through a filter. The amount of dissolved solids gives a general indication of the suitability of the water as a drinking source and for certain agricultural and industrial uses (Sheykhzadeh 2019). In today’s industrial world, most of the global natural water sources, including those in Iran, contain impurities such as the TDS and EC. Numerous factors that include cations such as sodium ion (Na\(^+\)), potassium ion (K\(^+\)), calcium ion (Ca\(^2+\)), and magnesium ion (Mg\(^2+\)) with sulfate ion (SO\(_4\)\(^2-\)) and bicarbonate ion (HCO\(_3\)) with chloride ion (Cl\(^-\)) affect the concentration of these parameters in natural water systems (Asadollahfardi et al. 2017).

In the last decades, artificial intelligence (AI) techniques, such as artificial neural networks (ANNs), adaptive neuro-fuzzy inference system (ANFIS), gene expression programming (GEP), support-vector machines (SVMs), and model tree (MT), have become viable and popular due to their advantages and have been widely developed in solving a variety of environmental engineering and water quality engineering problems (Solomatine & Dulal 2003; Shiri et al. 2011; Najafzadeh et al. 2016; Yassin et al. 2016; Rezaie-Balf et al. 2017; Hong et al. 2018; Mouatadid et al. 2018; Rezaie-Balf & Kisi 2018; Najafzadeh et al. 2019; Chen et al. 2020; Sattari et al. 2020).

For the estimation of WQPs, Singh et al. (2011) utilized the clustering method, or support-vector clustering, to optimize surface water quality monitoring in the city of Lucknow, India. The overall view of the water quality index of their study area revealed that most of the study area come under highly to very highly polluted zones. Tan et al. (2012) predicted phosphorous values in China with the least-square support-vector regression (LS-SVR) method. They compared the efficiency of the LS-SVR method with neural networks of the radial basis function (RBF) and back-propagation (BP). Experimental results showed that for the small sample case with noise, the LS-SVM method is better than multilayer BP and RBF neural networks and is able to better meet the requirements of water quality prediction. Liu et al. (2013) addressed WQPs’ prediction in aquaculture employing the grid partition (GP) and real-value genetic algorithm support-vector regression (RGA-SVR). They used the GA to modify the coefficients of the SVR method. The results showed the superiority of the RGA-SVR algorithm over other methods based on the root-mean-square error (RMSE) and mean absolute percentage error. Ghavide & Montaseri (2014) employed ANN, GEP, and ANFIS with GP as well as ANFIS with subtractive clustering (ANFIS-SC) to predict TDS values of the Zarinehroud basin, Iran. A comparison was made between the above AI approaches, and the results demonstrated the superiority of GEP over the other intelligent models. Abyaneh (2014) compared ANN with multivariate linear regression (MLR) for the prediction of BOD and COD in the wastewater treatment plant. In their study, ANN could predict BOD and COD parameters with higher precision than MLR.

Alizadeh & Kavianpour (2015) implemented the ANN and wavelet neural network (WNN) to predict the DO concentration. A comparison of the performance of the proposed models showed that WNN was an efficient technique for the DO-level prediction. GEP and ANN were recruited by Mohammadpour et al. (2016) to predict water quality indicators (WQIs). Their findings indicated that ANN and GEP could prognosticate WQIs with good precision. Asadollahfardi et al. (2017) employed two types of ANN, multilayer perceptron and RBF, to predict TDS at Karaj Dam in Iran. They used data normalization in order to improve the models’ accuracy and found that this technique was useful for the models’ performance in terms of TDS prediction. Also, by using sensitivity analysis, they concluded that magnesium and sulfate were the most effective WQPs for predicting TDS at the Karaj Dam. ANFIS was applied by Ahmed ad Shah (2017) in order to predict BOD. Results showed that the ANFIS model was a useful method for BOD prediction. Montaseri et al. (2018) investigated the accuracy of the AI models for TDS in different climatic situations of Iran. They applied ANN and ANFIS, including ANFIS with GP and ANFIS-SC, GEP, wavelet-ANN, wavelet-ANFIS, and wavelet-GEP, in predicting TDS at four regions over a period of 20 years. Based on the Pearson correlation coefficient, EC, Na, and Cl parameters were selected as input variables to forecast the amount of TDS in four basins under study. A comparison was made between these AI approaches, which emphasized the superiority of wavelet-GEP over the other intelligent models. Pan et al. (2019), by applying three different
modeling approaches, including dual-step multiple linear regression (MLR), hybrid principal component regression (PCR), and BP neural networks (BPNNs), predicted TDS concentrations in an urban aquifer. The results showed that the percentage error for hybrid PCR was inversely proportional to TDS concentrations, which was not observed for dual-step MLR. In addition, larger errors were obtained from the BPNN models, and higher percentage errors were observed in monitoring wells with lower TDS concentrations.

Due to complex characteristics of time-series WQPs, a standalone model can hardly satisfy the estimation accuracy requirements. Therefore, the hybrid models combined with different single models will be an effective way to improve WQPs’ estimation accuracy (Liu et al. 2018; Rezaie-Balf et al. 2019). In this sense, a large number of research studies have concentrated on employing hybrid approaches. For instance, Napolitano et al. (2011) carried out a research to explore several aspects of ANN for forecasting daily streamflow using empirical mode decomposition (EMD). They showed that the analysis using the EMD technique increases ANN prediction accuracy and reliability. Likewise, Zhang et al. (2017) investigated and forecasted short-term wind speed in the eastern coastal area of China using CEEMDAN with flower pollination algorithm. Prasad et al. (2018) used CEEMDAN and ensemble EMD (EEMD) integrated with ELM to improve the models’ performance for soil moisture forecasting. They found that the CEEMDAN–ELM model outperformed the other proposed models for upper layer soil moisture forecasting. Moreover, the two-phase extreme learning machines integrated with the variational mode decomposition (VMD) algorithms for a better frequency resolution of the original datasets in the Yingluoxia watershed, Northwestern China.

