Artificial Intelligence Review (2024) 57:85
https://doi.org/10.1007/s10462-024-10707-4

New formulation for predicting total dissolved gas supersaturation in dam reservoir: application of hybrid artificial intelligence models based on multiple signal decomposition

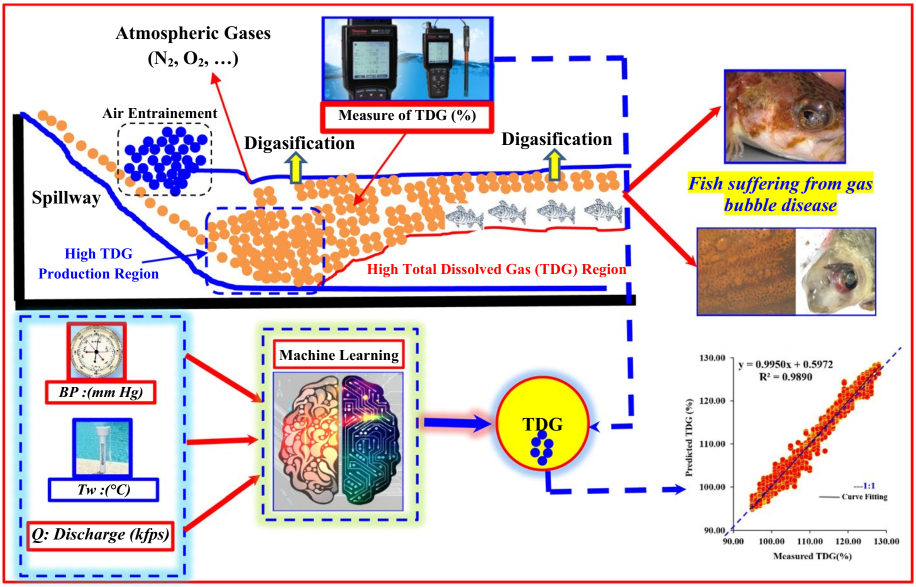
Salim Heddam · Ahmed M. Al-Areeq · Mou Leong Tan · Iman Ahmadianfar · Bijay Halder · Vahdettin Demir, et al. ***[full author details at the end of the article]***

Accepted: 18 January 2024 / Published online: 9 March 2024

Abstract

Total dissolved gas (TDG) concentration plays an important role in the control of the aquatic life. Elevated TDG can cause gas-bubble trauma in fish (GBT). Therefore, controlling TDG fluctuation has become of great importance for different disciplines of surface water environmental engineering.. Nowadays, direct estimation of TDG is expensive and time-consuming. Hence, this work proposes a new modelling framework for predicting TDG based on the integration of machine learning (ML) models and multiresolution signal decomposition. The proposed ML models were trained and validated using hourly data obtained from four stations at the United States Geological Survey. The dataset are composed from: (*i*) water temperature (T_w), (*ii*) barometric pressure (*BP*), and (*iii*) discharge (*Q*), which were used as the input variables for TDG prediction. The modelling strategy is conducted based on two different steps. First, six singles ML model namely: (*i*) multilayer perceptron neural network, (*ii*) Gaussian process regression, (*iii*) random forest regression, (*iv*) random vector functional link, (*v*) adaptive boosting, and (*vi*) Bootstrap aggregating (Bagging), were developed for predicting TDG using T_w, *BP*, and *Q*, and their performances were compared. Second, a new framework was introduced based on the combination of empirical mode decomposition (EMD), the variational mode decomposition (VMD), and the empirical wavelet transform (EWT) preprocessing signal decomposition algorithms with ML models for building new hybrid ML models. Hence, the T_w, *BP*, and *Q* signals were decomposed to extract the intrinsic mode functions (IMFs) by using the EMD and VMD methods and the multiresolution analysis (MRA) components by using the EWT method. Then after, the IMFs and MRA components were selected and regraded as new input variables for the ML models and used as an integral part thereof. The single and hybrid prediction models were compared using several statistical metrics namely, root mean square error, mean absolute error, coefficient of determination (R^2), and Nash–Sutcliffe efficiency (NSE). The single and hybrid models were trained several times with high number of repetitions, depending on the kind of modeling process. The obtained results using single models gave good agreement between the predicted TDG and the situ measured dataset. Overall, the Bagging model performed better than the other five models with R^2 and NSE values of 0.906 and 0.902, respectively. However, the extracted IMFs and MRA components using the EMD, VMD and the EWT have contributed to an improvement of the hybrid models' performances, for which the R^2 and NSE were significantly increased reaching the values of 0.996 and 0.995. Experimental results showed the superiority of hybrid models and more importantly the importance of signal decomposition in improving the predictive accuracy of TDG.

Graphical abstract



Keywords GBT · Hybrid machine learning · TDG supersaturation · EMD · VMD · EWT

Abbreviations

ANN	Artificial neural network
AdaBoost	Adaptive boosting
BP	Barometric pressure
Bagging	Bootstrap aggregating
CO_2	Dioxide de carbone
CART	Classification and regression tree
C_v	Coefficient of variation
DENFIS	Dynamic evolving neural-fuzzy inference system
DT	Decision tree
EWT	Empirical wavelet transform
ELM	Extreme Learning Machine
EMD	Empirical mode decomposition
GA	Genetic algorithm
GRNN	Generalized regression neural network
GPR	Gaussian process regression
GBT	Gas-bubble trauma
H-RSM	High-order response surface method (H-RSM)
IMF	Intrinsic mode functions
KIM	Kriging interpolation method

Kcfs	Thousands cubic foot by second
LSSVM	Least squares support vector machine
M5Tree	M5 model tree
MARS	Multivariate adaptive regression Spline
ML	Machine Learning
MLPNN	Multilayer perceptron neural network
MAE	Mean absolute error
MLPNN	Multilayer perceptron neural network
MRA	Multiresolution analysis
NSE	Nash–Sutcliffe efficiency
N2	Azote
O2	Oxygen
PC-ELM	Parallel chaos search based incremental extreme learning machine
Q	Discharge
RSM	Response surface method
RFR	Random forest regression
RVFL	Random vector functional link
RMSE	Root mean square error
R	Correlation coefficient
S_x	Standard deviation
SVR	Support vector regression
TDG	Total dissolved gas
Tw	Water temperature
X_{mean}	Mean value
X_{max}	Maximal value
X_{min}	Minimal value
USGS	United States Geological Survey
VMD	Variational mode decomposition

1 Introduction

Nowadays, high dam’s reservoir has become more numerous and play an important role for the production of hydropower energy (Qin et al. 2022). Another important role of high dams is the creation of artificial flood over the spillways of hydropower stations (Wang et al. 2019b), which is the major cause of the formation of total dissolved gas (TDG) along the river downstream of high dams (Yuan et al. 2018; Huang et al. 2021). TDG can be defined as the quantity of dissolved air available in water, and it is considered a natural phenomenon resulting from the interaction between air and water near surface water (Li et al. 2023a). From a computational point of view, the formation of TDG (i.e., N_2, O_2 and CO_2) persisted if pressure of TDG become superior to the atmospheric pressure (Yuan et al. 2022). Elevated TDG concentration affects the dynamics of aquatic life (Cheng et al. 2021), and thus an important ecological indicator that can be utilized for the evaluation of the state of the aquatic life downstream of high dam’s reservoirs (Yuan et al. 2023). However, water managers and decision makers are conducting an ongoing risk assessment of the elevated TDG concentration leading to the conclusion that; TDG supersaturation should not be exceeded the amount of 110% of saturation. If not, this will lead to many serious problems designated as the ‘‘*gas-bubble trauma in fish*’’, i.e., “GBT”, and a downward correction of the elevated concentration become imminent (Zeng et al. 2020). Downstream of the high dam's reservoirs, fish suffer from elevated concentration of TDG. Among other ecological indicators that can help in increasing TDG concentration, water temperature, barometric pressure and discharge are particularly important as they influence the fluctuation of TDG (Li et al. 2022; Chen et al. 2023). Furthermore, due to the limited in situ sampling sites for controlling TDG, knowledge on the formation, fluctuation, and spatiotemporal distribution of TDG along the rivers downstream of high dams is relatively sparse, except for some locations in USA and China. Nevertheless, at Columbia and Snake rivers, USA, increased in situ stations and continuous monitoring of TDG have enabled researchers to gain insight about formation and behaviors of TDG supersaturation.

The literature revealed many researchers have done several works on TDG concentration prediction (Li et al. 2009, 2022; Lu et al. 2019; Zhang et al. 2023). More precisely, previous investigations were mainly focused on understanding two important phenomena: (*i*) how TDG supersaturation is generated near downstream of dams, and (*ii*) how TDG can be dissipated to avoid the GBT (Ma et al. 2019). Furthermore, for achieving these two objectives, available research papers were focalized and mainly oriented toward the application of numerical models and simulation approaches. Understanding the fluctuation, dissipation and the overall behavior of TDG is of high importance for the protection and control of the fish and many other ecological phenomena's in rivers (Yuan et al. 2018). The complexity of TDG supersaturation formation depends mainly on the environmental factors responsible for the increase of the TDG rates, and recent recognition of the environmental causes are good steps towards a more comprehensive framework for the TDG simulation.

Nowadays, the role of numerical models seems to be crucial, and they significantly helped in improving our understanding of TDG formation and dissipation. Several numerical models have been developed and applied for predicting TDG dissipation (Lin et al. 2022). Among the proposed models, it is worth to point out: the 1-D unsteady TDG model proposed by (Ma et al. 2016), the depth averaged 2-D for the dissipation of TDG (Shen et al. 2016), the numerical 2-DTDG model for analyzing the link between water temperature and TDG (Feng et al. 2013), a basic prediction water renewal model for TDG proposed by (Peng et al. 2022a), the two phases flow model proposed by (Wang et al. 2019a). Several other models are available in the literature. For example; a study proposed a mass transfer for modelling TDG using water depth as predictor, and the unsteady 3D two-phase (Politano et al. 2007, 2009, 2012, 2017). Despite the benefits of numerical and fluid mechanic models for TDG simulation, they also have some disadvantages.

However, they have been faced to several particular difficulties mainly related to the high number of variables needed for calibrating the numerical models (Zhang et al. 2022). Certain TDG conditions can be ambiguous, hard to understand, and the environmental factors that contribute to their control and diagnosis cannot be accurately determined using single models. To solve the forgoing limitation, multi-resolution signal decomposition is an ensemble of methods that help to alleviate these underlying problems. Among the disadvantages of the numerical and fluid mechanic models, it can be highlighted: (*i*) the need of very large number of variables for model calibrations, (*ii*) the difficulties to use the calibrated models outside of the calibration sites, (*iii*) the need for a large dataset based on the laboratory experiment for which, the operator's intervention is required and the equipment shall be placed at the most adequate measuring point to ensure safe and timely access data.

Over the past few years, the exploration of newly developed modelling strategy based on the application of machine learning (ML) models has been established and successfully applied for modelling TDG concentration (Heddam 2017; AlOmar et al. 2020; Wang et al.

2022). Among the proposed models, extreme learning machine (ELM) and support vector regression models proposed by (AlOmar et al. 2020), ELM optimized genetic algorithm (GA-ELM) developed by (Wang and Sheng 2022), multilayer perceptron neural network (MLPNN) used by (Han et al. 2019), the generalized regression neural network (GRNN) proposed by (Heddam 2017). The parallel chaos search based incremental extreme learning machine (PC-ELM) developed by (Heddam 2023), dynamic evolving neural-fuzzy inference system (DENFIS) (Heddam and Kisi 2021), the kriging interpolation method (KIM) and response surface method (RSM) (Heddam and Kisi 2020), and the least squares support vector machine (LSSVM) used by (Keshtegar et al. 2019). All above reported models were successfully applied for modelling TDG and the obtained results were found to be very promising. However, previous ML approaches for TDG prediction have been faced to several particular difficulties mainly related to the high number of variables needed for calibrating the numerical models (Zhang et al. 2022). Certain TDG conditions can be ambiguous, hard to understand, and the environmental factors that contribute to their control and diagnosis cannot be accurately determined using single ML models. Multiresolution signal decomposition is an ensemble of methods that help to alleviate these underlying problems.

The aim of this paper is to continue the work already undertaken in the context of TDG prediction. Hence, this work seeks to propose new formulation for TDG estimation aiming at improving the prediction accuracy of TDG, and at the same time introducing a new modelling framework, that contributes to the improvement of standalone ML methods by reducing the error calculated between measured and predicted TDG data. For this purpose, we make use of different multiresolution signal decomposition techniques namely; the empirical mode decomposition (EMD), the variational mode decomposition (VMD), and the empirical wavelet transform (EWT) techniques. The basic steps involved in our new modelling framework can be summarized as follow. Initially, the EMD, VMD, and the EWT are performed on the available input variables. The *Tw*, *BP*, and *Q* input variables are then decomposed into various intrinsic mode functions (IMFs) by using the EMD and VMD methods and the multiresolution analysis (MRA) components by using the EWT method. After decomposition, the IMFs and the MRA were used as new input variables for the ML approaches. Compared to the single models, the use of the EMD, VMD, and EWT have the advantages of extracting and combining the multitude of nonlinear information available in the original signal, producing new input variables with greater useful information, making the establishment of an appropriate relationship between TDG and the *Tw*, *BP*, and *Q* more practical. For convenience, we refer to *Tw*, *BP*, and *Q* as the input variables for the single models, while the IMFs and MRA typically represent the input variables of the hybrid models.

2 Materials and methods

2.1 Study area and data

In this study, TDG (% saturation), water temperature (Tw: °C), barometric pressure (BP: mmHg), and discharge (Q: kcfs) data, were collected from four United States geological survey (USGS) stations. All data are available at (https://or.water.usgs.gov/cgi-bin/grapher/table_setup.pl). These stations were selected tacking into account continuous data availability. Supporting details characteristics of the stations are reported in Table 1. According

Table 1 Dataset presentations covering the period of the study

Description	USGS 14019240	USGS 13341000	USGS 14019220	USGS 13352950
River	Columbia River below Mcnary Dam near Umatilla, Oregon, USA	Clearwater River at Ahsahka Idaho, USA	Columbia River at Mcnary Dam lock near Umatilla, Oregon, USA	Lake Sacajawea Forebay at Ice Harbor Dam, Washington, USA
Latitude	45°56′00.96″	46°30′16″	45°56′29″	45°56′29″
Longitude	119°19′30.89″	116°19′10″	119°17′31″	119°17′31″
Begin date	01 January 2020	01 January 2020	26 Mars 2020	25 Mars 2020
End date	03 October 2022	03 October 2022	06 September 2022	01 September 2022
Total Pattern	23,908	23,908	9958	9821
Training (70%)	16,736	16,736	6970	6875
Validation (30%)	7172	7172	2988	2946

to Table 1, we can see that, the period of record varied between stations, and all available data have been used, however, the presence of the incomplete days with missing values in data records make the length of data varied between stations. All data were available at hourly time step and composed from TDG, Tw, BP, and Q. The study area and locations of in situ measured data are shown in Fig. 1, in which the different stations are represented by different colors. For each station, we split the dataset into training (70%) and validation (30%). Thus, TDG was predicted using three input variables namely, Tw, BP and Q, and all variables were standardized using the Z-score method (Eq. 1) by subtracting the mean and dividing by the standard deviation.

$$Z_n = \frac{x_n - x_m}{\sigma_x} \tag{1}$$

where: Z_n is the normalized value of the variable n; x_n is the measured value of the variable n; x_m and σ_x are the mean value and standard deviation of the variable x.

Table 2 lists a brief descriptive statistic for all data and for all stations. In the Table 2, the mean value (X_{mean}), the maximal value (X_{max}), the minimal value (X_{min}), the standard deviation (S_x), the coefficient of variation (C_v), and the coefficient of correlation calculated between TDG and the three input variables are summarized. Finally, it is noted that, TDG was predicted according to two different scenarios: (*i*) modelling TDG using three input variables (i.e., *Tw*, *BP* and *Q*), and (*ii*) the *Tw*, *BP* and *Q* variables were decomposed into several IMFs and MRA using the EMD, VMD and EWT.

2.2 Machine learning methods

In the present study, six ML models were developed for predicting TDG supersaturation namely: (*i*) multilayer perceptron neural network (MLPNN), (*ii*) Gaussian process regression (GPR), (*iii*) random forest regression (RFR), (*iv*) random vector functional link (RVFL), (*v*) adaptive boosting (AdaBoost), and (*vi*) Bootstrap aggregating (Bagg). The theoretical description of these models is given below.

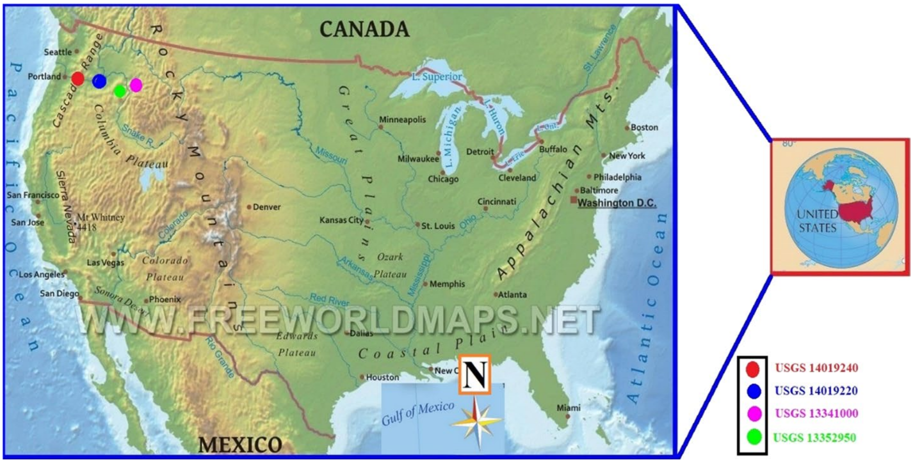


Fig. 1 The location of the studied sites “USGS stations” at the north-west of United States

Table 2 Summary statistics of total dissolved gas concentration and input variables

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R
USGS 14019240 Columbia River below Mcnary Dam near Umatilla, Oregon, USA								
TDG	Training	%	108.271	130.000	95.000	8.522	0.079	1.000
	Validation	%	108.246	128.000	95.000	8.557	0.079	1.000
	All data	%	108.264	130.000	95.000	8.533	0.079	1.000
T_w	Training	°C	12.248	22.400	2.300	6.263	0.511	0.380
	Validation	°C	12.095	22.400	2.300	6.255	0.517	0.396
	All data	°C	12.202	22.400	2.300	6.261	0.513	0.385
BP	Training	*mm Hg*	755.309	773.000	733.000	5.397	0.007	-0.328
	Validation	*mm Hg*	755.323	773.000	734.000	5.575	0.007	-0.354
	All data	*mm Hg*	755.313	773.000	733.000	5.451	0.007	-0.336
Q	Training	*kcfs*	169.403	472.000	54.000	74.106	0.437	0.746
	Validation	*kcfs*	169.413	471.000	53.900	73.806	0.436	0.747
	All data	*kcfs*	169.406	472.000	53.900	74.014	0.437	0.746
USGS 13341000 NF Clearwater River at Ahsahka Idaho, USA								
TDG	Training	%	100.893	122.000	92.000	3.685	0.037	1.000
	Validation	%	100.845	122.000	92.000	3.751	0.037	1.000
	All data	%	100.878	122.000	92.000	3.705	0.037	1.000
T_w	Training	°C	6.794	10.600	4.400	1.608	0.237	0.007
	Validation	°C	6.761	10.500	4.500	1.614	0.239	0.018
	All data	°C	6.784	10.600	4.400	1.609	0.237	0.353
BP	Training	*mm Hg*	736.238	752.000	719.000	5.276	0.007	-0.001
	Validation	*mm Hg*	736.222	752.000	719.000	5.258	0.007	0.001
	All data	*mm Hg*	736.233	752.000	719.000	5.270	0.007	-0.261
Q	Training	*kcfs*	5.090	25.200	1.300	3.950	0.776	0.001
	Validation	*kcfs*	5.127	25.200	1.300	3.976	0.775	-0.001
	All data	*kcfs*	5.101	25.200	1.300	3.958	0.776	0.036
USGS 14019220 Columbia River at Mcnary Dam lock near Umatilla, Oregon, USA								
TDG	Training	%	109.765	123.000	100.000	4.277	0.039	1.000
	Validation	%	109.694	122.000	100.000	4.312	0.039	1.000
	All data	%	109.744	123.000	100.000	4.290	0.039	1.000
T_w	Training	°C	15.308	22.800	5.600	4.853	0.317	0.009
	Validation	°C	15.438	22.700	5.600	4.828	0.313	-0.014
	All data	°C	15.347	22.800	5.600	4.847	0.316	0.002
BP	Training	*mm Hg*	751.239	767.000	733.000	3.968	0.005	-0.242
	Validation	*mm Hg*	751.255	767.000	734.000	3.986	0.005	-0.269
	All data	*mm Hg*	751.244	767.000	733.000	3.981	0.005	-0.250
Q	Training	*kcfs*	217.112	472.000	70.000	88.259	0.407	0.800
	Validation	*kcfs*	216.237	470.000	69.500	86.951	0.402	0.814
	All data	*kcfs*	216.849	472.000	69.500	87.905	0.405	0.804
USGS 13352950 Lake Sacajawea Forebay at Ice Harbor Dam, Washington, USA								
TDG	Training	%	113.388	126.000	100.000	5.832	0.051	1.000
	Validation	%	113.406	126.000	100.000	5.807	0.051	1.000
	All data	%	113.393	126.000	100.000	5.827	0.051	1.000

Table 2 (continued)

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R
T_w	Training	°*C*	15.553	22.800	5.600	5.093	0.327	-0.263
	Validation	°*C*	15.717	22.700	5.500	5.037	0.321	-0.289
	All data	°*C*	15.602	22.800	5.500	5.077	0.325	-0.271
BP	Training	*mm Hg*	748.616	764.000	732.000	3.896	0.005	-0.069
	Validation	*mm Hg*	748.584	764.000	732.000	3.916	0.005	-0.090
	All data	*mm Hg*	748.607	764.000	732.000	3.908	0.005	-0.075
Q	Training	*kcfs*	61.436	238.000	10.300	39.850	0.649	0.740
	Validation	*kcfs*	60.500	236.000	10.400	38.239	0.632	0.756
	All data	*kcfs*	61.155	238.000	10.300	39.415	0.645	0.744

X_{mean} mean, X_{max}, maximum, X_{min} minimum, S_x standard deviation, C_v coefficient of variation, *R* coefficient of correlation with *TDG*, *Tw* river water temperature, *TDG* total dissolved gas, *BP* Barometric pressure, *Q* discharge, *kcfs* thousand cubic foot by second

2.2.1 Multilayer perceptron neural network (MLPNN)

Artificial neural network (ANN) was inspired from the structure of the human brain. Thus, the basic component of the brain, i.e., the biological neurons were simulated to be artificial neurons arranged in several layers and dealing with specific tasks (Salman and Kadhum 2022). In the present study, the famous and well-known multi-layer perceptron neural network (MLPNN) was used for modelling total dissolved gas (TDG). The MLPNN includes several layers: (*i*) input layer with three input variables, i.e., *BP*, T_w and *Q*, each one designated as x_i, (*ii*) one or more hidden layers arranged in parallel and possess an ensemble of neurons, and (*iii*) on output layer with only one neuron, i.e., the TDG. Similar to the biological neuron, the information is disseminated rapidly from the input to the output layer using an ensemble of parameters called the weight factors and biases, which need to be updated during the training process. Consequently, we can summarize the mathematical formulas of the MLPNN as follow:

$$A_j = \left(\sum_{i=1}^{N} W_{ij} x_i\right) + b_j \tag{2}$$

where, A_j is the output of the hidden neuron *j*, W_{ij} is the weight linking the input variable x_i to the hidden neuron j, and b_j is bias of the hidden neuron *j*. Before passing the A_j to the next layer, the sigmoid activation function is used as follow:

$$k_j = f(A_j) = \mathrm{sigmoid}(A_j) = \frac{1}{1 + e^{-A_j}} \tag{3}$$

Finally, the output of the MLPNN is calculated as follow:

$$y = \left(\sum_{j=1}^{N} W_{jk} k_j\right) + b_o \tag{4}$$

where, y is the output of the MLPNN model, W_{jk} is the weight linking the hidden layer to the output layer, and b_0 is bias of the single output neuron. More details about the MLPNN can be found in large number of published papers.

2.2.2 Gaussian process regression (GPR)

The Gaussian Process Regression (GPR) proposed by (Williams and Rasmussen 2006) can be formulated as follow (Ouyang et al. 2022; Zhao et al. 2023):

$$\mathrm{f(x)} \sim \mathrm{GP}\left(\mathrm{m(x)}, \mathrm{k}\left(\mathrm{x}, \mathrm{x}'\right)\right) \tag{5}$$

In the above equation, x refers to the input vector; $m(x)$ refers to the mean function, while $\mathrm{k}\left(\mathrm{x}, \mathrm{x}'\right)$ corresponds to the covariance (kernel) function. For example, the Radial basis function (RBF) can be expressed as follow (He and Zhou 2022):

$$\mathrm{k_{SE}}\left(\mathrm{x_i}, \mathrm{x_j}\right) = \sigma_\mathrm{f}^2 \exp\left(-\frac{\left(\mathrm{x} - \mathrm{x}'\right)^2}{2\mathrm{l}^2}\right) \tag{6}$$

In the above equation, σ_f^2 is the variance of the dataset (i.e., the signal), and l, is the length scale of the uncertainty fluctuations. The GPR is derived from the standard linear model as follow:

$$\mathrm{y} = \mathrm{f(x)} + \epsilon \tag{7}$$

In the above equation, ε refers to the noise having a mean zero and variance σ_f^2 (He and Zhou 2022; Ouyang et al. 2022; Zhao et al. 2023).

2.2.3 Random forest regression (RFR)

Random forest regression (RFR) is an ensemble of classification and regression tree (CART) models built on a serie of trees with a training process made using the concept of Bagging by random smpaling with replacement, tacking into account the statbility and the imporvement in the terms of accuracy (Breiman 2001; Takoutsing and Heuvelink 2022). As the model is based on improving the performances of week learners and making a final decision using averaging or majority voting, the RFR use the ''*out-of-bag: OOB*'' for quantifying the calculted error and for ranking the variables in terms of importance using the permutation strategy. Building a RFR model can be achieved according to the follwing steps: (*i*) start by extracting an ensemble of subset randomly form the original dataet, (*ii*) growing a tree for each subset, (*iii*) repeat this until the constrution of k decision tree (DT), and (*iv*) the final response is than calculted based on averaging (AVG) the response of all DT (Zong and Zhang 2019; Giri et al. 2023). However, it is important to note that, RFR works only with two parameters: the total number of trees and the number of predictors at each node of the subset (Weiqi et al. 2022). Furthermore, the remaining paprt of the data not included in the subset, i.e., the OOB is used for chacking the regression or the classificiation performances (Zong and Zhang 2019). Finally, we can summarize the adavatage of the RFR as follow: (*i*) can not be affected by the nonlinearity between variables, (*ii*) high capability for avoinding over fitting problem, (*iii*) the gerated trees are uncorrelated, and (*iv*) the predicotors can be ranked in terms of their contribution (Chong et al. 2019).

2.2.4 Random vector functional link (*RVFL*)

The random vector functional link neural network (RVFL) can be viewed as an improved version of the original single layer neural network with an important difference (Pao et al. 1992; Pao et al. 1994): in addition to the transition from the input to the output layer through the hidden layer, there is a direct link between the input and output layer. Regarding the model parameters, the weights between the input neurons and the hidden neurons (also called enhancements neurons) were randomly assigned and remain unchangeable during the training process (W_{ij}), while the remaining weight (W_{jk}) should be updated during the training of the model. Given a set of data point (x_i, $i=1, \ldots, N$) with the corresponding output (y_i), the response of the enhancements, neurons can be calculated as follow (δ_j):

$$\delta_j\left[W_j x_i + \beta_j\right] = \frac{1}{1 + e^{-(W_{ij}x_i+\beta_j)}}, \beta_j \in [0, S], W_j \in [-S, +S], j = 1, 2, \ldots, N_h \tag{8}$$

In Eq. 8, *Wj* refers to the weight between the input and the hidden neurons, β_j is the bias of the hidden neuron *j*, and *S* refers to the scale factor determined during the training process. The final output can be calculated as follow (Jiao et al. 2023; Nabih et al. 2023):

$$Y = Bw, w \in R^{N+P}, \text{and } B = \left[B_1 B_2\right] \tag{9}$$

In Eq. 9, B_1 corresponds to the input data, and B_2 corresponds to the output of the enhancements neurons. Finally, the weight *w* is calculated using the Moore–Penrose pseudo inverse as follow (Jiao et al. 2023; Nabih et al. 2023):

$$w = B^{\dagger}Z \tag{10}$$

More details about the RVFL can be found in (Pao et al. 1992; Pao et al. 1994; Jiao et al. 2023; Nabih et al. 2023).