By reviewing the literature, it is very difficult to predict hydrological variables in a certain region due to the nonstationary and trends in hydrological time series that may arise from climatic changes, anthropogenic changes in river basins, and low-frequency climatic variability. Without considering these issues, target variables cannot be estimated with sufficient accuracy from time-series hydrological records in terms of WQPs’ modeling, especially TDS and EC parameters. To the best knowledge of the authors, although many previous studies have modeled the WQPs using AI techniques, none of them investigated and analyzed the complexity, nonlinearity, and nonstationary of the hydrological variables by data decomposition algorithms.

This study proposes a new and accurate hybrid model for predicting WQPs, TDS, and EC, using ions at Varand and Garmrood, two hydrometric stations of Tajan basin, Iran. The proposed WQP estimating framework was developed based on the combination of three data preprocessing algorithms (DPAs), EMD, EEMD, and VMD, with an AI-based MT model that was not addressed by the literature related to the WQPs modeling. The main contributions of the research are as follows:

(1) Presenting an accurate and stable formula for each TDS and EC using a classification-based AI model at both above-mentioned stations.

(2) Introducing a more reliable DPA coupled with MT in order to estimate monthly scaled EC and TDS at the Tajan basin. In this regard, the most effective DPAs are adopted for managing the volatility and randomness of WQP series data. After decomposition and reconstruction, the volatility and randomness of the original wind speed data were effectively reduced, and the forecasting performance was validly enhanced.

(3) Conducting a comprehensive and scientific evaluation to test the performance of the proposed system. So, all standalone and hybrid models, MT, EMD-MT, EEMD-MT, and VMD-MT, were compared with several performance metrics and graphical plots.

MATERIALS AND METHODS

Case study: Tajan basin

Tajan River flows south of Sari city located in Mazandaran, Iran. The length of the river is approximately 95 km, flowing westward to join the Tajan River, and ultimately discharging into the Caspian Sea. The Tajan watershed (53°56’–36°17’ north latitude and 53°7’–53°42’ east longitude) has experienced several floods with the most drastic one in 1997 (Rezaie-Balf & Kisi 2018). This basin is bordered in the
north by the Caspian Sea, in the south by the central Iran Basin, in the east by the Nekarood basin, and in the west by the Babolrood basin. The elevation of general landscapes of the area ranges from −26 to 1,600 m above the mean sea level (Aazami et al. 2015). Its climate is semi-humid and cold-humid. The average slope of the area, discharge of the river, and annual rainfall are 85%, 20 m³/s, and 539 mm, respectively. The catchment elevation varies between 26 and 3,728 m (Ahmadi-Mamaqani et al. 2011). The two most subtributary rivers, Zaremrood and Chahardangeh, with two measurement stations, namely Garmrood and Varand, in the Tajan basin were considered in order to investigate the application of the proposed models for suspended sediment loads using monthly discharge as the input variable (see Figure 1).

In this research, several WQPs on a monthly scale were selected from hydrometric stations. The data were obtained from the Meteorological Organization of Mazandaran Province. The lengths of the selected stations were 26 (from 1991 to 2017) and 24 (from 1993 to 2017) years, respectively, for Garmrood and Varand, based on the consistency and availability of the data during that time period. The data were used into two phases: approximately 75% of the datasets were used for the training phase, while the remained datasets (25%) were kept for testing purposes. The statistical analysis for training and testing periods for physiochemical data was conducted, and the values obtained for minimum, maximum, mean, skewness coefficient ($C_s$), and standard deviation (SD) are given in Table 1.

![Figure 1](https://example.com/figure1.png)  
**Figure 1** Tajan River basin and elevation of the area.

### Model tree

MT was initially implemented by Quinlan (1992) to attain a relationship between input–output parameters and was reintroduced and improved by Wang & Witten (1997), who presented physically meaningful insights of phenomena. The M5 model can solve highly complex problems by dividing the data space into several subproblems (subspaces) and building piecewise linear functions for each subdomain at its terminal nodes. Therefore, the process of this algorithm generally completes in two different steps. In the first step, datasets are divided into subsets to construct an initial tree through a recursive splitting process. To determine the best attribute for splitting the dataset at each node, the SD is utilized as a splitting criterion. Trees are constructed using the SD reduction schema, which maximizes the expected error reduction for each node as follows (Quinlan 1992):

$$S D R = S D (I) - \sum_{i} \frac{|I_i|}{|I|} S D (I_i)$$  \hspace{1cm} (1)

where $I_i$ refers to the set of instances that reach the leaf (node), $I_i$ represents a subset of input data to the parent node, and SD is the standard deviation.
In the second phase, after pruning an overgrown tree, linear regression models are created and presented for each of the subtrees of samples in the terminating nodes. The pruning process is performed to prune back the overgrown trees, which is a pivotal step to avoid the overfitting problem and gain better generalization in the second stage. The subtrees in this step are transformed into leaf nodes by replacing them with linear regression functions. Later, a smoothing procedure in which all the leaf models are amalgamated along the path back to the root is applied to realize a final model (Solomatine & Xue 2004; Bhattacharya & Solomatine 2005). In this study, for a better understanding of MT, a general tree structure of the MT technique as a simple example is presented. According to Figure 2(a), four linear regression models are built based on the tree-building process, and the structure of each subset is created by its knowledge. In the illustration (Figure 2(b)), if \( X_1 < 2 \) and \( X_2 > 2.5 \), then the third model should be selected in the form of \( Y = a_0 + a_1 X_1 + a_2 X_2 \), where \( a_0, a_1, \) and \( a_2 \) represent regression coefficients, \( X_1 \) and \( X_2 \) are inputs, and \( Y \) is the output. Readers can refer to Wang & Witten (1997) and Talebi et al. (2017) for more details on MT. In this study, the MT was constructed using the WEKA 3.7 software, and default parameters were employed in the model development stage.

**Empirical mode decomposition**

EMD is a new and innovative self-adaptive time–frequency method for signal processing proposed by Huang et al. (1998). The oscillatory modes can be expressed using intrinsic mode function (IMF) components embedded in the data. Using IMFs, EMD represents a signal as a sum of zero-mean well-behaved fast and slow oscillation modes (Huang et al. 2003; Wu & Huang 2009). In general, an IMF represents a simple oscillatory mode compared with the simple harmonic function. Based on its definition, a shifting process of the original time series can be briefly expressed as follows (Yeh et al. 2010):

**Step 1:** Identify all local maxima and minima in the original time series \( E(t) \). The upper and lower envelopes of the time series are obtained by cubic spline interpolation. The mean of the upper and lower enveloping lines is \( m(t) \):

\[
m(t) = \frac{(E_{\text{max}}(t) + E_{\text{min}}(t))}{2}
\]  

(2)

**Step 2:** A new series \( h(t) \) is calculated by subtracting the mean \( m(t) \) from the original series \( E(t) \):

\[
h(t) = E(t) - m(t)
\]

(3)

**Step 3:** The EMD stopping criteria for shifting determines whether shifting should stop. If the stopping condition is met, \( h(t) \) is the IMF, and the next step is executed. If the stopping condition is not met, then \( h(t) \) is used as the original series; steps 1 and 2 are repeated until the stopping condition is met; and the first IMF, IMF1 \( c_1(t) \), is calculated.

**Step 4:** The residual series \( r_1(t) \) is obtained by subtracting the IMF \( c_1(t) \) from the original series \( E(t) \):

\[
r_1(t) = E(t) - c_1(t)
\]

(4)

**Step 5:** The residual series \( r_1(t) \) is used as the new original series, and steps 1–4 are repeated. All the IMFs, \( c_1(t) \), \( c_2(t) \), \ldots, \( c_n(t) \), are decomposed until \( c_n(t) \) is a monotonic or single-extreme-point residual.

**Ensemble EMD**

The EEMD method, proposed by Wu & Huang (2009), is an empirical procedure used to represent a nonlinear and non-stationary signal from original data. This data preprocessing method is an improvement of the EMD that reduces mode...
mixing and obtains the actual time–frequency distribution of the signal.

The principle is to leverage the statistical features (uniform frequency distribution) of Gaussian white noise. When white noise is added to a signal, the signal becomes continuous on different scales to reduce mode mixing. Details of the decomposition principle and procedure are as follows (Rezaie-Balf et al. 2019):

Step 1: White noise $n(t)$ with a mean of 0 and the SD constant is added to the original signal $E(t)$ multiple times. The SD of the white noise is set to 0.1–0.4 times the SD of the original signal (0.2 in this study):

$$E_1(t) = E(t) - n_1(t)$$

where $X_i(t)$ represents the signal after the $i$th addition of Gaussian white noise.

Step 2: Each $X_i(t)$ undergoes the EMD procedure. The IMF component obtained is denoted by $c_{ij}(t)$, and the residual term is denoted by $r_i(t)$. Among them, $c_{ij}(t)$ represents the $j$th IMF from the decomposition of the signal after the $i$th addition of Gaussian white noise.

Step 3: Steps 1 and 2 are repeated $N$ times. Based on the principle that the statistical mean of an uncorrelated random series is 0, the IMFs are subjected to an overall averaging operation to eliminate the impact of adding Gaussian white noise to the actual IMF multiple times. Finally, the IMF obtained from EEMD is as follows:

$$C_j(t) = \frac{1}{N} \sum_{i=1}^{N} c_{ij}$$

where $C_j(t)$ represents the $j$th IMF of the original signal obtained by EEMD. As the value of $N$ increases, the sum of IMFs for the corresponding white noise approaches 0. At this point, the result of EEMD is as follows:

$$X(t) = \sum_j C_j(t) + r(t)$$

where $r(t)$ is the final residual, which represents the average trend of the signal. Any signal $E(t)$ can be decomposed into multiple IMFs and one residual via EEMD. IMF $c_{ij}(t)$ ($j = 1, 2, \ldots$) represents the signal’s components from high frequency to low frequency. Each frequency contains distinct components and varies with the signal $E(t)$.