2.2.5 Adaptive boosting (AdaBoost)

The adaptive Boosting (AdaBoost) developed by (Freund and Schapire 1997), is one of the most widely reported ensemble ML method. Similar to the ensemble models, AdaBoost is a series of weaker learners (i.e., trees) and each one take into account one subset. An important point to note is that, the data that are hardly predicted (i.e., complex) should weighted stronger, and during the training process of the AdaBoost, the weights are being reshuffled and consequently they are increased for the poor learned dataset: this is a ‘‘sequential procedure’’ (Truong et al. 2022; Saha et al. 2023; Zhu et al. 2023). This process is very important as it helps the poor learners improve their performances during the training and at the end, we can obtain a robust model by weighted averaging of the response of poor regressor. From a mathematical point of view, the AdaBoost model can be expressed as follow (Truong et al. 2022):

$$\delta_M(x) = \sum_{m=1}^{M} \varphi_m(x) \tag{11}$$

where, $\delta_M(x)$ is the form training of the AdaBoost model, M corresponds to the number of iterations, $\varphi_m(x)$ is the weak learner (Truong et al. 2022; Saha et al. 2023; Zhu et al. 2023).

2.2.6 Bootstrap aggregating (Bagg)

The bootstrap aggregating, i.e., Bagging (Bagg) is used for combining an ensemble of weak learners for composing a strong model which can help in decreasing the variance and avoiding the overfitting (Li et al. 2023b). However, an important point to note is, the Bagging model should be handled as being a suite of multiple similar learners arranged in parallel, and the final response is calculated as the average (Khozani et al. 2019). For simplicity, the initial dataset is divided into an ensemble of sub-training set using bootstrap sampling with replacement, this cause to have some sample available multiple time in the training set while other none, which involve looking to repeat this several times. From a mathematical point of view, if the training dataset is designated as δ and composed of an ensemble of pairs as follow (Nancy Jane et al. 2023; Sun 2023):

$$\delta = \left[\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_m, y_m\right)\right] \tag{12}$$

Each individual subset is affected by an equal (1/M) weight, and consequently a specific weak learner is attributed to one subset, and based on this the error is calculated. The procedure of updating the weight is then started. If the example is correctly predicted or classified, its weight is reduced and vice versa, until the end of the training. The final response is calculated as the average (AVG) (Nancy Jane et al. 2023; Sun 2023).

2.3 Signal decomposition methods

In the present study, three signal decomposition algorithms were used, namely, the variational mode decomposition (VMD) (Dragomiretskiy and Zosso 2014), the empirical mode decomposition (EMD) (Huang et al. 1998), and the empirical wavelet transform (EWT) (Gilles 2013). The three algorithms were used for decomposing the input variables, i.e., T_w, *BP*, and *Q*, into several subcomponents. Thus, in EMD and VMD, the components are called intrinsic mode functions (IMF), while the components of the EWT are called multiresolution analysis (MRA) components (Bokde et al. 2020). An example of *BP* decomposition is provided in Fig. 2. According to Fig. 2, the BP signal depicted at the top of the Figure was decomposed into nine IMFs using the VMD and EMD, and nine MRA using the EWT. Furthermore, the extracted subcomponent were arranged from high frequency to low frequency. The number of extracted IMF is determined by trial and error, and in our present study, nine subcomponent was found to be sufficient and their aggregation have helped in providing excellent predictive accuracies. This process of decomposition make a very complicated signal very simpler (Rezaie-Balf et al. 2020). When the whole process of decomposition is finished, the obtained IMFs and MAR were aggregated and used as input variables of the models. The example trend shown in Fig. 3 is considered to demonstrate the decomposition process. If we consider that each input variable, i.e., the *BP*, T_w, and *Q* was decomposed into nine subcomponents, then, in total twenty-seven new input variables are used by the ML models. Theoretical description of these algorithms is given below. The flowchart of the overall modelling framework is depicted in Fig. 4.

According to Fig. 4, our investigation were oriented toward a deeply comparison between standalone single models and hybrid models. Thus, our study originality can be summarized as follow:

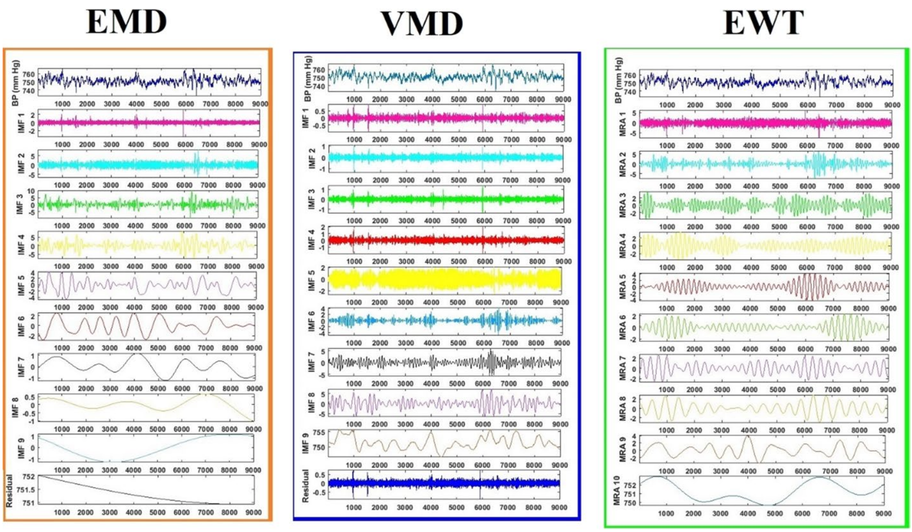


Fig. 2 Barometric pressure (BP) signal decomposition using the EMD, VMD, and EWT

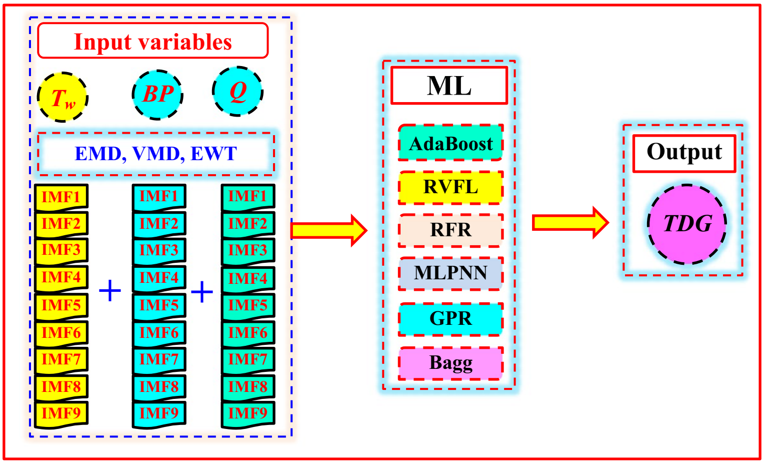


Fig. 3 The schematic diagram of the preprocessing signal decomposition with the aggregation of the intrinsic mode function (IMFs) with machine learning approaches

- To the best of the author's knowledge, this is the first study in the literature focused on the application of machine leaning combined with signal decomposition for modelling TDG in river.
- The primary objective of the present paper is to demonstrate whether signal decomposition can be presented as a good and robust tool for predicting TDG.
- Fewer variables are used for modelling TDG, i.e., T_w, *BP*, and *Q*.
- Three robust signal decomposition are selected and compared, i.e., VMD, EMD and EWT.

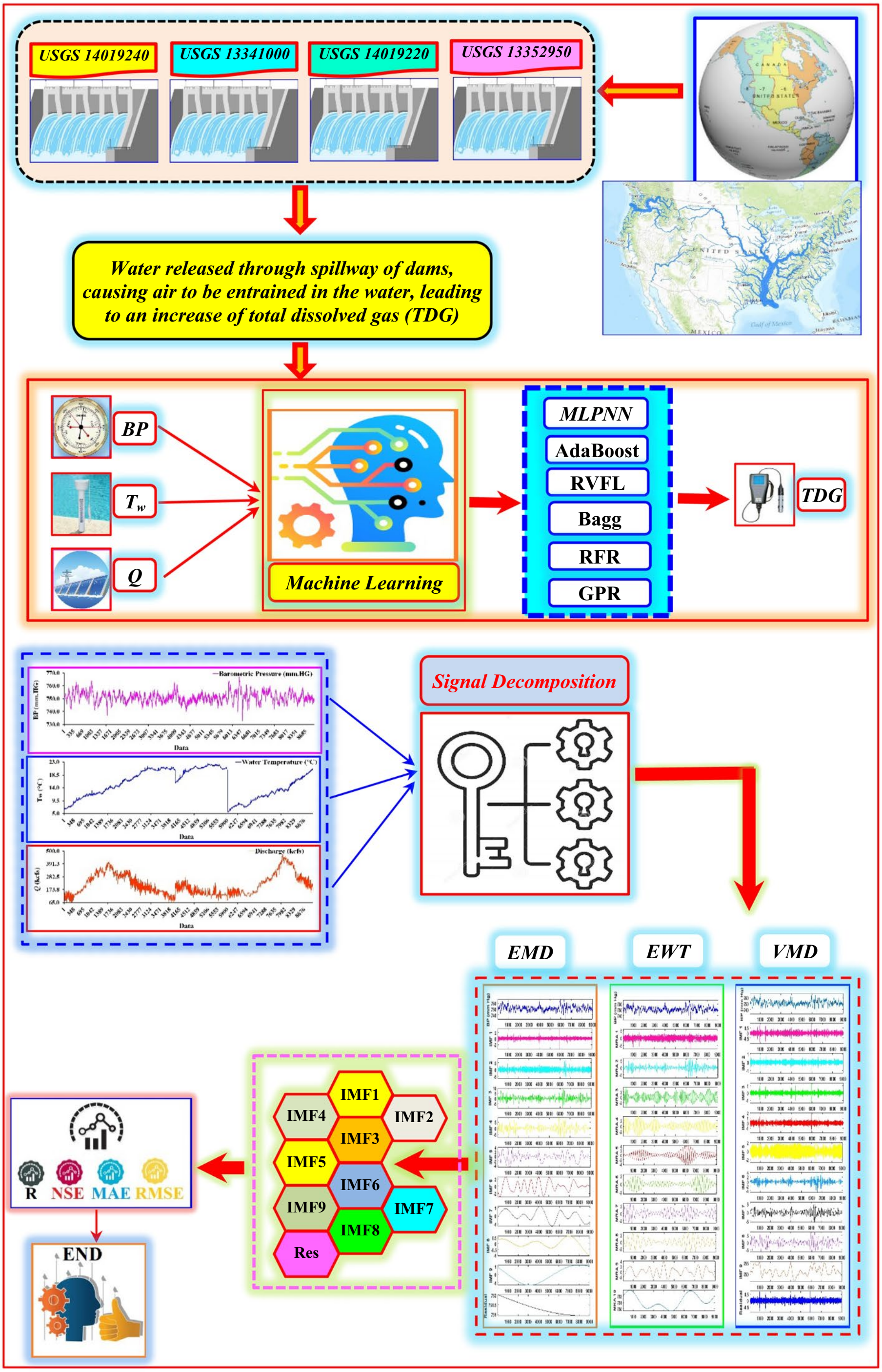


Fig. 4 Flowchart of the proposed modelling framework

2.3.1 Variational mode decomposition (VMD)

Variational mode decomposition (VMD) is a preprocessing signal decomposition developed by (Dragomiretskiy and Zosso 2014). The VMD is highly recognized by its capability to find the suitable nombre of modal decompositions tacking into account the actual situation (Qiao et al. 2022). The VMD is used for decomposing a specific signal into a series of intrinsic model function (IMF) components having a particular bandwidth and starting by the construction of a variational problem (Xiong et al. 2022). From a mathematical point of view, the VMD decomposes a signal $f(t)$ into M narrowband IMFs as follow:

$$f(t) = \sum_{m=1}^{M} u_m(t) \tag{13}$$

In the above equation, the IMFs have the following characteristics (Netsanet et al. 2022):

- Each IMF has an envelope $\varnothing_m(t)$ and phase $A_m(t)$ as follow:

$$u_m(t) = A_m(t)cos\left(\varnothing_m(t)\right) \tag{14}$$

- The envelopes of all IMF are positive and will progress more slowly.
- An instantaneous frequency $(\varnothing\prime_m(t))$ is attributed to each IMF and it should be concentrated around a central frequency (w_m).

More details about the VMD can be found in (Dragomiretskiy and Zosso 2014).

2.3.2 Empirical mode decomposition (EMD)

Empirical mode decomposition is a preprocessing signal developed by (Huang et al. 1998). The EMD is used for analyzing nonlinear and smooth signal, and it works by decomposing an original signal into a series of intrinsic mode function (IMF) components. Each component (i.e., each IMF) becomes himself a new signal with respect to a particular frequency, without requiring prior knowledge, and it works by supposing that: (*i*) signal and noise are ''uncorrelated'' and (*ii*) denoising the signal can be done by discarding lower order IMF signals (Li et al. 2023b). Suppose we have an original signal:

$$\mathrm{X_t = Signal(t) + Noise(t)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{IMF_k + r_t} \tag{15}$$

where K denotes the number of extracted *IMF* and r_t is the residual. The EMD works by applying a ''*sifting*'' process as follows (Li et al. 2023b; Shamaee and Mivehchy 2023):

- The mean envelope $\delta_{\mathrm{avg}}(\mathrm{t})$ is calculated as follow:

$$\delta_{\mathrm{avg}}(\mathrm{t}) = \frac{\delta_{\mathrm{max}}(\mathrm{t}) + \delta_{\mathrm{min}}(\mathrm{t})}{2} \tag{16}$$

  where $\delta_{\mathrm{max}}(\mathrm{t}), \delta_{\mathrm{min}}(\mathrm{t})$ are the maximum and minimum envelopes of the initial signal.
- Thus, the calculated mean envelope is then extracted from the original signal as follow:

$$\beta(\mathrm{t}) = \mathrm{X}(\mathrm{t}) - \delta_{\mathrm{avg}}(\mathrm{t}) \tag{17}$$

- Stop, control and check that the conditions are correctly satisfied. If ok, the first IMF is therefore obtained, if none, repeat:

$$\gamma(\mathrm{t}) = \beta(\mathrm{t}) \tag{18}$$

- The first residual is therefore calculated and considered to become the original signal ($X(t)$):

$$\mathrm{r}_1(\mathrm{t}) = \mathrm{X}(\mathrm{t}) - \gamma(\mathrm{t}) \tag{19}$$

This process of decomposition and extraction shall continue to grow until the function become a “*monotonic function*”. Finally, we obtain the function as follow:

$$\mathrm{x}(\mathrm{t}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \gamma(\mathrm{t}) + \mathrm{r}(\mathrm{t}) \tag{20}$$

More details about the EMD can be found in (Huang et al. 1998).