**Variational mode decomposition**

The VMD, as a newly developed algorithm for adaptive and quasi-orthogonal signal decomposition, was applied (Dragomiretskiy & Zosso 2014). The VMD algorithm has an ability in decomposing a signal $x(t)$ to $K$ discrete number of subsignals or modes $u_k$, where every component is selected compact around their respective center frequency $\omega_k$. The VMD is employed as a constrained optimization issue, which is expressed as (Dragomiretskiy & Zosso 2014) follows:

$$f = \min_{\{U_k, \{\omega_k\}\}} \left\{ \sum_k \left\Vert \delta(t) + \frac{j}{\pi t} u_k(t) \right\Vert e^{j\omega_k t} \right\}_2 \text{subject to } \sum_k U_k$$

where $\{U_k\}$: $\{U_1, U_2, \ldots, U_K\}$ and $\{\omega_k\}$: $\{\omega_1, \omega_2, \ldots, \omega_K\}$ indicate stenography impression for the set of all modes and their central frequencies, respectively. Moreover, $\delta(t)$ and $*$ denote the Dirac distribution and convolution, respectively. So, the Lagrangian multipliers and quadratic penalty term are defined in order to convert this constrained optimization problem into an unconstraint one (Wang et al. 2015):

$$L([U_k], \{\omega_k\}, \lambda) = \alpha \left\{ \sum_k \left\Vert \delta(t) + \frac{j}{\pi t} u_k(t) \right\Vert e^{j\omega_k t} \right\}_2^2 + \left\Vert f(t) - \sum_k U_k(t) \right\Vert_2^2 + (\lambda(t), f(t) - \sum_k U_k(t))$$

Equation (4) can be resolved with the alternative direction method of multipliers.

The optimization of Equation (8) has been shown as two stages:

1. $U_k$ minimization:

$$\dot{U}_k^{n+1} = \frac{f(\omega) - \sum_{i=k} U_i(\omega) + \frac{\lambda(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \quad \text{(10)}$$
(2) $\omega_k$ minimization:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{U}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{U}_k(\omega)|^2 d\omega}$$

(11)

where $\hat{U}_i(\omega)$, $\hat{U}_{k}^{n+1}$, $\hat{f}(\omega)$, and $\hat{\lambda}(\omega)$ indicate the Fourier transform of $U_i(\omega)$, $U_k^{n+1}$, $f(\omega)$, and $\lambda(\omega)$, respectively, and $n$ is the iteration number.

Development of the DPA-based models

Hydro-climatic series often include many IMFs with different frequencies and exhibit complex nonlinear characteristics (Rezaie-Balf et al. 2019; Torabi et al. 2020). So, it is difficult for a single model to predict WQPs accurately. Signal decomposition algorithms can decompose hydrological time series into a set of relatively stable sub-series and reduce the modeling difficulty. Hence, sub-series decomposed by DPAs, EMD, EEMD, and VMD only include one or a similar scale of hydrological variables and are easier to be predicted. EMD, EEMD, and VMD are first used to decompose WQP time series into a set of sub-series, and then standalone MT is applied to build a proper model for the prediction of each of the sub-series, according to its characteristics.

The preprocessing-based model (i.e., EMD-/EEMD-/VMD-MT) mainly includes the following steps. In the first step, the original data is divided into two parts, including the training part and the testing part. Secondly, EMD/EEMD/VMD procedures are utilized to decompose the original input and output observed time-series data $E(t)$ into several IMF components $C_j(t)$ $(i = 1, 2, 3, ..., n)$ and one residual component $r_n(t)$. In the next step, for each extracted IMF component and the residual component (e.g., IMF1), the MT model is established as a WQP-predicting tool to simulate the decomposed IMF and residual components and to calculate each component using the same sub-series (IMF1) of input variables. Finally, the predicted values of all extracted IMF and residual components using the MT models are aggregated to generate the TDS and EC, and then the prediction error is evaluated using the predicted dataset.

As the proposed algorithms are not model-adaptive in real applications, the best and optimal structure of hybrid models is chosen based on the problem. In other words, different problems need distinct optimal structures of WQP predictive models, and determining the structure of the network is an intellectual challenge for all researchers. As no solution was presented by the researchers to achieve the optimum values of the model’s parameters, this study employed the trial-and-error procedure to obtain the optimum user-defined parameters. Table 2 indicates design parameters of the MT, EEMD, and VMD algorithms for WQPs’ prediction. It is noticeable that if ‘Noise SD = 0’ and ‘Number of realization = 1’, then the EMD decomposition is obtained.

Model’s performance analysis

The acceptance and reliability of the proposed models need to be evaluated in order to assess the models’ performance. The evaluation criteria used in this study were the coefficient of determination ($R^2$), RMSE, mean absolute error

| Model | Design parameters | Unpruned | Number of decimal places |
|-------|-------------------|----------|--------------------------|
| MT    | Batch size        | Minimum number of instances | Unpruned | Number of decimal places |
|       | 100               | 4        | No                       | 2         |
| EEMD  | Noise SD          | Number of realization | Maximum number of iterations | –         |
|       | 0.1–0.4           | 200–500 | 2,000                    | –         |
| VMD   | Penalty           | Mode number | –                       | –         |
|       | 500–2,000         | 2–15     | –                        | –         |
(MAE), Nash–Sutcliffe efficiency ($E_{ns}$), and ratio of RMSE to SD (RSD), which can be formulated as follows:

$$R^2 = \left( \frac{\sum_{i=1}^{N} (O_i - \bar{O})(P_i - \bar{P})}{\left( \sum_{i=1}^{N} (O_i - \bar{O})^2 \right)^{1/2} \left( \sum_{i=1}^{N} (P_i - \bar{P})^2 \right)^{1/2}} \right)^2$$  \hfill (12)

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - O_i)^2}$$  \hfill (13)

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |O_i - P_i|$$  \hfill (14)

$$E_{ns} = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2}$$  \hfill (15)

$$\text{RSD} = \frac{\text{RMSE}}{\text{STDEV}_{obs}} = \sqrt{\frac{\sum_{i=1}^{N} (O_i - \bar{O})^2}{\sum_{i=1}^{N} (O_i - P_i)^2}}$$  \hfill (16)

where $N$ is the number of observations, $O_i$ and $P_i$ are the observed and predicted values of WQP, $\bar{O}$ and $\bar{P}$ are the mean of the observed and predicted data, and STDEV$_{obs}$ is the SD of the observed value of WQP.

**APPLICATION RESULTS AND DISCUSSION**

In this section, the performance of the proposed hybrid and standalone models in terms of accuracy and error will be investigated by several evaluation metrics and visual plots. Figure 3 shows eight decomposed time-series data of monthly TDS and EC by VMD for the Garmrood station. The original TDS and EC time-series data from January 1991 to December 2017 were transformed to a set of IMFs and a residue. Figure 4 illustrates the workflow of the present study step by step for the prediction of the WQPs at Tajan River.

**Model’s performance in terms of TDS estimation**

In this study, the capability of DPAs, EMD, EEMD, and VMD, integrated with the MT model, is investigated on monthly TDS estimation, and their performances are
compared with standalone MT. Accuracy and error estimation results of TDS for the proposed decomposition-based models are provided in Table 3 for both Garmrood and Varand stations at training (simulation) and testing (estimation) stages. It is apparent from the table that hybrid models provided better predictions for monthly TDS at both above-mentioned stations. In the training stage, the VMD-MT model with the lowest RMSE (18.739 mg/l), MAE (13.42 mg/l), and RSD (0.135) was compared with other hybrid and standalone MT models at the Garmrood station. In addition, the accuracy of the VMD-MT model (0.98 and 0.98) at this station is slightly better than EMD-MT (0.94 and 0.93) and EEMD-MT (0.96 and 0.96) in terms of $R^2$ and $E_{ns}$ respectively. For the Varand station at the training stage, the results are similar to Garmrood, and the MT model coupled with VMD had the best accuracy and the lowest error for TDS estimation.

It can be seen that there are distinct differences among all the estimating models, with the hybrid methods having the best performance with respect to five metrics in the testing stage as compared with the conventional MT model. In the testing stage, the results showed that models calibrated by the DPA and VMD-MT had invariably larger $R^2$ and $E_{ns}$ values for both Garmrood and Varand stations compared with their counterparts calibrated using EMD and EEMD algorithms. For example, VMD-MT outweighs EMD-MT, EEMD-MT, and MT with about 48, 33, and 41% for Garmrood and 64, 46, and 13% for Varand in the RMSE value, while the increases in the $E_{ns}$ value are by almost 16, 8, and 13% for Garmrood and 10%, less than 1%, and 5% for the Varand station. In Figure 5, the scatter plot and Pearson determination coefficients ($R^2$) for hybrid DPA-based models are compared for observed and estimated TDS values. It is apparent from the graphs that the VMD-MT has less scattered estimates compared with EMD-MT, EEMD-MT, and MT models.

Figure 6 shows the observed and simulated data obtained by different models in the training and testing phases. It can be observed that all the models were able to effectively approximate the target monthly TDS, demonstrating their strong trace-ability to the time-varying TDS. Besides, the simulated result of the hybrid VMD-MT method is better than the other hybrid and standalone methods because its trend line between the observed and predictive data is closer to the best one; the EEMD-MT performance is better than the EMD-MT model due to the higher determination coefficient; the MT performance is inferior to decomposition-based methods (EMD/EEMD/VMD-MT), demonstrating the positive role of DPA in TDS estimation.

### Table 3 | The performance results of predictive models in the estimation of monthly TDS for Garmrood and Varand

| Station | Model | $R^2$ | RMSE  | MAE   | $E_{ns}$ | RSD |
|---------|-------|-------|-------|-------|---------|-----|
|         |       | Training period |       |       |         |     |
| Garmrood | MT    | 0.89  | 50.085| 21.488| 0.86    | 0.363|
|         | EMD-MT| 0.94  | 33.864| 18.088| 0.93    | 0.245|
|         | EEMD-MT| 0.96 | 26.966| 19.025| 0.96    | 0.195|
|         | VMD-MT| 0.98  | 18.739| 13.426| 0.98    | 0.135|
|         |       | Testing period |       |       |         |     |
|         | MT    | 0.84  | 42.624| 21.737| 0.83    | 0.403|
|         | EMD-MT| 0.87  | 47.270| 34.461| 0.81    | 0.447|
|         | EEMD-MT| 0.91 | 36.646| 24.796| 0.87    | 0.346|
|         | VMD-MT| 0.96  | 24.412| 17.559| 0.94    | 0.231|
| Varand  | MT    | 0.91  | 78.662| 30.521| 0.91    | 0.303|
|         | EMD-MT| 0.93  | 67.356| 37.662| 0.93    | 0.262|
|         | EEMD-MT| 0.96 | 60.587| 47.688| 0.94    | 0.234|
|         | VMD-MT| 0.97  | 41.520| 20.645| 0.97    | 0.165|
|         |       | Testing period |       |       |         |     |
|         | MT    | 0.96  | 60.030| 45.796| 0.93    | 0.259|
|         | EMD-MT| 0.94  | 89.108| 75.799| 0.89    | 0.385|
|         | EEMD-MT| 0.98 | 36.912| 25.107| 0.98    | 0.137|
|         | VMD-MT| 0.99  | 31.857| 18.279| 0.98    | 0.133|