2.3.3 Empirical wavelet transform (EWT)

The empirical wavelet transform (EWT) was developed by (Gilles 2013). This algorithm of decomposition is based on ‘‘*revolutionary*’’ rather than the ‘‘*stochastic volatility*’’ available in the data (Karbasi et al. 2022). The EWT can be used as robust decomposition algorithm for any nonlinear and non-stationary signal because it can select the ideal value of frequency (Rout et al. 2022). Briefly, using the EWT we can obtain, and ensemble of subcomponents called the multiresolution analysis (MRA) components, to make a reasonable extraction efficiency. This can be achieved by dividing any signal *X (t)* into suite of MRA in the range of ‘‘*frequency*’’ domain, and keep by building wavelet ‘‘*band-pass*’’ filters for each sub-interval (Ren et al. 2022). From a mathematical point of view, the EWT extract the sub-components using two particular functions: the empirical wavelet functions ($\widehat{\gamma}_n(\omega)$) and the empirical scale function $\widehat{\theta}_n(\omega)$, expressed as follow (Peng et al. 2022b; Wang and Sheng 2022):

$$\widehat{\gamma}_n(\omega) = \begin{cases} 1 & if\ |\omega| \leq \omega_n - \sigma_n \\ cos\left[\frac{\pi}{2}\alpha\left(\frac{1}{2\sigma_n}|\omega| - \omega_n + \sigma_n\right)\right] & if\ \omega_n - \sigma_n \leq |\omega| \leq \omega_n + \sigma_n \\ 0 & otherwise \end{cases} \tag{21}$$

$$\widehat{\theta}_n(\omega) = \begin{cases} 1\ if \omega_n + \sigma_n \leq |\omega| \leq \omega_{n+1} - \sigma_{n-1} \\ cos\left[\frac{\pi}{2}\alpha\left(\frac{1}{2\sigma_{n+1}}|\omega| - \omega_{n+1} + \sigma_{n+1}\right)\right]\ if\ \omega_{n+1} - \sigma_{n+1} \leq |\omega| \leq \omega_{n+1} + \sigma_{n+1} \\ sin\left[\frac{\pi}{2}\alpha\left(\frac{1}{2\sigma_n}|\omega| - \omega_n + \sigma_n\right)\right]\ if\ \omega_n - \sigma_n \leq |\omega| \leq \omega_n + \sigma_n \\ 1\ otherwise \end{cases} \tag{22}$$

The function $\alpha(x) \in \mathrm{L}^{\mathrm{k}}$ ([0, 1]) is an arbitrary function and expressed as follow (Gilles 2013):

$$\alpha(x) = \begin{cases} 0\,x \le 0 \\ 1\,x \ge 1 \end{cases} and\alpha(x) + \alpha(1 - x) = 1 \forall x \in [0, 1] \tag{23}$$

where (ω) is the n-th maxima of the Fourier spectrum (Peng et al. 2022b; Ren et al. 2022; Wang and Sheng 2022). More details about EWT can be found in (Gilles 2013).

2.4 Performance assessment of the models

The performances of all models developed in the present study were evaluated using root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and Nash–Sutcliffe efficiency (NSE) (Yaseen 2021), calculated as follow:

$$RMSE = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}[(TDG_{obs})_i - (TDG_{est})_i]^2}, (0 \le RMSE < +\infty) \tag{24}$$

$$MAE = \frac{1}{N}\sum\nolimits_{i=1}^{N}|(TDG_{obs})_i - (TDG_{est})_i|, (0 \le MAE < +\infty) \tag{25}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N}\left(TDG_{obs,i} - \overline{TDG_{obs}}\right)\left(TDG_{est,i} - \overline{TDG_{est}}\right)}{\sqrt{\sum_{i=1}^{N}(TDG_{obs,i} - \overline{TDG_{obs}})^2 \sum_{i=1}^{N}(TDG_{est,i} - \overline{TDG_{est}})^2}}\right]^2 \tag{26}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}[TDG_{obs} - TDG_{est}]^2}{\sum_{i=1}^{N}[TDG_{obs} - \overline{TDG_{obs}}]^2}\right], (-\infty < NSE \le 1) \tag{27}$$

where $\overline{TDG_{obs}}$ and $\overline{TDG_{est}}$ are the mean measured and mean forecasted TDG, respectively, and TDG_{obs} and TDG_{est} specifies the observed and forecasted total dissolved gas for *ith* observations, and N shows the number of data points.

2.5 Models development

In this study, hybrid predictive models based on the EMD, VMD and EWT algorithms and ML methods are developed to predict the TDG in rivers. To build the models, different scenarios are considered. First, the six ML models, i.e., the MLPNN, RVFL, RFR, GPR, AdaBoost and Bagg were applied using three input variables (i.e., T_w, BP and Q). Second, the same models were combined with the EMD, VMD and EWT and further compared. The models' accuracies are assessed using model efficiency indices, including MAE, RMSE, R and NSE. Tables 3, 4, 5 and 6 show the performances of the developed models for the four stations. It is noteworthy that the models designated as MLPNN, RVFL, RFR, GPR, AdaBoost and Bagg correspond to the single models without decomposition. The MLPNN_EWT, RVFL_EWT, RFR_EWT, GPR_EWT, AdaBoost_EWT and Bagg_EWT correspond aux hybrid models based on the EWT signal decomposition. Similarly, the MLPNN_VMD, RVFL_VMD, RFR_VMD, GPR_VMD, AdaBoost_VMD and Bagg_VMD correspond aux hybrid models based on the VMD signal decomposition. Finally, the MLPNN_EMD, RVFL_EMD, RFR_EMD, GPR_EMD, AdaBoost_EMD and Bagg_EMD correspond aux hybrid models based on the EMD signal decomposition. Details of the obtained results are depicted and discussed hereafter, and only model's performances during the validation stage were highlighted and compared.

Table 3 Performances of different models at the USGS 14019240 station

Models	Training				Validation			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
Standalone models without decomposition								
MLPNN	0.861	0.862	3.167	2.281	0.861	0.862	3.182	2.319
AdaBoost	0.797	0.797	3.837	3.001	0.783	0.784	3.977	3.118
Bagg	0.947	0.945	2.005	1.408	0.903	0.900	2.699	1.911
GPR	0.867	0.867	3.105	2.205	0.863	0.864	3.158	2.275
RFR	0.947	0.944	2.024	1.427	0.901	0.900	2.706	1.922
RVFL	0.752	0.752	4.246	3.385	0.757	0.756	4.223	3.345
Models based on empirical wavelet transform (EWT)								
MLPNN_EWT	0.994	0.994	0.674	0.478	0.990	0.990	0.865	0.624
AdaBoost_EWT	0.976	0.975	1.336	0.987	0.953	0.952	1.871	1.351
Bagg_EWT	0.998	0.998	0.387	0.225	0.988	0.988	0.919	0.585
GPR_EWT	0.996	0.996	0.532	0.349	0.994	0.993	0.707	0.477
RFR_EWT	0.998	0.998	0.387	0.225	0.988	0.989	0.917	0.584
RVFL_EWT	0.794	0.793	3.876	3.171	0.794	0.795	3.876	3.163
Models based on variational mode decomposition (VMD)								
MLPNN_VMD	0.958	0.959	1.728	1.328	0.941	0.941	2.076	1.585
AdaBoost_VMD	0.901	0.900	2.693	2.072	0.872	0.872	3.055	2.394
Bagg_VMD	0.996	0.995	0.593	0.401	0.976	0.976	1.328	0.934
GPR_VMD	0.893	0.893	2.791	2.157	0.863	0.864	3.157	2.480
RFR_VMD	0.996	0.995	0.595	0.403	0.976	0.976	1.334	0.939
RVFL_VMD	0.729	0.729	4.438	3.656	0.731	0.731	4.435	3.639
Models based on empirical mode decomposition (EMD)								
MLPNN_EMD	0.992	0.993	0.738	0.519	0.988	0.989	0.900	0.653
AdaBoost_EMD	0.980	0.980	1.198	0.863	0.960	0.960	1.721	1.231
Bagg_EMD	0.998	0.998	0.397	0.229	0.986	0.985	1.044	0.655
GPR_EMD	0.994	0.993	0.711	0.477	0.990	0.990	0.875	0.598
RFR_EMD	0.998	0.998	0.396	0.229	0.986	0.986	1.030	0.650
RVFL_EMD	0.899	0.899	2.709	2.110	0.897	0.897	2.740	2.136

3 Experimental results

3.1 USGS 14019240 station modeling results

Table 3 presents the results of TDG prediction using all models at the USGS 14019240 station. The values stated in Table 3 clearly demonstrated that using single models, the RFR and Bagg achieved the high prediction efficiencies among all the developed models based on the four performances metrics (RMSE≈2.706, MAE≈1.922, R^2≈0.901, NSE≈0.900). In general, the two models MLPNN and GPR appear to have performed better than the RVFL and AdaBoost models based on the different efficiency indices and the differences between the two is negligible. However, it is worth nothing that the RVFL was the poorest one and they worked with a moderate degree of accuracy (RMSE≈4.223, MAE≈3.345, R^2≈0.757, and NSE≈0.756). In overall, obtained results reported in Table 3 confirm the good level of efficiency of the developed single ML models without including the signal decomposition algorithms and more importantly, using only fewer inputs variables, i.e., T_w, BP, and Q. Figure 5 displays the scatterplot of measured and predicted TDG data for the USGS 14019240 station. In accordance with the depicted data points, it is evident that the two models RFR and Bagg appear to predict TDG more accurately and with more precision than the other models, and this is reflected by the data, which are less scattered, while the RVLF model was characterized by a high-scattered data. According to Table 3 and Fig. 5 (i.e., the scatterplot), there is a remarkable improvement of the model's performances gained using the three signals decomposition algorithms. Using the EWT algorithm, it is clear that all models show their numerical performances improved by an increase of the R^2 and NSE values and a decrease of the RMSE and MAE indices. Using the R^2-values as a basis for comparison, it is clear that the MLPNN_EWT, Bagg_EWT, GPR_EWT, and RFR_EWT model's performances were similar with negligible difference, and in the case of Bagg_EWT, and RFR_EWT, the equality between the two is obvious. The comparison between single and hybrids models is further highlighted using Taylor diagram (Fig. 6), for which the superiority of the hybrid models is obvious.

Although the hybrid models exhibit a statistically similar performance (e.g., statistically similar values for the RMSE, MAE, R^2 and NSE), the GPR_EWT model seems to be slightly more accurate showing the biggest R^2 (≈0.994) and NSE (≈0.993) values, and the lowest RMSE (≈0.707) and MAE (≈0.477) values. Overall, the EWT algorithm helped in improving the performances of MLPNN_EWT, AdaBoost_EWT, Bagg_EWT, GPR_EWT, and RFR_EWT by ≈72.816%, ≈52.954%,≈65.950%,≈77.612%, and ≈66.112% in terms of RMSE metric, and by ≈73.092%, ≈56.671%,≈69.388%,≈79.033%, and ≈69.615% in terms of MAE metric, respectively. However, the improvement of the RVFL_EWT model is less than the other models and does not exceed the ratios of ≈8.217% and ≈5.441% in terms of RMSE and MAE metrics, respectively. Using the VMD algorithm as reported in Table 3 and Fig. 5 (i.e., the scatterplot), it is clear that, for all models, the improvements gained are less than those obtained from the EWT. Furthermore, using the VMD, it clear that, only the Bagg_VMD and the RFR_VMD have guaranteed great improvement (e.g., ≈50.797%, ≈50.702% for the RMSE, and ≈51.125%, ≈51.145% for the MAE). No improvement was recorded for the GPR_VMD model, while using the RVFL_VMD, a decrease in terms of models performances was recorded. More precisely, using the EWT algorithm, the performances of the RVFL model were decreased by ≈4.780% and ≈8.079% in terms of RMSE and MAE metrics, the performances of the GPR model

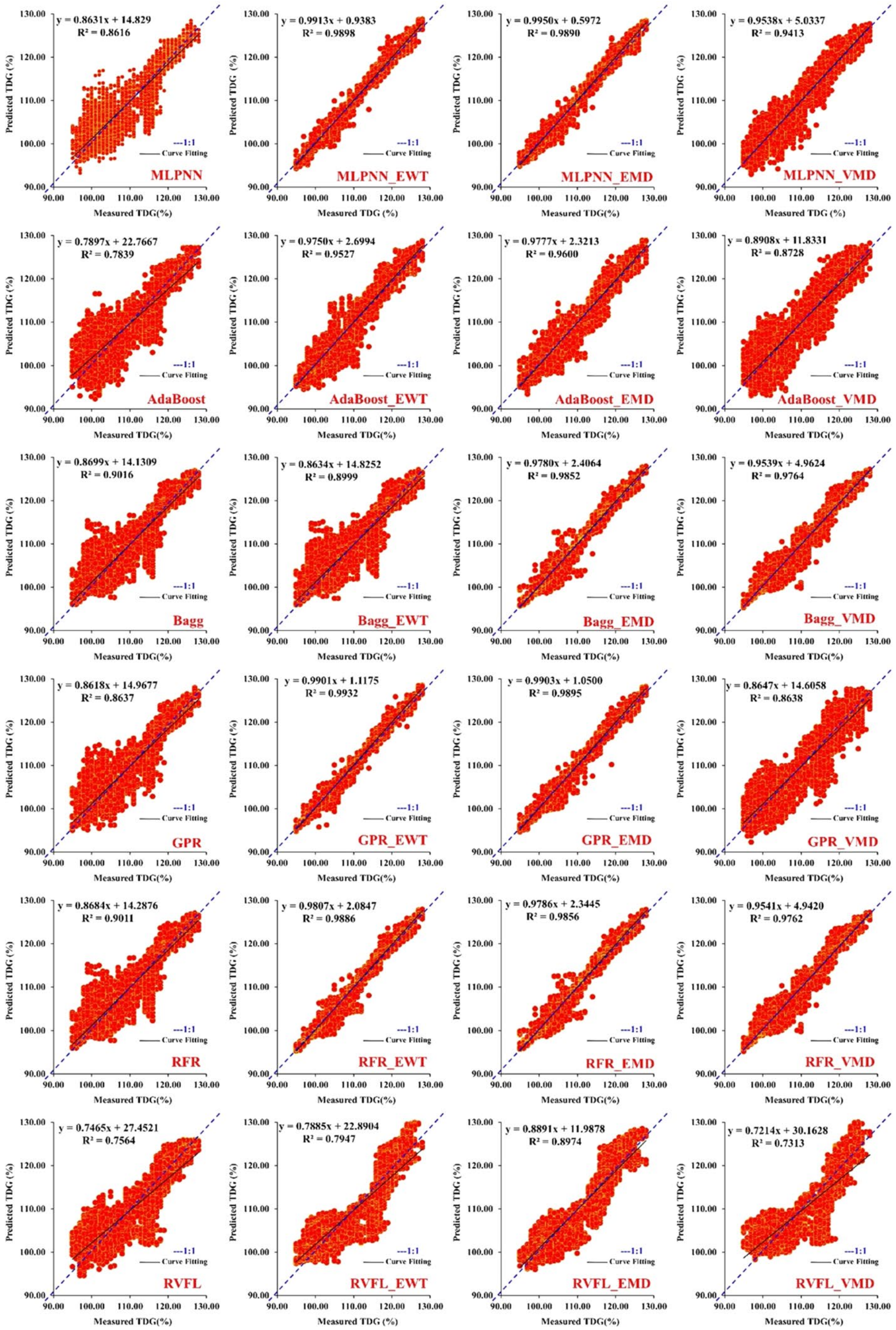
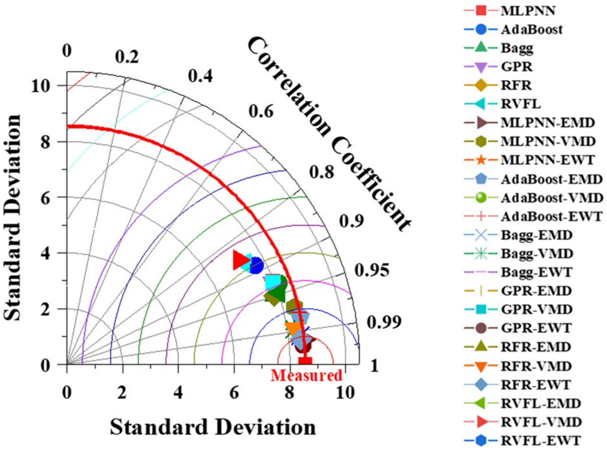