In terms of EC estimation, evaluation metrics $R^2$, RMSE, MAE, $E_{ns}$, and RSD were employed for EC estimation for
training and testing datasets at two aforementioned stations, Garmrood and Varand. According to the training results, among the proposed models, the integration of DPAs with the MT model caused an increase in the models’ accuracy for monthly EC estimation. On the other hand, the MT model, since coupled with VMD, provided the lowest values of RMSE (57.43 and 27.27 mg/l), MAE (28.11 and 18.15 mg/l), and RSD (0.14 and 0.12) for EC estimation at Garmrood and Varand stations, respectively. After that, EEMD-ELM with higher RMSE (64.65%) and the RSD (64.56%) presented better results than other models (see Table 4).

A similar trend is also observed for the testing stage. The VMD-MT model generally estimated monthly EC better than the MT, EMD-MT, and EEMD-MT; for instance, the decreases in RSD for the VMD-MT at Garmrood and
Figure 6 | Time-series plots of measured and estimated monthly TDS for Garmrood and Varand.

Table 4 | The performance results of predictive models in the estimation of monthly EC for Garmrood and Varand

| Performance metrics |
|----------------------|
| Station | Model | $R^2$ | RMSE | MAE | $E_{ns}$ | RSD |
|----------|-------|-------|------|-----|---------|-----|
| **Garmrood** |       |       |      |     |         |     |
|          | Training period |       |       |     |         |     |
|          | MT | 0.91 | 120.602 | 42.861 | 0.90 | 0.303 |
|          | EMD-MT | 0.93 | 98.621 | 54.624 | 0.93 | 0.247 |
|          | EEMD-MT | 0.97 | 88.429 | 69.646 | 0.95 | 0.222 |
|          | VMD-MT | 0.97 | 57.431 | 28.111 | 0.97 | 0.144 |
|          | Testing period |       |       |     |         |     |
|          | MT | 0.92 | 156.554 | 133.181 | 0.79 | 0.451 |
|          | EMD-MT | 0.96 | 84.954 | 65.786 | 0.93 | 0.244 |
|          | EEMD-MT | 0.98 | 40.281 | 22.294 | 0.98 | 0.116 |
|          | VMD-MT | 0.99 | 38.633 | 29.441 | 0.98 | 0.111 |
| **Varand** |       |       |      |     |         |     |
|          | Training period |       |       |     |         |     |
|          | MT | 0.93 | 62.411 | 26.357 | 0.91 | 0.293 |
|          | EMD-MT | 0.95 | 49.695 | 26.654 | 0.94 | 0.233 |
|          | EEMD-MT | 0.96 | 42.851 | 32.258 | 0.95 | 0.201 |
|          | VMD-MT | 0.98 | 27.271 | 18.151 | 0.98 | 0.128 |
|          | Testing period |       |       |     |         |     |
|          | MT | 0.88 | 53.705 | 36.226 | 0.86 | 0.346 |
|          | EMD-MT | 0.94 | 50.393 | 31.981 | 0.88 | 0.296 |
|          | EEMD-MT | 0.95 | 48.736 | 29.244 | 0.90 | 0.254 |
|          | VMD-MT | 0.97 | 32.801 | 25.546 | 0.95 | 0.211 |
Varand were 5.17 and 25.05%, respectively. Comparing empirical decomposition algorithms (EMD and EEMD), it should be stated that the EEMD-MT model had lower RMSE (40.281 mg/l for Garmrood and 48.801 mg/l for Varand) and higher $E_{ns}$ (0.98 for Garmrood and 0.90 for Varand).

Figure 7 illustrates the scatter graphs of the train and test results provided by the proposed standalone and DPA-based models for EC estimation. As clearly observed from the figure, the EEMD-MT and VMD-MT methods have less scattered estimates with higher $R^2$ and fit the line equation to the exact line ($y = x$) compared with those of the MT and EMD-MT. However, in the testing stage, it can be clearly observed that VMD DPA could successfully be integrated with the MT model in order to estimate EC parameter of the Tajan river.

Figure 8 demonstrates the line plots of observed versus forecasted monthly EC time series of the DPA hybrid MT
model and the standalone model of the train and test time period. Time-series plots of two stations depict the capabilities of both standalone and DPA-hybridized AI models in generating unbiased estimates of monthly EC time series, taking into account random and seasonal patterns. The VMD-MT model estimates appeared to almost catch/capture the cyclic, irregular, random, and extreme variations of the observed EC time series on a monthly scale. In addition, it can be seen from the Garmrood station that EEMD-based model almost overestimated EC and its performance was not reliable.

The VMD usage causes no residual noise in the modes, and it may avoid the redundant modes for decomposition purpose compared with the EMD and EEMD algorithms. Moreover, the VMD is an adaptive signal decomposition technique which nonrecursively decomposes a multicomponent signal into some mode functions of quasi-orthogonal IMFs (Liu et al. 2016). Theoretically, the VMD is the most promising solution for identifying complexity, nonlinearity, as well as irregular, random, and extreme variations of the observed WQP time-series data compared with EMD and EEMD algorithms. To sum up, the proposed VMD-MT method proves to be an effective tool to provide strong technical support for the monthly WQPs’ estimation in the Tajan basin.

The authors, according to the results and as the generalizability of the proposed hybrid models, recommend using decomposition algorithms for other WQPs’ assessment with the same scale of input/output parameters as well as watershed physical characteristics in order to assess the generalization of the hybrid proposed models. Furthermore, nonlinear and/or dynamic AI programming based on simulation models could be used to find the contributors to the WQPs in the river system; however, this type of assessment typically imposes a prohibitive computational burden, especially for the prediction of large and complex river systems.
Generally speaking, due to the correlations and interactions between WQPs, it is interesting to investigate whether a domain-specific mechanism governing observed patterns exists to prove the predictability of these variables. The identification of such forecasting models is particularly useful for ecologists and environmentalists, since they will be able to predict water pollution levels and take necessary precautionary measures in advance.

On the other hand, hydrological processes can be regarded as a nonlinear process in the nature. At this point, it can be said that some variables such as streamflow and precipitation at different places have time-dependent parameters. Therefore, it is required to describe, analyze, and interpret these nonlinear processes. So, it is crucial to overcome the complexity, nonlinearity, and random distribution of the hydrological datasets before using them to develop AI models. In this study, three DPAs were recruited to overcome these drawbacks. Therefore, a hybrid model that needs no more information about parameters and removes the noise trends of the WQPs datasets can be fruitful.

SUMMARY AND CONCLUSION

Over the past few decades, the approaches for improving hydrological estimating accuracy have received significant attention from scholars and engineers in the water resources field. To enrich this theory, the application of three DPAs, EMD, EEMD, and VMD, was considered to overcome time-series data: nonstationarity, nonlinearity, and complexity. To develop an ensemble hybrid model to estimate two crucial factors of water quality, namely TDS and EC, the MT technique was applied. Two stations, i.e., Garmrood and Varand, located at Tajan River for the time period of 1991–2007 were collected to test the performance of the developed models. The results revealed that the proposed DPA-based models can provide better estimating results than the standalone MT model with respect to five statistical indexes. For TDS and EC, compared with the standalone MT model at Garmrood, the VMD-MT model could enhance prediction accuracy by 14 and 7%, respectively. For Varand station, this metric was enhanced by 10 and 3% for TDS and EC parameters, respectively. Among these DPAs, VMD was a promising alternative for improving results of estimation for both TDS and EC on a monthly scale. However, further studies might be required to improve the proposed models by introducing more input variables that have not been investigated due to the lack of such information. In addition, presenting an advanced optimizer to be incorporated with the data-driven models might lead to more improvement in predicting WQPs.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Aazami, J., Sari, A. E., Abdoli, A., Sohrabi, H. & Van den Brink, P. J. 2015 Assessment of ecological quality of the Tajan River in Iran using a multimetric macroinvertebrate index and species traits. Environmental Management 56 (1), 260–269.

Ahyan, H. Z. 2014 Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering 12 (1), 40.

Ahmadi-Mamaqani, Y., Khorasani, N., Talebi, K., Hashemi, S. H., Rafee, G. & Bahadori-Khosroshahi, F. 2011 Diazinon fate and toxicity in the Tajan River (Iran) ecosystem. Environmental Engineering Science 28 (12), 859–868.

Ahmed, A. M. & Shah, S. M. A. 2017 Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. Journal of King Saud University-Engineering Sciences 29 (3), 237–243.

Alizadeh, M. J. & Kavianpour, M. R. 2015 Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean. Marine Pollution Bulletin 98 (1–2), 171–178.

Asadi, E., Isazadeh, M., Samadianfard, S., Ramli, M. F., Mosavi, A., Nabipour, N. & Chau, K. W. 2020 Groundwater quality assessment for sustainable drinking and irrigation. Sustainability 12 (1), 177.

Asadollahfard, G., Zangooei, H., Aria, S. H. & Danesh, E. 2017 Application of artificial neural networks to predict total dissolved solids at the Karaj Dam. Environmental Quality Management 26 (3), 55–72.

Bhattacharya, B. & Solomatine, D. P. 2005 Neural networks and M5 model trees in modelling water level – discharge relationship. Neurocomputing 63, 381–396.
Chen, H., Xu, L., Ai, W., Lin, B., Feng, Q. & Cai, K. 2020 Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy. *Science of The Total Environment* 714, 136765.

Dragomiretskiy, K. & Zosso, D. 2014 Variational mode decomposition. *IEEE Transactions on Signal Processing* 62 (3), 531–544.

Ghavidel, S. Z. Z. & Montaseri, M. 2014 Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. *Stochastic Environmental Research and Risk Assessment* 28 (8), 2101–2118.

Hong, H., Panahi, M., Shirzadi, A., Ma, T., Liu, J., Zhu, A. X. & … Kazakia, N. 2018 Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Science of The Total Environment* 621, 1124–1141.

Hosseini, F. S., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H. & Haghighi, A. T. 2020 Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: application of the simulated annealing feature selection method. *Science of the Total Environment* 711, 135161.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q. & … 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 (1971), 903–995.

Huang, N. E., Wu, M. L. C., Long, S. R., Shen, S. S., Qu, W., Gloersen, P. & Fan, K. L. 2005 A confidence limit for the empirical mode decomposition and Hilbert spectral analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 459 (2037), 2317–2345.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L. & Wei, Y. 2013 A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Mathematical and Computer Modelling* 58 (3–4), 458–465.

Liu, Y., Yang, G., Li, M. & Yin, H. 2016 Variational mode decomposition denoising combined the detrended fluctuation analysis. *Signal Processing* 125, 349–364.