Fig. 5 Scatterplot of measured versus predicted TDG concentration using all models for the USGS 14019240 station

Fig. 6 Taylor diagram for models comparison using the validation dataset: USGS 14019240 station



remains constant, the performances of the MLPNN_VMD and AdaBoost_VMD were improved by ≈34.758%, ≈23.183% for the RMSE and by ≈23.183%, ≈23.220% for the MAE, respectively. Finally, in the case of the EMD signal decomposition algorithm as reported in Table 3 and Fig. 5 (i.e., the scatterplot), the performances evaluation metrics RMSE and MAE showed an improvement (i.e., decrease) by more than ≈60% and ≈65% for all models except the RVFL_EMD for which the improvement rate has note exceeded ≈35% and ≈36%, respectively. Furthermore, the higher improvement rates were achieved using the GPR_EMD model with ≈72.293% and ≈73.714% in term of RMSE and MAE, and it is clear that, the GPR_EMD model was the only one for which high R^2 and NSE values were obtained, i.e., ≈0.995 and ≈0.990, respectively. It can be seen that, beyond the RVFL model for which a negligible improvement was gained using the signal decomposition algorithms, all other models were significantly improved and the percentage of improvement is higher for the EMD than that for the VMD and EWT. Furthermore, there is not yet any clear superiority concerning all performances metrics.

3.2 USGS 13341000 station modeling results

Table 4 presents the results of TDG prediction using all model's at the USGS 13341000 station. The single RVFL model tend to have moderate performances having an R^2 and NSE values of approximately ≈ 0.656 and ≈ 0.626, respectively, while the RVFL_VMD tend to have the worst performances with R^2, NSE, RMSE, and MAE of approximately ≈ 0.335, ≈ 0.336, ≈ 3.050 and ≈ 2.373, respectively. More in depth analysis of the numerical results reported in Table 4 revealed that, the Bagg and the RFR models show the same numerical performances, and they were more accurate compared to the other models exhibiting an R^2 and NSE values of approximately ≈ 0.906 and ≈ 0.902. Furthermore, the RMSE and MAE values for the single models ranged from ≈ 1.156 to ≈ 2.269 and from ≈ 0.818 to ≈ 1.726, respectively, while the R^2 and NSE values were in the range of ≈ 0.656 to ≈ 0.906 and ≈ 0.626 to ≈ 0.903, respectively. Among all three decomposition algorithms, it is clear that, the most accurate simulation of the TDG was obtained using the EWT algorithm, followed by the EMD, while the VMD was ranked last. The comparison between the singles and hybrids models was also conducted using the scatterplot (Fig. 7)

Table 4 Performances of different models at the USGS 13341000 station

Models	Training				Validation			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
Standalone models without decomposition								
MLPNN	0.857	0.857	1.399	1.019	0.863	0.863	1.374	1.008
AdaBoost	0.824	0.825	1.546	1.156	0.826	0.826	1.548	1.165
Bagg	0.916	0.911	1.103	0.782	0.906	0.902	1.161	0.823
GPR	0.880	0.880	1.281	0.912	0.880	0.879	1.291	0.929
RFR	0.916	0.912	1.099	0.779	0.906	0.903	1.156	0.818
RVFL	0.651	0.627	2.260	1.721	0.656	0.626	2.269	1.726
Models based on empirical wavelet transform (EWT)								
MLPNN_EWT	0.982	0.981	0.504	0.371	0.972	0.972	0.616	0.454
AdaBoost_EWT	0.904	0.905	1.141	0.864	0.832	0.832	1.520	1.129
Bagg_EWT	0.992	0.992	0.330	0.215	0.964	0.963	0.712	0.477
GPR_EWT	0.984	0.984	0.470	0.334	0.978	0.978	0.548	0.392
RFR_EWT	0.992	0.992	0.329	0.215	0.966	0.964	0.704	0.474
RVFL_EWT	0.529	0.528	2.540	1.990	0.531	0.531	2.540	1.992
Models based on variational mode decomposition (VMD)								
MLPNN_VMD	0.958	0.958	0.752	0.564	0.897	0.893	1.223	0.851
AdaBoost_VMD	0.815	0.814	1.587	1.178	0.766	0.764	1.819	1.345
Bagg_VMD	0.990	0.989	0.385	0.257	0.956	0.952	0.816	0.573
GPR_VMD	0.895	0.894	1.200	0.883	0.812	0.811	1.625	1.193
RFR_VMD	0.990	0.989	0.389	0.259	0.956	0.952	0.818	0.573
RVFL_VMD	0.339	0.339	2.996	2.320	0.335	0.336	3.050	2.373
Models based on empirical mode decomposition (EMD)								
MLPNN_EMD	0.980	0.981	0.512	0.374	0.970	0.969	0.650	0.485
AdaBoost_EMD	0.916	0.916	1.071	0.778	0.835	0.836	1.502	1.044
Bagg_EMD	0.992	0.992	0.332	0.215	0.912	0.908	1.127	0.708
GPR_EMD	0.988	0.988	0.404	0.280	0.976	0.976	0.579	0.401
RFR_EMD	0.992	0.992	0.332	0.216	0.912	0.906	1.136	0.708
RVFL_EMD	0.590	0.589	2.370	1.870	0.585	0.586	2.388	1.879

and the Taylor diagram (Fig. 8). It is clear from the figures that, the hybrid models exhibited high numerical performances.

Comparison of the models one by one revealed that, the GPR_EWT was the most accurate in terms of R^2, NSE, RMSE, and MAE, with values of ≈ 0.978, ≈ 0.978, ≈ 0.548 and ≈ 0.392, respectively. The GPR_EWT was slightly higher than the GPR_EMD who provided the values of ≈ 0.976, ≈ 0.976, ≈ 0.579 and ≈0.401, respectively, while the RVFL_VMD model was the worst among all hybrid models for which the values of ≈ 0.335, ≈ 0.336, ≈ 3.050 and ≈ 2.373 were obtained. In this regard, it is important to note that, the performances of the RVFL model were significantly decreased using the EMD, VMD, and EWT algorithms, and the AdaBoost benefits less from the these three algorithms. Regarding the RVFL model, it is clear that, the VMD algorithm leads to a slightly decrease of the model's performances for which the R^2 and NSE values were dropped from (≈ 0.656 and ≈ 0.626) to (≈ 0.335 and ≈ 0.336), while the RMSE and MAE were raised from (≈ 2.269 and ≈ 1.726) to (≈ 3.050 and ≈ 2.373), respectively. Among

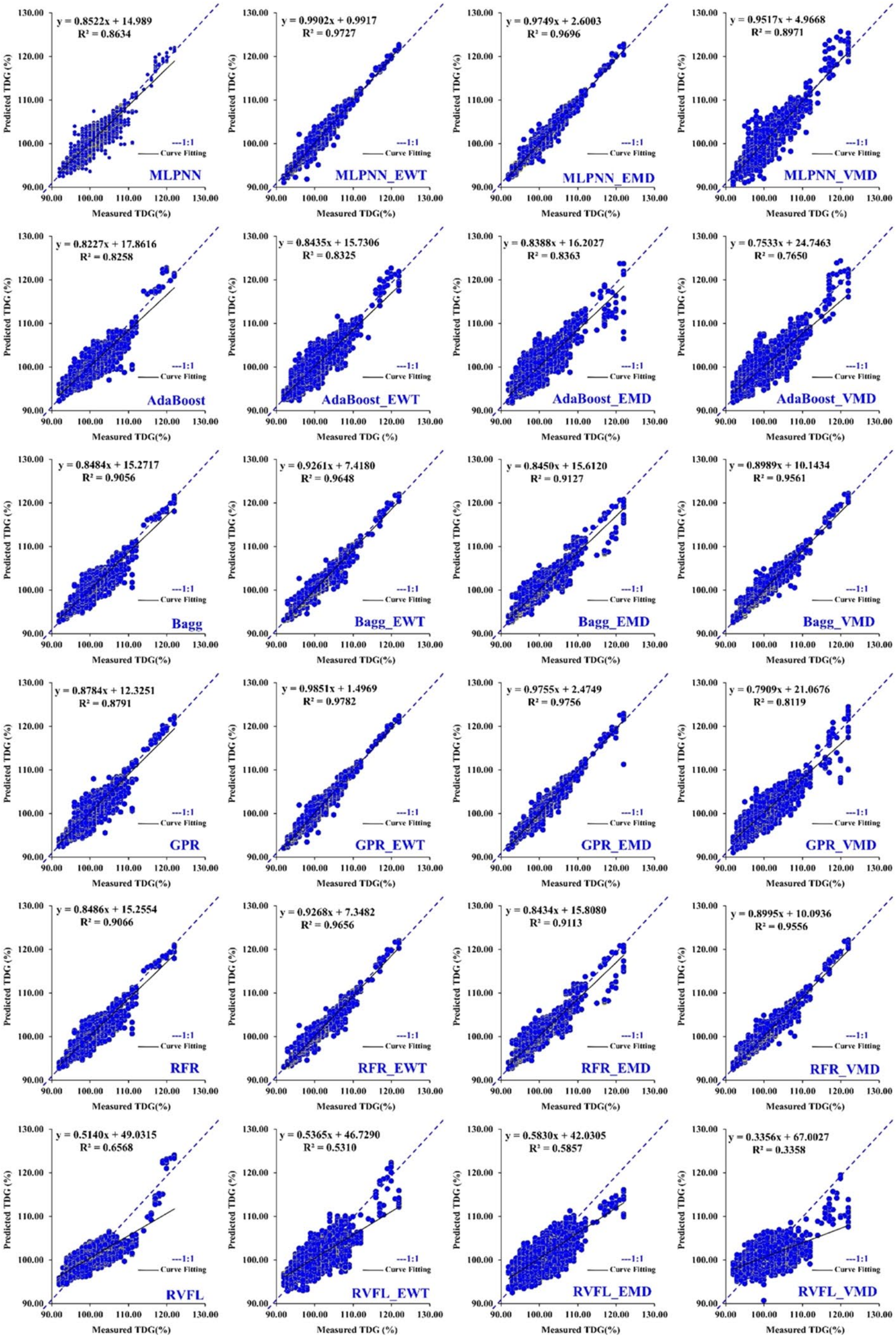


Fig. 7 Scatterplot of measured versus predicted TDG concentration using all models for the USGS 13341000 station

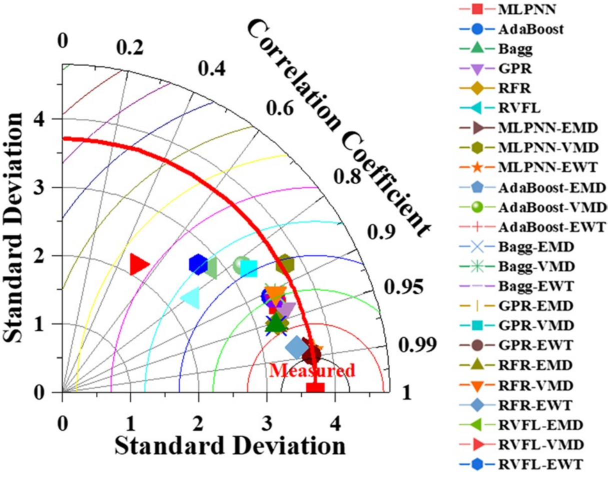


Fig. 8 Taylor diagram for models comparison using the validation dataset: USGS 13341000 station

all hybrid model's as stated in Table 4, it is clear that, the best predictive accuracies was gained using the hybrid GPR_EWT for which the R^2, NSE, RMSE, and MAE were immediately increased to reach their extreme values with improvement rates of ≈ 10.020%, ≈10.123%, ≈ 57.552%, and ≈ 57.804%, respectively. The GPR_EWT was followed by the MLPNN_EWT with R^2, NSE, RMSE, and MAE values of ≈ 0.972, ≈ 0.972, ≈ 0.616, and ≈ 0.454, respectively. Thus, the improvement observed using the EWT was more obvious, followed by the EMD and the VMD in the last round.

3.3 USGS 14019220 station modeling results

Table 5 presents the results of TDG prediction using all model's at the USGS 14019220 station. As clearly stated in Table 5, as we combine the signal decomposition with the ML models, an improvement can be seen in all models except: the AdaBoost_EWT, the GPR_VMD and the RVFL_VMD. More precisely, using the EWT algorithm, the R^2 and NSE values of the AdaBoost were dropped from (≈ 0.827 and ≈ 0.836) to (≈ 0.815 and ≈ 0.810), while the RMSE and MAE were raised from (≈ 1.745 and ≈1.377) to (≈ 1.881 and ≈ 1.340). According to Table 5, using single model's it is clear that the Bagg and the RFR were the most accurate model's exhibiting the higher R^2 and NSE values of ≈ 0.901 and ≈ 0.899, respectively, and the lowest RMSE and MAE value of ≈ 1.360 and ≈ 1.021, respectively. Among all single model's, the RVFL was found to be the poorest one having the lowest numerical performances with R^2 ≈ 0.773, NSE ≈ 0.772, RMSE ≈ 2.060 and MAE≈1.629, respectively.

Further comparison between the hybrid model's based on the EWT algorithm revealed that, the GRP_EWT was the most accurate model having the greatest numerical performances (R^2 ≈ 0.992, NSE ≈ 0.992, RMSE ≈ 0.394, MAE ≈ 0.297), respectively, slightly higher than the MLPNN_EWT, while the RVFL_EWT (R^2 ≈ 0.808, NSE ≈ 0.808, RMSE ≈ 1.889, MAE ≈ 1.502) was the less accurate model. Using the VMD algorithm, it is clear from the results reported in Table 5 that, the Bagg_VMD and the RFR_VMD were the most accurate model's compared to the all other and they exhibited high numerical performances of approximately (R^2 ≈ 0.972, NSE ≈ 0.972, RMSE ≈ 0.723, MAE ≈ 0.507).

Table 5 Performances of different models at the USGS 14019220 station

Models	Training				Validation			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
Standalone models without decomposition								
MLPNN	0.870	0.870	1.544	1.207	0.861	0.859	1.618	1.259
AdaBoost	0.852	0.851	1.648	1.295	0.837	0.836	1.745	1.377
Bagg	0.943	0.939	1.053	0.792	0.901	0.899	1.369	1.033
GPR	0.885	0.886	1.446	1.105	0.874	0.873	1.538	1.185
RFR	0.943	0.940	1.045	0.783	0.901	0.900	1.360	1.021
RVFL	0.760	0.759	2.101	1.670	0.773	0.772	2.060	1.629
Models based on empirical wavelet transform (EWT)								
MLPNN_EWT	0.996	0.996	0.284	0.218	0.990	0.991	0.412	0.321
AdaBoost_EWT	0.943	0.943	1.019	0.785	0.815	0.810	1.881	1.340
Bagg_EWT	0.996	0.997	0.251	0.172	0.953	0.949	0.973	0.558
GPR_EWT	0.996	0.996	0.262	0.195	0.992	0.992	0.394	0.297
RFR_EWT	0.996	0.997	0.251	0.172	0.953	0.949	0.976	0.556
RVFL_EWT	0.808	0.808	1.876	1.508	0.808	0.808	1.889	1.502
Models based on variational mode decomposition (VMD)								
MLPNN_VMD	0.988	0.989	0.450	0.347	0.960	0.959	0.875	0.607
AdaBoost_VMD	0.945	0.944	1.009	0.782	0.901	0.899	1.370	1.036
Bagg_VMD	0.996	0.995	0.292	0.206	0.972	0.972	0.727	0.509
GPR_VMD	0.908	0.908	1.295	1.003	0.869	0.868	1.564	1.198
RFR_VMD	0.996	0.995	0.293	0.207	0.972	0.972	0.723	0.507
RVFL_VMD	0.753	0.753	2.127	1.683	0.764	0.763	2.097	1.649
Models based on empirical mode decomposition *(*EMD*)*								
MLPNN_EMD	0.994	0.994	0.325	0.253	0.980	0.981	0.599	0.403
AdaBoost_EMD	0.978	0.977	0.642	0.494	0.869	0.858	1.626	1.146
Bagg_EMD	0.996	0.997	0.244	0.168	0.955	0.954	0.923	0.670
GPR_EMD	0.996	0.996	0.280	0.206	0.984	0.984	0.540	0.370
RFR_EMD	0.996	0.997	0.243	0.167	0.955	0.953	0.933	0.666
RVFL_EMD	0.869	0.869	1.548	1.221	0.863	0.862	1.600	1.246

Finally, using the EMD algorithm, it is clear that the GPR_EMD was the most accurate with the biggest numerical performances of approximately ($R^2 \approx 0.984$, NSE ≈ 0.984, RMSE ≈ 0.540, MAE ≈ 0.370).

The Scatterplot (Fig. 9) and Taylor diagram (Fig. 10) have helped in providing a concise and solid comparison between the singles and hybrids models. According to Fig. 9, it is clear that the comparison between measured and predicted data demonstrated that the hybrids models exhibited less scattered data compared to the single models.

3.4 USGS 13352950 station modeling results

Table 6 presents the results of TDG prediction using all models at the USGS 13352950 station. The performances of the singles and hybrid models were assessed through a comparative analysis based on the performances metrics reported in Table 6. First, the

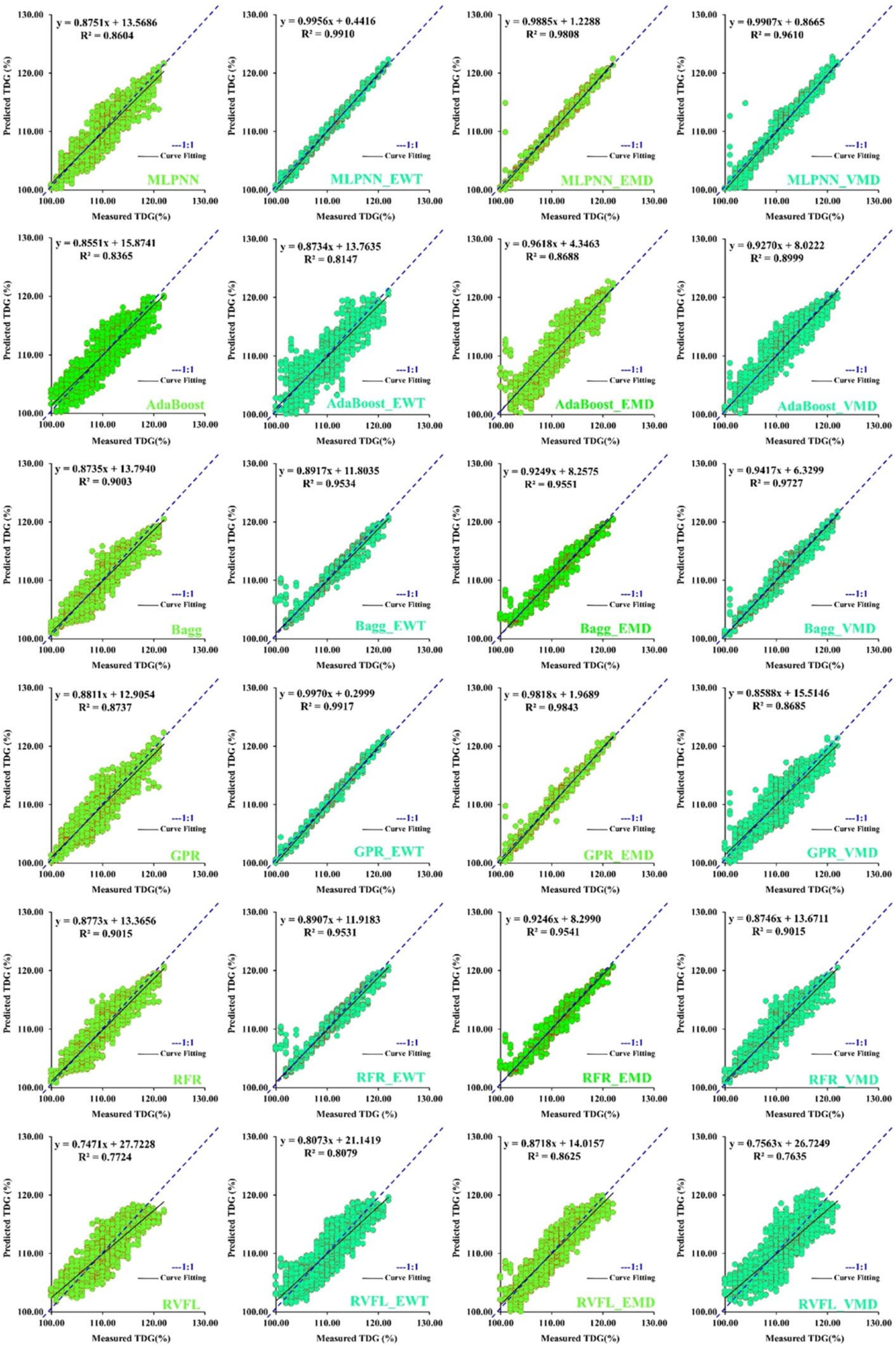
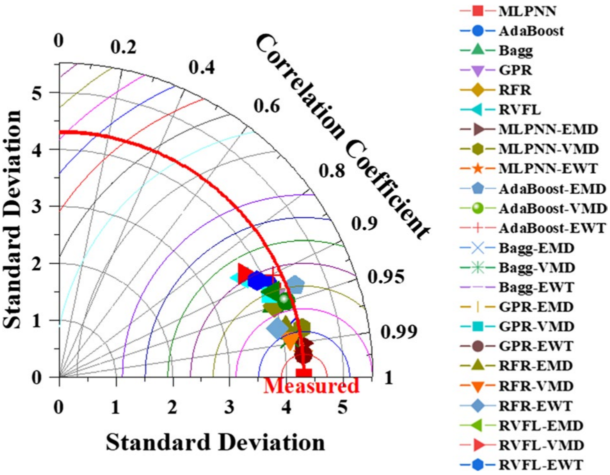


Fig. 9 Scatterplot of measured versus predicted TDG concentration using all models for the USGS 14019220 station

Fig. 10 Taylor diagram for models comparison using the validation dataset: USGS 14019220 station



models were arranged from the highest (R^2 and NSE) to the lowest (RMSE and MAE), and consequently the models were arranged as follow: RFR(1), Bagg(2), GPR(3), MLPNN(4), AdaBoost(5) and RVFL(6). It is clear that, the RFR and Bagg revealed the best performances by presenting the lowest RMSE and MAE values of ≈ 1.914 and ≈ 1.400, respectively, and the highest R^2 and NSE values of ≈ 0.891 and ≈ 0.891, respectively. The RVFL model presented the worst results compared to all other models with R^2, NSE, RMSE, and MAE values of ≈ 0.724, ≈ 0.722, ≈ 3.061 and ≈ 2.410, respectively. Moreover, according to the previous results, we can conclude that, obtained results demonstrated that RFR and Bagg would be more feasible for modelling TDG and possess more capability of nonlinear mapping, followed by the GPR and the MLPNN models.

The previous statement was further confirmed by analyzing the results obtained using the VMD, EMD and EWT signal decomposition. According to Table 6, the VMD, EMD and EWT further improve the model's performances, which lead to a great improvement with the involvement of these techniques in overcoming some limitations of the standalone models, especially, the capabilities in capturing the nonlinearity present in the input variables. As results of this, we can see that, the R^2 and NSE metrics reached extreme values of ≈ 0.996 and ≈ 0.995 (i.e., the maximal values obtained at the 13,352,950 station) obtained using the GPR_EWT accompanied by an improvements rate of ≈ 80% and ≈ 80.44% in terms of RMSE (≈ 0.405) and MAE (≈ 0.303) reduction compared to the single GPR model. Notably, the EWT is the sole algorithm for which the performances of all-single models were improved. We can note that the MLPNN_EWT was slightly lower than the GPR_EWT with negligible difference in terms of model's performances, while the Bagg_EWT and RFR_EWT worked equally. By further analysis of the obtained results, the following conclusions can be drawn. First, the GPR_EMD yielded the best performances (equally with the MLPNN_EMD) compared to the other models, while the AdaBoost_EMD was the worst model. More precisely, the R^2 and NSE values of ≈ 0.992 and ≈ 0.990 were obtained using the GPR_EMD and MLPNN_EMD against ≈ 0.839, and ≈ 0.828 obtained using the AdaBoost_EMD. Second, the improvement gained using the VMD was the least significant and two models have shown their performances decreased compared to the single models, i.e., the GPR_VMD and the RVFL_VMD, for which the R^2 and NSE values were dropped from (≈ 0.880 and ≈ 0.879) to (≈ 0.859 and ≈ 0.858), and from

Table 6 Performances of different models at the USGS 13352950 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	R	NSE	RMSE	MAE
Standalone models without decomposition								
MLPNN	0.878	0.877	2.042	1.595	0.861	0.859	2.182	1.651
AdaBoost	0.891	0.891	1.929	1.521	0.857	0.855	2.209	1.730
Bagg	0.956	0.954	1.246	0.921	0.891	0.891	1.914	1.401
GPR	0.899	0.898	1.861	1.434	0.880	0.879	2.024	1.551
RFR	0.956	0.954	1.248	0.922	0.891	0.891	1.914	1.399
RVFL	0.717	0.715	3.112	2.469	0.724	0.722	3.061	2.410
Models based on empirical wavelet transform (EWT)								
MLPNN_EWT	0.998	0.998	0.271	0.202	0.994	0.994	0.437	0.341
AdaBoost_EWT	0.982	0.982	0.785	0.607	0.939	0.937	1.455	1.036
Bagg_EWT	0.998	0.998	0.246	0.160	0.976	0.976	0.902	0.526
GPR_EWT	0.998	0.999	0.220	0.156	0.996	0.995	0.405	0.303
RFR_EWT	0.998	0.998	0.245	0.160	0.974	0.973	0.952	0.530
RVFL_EWT	0.792	0.792	2.659	2.114	0.789	0.788	2.672	2.114
Models based on variational mode decomposition (VMD)								
MLPNN_VMD	0.994	0.993	0.479	0.369	0.974	0.971	0.984	0.723
AdaBoost_VMD	0.972	0.973	0.959	0.739	0.925	0.921	1.634	1.154
Bagg_VMD	0.998	0.998	0.266	0.184	0.978	0.977	0.878	0.573
GPR_VMD	0.916	0.915	1.701	1.296	0.859	0.858	2.189	1.688
RFR_VMD	0.998	0.998	0.264	0.183	0.978	0.978	0.865	0.568
RVFL_VMD	0.691	0.690	3.248	2.625	0.694	0.693	3.216	2.580
Models based on empirical mode decomposition (EMD)								
MLPNN_EMD	0.998	0.998	0.284	0.216	0.992	0.990	0.570	0.429
AdaBoost_EMD	0.986	0.986	0.687	0.513	0.839	0.828	2.408	1.344
Bagg_EMD	0.998	0.998	0.229	0.152	0.972	0.970	1.011	0.710
GPR_EMD	0.998	0.998	0.251	0.185	0.992	0.991	0.549	0.394
RFR_EMD	0.998	0.998	0.230	0.152	0.972	0.969	1.022	0.712
RVFL_EMD	0.869	0.869	2.111	1.686	0.874	0.873	2.065	1.646

(≈ 0.724 and ≈ 0.722) to (≈ 0.694 and ≈ 0.693), respectively. Yet, it is important to note that, high accuracy was obtained using the MLPNN_VMD, RFR_VMD, and Bagg_VMD, for which the R^2 and NSE values have raised the values of ≈ 0.978 and ≈ 0.970, respectively. The scatterplot of measured and calculated TDG using all models are depicted in Fig. 11. The Taylor diagram in Fig. 12 is presented for further highlighting the superiority of the hybrid models compared to the single models.