Liu, H., Duan, Z., Han, F. Z. & Li, Y. F. 2018 Big multi-step wind speed forecasting model based on secondary decomposition, ensemble method and error correction algorithm. *Energy Conversion and Management* 156, 525–541.

Mohammadpour, R., Shaharuddin, S., Zakaria, N. A., Ghani, A. A., Vakili, M. & Chan, N. W. 2016 Prediction of water quality index in free surface constructed wetlands. *Environmental Earth Sciences* 75 (2), 139.

Montaseri, M., Ghavidel, S. Z. Z. & Sanikhani, H. 2018 Water quality variations in different climates of Iran: toward modeling total dissolved solid using soft computing techniques. *Stochastic Environmental Research and Risk Assessment* 32 (8), 2253–2273.

Mouatadid, S., Raj, N., Deo, R. C. & Adamowski, J. F. 2018 Input selection and data-driven model performance optimization to predict the Standardized Precipitation and Evaporation Index in a drought-prone region. *Atmospheric Research* 212, 130–149.

Najafzadeh, M., Rezaie-Balf, M. & Rashedi, E. 2016 Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models. *Journal of Hydroinformatics* 18 (5), 867–884.

Najafzadeh, M., Ghaemi, A. & Emamgholizadeh, S. 2019 Prediction of water quality parameters using evolutionary computing-based formulations. *International Journal of Environmental Science and Technology* 16 (10), 6577–6596.

Napolitano, G., Serinaldi, F. & See, L. 2011 Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: an empirical examination. *Journal of Hydrology* 406 (3–4), 199–214.

Olson, J. R. & Hawkins, C. P. 2017 Effects of total dissolved solids on growth and mortality predict distributions of stream macroinvertebrates. *Freshwater Biology* 62 (4), 779–791.

Pan, C., Ng, K. T. W., Fallah, B. & Richter, A. 2019 Evaluation of the bias and precision of regression techniques and machine learning approaches in total dissolved solids modeling of an urban aquifer. *Environmental Science and Pollution Research* 26 (2), 1821–1833.

Prasad, R., Deo, R. C., Li, Y. & Maraseni, T. 2018 Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. *Soil and Tillage Research* 181, 63–81.

Quinlan, J. R. 1992 Learning with continuous classes. In: *Proceedings of AI’92* (Sterling Adams, ed.). World Scientific, pp. 343–348.

Rezaie-Balf, M. & Kisi, O. 2018 New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine. *Hydrology Research* 49 (3), 939–953.

Rezaie-Balf, M., Naganna, S. R., Ghaemi, A. & Deka, P. C. 2017 Wavelet coupled MARS and M5 model tree approaches for groundwater level forecasting. *Journal of Hydrology* 553, 356–373.

Rezaie-Balf, M., Kim, S., Fallah, H. & Alaghmand, S. 2019 Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: application on the perennial rivers in Iran and South Korea. *Journal of Hydrology* 572, 470–485.

Sakizadeh, M. 2019 Spatial analysis of total dissolved solids in Dezful Aquifer: comparison between universal and fixed rank kriging. *Journal of Contaminant Hydrology* 221, 26–34.

Shiri, J., Makarynskyy, O., Kisi, O., Dierickx, W. & Fard, A. F. 2017 Evaluation of short-term operational water levels using an adaptive neuro-fuzzy inference system. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 137 (6), 344–354.

Singh, K. P., Basant, N. & Gupta, S. 2011 Support vector machines in water quality management. *Analytica Chimica Acta* 703 (2), 152–162.

Solomatine, D. P. & Dulal, K. N. 2003 Model trees as an alternative to neural networks in rainfall – runoff modelling. *Hydrological Sciences Journal* 48 (3), 599–611.
Solomatine, D. P. & Xue, Y. 2003 M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. Journal of Hydrologic Engineering 9 (6), 491–501.

Tan, G., Yan, J., Gao, C. & Yang, S. 2012 Prediction of water quality time series data based on least squares support vector machine. Procedia Engineering 31, 1194–1199.

Torabi, S. A., Mastouri, R. & Najarchi, M. 2020 Daily flow forecasting of perennial rivers in an arid watershed: a hybrid ensemble decomposition approach integrated with computational intelligence techniques. Journal of Water Supply: Research and Technology-Aqua.

Wang, Y. & Witten, I. H. 1997 Induction of model trees for predicting continuous lasses. In: Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague.

Wang, Y., Markert, R., Xiang, J. & Zheng, W. 2015 Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system. Mechanical Systems and Signal Processing 60, 243–251.

Wen, X., Feng, Q., Deo, R. C., Wu, M., Yin, Z., Yang, L. & Singh, V. P. 2019 Two-phase extreme learning machines integrated with complete ensemble empirical mode decomposition with adaptive noise for multi-scale runoff prediction. Journal of Hydrology.

Wu, Z. & Huang, N. E. 2009 Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis 1 (01), 1–41.

Yassin, M. A., Alazba, A. A. & Mattar, M. A. 2016 Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. Agricultural Water Management 163, 110–124.

Yeh, J. R., Shieh, J. S. & Huang, N. E. 2010 Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. Advances in Adaptive Data Analysis 2 (02), 135–156.

Zhang, W., Qu, Z., Zhang, K., Mao, W., Ma, Y. & Fan, X. 2017 A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. Energy Conversion and Management 136, 439–451.

First received 5 September 2020; accepted in revised form 30 December 2020. Available online 25 January 2021