4 Summary and remarks

It is clear from the above discussed results that the proposed hybrid models based on signal decomposition are valuable for modelling TDG concertation. Meanwhile, the superiority, robustness and effectiveness of the VMD, EMD and EWT algorithms were justified and

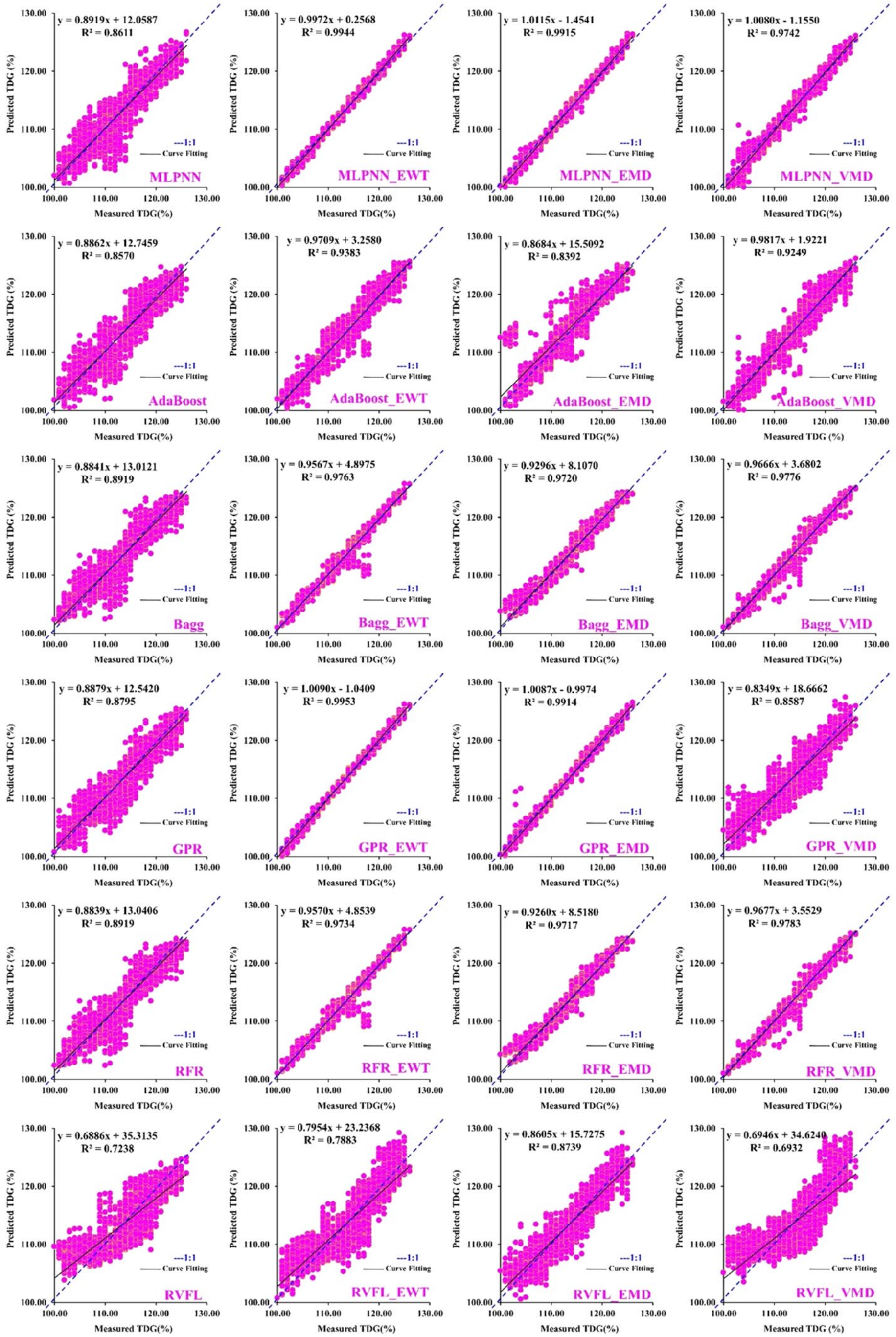
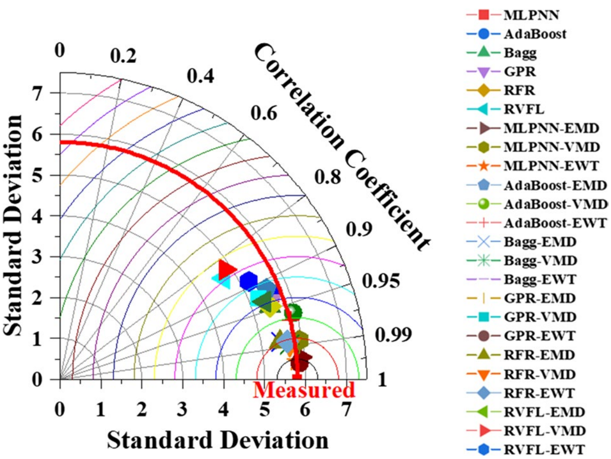


Fig. 11 Scatterplot of measured versus predicted TDG concentration using all models for the USGS 13352950 station

Fig. 12 Taylor diagram for models comparison using the validation dataset: USGS 13352950 station



validated by a deeply comparison of the singles and hybrid models performances. For further discussion and comparison of the efficiency of the proposed modelling framework based on signal decomposition, previous models available in the literature were compared with our models and the comparison is presented hereafter.

A research study used the ELM and SVR models for predicting TDG concentration using several input variables namely, T_w, *BP, Q* and gage height (*GH*) (AlOmar et al. 2020). However, in their study, they have forecasted the TDG one hour in advance ($t+1$), using the input variables measured at the previous lag time ($t - 1$). From the obtained results, the SVR was slightly more accurate and exhibiting an R-value of ≈ 0.992, compared to the values of ≈ 0.990 obtained using the ELM model. In comparison to our results, it is clear the SVR and ELM have provided the same performances obtained in our present study, however, the inclusion of the TDG measured at (t − 1) have certainly contributed to the improvement of the ELM and SVR models performances. A research used the SVR, ELM, genetic algorithm ELM (GA-ELM), and GA-SVR for modelling TDG measured at three dams' reservoir stations (Wang and Sheng 2022). They used a large number of input variables namely, *Q, BP,* T_w, discharge per unit width (q), upstream and downstream water level difference (ΔH), bubble pressure(F) in stilling basin, and retention time (*TS*). Form the obtained results, it was found that, GA-SVR (RMSE ≈ 1.43, MAE ≈ 0.95) was more accurate compared to GA-ELM (RMSE ≈ 1.48, MAE ≈ 1.31). However, the two models, i.e., the GA-ELM and GA-SVR were significantly less accurate compared to the models proposed in the present study, for which an RMSE and MAE of ≈ 0.394 and ≈ 0.297 were obtained using the GPR combined with the EWT, EMD and VMD algorithms. Another study used the MLPNN model for modelling TDG using sensor depth (SD), *Q, BP,* T_w, spill from dam (SFD), and water elevation (WL) (Han et al. 2019). It was found that; the MLPNN was more accurate compared to the MLR model exhibiting R, NSE, RMSE, and MAE values of ≈ 0.970, ≈ 0.930, ≈ 2.05 and ≈ 1.22, respectively, which were less accurate compared to models developed in the present study (R ≈ 0.996, NSE ≈ 0.992, RMSE ≈ 0.394 and MAE ≈ 0.297).

Recently, a scientific research introduced a new modelling strategy using the parallel chaos search based incremental extreme learning machine (PC-ELM) for predicting hourly TDG using only water temperature (T_w) as predictor (Heddam 2023). The PC-ELM model was tested using data from four station located at Snake River, USA, and operated by the USGS. For improving the performance of the PC-ELM, the T_w was combined with the periodicity, i.e., the year, month, day, and hour number. It was found that; the PC-ELM was slightly more accurate compared to the standalone ELM, exhibiting an R^2 value of approximately ≈ 0.965, however, the PC-ELM is less accurate compared to the hybrid models reported in our present study (R^2 ≈ 0.990). In an investigation a research article compared between adaptive neuro-fuzzy inference systems (ANFIS) and dynamic evolving neural-fuzzy inference system (DENFIS) for modelling hourly TDG measured in two USGS stations (Heddam and Kisi 2021). The two ANFIS and DENFIS models were developed using *Q, BP, T_w,* and SFD. The ANFIS model was found to be more accurate with R, NSE, RMSE and MAE values of ≈ 0.977, ≈ 0.954, ≈ 1.084 and ≈ 0.773, compared to the values of ≈ 0.968, ≈ 0.936, ≈ 1.271 and ≈ 0.868, obtained using the DENFIS model. By comparison with the present research, the DENFIS and ANFIS models were less accurate compared to the hybrid models proposed in our study. Another study used the generalized regression neural network (GRNN) with the MLR model for predicting TDG using a large number of predictors namely, *Q, BP, T_w,* SFD, SD (Heddam 2017). Good performances were obtained using the GRNN model with R, NSE, RMSE and MAE values of ≈ 0.946, ≈ 0.895, ≈ 0.995 and ≈ 0.593, respectively, however, they are significantly lower than the values obtained in this study. In another study, the authors compared between kriging interpolation method (KIM), response surface method (RSM), and the MLPNN models for predicting TDG using *Q, BP, T_w,* and *SFD* (Heddam and Kisi 2020). It was found that he proposed KIM was more accurate compared to the MLPNN and RSM, and excellent performances were obtained with R, NSE, RMSE and MAE values of ≈ 0.973, ≈ 0.941, ≈ 1.462 and ≈ 1.122, respectively, always less than the values obtained using the hybrid models proposed in the present study. Finally, in a research the authors applied the high-order response surface method (H-RSM), M5Tree, least squares support vector machine (LSSVM), and multivariate adaptive regression spline (MARS) for modelling TDG, showing the superiority of the H-RSM model with R, NSE, RMSE and MAE values of ≈ 0.965, ≈ 0.931, ≈ 1.456 and ≈ 1.022, respectively [31].

5 Conclusion

Total dissolved gas (TDG) produced at high dam reservoir was regressed against water temperature, barometric pressure and discharge measurements from four USG stations to explore whether ML algorithms were able to accurately predict TDG. In addition to this, the objective was to determine how signal decomposition approaches and conventional ML could be combined to improve the predictive accuracy. The results showed that the modelling using single models without signal decomposition results in TDG estimates with a good predictive accuracy for all models. However, the RVFL and AdaBoost have provided moderate results. In addition, the performances of the models varied from one station to another and from one model to another and no general conclusion could be drawn. The second part of the study was mainly motivated by the requirement of high-quality TDG estimations, which is crucial to efficiently control water resources and aquatic life. We looked at where could make the greatest gains on reducing the predictive errors between measured and predicted TDG. For answering this query, we have deeply analyzed the potential that can be gained from the signal decomposition algorithms. Three algorithms were then tested, i.e., the VMD, EMD, and EWT. Thus, hybrid models were used and compared to the single models, providing consistent estimations for the major's cases and for all stations. We argued that it was essential to introduce new robust tools for better predicting of TDG, which is successfully done in the present study, the combined ML and signal decomposition reveals reliability and relevance of the TDG estimations. Future research may be focused on the application of the proposed models for large dataset and by testing other ML models; in addition, testing the models with other input variables could be an innovative idea.

Author contributions Salim Heddam: conceptualization, modelling, methodology, writing up, revision and edits, software, analysis, supervision. Ahmed M. Al-Areeq: Writing up, review and edit, analysis, investigation, visualization, validation. Mou Leong Tan: writing up, review and edit, analysis, investigation, visualization, validation. Iman Ahmadianfar: writing up, review and edit, analysis, investigation, visualization, validation. Bijay Halder: writing up, review and edit, analysis, investigation, visualization, validation. Vahdettin Demir: writing up, review and edit, analysis, investigation, visualization, validation. Huseyin Cagan Kilinc: writing up, review and edit, analysis, investigation, visualization, validation. Sani I. Abba: writing up, review and edit, analysis, investigation, visualization, validation. Atheer Y Oudah: writing up, review and edit, analysis, investigation, visualization, validation. Zaher Mundher Yaseen: conceptualization, modelling, methodology, writing up, revision and edits, software, analysis, supervision, project leader. All authors have read and agreed to the published version of the manuscript.

Funding Not applicable.

Data availability The data presented in this study will be available on interested request from the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

Informed consent Not applicable.

Consent to participate Not applicable.

Consent for publication All the authors have declared their consent to publish the manuscript.

References

AlOmar MK, Hameed MM, Al-Ansari N, AlSaadi MA (2020) Data-driven model for the prediction of total dissolved gas: robust artificial intelligence approach. Adv Civ Eng 2020:6618842. https://doi.org/10.1155/2020/6618842

Bokde N, Feijóo A, Al-Ansari N et al (2020) The hybridization of ensemble empirical mode decomposition with forecasting models: application of short-term wind speed and power modeling. Energies 13:1666

Breiman L (2001) No title. Mach Learn 45:5–32. https://doi.org/10.1023/a:1010933404324

Chen Y, Wu X, Liu X et al (2023) Biochemical, transcriptomic and metabolomic responses to total dissolved gas supersaturation and their underlying molecular mechanisms in Yangtze sturgeon (*Acipenser dabryanus*). Environ Res. https://doi.org/10.1016/j.envres.2022.114457

Cheng X, Lu J, Li R et al (2021) Experimental study of the degasification efficiency of supersaturated dissolved oxygen on stepped cascades and correlation prediction model. J Clean Prod. https://doi.org/10.1016/j.jclepro.2021.129611

Chong D, Zhu N, Luo W, Pan X (2019) Human thermal risk prediction in indoor hyperthermal environments based on random forest. Sustain Cities Soc 49:101595. https://doi.org/10.1016/j.scs.2019.101595

Dragomiretskiy K, Zosso D (2014) Variational mode decomposition. IEEE Trans Signal Process 62:531–544. https://doi.org/10.1109/tsp.2013.2288675

Feng J, Li R, Yang H, Li J (2013) A laterally averaged two-dimensional simulation of unsteady supersaturated total dissolved gas in deep reservoir. J Hydrodyn 25:396–403. https://doi.org/10.1016/s1001-6058(11)60378-9

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139

Gilles J (2013) Empirical wavelet transform. IEEE Trans Signal Process 61:3999–4010. https://doi.org/10.1109/tsp.2013.2265222

Giri S, Kang Y, MacDonald K et al (2023) Revealing the sources of arsenic in private well water using random forest classification and regression. Sci Total Environ 857:159360. https://doi.org/10.1016/j.scitotenv.2022.159360

Han L, Cai S, Gao M et al (2019) Selective catalytic reduction of NOx with NH_3 by using novel catalysts: State of the art and future prospects. Chem Rev 119:10916–10976

He Z, Zhou W (2022) Improvement of burst capacity model for pipelines containing dent-gouges using Gaussian process regression. Eng Struct 272:115028. https://doi.org/10.1016/j.engstruct.2022.115028

Heddam S (2017) Generalized regression neural network based approach as a new tool for predicting total dissolved gas (TDG) downstream of spillways of dams: a case study of Columbia river basin dams, USA. Environ Process 4:235–253

Heddam S (2023) Parallel chaos search-based incremental extreme learning machine. Handbook of hydroinformatics. Elsevier, Amsterdam

Heddam S, Kisi O (2020) Evolving connectionist systems versus neuro-fuzzy system for estimating total dissolved gas at forebay and tailwater of dams reservoirs. Springer, Berlin, pp 109–126

Heddam S, Kisi O (2021) Evolving connectionist systems versus neuro-fuzzy system for estimating total dissolved gas at forebay and tailwater of dams reservoirs. Intell Data Anal Decis Syst Hazard Mitig Theory Pract Hazard Mitig. https://doi.org/10.1007/978-981-15-5772-9_6

Huang NE, Shen Z, Long SR et al (1998) The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc A Math Phys Eng Sci. https://doi.org/10.1098/rspa.1998.0193

Huang J, Li R, Feng J et al (2021) The application of baffle block in mitigating TDGS of dams with different discharge patterns. Ecol Indic. https://doi.org/10.1016/j.ecolind.2021.108418

Jiao W, Song S, Han H et al (2023) Artificially intelligent differential diagnosis of enlarged lymph nodes with random vector functional link network plus. Med Eng Phys 111:103939. https://doi.org/10.1016/j.medengphy.2022.103939

Karbasi M, Jamei M, Ali M et al (2022) Developing a novel hybrid auto encoder decoder bidirectional gated recurrent unit model enhanced with empirical wavelet transform and Boruta-Catboost to forecast significant wave height. J Clean Prod 379:134820. https://doi.org/10.1016/j.jclepro.2022.134820

Keshtegar B, Heddam S, Kisi O, Zhu SP (2019) Modeling total dissolved gas (TDG) concentration at Columbia river basin dams: high-order response surface method (H-RSM) vs. M5Tree, LSSVM, and MARS. Arab J Geosci. https://doi.org/10.1007/s12517-019-4687-3

Khozani ZS, Khosravi K, Pham BT et al (2019) Determination of compound channel apparent shear stress: application of novel data mining models. J Hydroinformatics. https://doi.org/10.2166/hydro.2019.037

Li R, Li J, Li KF et al (2009) Prediction for supersaturated total dissolved gas in high-dam hydropower projects. Sci China, Ser E Technol Sci. https://doi.org/10.1007/s11431-009-0337-4

Li P, Zhu DZ, Li R et al (2022) Production of total dissolved gas supersaturation at hydropower facilities and its transport: a review. Water Res 223:119012. https://doi.org/10.1016/j.watres.2022.119012

Li Y, Alameri AA, Farhan ZA et al (2023a) Theoretical modeling study on preparation of nanosized drugs using supercritical-based processing: Determination of solubility of Chlorothiazide in supercritical carbon dioxide. J Mol Liq 370:120984. https://doi.org/10.1016/j.molliq.2022.120984

Li Y, Luo J, Dai Q et al (2023b) A deep learning approach to cardiovascular disease classification using empirical mode decomposition for ECG feature extraction. Biomed Signal Process Control 79:104188. https://doi.org/10.1016/j.bspc.2022.104188

Lin L, Li R, Feng J et al (2022) Experimental study of the growth period of wall-attached bubbles. Water Supply 22:4769–4780. https://doi.org/10.2166/ws.2022.168

Lu J, Li R, Ma Q et al (2019) Model for total dissolved gas supersaturation from plunging jets in high dams. J Hydraul Eng 145:4018082

Ma Q, Liang R, Li R et al (2016) Operational regulation of water replenishment to reduce supersaturated total dissolved gas in riverine wetlands. Ecol Eng 96:162–169. https://doi.org/10.1016/j.ecoleng.2016.03.019

Ma Q, Li R, Feng J et al (2019) Ecological regulation of cascade hydropower stations to reduce the risk of supersaturated total dissolved gas to fish. J Hydro-Environment Res 27:102–115. https://doi.org/10.1016/j.jher.2019.10.002

Nabih M, Ghoneimi A, Bakry A et al (2023) Rock physics analysis from predicted Poisson's ratio using RVFL based on wild geese algorithm in scarab gas field in WDDM concession. Egypt Mar Pet Geol 147:105949. https://doi.org/10.1016/j.marpetgeo.2022.105949

Nancy Jane Y, Charanya SK, Amsaprabhaa M et al (2023) 2-HDCNN: a two-tier hybrid dual convolution neural network feature fusion approach for diagnosing malignant melanoma. Comput Biol Med 152:106333. https://doi.org/10.1016/j.compbiomed.2022.106333

Netsanet S, Zheng D, Zhang W, Teshager G (2022) Short-term PV power forecasting using variational mode decomposition integrated with Ant colony optimization and neural network. Energy Rep 8:2022–2035. https://doi.org/10.1016/j.egyr.2022.01.120

Ouyang Z-L, Liu S-Y, Zou Z-J (2022) Nonparametric modeling of ship maneuvering motion in waves based on Gaussian process regression. Ocean Eng 264:112100. https://doi.org/10.1016/j.oceaneng.2022.112100

Pao Y-H, Phillips SM, Sobajic DJ (1992) Neural-net computing and the intelligent control of systems. Int J Control 56:263–289. https://doi.org/10.1080/00207179208934315

Pao Y-H, Park G-H, Sobajic DJ (1994) Learning and generalization characteristics of the random vector functional-link net. Neurocomputing 6:163–180. https://doi.org/10.1016/0925-2312(94)90053-1

Peng L, Wang L, Xia D, Gao Q (2022a) Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. Energy 238:121756. https://doi.org/10.1016/j.energy.2021.121756

Peng Y, Lin Y, Zeng C et al (2022) Improved model for predicting total dissolved gas generation with the residence time of the water in the stilling phase. Front Environ Sci. https://doi.org/10.3389/fenvs.2021.770187

Politano MS, Carrica PM, Turan C, Weber L (2007) A multidimensional two-phase flow model for the total dissolved gas downstream of spillways. J Hydraul Res 45:165–177. https://doi.org/10.1080/00221686.2007.9521757

Politano M, Carrica P, Weber L (2009) A multiphase model for the hydrodynamics and total dissolved gas in tailraces. Int J Multiph Flow 35:1036–1050. https://doi.org/10.1016/j.ijmultiphaseflow.2009.06.009

Politano M, Arenas Amado A, Bickford S et al (2012) Evaluation of operational strategies to minimize gas supersaturation downstream of a dam. Comput Fluids 68:168–185. https://doi.org/10.1016/j.compfluid.2012.08.003

Politano M, Castro A, Hadjerioua B (2017) Modeling total dissolved gas for optimal operation of multireservoir systems. J Hydraul Eng. https://doi.org/10.1061/(asce)hy.1943-7900.0001287

Qiao Z-K, Yuan P, Hu R et al (2022) Research on aeromagnetic data error analysis and processing of multirotor UAV based on variational mode decomposition algorithm. Heliyon 8:e11808–e11808. https://doi.org/10.1016/j.heliyon.2022.e11808

Qin Y, Wei Q, Ji Q et al (2022) Determining the position of a fish passage facility entrance based on endemic fish swimming abilities and flow field. Environ Sci Pollut Res 30:6104–6116. https://doi.org/10.1007/s11356-022-22581-0

Ren X, Zhang X, Yan C, Gozgor G (2022) Climate policy uncertainty and firm-level total factor productivity: evidence from China. Energy Econ 113:106209

Rezaie-Balf M, Attar NF, Mohammadzadeh A et al (2020) Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. J Clean Prod 271:122576

Rout SK, Sahani M, Dora C et al (2022) An efficient epileptic seizure classification system using empirical wavelet transform and multi-fuse reduced deep convolutional neural network with digital implementation. Biomed Signal Process Control 72:103281. https://doi.org/10.1016/j.bspc.2021.103281

Saha S, Bera B, Shit PK et al (2023) Modelling and predicting of landslide in Western Arunachal Himalaya. India Geosyst Geoenviron 2:100158. https://doi.org/10.1016/j.geogeo.2022.100158

Salman B, Kadhum MM (2022) Predicting of load carrying capacity of reactive powder concrete and normal strength concrete column specimens using artificial neural network. Knowledge-Based Eng Sci 3:45–53

Shamaee Z, Mivehchy M (2023) Dominant noise-aided EMD (DEMD): Extending empirical mode decomposition for noise reduction by incorporating dominant noise and deep classification. Biomed Signal Process Control 80:104218. https://doi.org/10.1016/j.bspc.2022.104218

Shen X, Li R, Huang J et al (2016) Shelter construction for fish at the confluence of a river to avoid the effects of total dissolved gas supersaturation. Ecol Eng 97:642–648. https://doi.org/10.1016/j.ecoleng.2016.10.055

Sun H (2023) Construction of integration path of management accounting and financial accounting based on big data analysis. Optik (Stuttg) 272:170321. https://doi.org/10.1016/j.ijleo.2022.170321

Takoutsing B, Heuvelink GBM (2022) Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma 428:116192. https://doi.org/10.1016/j.geoderma.2022.116192

Truong GT, Choi K-K, Kim C-S (2022) Implementation of boosting algorithms for prediction of punching shear strength of RC column footings. Structures 46:521–538. https://doi.org/10.1016/j.istruc.2022.10.085

Wang M, Sheng X (2022) Combining empirical wavelet transform and transfer matrix or modal superposition to reconstruct responses of structures subject to typical excitations. Mech Syst Signal Process 163:108162

Wang Y, Politano M, Weber L (2019a) Spillway jet regime and total dissolved gas prediction with a multiphase flow model. J Hydraul Res 57:26–38

Wang Z, Lu J, Yuan Y et al (2019b) Experimental study on the effects of vegetation on the dissipation of supersaturated total dissolved gas in flowing water. Int J Environ Res Public Health 16:2256. https://doi.org/10.3390/ijerph16132256

Wang Z, Feng J, Liang M et al (2022) Prediction model and application of machine learning for supersaturated total dissolved gas generation in high dam discharge. Water Res. https://doi.org/10.1016/j.watres.2022.118682

Weiqi K, Weisong W, Maoxing Z (2022) Integrated learning algorithms with Bayesian optimization for mild steel mechanical properties prediction. Knowledge-Based Eng Sci 3:101–112

Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning Vol. 2, No. 3: p. 4. Cambridge MA: MIT press

Xiong B, Meng X, Xiong G et al (2022) Multi-branch wind power prediction based on optimized variational mode decomposition. Energy Rep 8:11181–11191. https://doi.org/10.1016/j.egyr.2022.08.271

Yaseen ZM (2021) An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere 277:130126. https://doi.org/10.1016/j.chemosphere.2021.130126

Yuan Y, Feng J, Li R et al (2018) Modelling the promotion effect of vegetation on the dissipation of supersaturated total dissolved gas. Ecol Modell 386:89–97. https://doi.org/10.1016/j.ecolmodel.2018.08.016

Yuan Y, Wang C, Feng J et al (2022) Mortality risk evaluation methods for total dissolved gas supersaturation to fish based on a mitigation measure of utilizing activated carbon. Water Res 225:119157. https://doi.org/10.1016/j.watres.2022.119157

Yuan Y, Chen Z, Feng J et al (2023) Research on the dissipation framework and dissipation coefficient prediction model of the supersaturated dissolved gas in solid media containing water. Process Saf Environ Prot 170:921–934. https://doi.org/10.1016/j.psep.2022.12.065

Zeng C, Mo K, Chen Q (2020) Improvement on numerical modeling of total dissolved gas dissipation after dam. Ecol Eng 156:105965. https://doi.org/10.1016/j.ecoleng.2020.105965

Zhang P, Liu Q, Wang Y et al (2022) River habitat assessment and restoration in high dam flood discharge systems with total dissolved gas supersaturation. Water Res. https://doi.org/10.1016/j.watres.2022.118833

Zhang D, Yang H, Ou Y et al (2023) Experimental and simulation investigation of total dissolved gas prediction in supersaturated water treatment: focusing on source calibration and combining with bubble coalescence. Environ Eng Sci. https://doi.org/10.1089/ees.2022.0345

Zhao J, Xuebin L, Daiwei Y et al (2023) Lithium-ion battery state of health estimation using meta-heuristic optimization and Gaussian process regression. J Energy Storage 58:106319. https://doi.org/10.1016/j.est.2022.106319

Zhu Z, Zhou M, Hu F et al (2023) A day-ahead industrial load forecasting model using load change rate features and combining FA-ELM and the AdaBoost algorithm. Energy Rep 9:971–981. https://doi.org/10.1016/j.egyr.2022.12.044

Zong W, Zhang J (2019) Use of smartphone applications and its impacts on urban life: a survey and random forest analysis in Japan. Sustain Cities Soc 49:101589. https://doi.org/10.1016/j.scs.2019.101589

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Salim Heddam[1] · Ahmed M. Al-Areeq[2,10] · Mou Leong Tan[3] · Iman Ahmadianfar[4] · Bijay Halder[5] · Vahdettin Demir[6] · Huseyin Cagan Kilinc[7] · Sani I. Abba[2] · Atheer Y. Oudah[8,9] · Zaher Mundher Yaseen[10,2]

✉ Salim Heddam
heddamsalim@yahoo.fr

✉ Zaher Mundher Yaseen
z.yaseen@kfupm.edu.sa

Ahmed M. Al-Areeq
ahmed.areeq@kfupm.edu.sa

Mou Leong Tan
mouleong@gmail.com

Iman Ahmadianfar
Im.ahmadian@gmail.com

Bijay Halder
halder06bijay@gmail.com

Vahdettin Demir
vahdettin.demir@karatay.edu.tr

Huseyin Cagan Kilinc
huseyincagankilinc@aydin.edu.tr

Sani I. Abba
saniisaabba86@gmail.com

Atheer Y. Oudah
atheer@alayen.edu.iq

[1] Agronomy Department, Faculty of Science, University 20 Août 1955, Skikda, Algeria

[2] Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia

[3] GeoInformatic Unit, Geography Section, School of Humanities, University Sains Malaysia, 11800 Minden, Penang, Malaysia

[4] Department of Civil Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran

[5] Department of Earth Sciences and Environment, Faculty of Sciences and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

[6] Department of Civil Engineering, KTO Karatay University, Konya 42020, Turkey

[7] Department of Civil Engineering, İstanbul Aydın University, Istanbul, Turkey

[8] Department of Computer Sciences, College of Education for Pure Science, University of Thi-Qar, Nasiriyah 64001, Iraq

[9] Information and Communication Technology Research Group, Scientific Research Centre, Al-Ayen University, Thi-Qar, Iraq

[10] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, 31261 Dhahran, Saudi Arabia